

Bangla Speech Baesd Emotion Detection

*Capstone project report to be submitted in partial fulfillment of the
requirements for the degree*

of

Bachelor of Science in Computer Science and Engineering

by

**Sadia Hasan Hima
161014008**

A. S. M Lutful Kabir	Md Mehedi Hasan
161014012	161014023

Under the supervision of

Mohammad Rifat Ahmed Rashid, PhD



**COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF LIBERAL ARTS BANGLADESH**

SPRING 2021

© *Sadia Hasan Hima, A. S. M Lutful Kabir, and Md Mehedi Hasan*
All rights reserved

DECLARATION

Project Title Bangla Speech Baesd Emotion Detection
Authors *Sadia Hasan Hima, A. S. M Lutful Kabir, and Md Mehedi Hasan*
Student IDs 161014008, 161014012, and 161014023
Supervisor Mohammad Rifat Ahmed Rashid, PhD

We declare that this capstone project report entitled *Bangla Speech Baesd Emotion Detection* is the result of our own work except as cited in the references. The capstone project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Sadia Hasan Hima
161014008

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

A. S. M Lutful Kabir
161014012

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Md Mehedi Hasan
161014023

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Date: June 13, 2021



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the capstone project report entitled **Bangla Speech Based Emotion Detection**, submitted by **Sadia Hasan Hima** (Student ID: 161014008), **A. S. M Lutful Kabir** (Student ID: 161014012) and **Md Mehedi Hasan** (Student ID: 161014023) are undergraduate students of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

Mohammad Rifat Ahmed Rashid, PhD
Assistant Professor

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Dr. Syed Akhter Hossain
Professor and Head

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Place: Dhaka

Date: June 13, 2021

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to our research supervisor, *Mohammad Rifat Ahmed Rashid, PhD*, for giving us the opportunity to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh who taught us during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends and relatives for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

Sadia Hasan Hima, A. S. M Lutful Kabir and Md Mehedi Hasan

University of Liberal Arts Bangladesh

Date: June 13, 2021

To my mother
Rashida Begum
and father
Md Kamrul Hasan

–Sadia Hasan Hima

To my mother
Momtaz Begum
and father
Late Golam Rosul Sikder

–A. S. M Lutful Kabir

To my mother
Shamsunnahar
and father
Md. Anisur Rahman Khan

–Md Mehedi Hasan

ABSTRACT

When humans communicate with each other, emotion plays quite a major role. It allows us to consider the emotions of others and give them meaningful responses. Since it is a subjective matter, identification of emotion is very difficult. Here, we will be trying to develop a system called "Bangla Speech Based Emotion Detection", which will analyze and classify the embedded emotions. Our system will convert audio data to text and perform emotion detection on the text. As we can categorize the human emotion state in 5 different categories: Excited, Happy, Neutral, Sad and Angry our system will predict the emotional state from the textual data and categorize them accordingly. For that reason, we are creating datasets based on audio files which contain Bangla sentences. We will convert those audio files into textual data. After the conversion, we will apply Bert algorithm to it for our emotion detection.

Keywords: emotion detection, Bert

Contents

1	Introduction	1
1.1	Problem Statement	1
1.2	Objective	2
1.3	Motivation	2
2	Literature Review	3
2.1	Sentiment Analysis	3
2.2	Speech Sentiment Analysis	4
3	Project Specification	7
3.1	Dataset Description	7
3.2	Deep Learning Approach	7
3.3	Application Specification	8
3.4	Specific Requirements	9
3.4.1	Software Requirements for User	9
3.4.2	Hardware Requirements for User	9
3.4.3	System Attribute	9
4	System Design	10
4.1	Use-Case Diagram	10
4.2	Sequence Diagram	12
4.3	Data Flow Diagram	14
4.4	Class Diagram	15
5	Proposed Methodology	17
5.1	Overview	17
5.2	Deep Learning Architecture	17

5.3	Software Development Process	19
5.3.1	Incremental Process	19
5.3.2	SDLC of Incremental Model	20
5.3.2.1	Requirement Analysis	21
5.3.2.2	Design & Development	21
5.3.2.3	Testing	21
5.3.2.4	Implementation	21
5.4	Design Pattern	22
5.5	Technology Used	22
6	Project Contribution	23
6.1	Overview	23
6.2	Proposed System	23
6.2.1	Training	23
6.2.2	Deploy	24
6.3	User Interface	24
7	Result and Analysis	27
7.1	Deep learning Model Performance Analysis	27
7.2	Security Analysis	28
7.2.1	Privacy	28
7.2.2	Integrity	29
7.2.3	Security	29
7.3	Evaluation	29
7.3.1	Interference Time	29
7.3.2	Application Performance(Time Complexity)	29
7.4	Cost Analysis	30
7.5	Social, Legal and Ethical Issues	30
8	Conclusion and Future Work	31
8.1	Conclusion	31
8.2	Scope of future work	31
	Bibliography	33

List of Figures

4.1	Use-case Diagram	10
4.2	Sequence Diagram	12
4.3	Data Flow Diagram	14
4.4	Class Diagram	15
5.1	Bert Model	17
5.2	Encoder Layer	18
5.3	Bert Model Encoder Layer	18
5.4	Incremental Model Step	20
5.5	Incrimental Model	20
5.6	SDLC of Incrimental Model	21
6.1	Sign Up	24
6.2	Sign In	25
6.3	Input	25
6.4	Output	26
6.5	Multiple File Output	26
7.1	Model Performance Analysis	28
8.1	Gantt-Chart	32

List of Tables

7.1	Total Cost	30
-----	----------------------	----

Chapter 1

Introduction

In the era of science and technology, researchers are trying to improve the interaction between machine and human. Though it is challenging for a machine to perceive emotions that are underlying speech, researchers have done so many works in this field. There are so many works on English language but compared to English, in Bangla language, it is quite low [1]. As a native speaker of Bangla, we are building a model that will recognize Bangla speech and extract the sentiment from it. It will be a web-based application. Researchers have been doing many works on Bangla language. Most of the works of sentiment analysis are based on word or text [2] and isolated speech [3]. The work on continuous speech is still low [4]. Researchers have been using many approaches such as Naïve Bayes Classification Approach, Topical Approach, HMM, DNN etc. to improve accuracy but there are still some accuracy issues. This project aims to achieve the maximum accuracy. We are using a self-created dataset to train our model. The dataset has been categorized into 5 classes such as excited, happy, neutral, sad and angry. This system will convert the audio data into text and then classify the emotion from the text data. It will classify whether the speech is happy or excited or neutral or sad or angry.

