

# Emotion Detection for Bangla Language

by

Rowshan Rahman Rushan

19301210

Sazid Hossain

19301224

Shams Shahariar Shovon

19301081

Md Arafat Rahman

20101121

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
School of Data and Sciences  
Brac University  
January 2024

© 2024. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

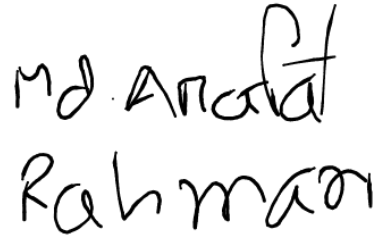
**Student's Full Name & Signature:**



---

Rowshan Rahman Rushan

19301210



---

Arafat Rahman  
20101121

Shams Shahriar Shovon

---

Shams Shahriar Shovon

19301081



---

Sazid Hossain

19301224

# Approval

The thesis titled “Emotion Detection for Bangla Language” submitted by:

1. Rowshan Rahman Rushan (19301210)
2. Sazid Hossain (19301224)
3. Shams Shahariar Shovon (19301081)
4. Md Arafat Rahman (20101121)

Of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January, 2024.

## Examining Committee:

Supervisor:  
(Member)



---

**Najeefa Nikhat Choudhury**  
Lecturer  
Department of Computer Science  
and Engineering  
Brac University

Co-Supervisor:  
(Member)



---

**Dewan Ziaul Karim**  
Lecturer  
Department of Computer Science  
and Engineering  
Brac University

Program Coordinator:  
(Member)

---

**Dr. Md. Golam Rabiul Alam**  
Associate Professor  
Department of Computer Science  
and Engineering  
Brac University

Head of Department:  
(Chair)

---

**Sadia Hamid Kazi, PhD**  
Chairperson and Associate Pro-  
fessor  
Department of Computer Science  
and Engineering  
Brac University

## **Ethics Statement**

We hereby declare and confirm that all the work done in preparing this thesis 'Emotion Detection for Bangla Language', from the beginning of research to implementation of the proposed models are our own. All sources, datasets, and external resources that we used have all been cited properly. We have also abstained from employing any immoral methods in collecting any of the resources. We hereby declare that this paper has not been previously submitted to any other universities or institutes to complete a degree.

## Abstract

This project aims to identify emotions and generate Bangla-language emojis. Emojis have become essential to digital communication, transcending language borders and improving textual exchanges. This study identifies emotions as Emojis from Bangla text rather than directly detecting them. This study aims to develop a system that uses machine learning and NLP to recognize Bangla text feelings and correlate them to relevant emojis. Using a large sample of Bangla texts, the approach maps emotional content to emojis. Innovative NLP methods, including sentiment analysis, contextual embeddings, and deep learning algorithms, identify emotions in the framework. It's difficult to effectively read Bengali, a language full of expressions and strong emotions, and produce contextually suitable and culturally informed emojis. An Emotion Detection system for Bangla will improve digital communication in Bangla-speaking populations by making texting more expressive and emotionally resonant. It will advance computational linguistics and human-computer interaction, especially for Bangla. The method integrates smoothly with internet platforms to revolutionize the Bangla-speaking digital exchange of emotions.

**Keywords:** Emotion Detection, Emoji Generation, Bangla Language, Digital Communication, NLP, Machine Learning, Deep Learning, Cultural Sensitivity.

## Dedication

This thesis is dedicated to our parents and well-wishers to acknowledge their encouragement, and affection throughout our lives. Their sacrifices and unwavering faith in us have been the foundation of our accomplishments.

## **Acknowledgement**

To our supervisor Ms.Najeefa Nikhat Choudhury miss, for her expertise, insightful feedback invaluable and support throughout this research journey. Also our co-supervisor Dewan Ziaul Karim sir, for his guidance.

And finally to our parents, without their support, it may not be possible and with their kind support and prayer, we are now on the verge of our graduation.



# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Ethics Statement</b>	<b>iv</b>
<b>Abstract</b>	<b>v</b>
<b>Dedication</b>	<b>vi</b>
<b>Acknowledgment</b>	<b>vii</b>
<b>Table of Contents</b>	<b>viii</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xi</b>
<b>Nomenclature</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Problem . . . . .	2
1.2 Our Contributions . . . . .	4
1.3 Research Objectives . . . . .	5
<b>2 Literature Review</b>	<b>6</b>
2.1 Comparative Analysis Of Related Work . . . . .	13
2.2 Visual representation of papers . . . . .	14
<b>3 Description of Model, Data and Preliminary Analysis</b>	<b>16</b>
3.1 Description of Model . . . . .	16
3.1.1 Random Forest Classifier . . . . .	16
3.1.2 BERT-based Deep Learning . . . . .	17
3.1.3 Linear SVC(SVM) . . . . .	18
3.1.4 Naive Bayes Classifier . . . . .	19
3.1.5 Enhanced Ensemble Method: Integrating RandomForest and SVM . . . . .	19
3.1.6 SGD Classifier: . . . . .	21
3.2 Model Implementation . . . . .	21
3.3 Dataset Description . . . . .	22

3.4	Preliminary Analysis . . . . .	26
3.5	Dataset Comparative Analysis . . . . .	27
<b>4</b>	<b>Work Plan</b>	<b>29</b>
4.1	Methodology . . . . .	31
<b>5</b>	<b>Result and Analysis</b>	<b>34</b>
<b>6</b>	<b>Conclusion</b>	<b>37</b>
	<b>Bibliography</b>	<b>39</b>

# List of Figures

2.1	Architecture from the study [[10]] . . . . .	14
2.2	Architecture from the study [[17]] . . . . .	15
2.3	Architecture from the study [[12]] . . . . .	15
3.1	Random Forest . . . . .	17
3.2	SVC . . . . .	18
3.3	Ensemble Model . . . . .	20
3.4	Dataset . . . . .	22
3.5	DatasetTest . . . . .	23
3.6	DatasetTain . . . . .	24
3.7	Prediction Label . . . . .	26
4.1	Work Plan . . . . .	30
4.2	Tokenization . . . . .	31
4.3	FastText . . . . .	32
5.1	Model's Comparison . . . . .	36

# List of Tables

2.1	Comparative Study of Models . . . . .	13
5.1	Accuracy Scores . . . . .	35

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

*BERT* (Bidirectional Encoder Representations from Transformers)

*EDA* (Exploratory Data Analysis)

*LSTM* (Long Short Term Memory)

*ML* (Machine Learning)

*NLP* (Natural Language Processing)

*RNN* (Recurrent Neural Network)

*SVC* (Support Vector Classifier)

# Chapter 1

## Introduction

Digital platforms revolutionized communication. Natural Language Processing has possibilities and problems with this transition. Bangla-language digital communication relies on emojis. Visual and emotional factors expand language boundaries. The COVID-19 pandemic has increased their use, which has changed how emotions, humor, and subtlety are expressed online. However, using emojis and understanding their emotions in Bangla writing is challenging. Bangla's cultural and linguistic complexities complicate this issue. The study of emoji emotion recognition is vital to improving Bangla digital communications' understanding and accurate depiction of emotions. This research uses Random Forest, SVM, Ensemble, BERT-based deep learning, and SGD Classifier. They were chosen for their pattern identification and predictive analytic skills across multiple datasets. Accuracy ratings and classification reports will evaluate the model. These assessments will reveal the model's performance and inform improvements. Additionally, grid search will be used to fine-tune model parameters to maximize system efficiency. This research explores Bangla's complex linguistics and creates a system that can reliably discern emotions and improve digital communication by suggesting emojis. This initiative is crucial because it intends to make Bangla-speaking digital interactions more engaging and emotive. It will meet a major requirement in Natural Language Processing (NLP) and set a new standard for Bangla-language emotion detection and emoji association models. The goal is to analyze Bangla-language literature and create a model that can predict emojis. Text conversation will become more vivid and emotional.

## 1.1 Research Problem

Researching and building a Bangla-language emotion detection system to produce text emojis manually is complex. We seek to accurately recognize Bangla language emotions and match them with appropriate emojis. Natural Language Processing (NLP) is needed to recognize emotions in Bangla. This study addresses three emotions: Happiness, Sadness, and Anger, which is a limitation [14]. Fear, surprise, and revulsion are ignored. The study also uses just the CNN-BiLSTM model for classification. Alternative models may improve performance.

The study at hand [18] is limited by certain conditions. Hinglish Emoji Prediction (HEP), a Twitter-based dataset, was used to evaluate the suggested model. Due to the model's performance on previous datasets, the dataset may not represent all Hinglish text. However, the recommended method ignores emoji order, which may affect message interpretation.

In addition, the work does not examine the model's mistakes, which might help find areas for improvement. We propose a machine learning method combining Neural Networks and sophisticated NLP methods to recognize emotions in the Bangla language, addressing these challenges. Despite this, this project faces several major obstacles. Our primary goal is to collect Bangla texts and emoticons. We will carefully clean and prepare the data to remove unnecessary information and maintain key language components for the study. Feature engineering, where we develop new methods for discovering features that capture Bangla text's nuanced linguistic qualities, follows. Building a solid basis for the prediction model.