1.1 Problem Statement

There are many works in the field of sentiment analysis using Bangla speech but there is a lack of proper dataset of Bangla speech and it causes lower accuracy. We have created a dataset which has five categories: Excited, Happy, Neutral, Sad and Angry. People also have different accents. It has created some issues in recognition. To solve this problem, we will convert the audio data into text first.

1.2 Objective

There is no common consensus on how to categorize emotions. Our aim is to develop a web app that can convert audio into text & apply sentiment analysis on the textual data. For those purposes we are creating a dataset based on a Bangla audio file. Our main objective is to develop a system that can predict emotional state from Bangla audio speech data and achieve maximum accuracy

- We will convert the audio speech data to textual data.
- We will apply our machine learning algorithms for building up a model. This will be our sentiment classifier model that will predict emotional state of the textual data.
- this sentiment will be shown through a web application

1.3 Motivation

There are huge works on English language to detect emotion but the number of works has still low compared to that. As a native speaker of Bangla, we want to build an application in our mother tongue that will detect emotion. As we know it is more important to perceive the emotion in the field of business, politics, education etc. For an organization the most important thing is understanding the customer's feelings and opinion in real time. It is also important to analyze people's opinion for a political leader. This application will detect the emotion automatically which will be beneficial for them. There are many people who are physically challenged in our society, who can't hear and because of that they are unable to understand others feelings. This application will help them to understand others' opinions.

Chapter 2

Literature Review

As we have already mentioned that detecting emotion using speech has become a popular topic for research it is not entirely surprising that there are many similar studies that might bear resemblance with the concept of this study. Following are some of the related works that have been conducted earlier.

2.1 Sentiment Analysis

Sentiment analysis is a form of analytical technique of determining the emotional state of human or attitude of human towards something. It can take many form. Can be a collection of sentences in text form, can be a voice clip in audio form. In this modern era understanding the human feedback is very important for the small or large business organization. So organizations are using sentiment analysis effectively for understanding the true insights of their customer. In the study of L Kerkeni, Y Serrestou, et al. [5] have compared different approaches for emotion recognition. Based on the combination of the approaches they have proposed an efficient solution. They have used two types of databases such as Berlin database and Spanish database. They have classified seven emotions which are found in those databases by RNN classifier and compared with MLR and SVM classifier. 70% of data had been used for training and 30% had been used for testing in their investigation. They have combined MFCC and MS features using RNN and achieve the best recognition rate which is 90.05%. [5] Another related work conducted by P Badjatiya, S Gupta, et al. [6] have experimented on a dataset containing 16,000 tweets to define their task of hate speech detection as categorizing whether a par-

ticular tweeter post is racist or not, sexist or not. The results of this experiment are-

- 1,972 tweets have been classified as racist
- 3,383 tweets have been classified as sexist
- The rests have been labeled as neither racist nor sexist

By this research, it is shown that deep learning is more than efficient than the current n-grams method for detection of hate speech. [6]

2.2 Speech Sentiment Analysis

In the study conducted by Md Tahsin Tousif, Sayonton Chowdhury, et al. [7] have used Deep Speech Neural Network Architecture for mapping frames of a sequence of audio to characters. Though there are some issues in Bangla language such as implicit vowel, vowel modifiers, silent letter and joint letter. They have converted the transcript into Broken Language Format (BLF) to deal with these issues. To tackle implicit vowel and silent letters they have explicitly added the implicit vowel between the consonants. The proposed system will convert the vowels by their respective modifiers. It will reduce ambiguity for the network. Instead of 'kar' the vowels and consonants will be mapped to their respective letters. To predict, they have used DNN and N-grams and for post-processing, they have used Levenshtein Distance. Edit distance has been used after getting the BLF from the model. The model is trying to predict possible words that have the minimum edit distance from the word predicted. After this prediction the broken word will be replaced by the normal word. The prediction of their training dataset has 95% accuracy and test data has 50% accuracy.[7]

Another related work conducted by Md Akkas Ali, Manwar Hossain, et al.[8] have represented a speech detection technique for Bangla language. Since the pronunciation and accent of the Bangla words depends on the locality of the people, so based on that they have created a dataset. Here the approach is divided into two phases (Training & Recognition). Training phase includes the creation of datasets and the feature extraction process. In the recognition phase, the cross-matching process where the inputted test voice matched with the stored reference model

and gave out results. To detect and process the signal there are lots of algorithms and filters. In the processing firstly the recording needs to divide in frame, so that the analyzing procedure can be done frame by frame. For the filtering process here they have used the high pass filter called Pre-emphasis filter. After applying the filter to remove the ripples in the signal here they have a process called Hamming Window. Then they have applied Fourier Transformation to represent it in the frequency domain. Then from the speech signal waveform begins the feature extraction process. They have used different mathematical models MFCC, GMM and LPC. In this paper they have proposed 4 different model for speech recognition, model 1: Mel Frequency Cepstral Coefficient for feature extraction and Dynamic Time Warping, model 2 :Linear Predictive Coding and calculate linear predictive coefficients for extract feature and Dynamic Time Warping for matching , model 3: Mel Frequency Cepstral Coefficient and Gaussian Mixture Model for feature extraction and posterior probability function for matching, model 4: Mel Frequency Cepstral Coefficient and compress it by Linear Predictive Coding for extract feature and Dynamic Time Warping for matching. [8] In another study conducted by A Feroze, Ms Arifin, et al. [9] have proposed a platform that will translate Bengali speech to Bangla text by recognition and fuzzy logic. There are two major phase Training and recognition. Training phase includes creation of datasets, then they represent it in a time domain. They have used mono audio files. The technique they have used for the segmentation process is end to end, they have maintained threshold value. The system will seek for a greater amplitude value from threshold. To generate spectrum, they have used short term Fourier Transformation. To get better approximation they have filtered and down-sampled the spectrum.