Some limitations of this model include:

- There is a noticeable lack of extensive and publicly available datasets for the Bangla language that cover a wide range of emotions. Creating such a dataset from scratch is crucial but also demanding, requiring meticulous data collection, annotation, and preprocessing.
- The rich morphological structure, diverse syntax, and unique script of Bangla present significant challenges in processing text and detecting emotions. Specialized NLP techniques are necessary to accurately interpret and analyze the nuances of the language.
- Understanding emotions in Bangla texts is closely connected to cultural and contextual factors. Emotions are often conveyed subtly and can vary significantly in different contexts, making their detection and interpretation a complex task.

- It is crucial to employ and adapt advanced machine learning and deep learning methods to suit the specific characteristics of the Bangla language. This not only involves selecting appropriate algorithms but also customizing them to handle the unique features of the language.
- Establishing effective evaluation metrics for emotion detection and conducting thorough hyperparameter tuning is essential to optimize the performance of the model and ensure its accuracy and reliability.
- Conducting a detailed analysis of the data, including error analysis and data visualization, is critical for understanding the characteristics of the data and addressing any potential issues in the model's predictions.

This prompts the inquiry - *How can we effectively develop an Emotion Detection system for the Bangla language that accurately identifies a range of emotions from text and facilitates the manual generation of corresponding emojis, thereby enhancing the richness and expressiveness of digital communication in Bangla?*



## 1.2 Our Contributions

The previous research works have significantly improved about the Emotion detection from the Bangla text using different NLP and Machine learning approaches. Still, there are some important shortcomings in the process of Emotion detection for Bangla Language in the form of Emojis. The core of our research lies in creating a system that can accurately detect a range of emotions from Bangla text and facilitate the manual generation of corresponding emojis. We make the following contributions to this study based on our research:

- The first goal is to gather a diverse dataset of Bangla text with emojis from various sources. The dataset will undergo preprocessing to handle noise, eliminate irrelevant information, and address specific text challenges, such as language identification and transliteration.
- Developed the emotion detection system with a user-centric approach, ensuring that it meets the practical needs of Bangla-speaking communities for emotional expression in digital platforms.
- Extended the boundaries of sentiment analysis by categorizing sentiments and associating them with the appropriate emojis, adding a layer of emotional intelligence to digital interactions.
- Addressed computational efficiency by optimizing models to perform with lower computational costs, making them accessible for implementation in environments with limited resources.
- Conducted a deep linguistic analysis of the Bangla language, identifying key emotional indicators within the lexicon critical for accurate emotion detection.
- Addressed the challenge of cultural sensitivity in NLP applications by adapting models to the unique linguistic features of Bangla.

## 1.3 Research Objectives

This research aims to develop a sophisticated Emotion Detection system for the Bangla language, which will subsequently enable the manual generation of emojis based on the detected emotions in text. To achieve this, several specific objectives have been outlined, leveraging Natural Language Processing (NLP) techniques:

- The primary objective is to collect a varied collection of Bangla text including emotional content from reliable sources. The dataset will be subjected to preprocessing techniques to mitigate noise, remove extraneous information, and tackle specific difficulties.
- Conducting a detailed linguistic analysis of Bangla texts to understand how various emotions are expressed. This involves identifying key linguistic features and patterns that are indicative of different emotions, thus providing a foundational understanding for accurate emotion detection.
- Conducting research and applying sophisticated NLP methodologies, such as sequence models and transformer models, specifically designed to capture the intricacies of the Bangla language. The purpose of these models is to precisely detect emotions from textual input.
- The models will undergo thorough training and evaluation utilising reliable metrics and cross-validation approaches. The purpose of this step is to verify the dependability and efficacy of the models in detecting emotions. We will conduct comparative assessments with existing models in the field to assess and compare performance.
- Performing a comprehensive analysis of the linguistic features contributing to the accuracy of emotion detection in Bangla text. This analysis will help in understanding the relationship between text characteristics and emotional expression.
- Undertaking a detailed error analysis to identify common errors and limitations in the model. Insights from this analysis will be used to refine and enhance the model's performance.

This project aims to make important advancements in the field of Natural Language Processing (NLP) by providing novel techniques and models for identifying emotions in Bangla text. Additionally, it aims to enable the manual production of related emojis. The ramifications of this research are far-reaching, with possible implementations in sentiment analysis, heightened comprehension of Bangla language communication, and the enhancement of digital communication within Bangla-speaking groups.

# Chapter 2

## Literature Review

The exploration of emotion recognition using Emojis for the Bangla language is a recent endeavor, as much progress in emoji detection and sentiment analysis has mostly focused on the English language. The contemporary scholarly literature has focused extensively on the analysis of sentiments and the classification of sentiment polarity. This literature review seeks to offer a thorough and detailed examination of the current cutting-edge advancements in this field, with a specific emphasis on commonsense and knowledge-based conceptual and effective sentiment analysis.

The research work of [22] expanded on the profound impact that the World Wide Web has had on a variety of online platforms. It underscores the significance of textual content and emojis in facilitating online communication, along with a research article that posits an approach for sentiment polarity computing that amalgamates both text and emojis. The research objectives encompass formulating a framework for sentiment polarity computing and proposing a cognitive method for identifying sentence-level polarity by utilizing intricate pattern rules. The article endeavors to equip readers with an appreciation of pattern-based structures in online natural language sentiment polarity detection.

Emoji Detection’s significance in AI and NLP [8] The study underscores its potential to enhance human-computer interaction and its relevance for applications such as product reviews and market analysis. This research paper centers its attention on detecting emotions in Bangla text by utilizing the Multinomial Naïve Bayes classifier and various features. It accentuates the significance of scrutinizing human emotions that are conveyed through social media and blogs. The study has obtained an accuracy of 78.6 percent in categorizing Bangla text into different emotions. Therefore, this research significantly contributes to comprehending Emoji Detection in Bangla and its implications in market analysis and predicting public reactions.

According to this research [14] ”Emotion Detection in Hinglish/Hindi-English Code-Mixed Social Media Text” by Goyal and Singh (2020) presents a deep learning approach for detecting emotions in Hindi-English code-mixed social media texts. The authors explore the challenges of detecting emotions in code-mixed languages and propose a methodology that uses a pre-trained bilingual model and a combination of a character-level convolutional neural network (CNN) and a bidirectional long short-term memory (BiLSTM) network. Furthermore, the authors comprehensively

review related research on emotion detection in code-mixed languages. They discuss the various approaches employed, including lexicon-based, rule-based, and machine learning-based approaches. Based on the related research, the author proposes to use CNN, LSTM [14], BiLSTM [15], CNN-LSTM, and CNN-BiLSTM as classification models. Additionally, they highlight the limitations of existing work and the unavailability of code-mixed datasets. Finally, the authors present the results of their experiments, showing that the CNN-BiLSTM model achieved an accuracy of 83.21 percent in detecting emotions in Hindi-English code-mixed texts.

This research [18] focuses on Hinglish, a code-mixed language that combines Hindi and English, which is the fourth most used code-mixed language globally. The authors propose a new dataset called Hinglish Emoji Prediction (HEP) that was created using Twitter as a corpus. The author also proposes a hybrid emoji prediction model called BiLSTM attention random forest (BARF) that combines deep learning and machine learning algorithms with attention mechanisms to predict emojis in Hinglish. The author proposes creating a new Hinglish Emoji Prediction (HEP) dataset for the Hinglish emoji prediction task. The dataset comprises plain text as input and emojis as labels, and it was created by retrieving tweets containing emojis using the Twitter API. The authors used the emojis extracted from tweets as labels since manual annotation can introduce bias. The proposed model outperformed previous multilingual and baseline emoji prediction models, achieving an accuracy of 61.14 percent, precision of 0.66, recall of 0.59, and F1 score of 0.59 on the Hinglish Emoji Prediction (HEP) dataset created using Twitter as a corpus.

The author of this research [12] discusses the task of emoji prediction, which involves predicting the proper set of emojis associated with a piece of text, particularly in social media posts. They also provide a literature review of existing research on the topic, highlighting the limitations of prior work and the need for more extensive emojis and multi-label classification. To effectively tackle these concerns, the author focuses on the cleansing and labeling of emoji prediction datasets derived from Twitter posts. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for emoji prediction have achieved good performance in prior work. Still, they are limited in their ability to capture long-range dependencies and lack interpretability as per the analysis conducted by the author. To overcome this challenge, the authors propose using transformer networks, which are more appropriate for capturing long-range dependencies and possess better interpretability. They have made further adjustments to an already pre-existing BERT model by utilizing their own data set, as the BERT model performs well on emoji prediction. Prior research about the prediction of emojis has predominantly centered on datasets sourced from social media postings.

The present study [17] introduces an innovative method for predicting emojis through the utilization of a conversational dataset. The present study suggests a method for providing customized emoji suggestions based on temporal and spatial variables. The study introduces a novel annotated conversational dataset and examines the influence of temporal and spatial factors on the prediction of emojis. This paper introduces a new approach to forecast emojis by utilizing a hybrid algorithm that integrates a deep learning model incorporating text, time, and location, as well as

two distinct algorithms that utilize NLP-based techniques to associate emojis with input text. The methodology employed in this study involves a hybrid model that utilizes both neural networks and score-based metrics, specifically semantic and cosine similarity. The utilization of BERT has resulted in a notable enhancement in the precision of emoji prediction, with an increase of up to 73.32 percent. The findings of the qualitative analysis indicate that the utilization of the three channels for emoji predictions yielded outcomes that were deemed acceptable. The selection of three emojis for a given text message, location, and time input is determined through the utilization of various metrics, including the semantic similarity score, cosine similarity score, and deep learning model.