They have divided the frequency domain in 30 windows and the time domain into 40 windows. Now the logic implementation process, here they have used Fuzzy logic or FIS to implement the Comparison of matrix. This proposed works based on 50 Bangla words description including 3 cases: monosyllabic, bi-syllabic and polysyllabic. This proposed system has an adaptive capability according to the user input, and the user of the system gets the liberty of changing the description based on voice, tone and stress level. If a particular word – spoken by the user is recognized incorrectly, then the system asks for the correct word from the user.[9] So we can understand from the previous works of sentiment analysis on Bangla language have few issues such as lower accuracy, variation of accents. So,

we are going to develop a system that will recognize Bangla language and apply sentiment analysis on it.

Chapter 3

Project Specification

3.1 Dataset Description

After visualizing a few open source datasets, we have decided to create our own dataset for the system. Since we are creating a dataset for Bangla language, for that we need a source of Bangla speech. We have collected Bangla speech from YouTube. After clipping the Bangla speech sentences from the source we put them through into our preprocessing steps. Here clean noise from the speech data, then labeling them according to our dataset index. After that we normalize the speech data, now our curated dataset is ready for conversion. The number of datasets is 1900. We have divided our dataset into five different categories. The ratio of the categories are:

- Excited- 12.2%
- Happy- 17%
- Neutral- 23%
- Sad- 22%
- Angry- 26.5%

3.2 Deep Learning Approach

Deep learning is a technique of computer vision where computer model learns to execute a task from a given text data, audio data, image data. This computer model

trains based large number of curated dataset and NN architecture. Traditionally deep learning models have huge number of hidden layers of neural network. As we discussed earlier we created for our system, we have categorized it and labeled it by analyzing 'Russel's Circumplex' model. In this system we have experimented two different model of Natural language processing. One is BERT (Bidirectional Encoder Representations from Transformers) Model and another is LSTM (Long Short-Term Memory). BERT model is a transformer architecture which contains encoder stacks. This transformer architecture is a network of two side that contains encoder and decoder where encoder uses self-attention and decoder uses attention. [10] In this system BERT model have been used for developing the system.

3.3 Application Specification

Here we will be discussing about our total project specification:

- The system is developed using python language. The system is going to work through API.
- We are using FAST API for the system and for conversion of audio data to text data we are using Google's API.
- For any kinds of interaction between the data and trained model, the system will call the API.
- After that the user will be able to upload the audio file into the system.
- Then the uploaded audio data file will be converted into text data using Google's API.
- Now the system will send the text data via Fast API to the sentiment analysis model for analysis, classification and prediction.
- After that, the predicted output will be sent using the micro-servicing to the interface, the interface will show the result to the user which was given by the sentiment analysis model of the system.
- User needs to register and login to the system

- The user can upload multiple files into the system for analysis and classification.
- For maintenance, there will be an admin, who will monitor the working flow of the system.

3.4 Specific Requirements

As we already know this is a web-based application users just have to have a computer. A stable internet connection. Then they are ready to use the web application.

3.4.1 Software Requirements for User

This is a web based application. So there are no specific software requirements for the user. Any browser will be suitable for using the system,

3.4.2 Hardware Requirements for User

PC, routers etc.

3.4.3 System Attribute

- In this system user can register
- user can login to the system
- In this system user can upload file into the system
- User can upload multiple files at a time for sentiment analysis
- This system can count the multiple files and can categorize accordingly.
- In the system the speech recognizer API convert audio speech data to text data.
- After the preprocessing, the data fed into the trained model and the model processed it and gave output to the user via web app.

Chapter 4

System Design

4.1 Use-Case Diagram

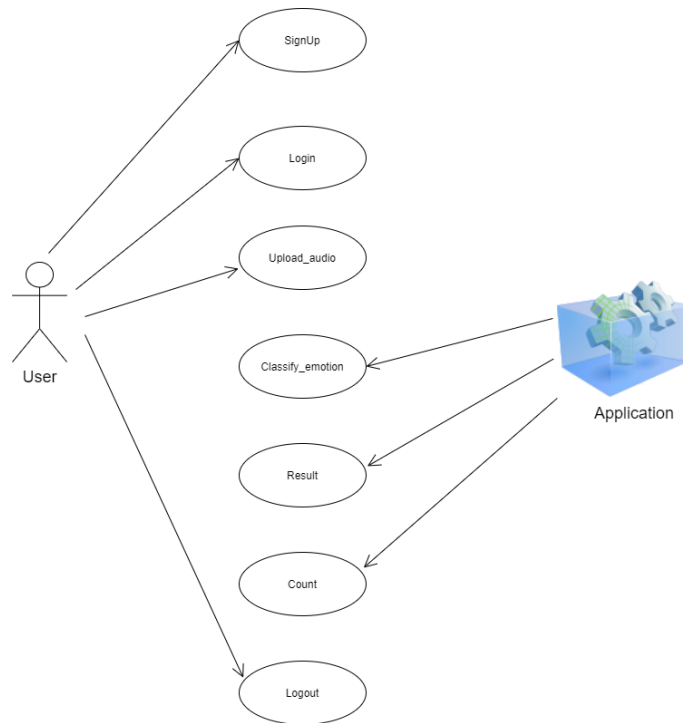


Figure 4.1: Use-case Diagram

There are several diagram in UML. Use-Case diagram is one of them. It is also known as behavior diagram. A system which is underdeveloped can be summarized by the use-case diagram. Use-case diagram has four major components. These are-

- **Actor:** Actor is a person who will interact with the system.
- **System:** System is the application itself for which we made the diagram.
- **Services:** A system has some functionalities that are known as service or use case.
- **Lines:** There is a interaction between system and user. These lines symbolizes the relationship between the components.

Here the figure which symbolizes people is the actor. The actor is the user who applying the services. These Oval shape represents the use-cases which are the functionality of the system. In this system the user can perform sign up, login, upload audio and logout. The application or system will convert audio to text, classify emotion from text and count according to the categories.