The article [5] proposes a methodology for forecasting Instagram emojis by integrating textual and visual data. The findings of the study demonstrate that textual and visual elements possess distinct yet complementary characteristics of emojis, thereby enhancing the precision of emoji prediction. The article proposes a methodology for forecasting Instagram emojis by utilizing both textual and visual data. A Bidirectional Long Short-Term Memory (LSTM) model with character representation was utilized to address orthographic variations in social media. The model in question exhibited superior performance compared to the baseline model proposed by Barberi et al. (2017). The researchers additionally noted that the textual features most frequently predicting the US flag and the emoji with the highest F1 score are prevalent. The researchers employed Instagram images that were geotagged in the United States between July and October of 2016. The images were required to have a user description consisting of at least four words and one emoji. The evaluation process only considered posts that contained one of the top 20 most frequently used emojis. The dataset in question consists of a total of 299,809 entries, each of which includes visual content in the form of pictures, textual information, and a singular instance of an emoji. The researchers additionally analyzed the most frequently used emojis, specifically the top 10 (238,646 posts) and top 5 (184,044 posts).

The research [2] elucidates how cognitive processes and intrapersonal experiences impact affective states. Text emotion recognition improves human-computer interaction and decision-making. The study suggests using keyword analysis, keyword negation analysis, proverbs, emoticons, short words, and exclamatory phrases to identify emotions in text and emoticons. The proposed methodology employs a set of 25 distinct emotion categories to ascertain affective states with a precision rate of 80 percent. This research employs various linguistic tools such as keyword analysis (KA), keyword negation analysis (KNA), proverbs, emoticons, short words, and exclamatory terms. A total of 25 distinct categories of emotions were produced. The study proposed utilizing various techniques such as keyword analysis, keyword negation analysis, proverbs, emoticons, short words, and exclamatory phrases to ascertain the emotional content of text and emoticons. The proposed methodology employs a set of 25 distinct emotion categories to ascertain effective states with a precision rate of 80 percent. The article asserts that the detection of emotions at the level of individual sentences and through the use of emoticons is of paramount importance for facilitating effective interaction between humans and computers.

The present article [15] scrutinizes the correlation between lexemes and emoticons

and prognosticates the specific emojis that are likely to be elicited by text-based tweets. The researchers employed Multinomial Naive Bayes and Long Short-Term Memory (LSTM) models to forecast emojis in tweets. The present systems for predicting emojis rely on keywords and tags, whereas the authors aim to investigate the semantics of emojis through the utilization of Natural Language Processing. The article employs word embedding and machine learning techniques, such as Multinomial Naive Bayes, SVM, and Bi-LSTM, to forecast tweet emojis. The researchers conducted training of Multinomial Naive Bayes and LSTM models to predict emojis in tweets. The article utilizes the tweet status ID and emoji annotation provided by Twemoji, which offers a vast selection of over 13 million options. The authors utilized the Twitter API to retrieve tweets associated with the provided tweet ID to consolidate their sentence-emoji ID pairings. The present article scrutinizes the correlation between lexical units and emoticons and prognosticates the specific emojis that are likely to be elicited by textual tweets. The model can predict emojis used in tweets. The results of the trials above indicate that the multinomial naive Bayes approach exhibited superior performance compared to all other strategies. The potential superiority of deep learning over conventional methods may be limited due to insufficient generalizability of the data.

The academic paper [6] discusses the development of a multilingual sentiment analysis and emoji prediction corpus in Hindi, Bengali, and Telugu by utilizing Twitter data. The author explores the importance of Twitter as a valuable source of data for various natural language processing (NLP) studies, such as sentiment analysis, polarity detection, and emoji prediction. According to the author, most of the existing research in this area focuses on English tweets, as English is the dominant language on Twitter. However, the authors emphasize the importance of resource-poor languages that are widely spoken but often ignored. Moreover, the author uses Twitter Application Programming Interface (API) to build corpora for multilingual sentiment analysis and emoji prediction. They analyzed 500 tweets for each language, manually annotated them with emojis, and categorized them into three sentiment classes - Positive, Negative, and Neutral - to explore the relation between emojis and the annotated sentiments.

The usage of emojis in social media [9] has experienced rapid growth, and as a result, algorithms and models capable of efficiently managing them in online communication have become necessary. Forecasting emojis based on text can positively impact various NLP tasks. Previous studies have shown the potential of emoji prediction, with Barbieri et al. (2017) achieving better performance than humans by utilizing a neural network model on Twitter data. These results have stimulated subsequent inquiry into expanding the emoji forecasting task to additional languages, such as French, German, Chinese, Arabic, Russian, and now Swahili. The selection of social media platforms for data gathering is crucial, and in this study, data was collected from Facebook, considering its popularity among Hebrew speakers in Israel. The analysis centers around two aims: emoji recognition (determining if a text contains emojis) and emoji forecast (categorizing emojis in a text). A supervised ML approach is utilized, comparing n-grams and character n-grams as text representation techniques. The impact of additional metadata features on classification is also analyzed. The results indicate that metadata features significantly improve emoji

identification accuracy. However, for emoji prediction, feature selection is more advantageous, and the best-performing character n-grams representation outperforms the FastText baseline.

Irony serves as a linguistic tool [10] for conveying ideas through the use of expressions that contradict their intended meaning. Its implementation poses an intricate challenge for text analytic algorithms, particularly in their ability to extract emotions and analyze sentiments, as these algorithms often struggle with accurately interpreting ironic statements. Due to language attributes and cultural reasons, Persian language speakers tend to utilize irony more frequently. The paper offers a thorough analysis of the ransomware attacks that took place in 2016. The authors have conducted a review and cited several pertinent studies and reports related to cyber-attacks and ransomware. Furthermore, the authors have provided an extensive analysis of the various techniques implemented by attackers to initiate ransomware attacks and the resulting impact on organizations. The paper emphasizes the significance of possessing a robust backup system and the necessity for organizations to adopt proactive measures to prevent such attacks. This paper provides a comprehensive understanding of the current state of research in the field of ransomware attacks.

The paper [13] discusses the annotation of an emoji prediction dataset. The annotations include passage-level, multi-class/multi-label annotations and aspect-level multiclass annotations. In addition to that, the authors propose an innovative annotation approach that they use to produce aspect-level annotations. Heuristically produced annotations are created by utilizing the self-attention mechanism that is present in Transformer networks. The validation process is performed automatically and manually to guarantee the annotations' high level of excellence. The dataset is evaluated by applying a BERT model that has already been pre-trained. The authors base their study on a dataset from Twitter that includes balanced emoji usage and more objective points of view. The following are some of the practical consequences of this paper: (1) The annotated dataset has the potential to be utilized in the training and evaluation of algorithms for emoji prediction tasks. (2) The aspect-level annotations may be used to assist in locating the portion of the text that corresponds to each emoji, which can be beneficial when attempting to summarize the text correctly. (3) The approach of annotating that takes advantage of the self-attention mechanism in Transformer networks has the potential also to be utilized in other activities.

This article [11] proposed a discussion on the significance of comparing and sharing different corpora for research. The amount of work that researchers have to put in is cut down by corpora that are easily accessible to the public and that can be compared to other corpora. However, Twitter has various terms and services that prohibit the sharing of aggregated tweets. As a result, some researchers' Twitter corpora have been destroyed as a result of these restrictions. This work offers a methodology for extracting mixed-language Hindi and English data from social media sites such as Facebook and Twitter. In this work, we outline the approach that was utilized to acquire the data and carry out the different NLP jobs on the corpus. It is suggested in this research that the technique utilized in this study may also be applied to other resource-deficient languages, given that people who are

qualified to do so are active on social media sites. The article draws attention to the significance of languages with limited resources, such as Hindi, and the dearth of substantial corpus for such languages. Furthermore, some of the practical ramifications of this study include providing a system to extract mixed-language data in Hindi and English from social media platforms such as Facebook and Twitter, which is one of the implications. This can be helpful for various tasks involving natural language processing, including identifying languages, sentiment analysis, and machine translation. The approach used in this study can be extended to other resource-deficient languages as well, given that relevant users are active on social media platforms, as the paper underlines the significance of resource-poor languages like Hindi and recommends that the methodology used in this paper can be applied to other resource-deficient languages as well.

The purpose of this work [4] is to get a deeper understanding of the intricacies that lie behind emoji prediction by proposing a label-wise attention mechanism. The suggested architecture outperforms the traditional baselines in emoji prediction, and it is especially effective in predicting rare emojis. The authors also compute the Classification Error (CE), which is the average number of labels that need to be in the predictions for all true labels to be predicted. This is among the 10 percent of the sample that contains the most rare emojis. The report presents a comprehensive analysis of the suggested architecture and the tests carried out to evaluate the effectiveness of the design. The research presented in this study offers a first step in improving the resilience of recurrent neural networks in datasets with uneven distributions. The label-wise attention mechanism may be utilized to gain an understanding of other fascinating linguistic aspects of human-generated material seen in social media as well as other multi-class or multilabel classification issues.