4.2 Sequence Diagram

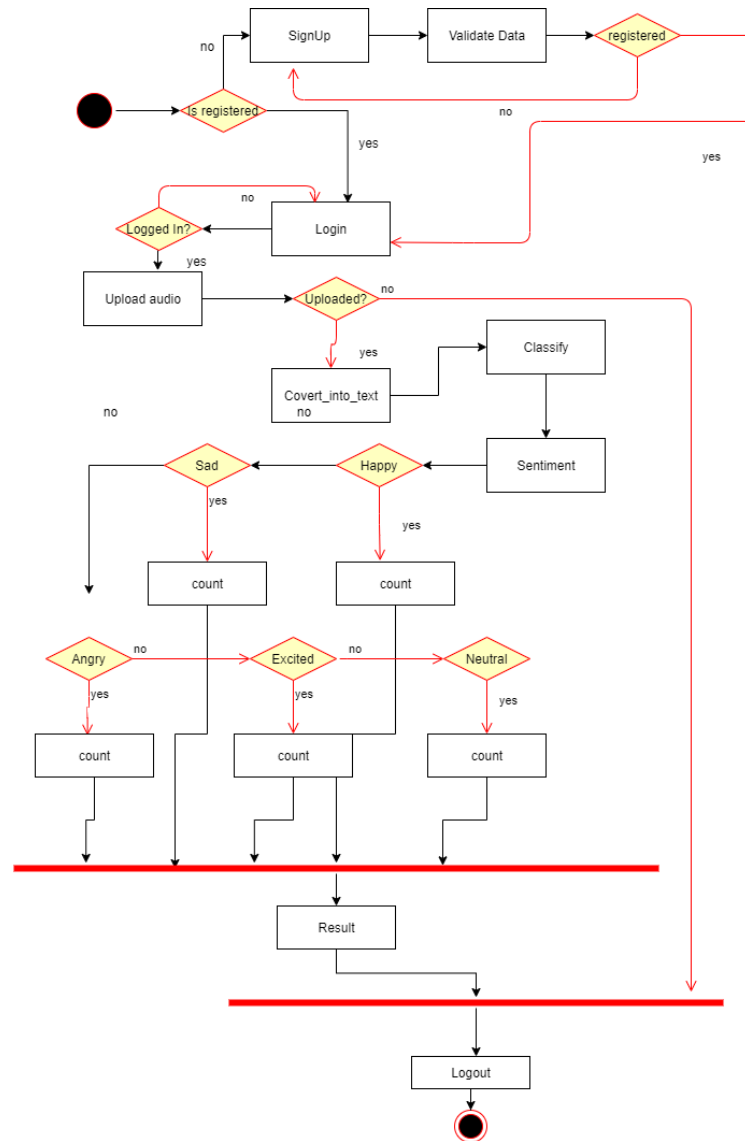


Figure 4.2: Sequence Diagram

Sequence diagram is another behavioural diagram which represents the system process flow from start to end. It also shows the decision paths while executing an activity. It shows how the system actually works based on condition.

In this diagram:

- the black circle represents the starting point.
- the diamond shape represents the condition.
- the rectangle shape represents the services or functionalities.
- the arrows represent the working flow
- the circle surrounded by the red line represents the finishing point.

Working flow:

- starts the application
- check whether the user is registered or not. If yes, then the next process is Login. If no, then the next process is Signup. there is a validation process. if the data is validated then the next process is Login. if not, then back to Signup.
- After successfully login user can upload audio. If the audio is uploaded system will convert it into text. if not, user needs to upload again.
- the system then classify the sentiment according to the categories such as Happy, Excited, Neutral, Sad and Angry and count them. then the system shows the result.
- the next process is logout
- end

4.3 Data Flow Diagram

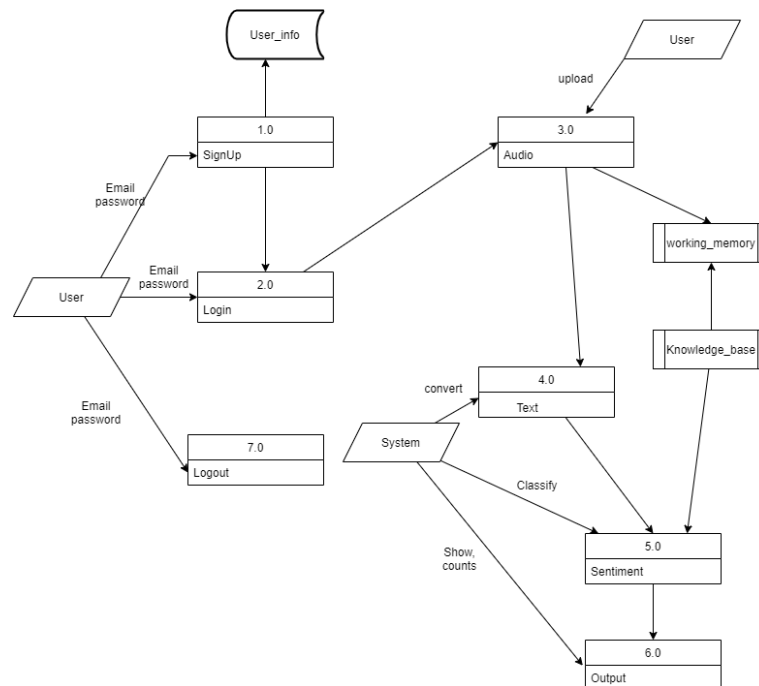


Figure 4.3: Data Flow Diagram

The Data Flow Diagram called as DFD used to represents the flow of the data that depicts the processes that are included to the services or processes to trade data from the input of external entity and store it into data store. There are 4 major components:

- **External entities:** Parallelogram shape represents the external entity.
- **Activity:** The rectangle represents the process or activity that can take input from entities and produce output.
- **Data store:** it stores the output that is produced by process or give the information to process.
- **Data flow:** The arrow symbolizes the path for a data for moving is called data flow.

4.4 Class Diagram

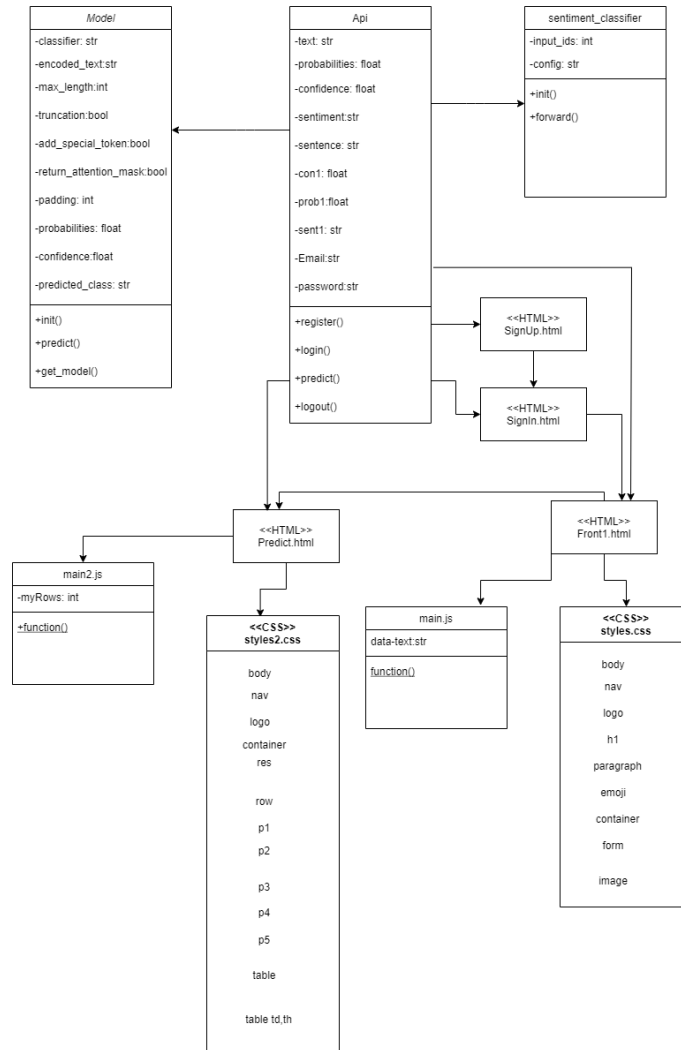


Figure 4.4: Class Diagram

Class diagram is UML diagram which is the building block of UML and represent the static view of executable code that how the modelling process has been done.

There are the classes named Model, sentiment_classifier, api, four html associated with CSS and two javascript class named main and main2.

Model class has some variables such as classifier, encoded_text, max_length, truncation, add_special_token, return_attention_mask, padding, probabilities, predicted_class and three methods such as init(), predict(), get_model().

api classes variables are text, probabilities, confidence, sentiment, sentence, con1, prob1, sent1, Email, password and methods are register(), login(), predict(), logout().

sentiment_classifier has some variable these are input_ids, config and has two methods such as init() and forward().

main has one variable name data-text and a function named function().

main2 has one variable name myRows and a function named function().

all the classes, HTML and CSS have been joined with the api class.

Chapter 5

Proposed Methodology

5.1 Overview

We have developed a system of sentiment classifier for Bangla language. We have created a dataset for the system's training and testing purpose. The curated dataset is in audio (FLAC) format. This System will get the input in audio data form. It will convert the audio data into text data using speech recognizer API. After the conversion the data will put through the BERT model for training. Once the BERT model have been trained we test and validate the model. After that the model has been deployed using fast API.

5.2 Deep Learning Architecture

As we have already discussed we have experimented two different NLP model one is BERT model. This are different one from another. BERT model works based on Transformer architecture of a stack of encoder decoder. BERT model has two sizes, one is 'BERT BASE' and other is 'BERT LARGE'.[\[10\]](#)

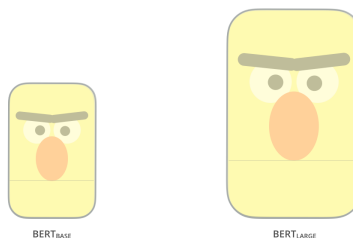


Figure 5.1: Bert Model

In their encoder stack BERT BASE has 12 layer and BERT LARGE has 24 layer. Both of them have large number of feed forward network (768 and 1024). This feed forward network is called hidden units. Both BERT BASE and BERT large have a parameter of 110M and 340M respectively. For this project BERT BASE has been used.[11]

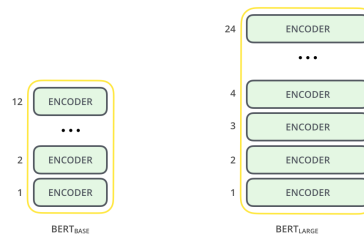


Figure 5.2: Encoder Layer

When we input a text into the model to work with it firstly it will tokenize the input, it is known classification token. This [CLS] token associate with embedding vectors. Then the model pass this classification token in the above encoder for applying the self-attention and get the result, this result will push through the feed-network. It hands it to the next encoder layer.[11]

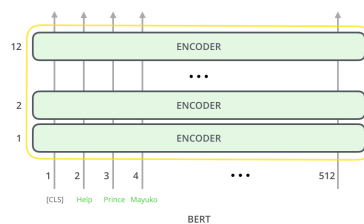


Figure 5.3: Bert Model Encoder Layer

After that it performs Masked Language Modelling (MLM). In MLM 15% of words of the input sequence have been masked and try to predict the masked words. From these masked tokens 80%-time masked token is replaced with [Mask] token, 10% time replaced with random words and 10% time replaced with the actual words. That's how it learns to understand the context of the word. After the MLM it takes the output of [CLS] token that is used for getting binary loss. Then, the binary loss is back propagated.[12]

5.3 Software Development Process

Software development process is a set of logical sequential work for brainstorming, designing, developing, implementation and testing phase of the software system. There are quite a number of software process models:

- Waterfall model
- Incremental model
- V model
- Iterative model
- Rapid Application Development model
- Spiral model
- Agile model
- Prototyping

In this project incremental model has been used as the software development process model.

5.3.1 Incremental Process

Incremental model split up the projects functionality into small increments that are given in rapid succession one after another. Here, in the first couple of increments the most important functionality has implemented. The following increments build on the prior ones until everything is updated and implemented. Every iteration will go through the phases of requirements, design, coding, and testing and validation process.