The work that was done by prior competitors of the SemEval competition is mentioned in this article [7]. The SVM classifier was the key to success for C’oltekin and Rama, who emerged victorious in the competition. To extract the bag-of-n-gram properties, they considered both the character and word levels. Baziotis et al. employed a Bidirectional Long-Term Memory (Bi-LSTM) with attention and pre-trained word2vec vectors, achieving the maximum recall possible. This work contributes to constructing a machine-learning model that, given the content of a tweet, can predict the emoji that is most likely linked with it. The purpose of this article is to investigate the relationship between the text that is written by the user and the emojis that are associated with it. This will be accomplished by implementing a tool that seeks to emulate the logic according to which a particular emoji is evoked by the written message.

This will allow the paper to analyze this relationship. This study presents numerous different models based on Neural Networks, such as Convolutional Neural Networks and Bidirectional Long Short-Term Memory Recurrent Neural Networks. The authors discovered that the latter method was the most successful since it exceeded all of their models and placed sixth out of a total of 47 participants in the study. In this study, the promise of simpler models is demonstrated. With some fine-tuning of their hyper-parameters, these models might attain an accuracy equivalent to that of the more sophisticated models of the challenge. Also, we can get a better level of empathy by using this tool to enhance textual language with graphical content that is supplementary to the text. The approach has several potential uses, including monitoring social media, marketing, and analysis of sentiment from online sources. The research also demonstrates the possibility of simpler models, which, with some fine-tuning of their hyper-parameters, attain an accuracy equivalent to that of the more complicated models being used in the challenge.

## 2.1 Comparative Analysis Of Related Work

The table below compares the model, method, dataset used, and the accuracy of 5 papers where Emotion detection is done.

Table 2.1: Comparative Study of Models

Ref No	Method	Architecture/ Model	Dataset	Accuracy
[12]	Emoji Extension and benching(ML approach)	Bert	Celebrity Profiling Corpus	52.65%
[9]	Identification (ML)	n-grams	Political and Facebook API	38.5%
[14]	Emotion Detection (ML,NLP)	LSTM	Social Media API	80%
[8]	Emotion detection from Bangla text	Naive Bayes Classifier	Social Media ,Blogs	78.6%
[22]	Sentiment Polarity (ML,NLP)	SVM	Twitter API	82.8%

After using each model, it contains the comparative study of methods, architecture, data sets, and accuracy.

## 2.2 Visual representation of papers

For better understanding, we visualized and compared the papers we studied. We studied a few papers, and below are their architectural flowcharts.

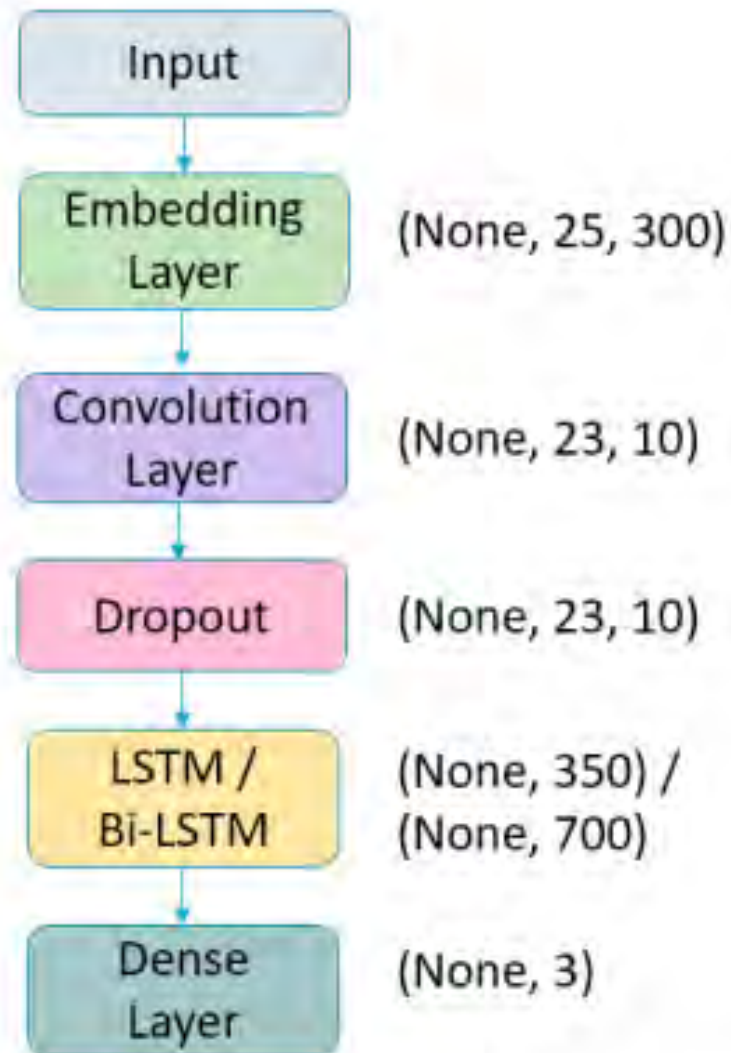


Fig. 2. Overview of Neural Network Model

Figure 2.1: Architecture from the study [[10]]