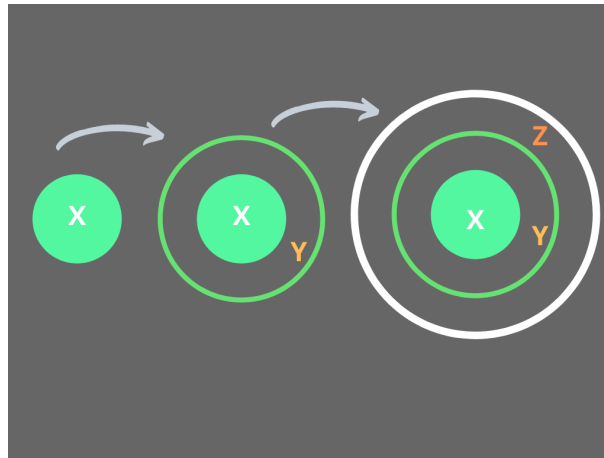


Figure 5.4: Incremental Model Step

In the given figure X, Y and Z software product modules are that developed following incremental model. So, we can say that incremental model depends on initial implementation testing it by user and iterating through newer versions. [13]

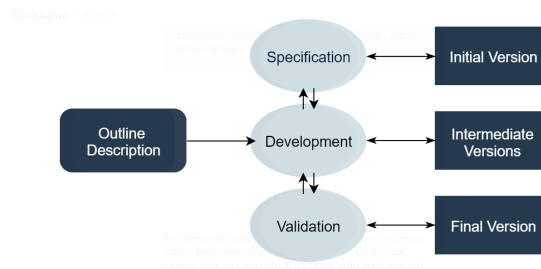


Figure 5.5: Incremental Model

5.3.2 SDLC of Incremental Model

In the incremental model requirements of the project broken down into various different segments which will build incrementally. Here, the plan is only for the next increment and not for any long-term goals, as a result of that it is simpler to update or change the version to meet the need of the customer. Following the collection and specification of needs, the requirements are spitted into various different versions, beginning with version-1, with each succeeding increment (version-n), the next version is produced and then deployed at the customer's end point.[14]

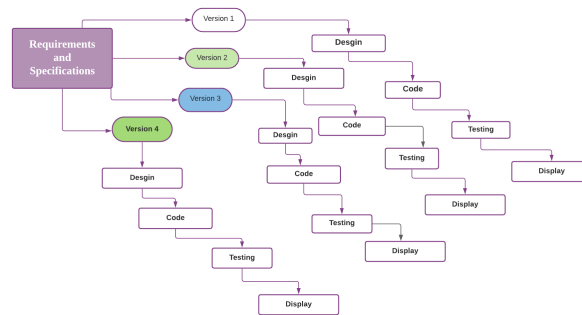


Figure 5.6: SDLC of Incremental Model

5.3.2.1 Requirement Analysis

In this step the experts related to project requirements, analyze the necessary functionality of the project. This phase plays a significant role for a project to succeed [15]

5.3.2.2 Design & Development

Firstly, the developers try to design and build the basic and main functionalities of the system. After completing the task developers look for refinement and as well as adding new features in the next update.[15]

5.3.2.3 Testing

In this phase here, evaluation of the performance of the project is going to take place. Also how the systems functionality is working or not working, methods, these are the things tested in this phase.[15]

5.3.2.4 Implementation

It entails the final stage coding, what was planned and designed and tested during the development stage. Following that the number of working products is increased and update all the way towards the final output of the product.

Every new incremental version is another iterative waterfall model of software development process[15]

5.4 Design Pattern

In this project there is no use of any specific design pattern

5.5 Technology Used

The list of technological tools that have used for implementation of the project is given bellow:

- PL: Python, HTML, CSS, JQuery.
- IDE: Google Colab, VS Code.
- Framework: Fast API, Google speech conversion API (audio to text conversion).
- Software: WavePad (audio preprocessing)

Chapter 6

Project Contribution

6.1 Overview

Sentiment classifier is a common project these days, we have discussed earlier that for dedicated to Bangla language there is not a lot of project on that. In this project we have a dataset based on Bangla language and divided into different five categories. We have labeled the dataset following Russell's Circumplex model.

6.2 Proposed System

We divided the project into few phases: As we have discussed earlier before starting the project work, at first, we did research thoroughly on our topic, we have analyzed and visualized a few datasets and have tried to build up our concept from that.

6.2.1 Training

In this phase we converted our curated audio dataset into text data using Google's API.

Then we applied our machine learning algorithm on that converted text data. We are using two state of the art machine learning techniques, those are BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short Term Memory). After training the model using BERT and LSTM, we have analyzed the performance of our trained model. According to that we will optimize our model for better performance.

6.2.2 Deploy

In the deploy phase, here we have worked on the deployment of our system. We have used the FAST API to serve our trained model for real time use. Firstly we have designed and built our user interface, which will help to interact between the user and system. The system is developed based on the FAST API. This framework will help to get the data from the user and will send it to our trained model for classification, after successfully classifying the data, the result will be in the interface via micro-servicing. For micro-servicing of the system Swagger and Docker has been used. The classifying result will be sent to the web application.

6.3 User Interface

This is the user sign up form for the user where user can sign up themselves by using their email address and a user generated password.

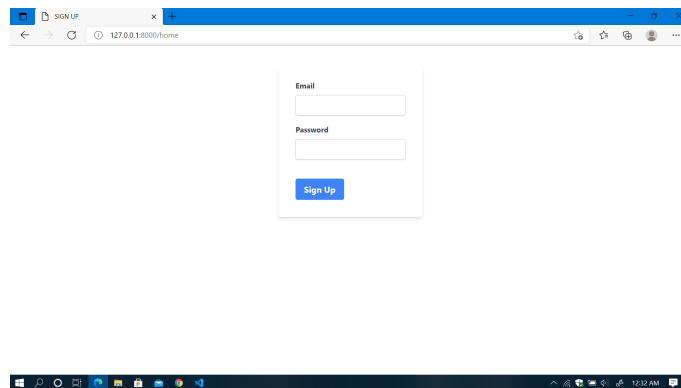
A screenshot of a web browser window. The browser's address bar shows the URL "127.0.0.1:8000/home". The page content is a simple sign-up form with two input fields: "Email" and "Password". Below these fields is a blue button labeled "Sign Up". The browser window has a single tab titled "SIGN UP". The operating system's taskbar is visible at the bottom of the screen.

Figure 6.1: Sign Up

This is user sign in panel where user can log into the system by using their USER Name and Password

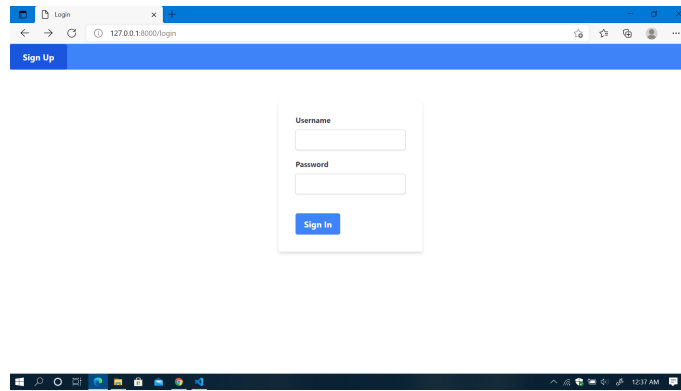


Figure 6.2: Sign In

This is the user input panel, where user can upload a file into the system for processing and classification

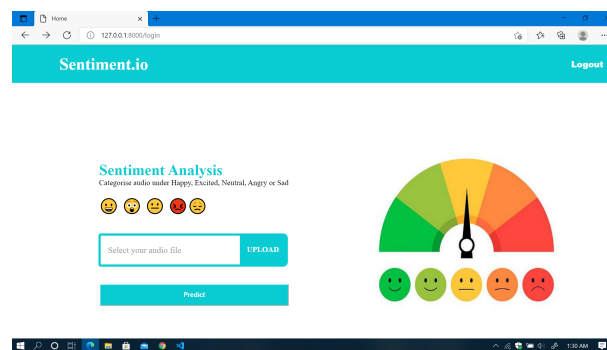


Figure 6.3: Input

This is where the output will show to user with few other statistics.

Sentence	Sentiment	Confidence	Probabilities
তুমি এখন কি করবে?	Angry	0.82299	{ "Excited": 0.0479176124268294, "Happy": 0.02491864262275, "Sad": 0.07148817986244, "Neutral": 0.360791015058604, "Angry": 0.52991513750495 }

Figure 6.4: Output

Output for multiple files is shown in this figure

Sentence	Sentiment	Confidence	Probabilities
তুমি এখন কি করবে?	Angry	0.82299	{ "Excited": 0.0479176124268294, "Happy": 0.02491864262275, "Sad": 0.07148817986244, "Neutral": 0.360791015058604, "Angry": 0.52991513750495 }
আমি যখন বাবা থেকে বের হই	Angry	0.6953	{ "Excited": 0.0296721615013166, "Happy": 0.021001428226044, "Sad": 0.0411130609112168, "Neutral": 0.4113497996461034, "Angry": 0.47261466427867 }
আমাকে খোব করে কি বলছিল	Neutral	0.8441	{ "Excited": 0.057116091912114, "Happy": 0.038145851740616, "Sad": 0.0623191718811976, "Neutral": 0.54408588391098, "Angry": 0.302241511416887 }
বলছি না তুমি আমার জন্য কিছর উপর থাকবে	Angry	0.6944	{ "Excited": 0.0397221917804472, "Happy": 0.039818290690916, "Sad": 0.039818290690916, "Neutral": 0.208678422615926 }

Figure 6.5: Multiple File Output

Chapter 7

Result and Analysis

7.1 Deep learning Model Performance Analysis

Before analyzing this systems performance we have to understand few terminologies, For classifier prediction there are mainly outcomes, there are True positive, True negative False positive and false negative. True positive, prediction of an observation belongs to a certain class and it actually belongs to that class. True negative prediction of an observation does not belong to a certain class and in real time it doesn't belong to that certain class. In terms of False positive , prediction of an observation belongs to a class but in real time it doesn't and False negative observation outcome does not to a certain class but in real time it does belong to that class.[16] This is the base of performance analysis.

Precision = the number of correct positive result divided by the number of positive results predicted by the classifier

$$precision = \frac{truepositives}{truepositives + falsepositives} \quad (7.1)$$

Recall: Number of true positive results divided by the sum of true positive and false negative

$$recall = \frac{truepositives}{truepositives + falsenegatives} \quad (7.2)$$

F1 score = It is the test accuracy score of the trained model

In the given bellow equation we can see the results of the performance of this model.

$$F1 = 2 \frac{precision \cdot recall}{precision + recall} \quad (7.3)$$

Due to the lower number of datasets, the accuracy of this model is on the lower side.

	precision*	recall	f1-score	support
Excited	0.12	0.20	0.15	5
Happy	0.75	0.43	0.55	7
Sad	0.29	0.33	0.31	6
Neutral	0.46	0.50	0.48	12
Angry	0.40	0.29	0.33	7
accuracy			0.38	37
macro avg	0.40	0.35	0.36	37
weighted avg	0.43	0.38	0.39	37

Figure 7.1: Model Performance Analysis

7.2 Security Analysis

Machine learning systems generally have these five types of security issues mainly:

- Data poisoning
- System manipulation
- Data confidentiality
- Transfer learning
- Adversarial examples

7.2.1 Privacy

This system provides authentication process for user. This system does not share any sort's user data. The system doesn't store any file that is uploaded by the user. So, this system ensures the data privacy and confidentiality of its user. [17]

7.2.2 Integrity

This systems model is trained by a new dataset which have been for the purpose of training, testing and validation. Its dataset is not open source for the time being. So, type of data that have fed into the system for training that is only known the systems developer and related people. It's a potential risk that if anyone get idea of training data type the attacker can invert it. It can affect the output of the system, can generate false output. It also can give out the information of the model that was fine tuned for the purpose. From that significant information attacker can also exposed the training dataset [18]

7.2.3 Security

We have safeguarded our system from SQL injection attacks. SQL injection is a technique of injecting maleficent code for revealing user information. We have remove quote mark form the each and every text field. [19]

7.3 Evaluation

7.3.1 Interference Time

In this application user first need to upload audio to get sentiment. This application will take 2.08s for upload to predict for a single audio.

7.3.2 Application Performance(Time Complexity)

This application is used Bert algorithm for prediction. In Bert every layer has used self-attention mechanism. Each word token has embedding dimension and attention weights are calculated for each word by their respective words.

Let's,

Embedding dimension= d

So, operation for each word= $O(n)$

$$\text{Operation for all of the word} = O(n * n) = O(n^2) \quad (7.4)$$

$$\text{The time complexity} = O(d * n^2) \quad (7.5)$$

7.4 Cost Analysis

We have divided our costing into few different categories:

Category	month1	month2	month3	Total
Software				
Wave-pad	8148/-			8148/-
Developer				
Front end-Developer(1)	18000/-	18000/-	18000/-	54000/-
Back end developer(2)	60000/-	60000/-	60000/-	180000/-
Hardware				
Router(1)	2000/-			2000/-
Laptop(3)	176000/-			176000/-
Maintenance	3000/-	3000/-	3000/-	9000/-
Total				429184/-

Table 7.1: Total Cost

7.5 Social, Legal and Ethical Issues

- This system doesn't store any data which is related to sentiment analysis, that are uploaded by the user. So, there is no manipulation of the data from the system.
- The source of the data that has been used to build the system is YouTube. We collected our data from YouTube by following YSL.
- This system does not unethical use of data, for any kinds of commercial gain.
- The system's accuracy rate of sentiment analysis can create a negative impact on individuals.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

The system that we have build is able to classify the emotional state from speech data. The speech data will converted into text and then classify according to the classification categories. Our initial goal is to classify emotion from converted text data. We have used self created dataset and Bert model to train. The system is properly functioning and classifying data.

Drawbacks: As our dataset size is not large, the accuracy of the model is not on the higher side. The system sometimes gives the wrong output.

There are some speech data that represented as angry but when it converted into text it acted like neutral. That is one of the issue as our system classify emotion from converted text.

this system has no functionality to classify emotion from real time speech.

8.2 Scope of future work

We have future plans for this project.

- We will develop more user-friendly interface for the system.
- We will introduce Multilanguage support in the system.
- We will develop a user profile management feature through admin.
- We will be working on safety and security features of the system.

- We will develop feature for detecting sentiment from real time speech data.

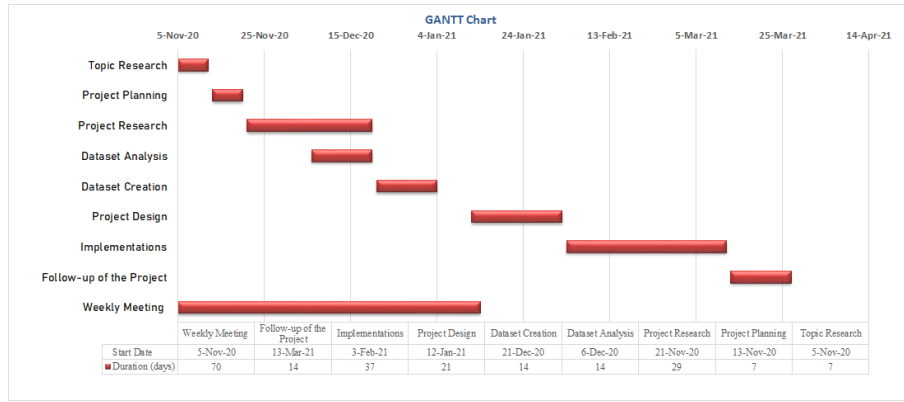


Figure 8.1: Gantt-Chart

Bibliography

- [1] Mohit, Anurag, and Prateek Kumar. Speech emotion recognition (ser) through machine learning, Jul 2020.
- [2] Sara Azmin and Kingshuk Dhar. Emotion detection from bangla text corpus using naïve bayes classifier. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–5, 2019.
- [3] Md Mijanur Rahman and Fatema Khatun. Development of isolated speech recognition system for bangla words. *Daffodil International University Journal of Science and Technology*, 6(1):30–35, 1970.
- [4] Biswajit Das, Sandipan Mandal, and Pabitra Mitra. Bengali speech corpus for continuous automatic speech recognition system. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 51–55, 2011.
- [5] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, and Mohamed Ali Mahjoub. Speech emotion recognition: Methods and cases study. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2:175–182, 2018.
- [6] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW 17 Companion*, page 759–760, Apr 2017.
- [7] Md. Tahsin Tausif, Sayontan Chowdhury, Md. Shiplu Hawlader, Md. Hasanuzzaman, and Hasnain Heickal. Deep learning based bangla speech-to-text conversion. In *2018 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*, pages 49–54, 2018.

- [8] Akkas Ali, Manwar Hossain, and M. N. Bhuiyan. Automatic speech recognition technique for bangla words. volume 50, pages 51–59, 2013.
- [9] A. Firoze, M. Arifin, R. Quadir, and R. Rahman. Bangla isolated word speech recognition. In *ICEIS*, volume 2, pages 73–82, 2011.
- [10] What is deep learning?: How it works, techniques & applications.
- [11] Jay Alammar. The illustrated transformer, Jun 2018.
- [12] Explanation of bert model - nlp, May 2020.
- [13] Maryam Sulemani. What is a software process model? top 7 models explained, Jan 2021.
- [14] Ritwik Shanker and Naikhrushikeshn. Software engineering: Incremental process model, Dec 2019.
- [15] Extended Learning Institute (ELI) at Northern Virginia Community College (NOVA). Introduction to computer applications and concepts.
- [16] Jeremy Jordan. Evaluating a machine learning model., Aug 2018.
- [17] Maria Korolov. How secure are your ai and machine learning projects?, Nov 2020.
- [18] Gary McGraw Ph.D. How to secure machine learning, Jun 2020.
- [19] Irfan Ahmed Khan. 5 common machine learning security risks and how to overcome them, Jul 2020.