## 4.2. Multinomial Naive Bayes Classifier

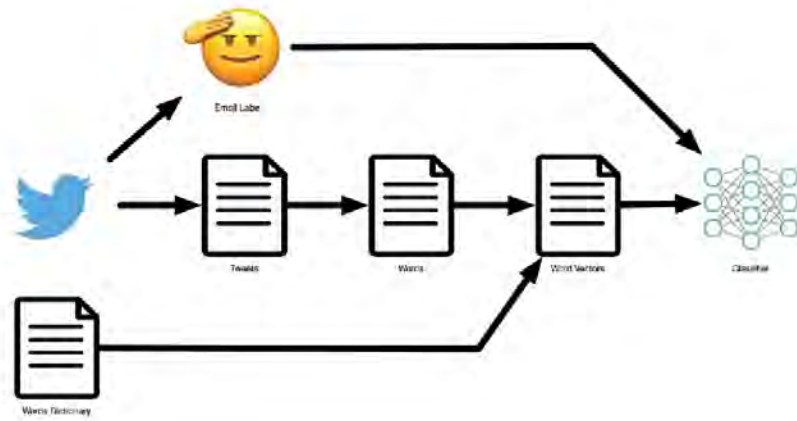
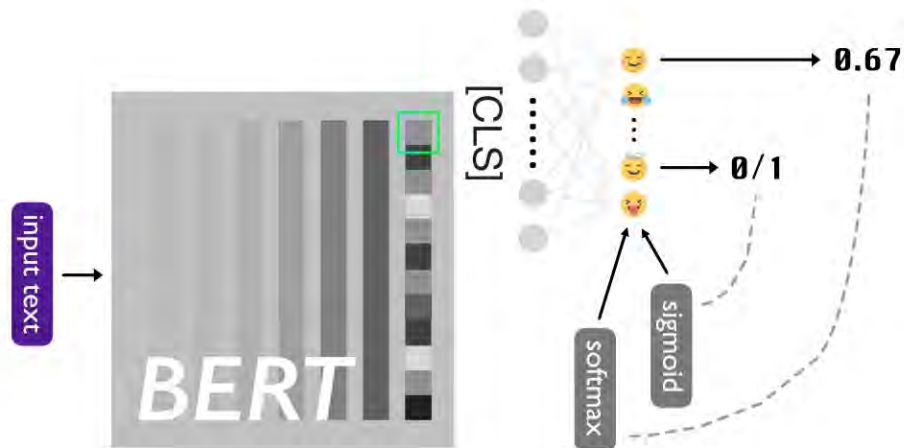


Figure 2.2: Architecture from the study [[17]]



**Figure 1: The BERT model architecture. The multi-class classification model applies the softmax activation function on top of the dense layer while the multi-label classification uses the sigmoid activation function.**

Figure 2.3: Architecture from the study [[12]]

# Chapter 3

## Description of Model, Data and Preliminary Analysis

### 3.1 Description of Model

This study presents a novel method for detecting emotions in Bangla language texts, to facilitate the manual creation of emojis that correspond to the discovered emotions. The system employs a diverse range of sophisticated machine learning classifiers, such as RandomForestClassifier, SVM, Ensemble, BERT-based deep learning, Naive Bayes, and SGD Classifier, to precisely analyse a broad spectrum of emotions from Bangla text. This model incorporates customized preprocessing and feature extraction designed specifically for the Bangla language, in addition to advanced NLP techniques such as sentiment analysis and contextual embeddings. The model undergoes rigorous training and validation, as well as a manual process of mapping emotions to emojis, to ensure that it effectively captures the subtle emotional aspects of the text while taking into account cultural and contextual relevance. The system undergoes continuous error analysis and optimization, resulting in a refined state. This research significantly contributes to the improvement of digital communication and emotional expression within the Bangla-speaking community.

#### 3.1.1 Random Forest Classifier

The Random Forest Classifier is a versatile machine-learning model employed for classification applications. The system exhibits exceptional accuracy and is capable of efficiently handling extensive datasets containing various dimensions. During the training phase, it operates by creating a "forest" consisting of many decision trees. In this discussion, we will explore the functionality and utilization of the emoji prediction model.

#### Working Process

The random forest classifier initiates the process by extracting features. In this process, the Bangla text input is converted into a structured format that is comprehensible to machine learning algorithms. This is achieved by employing NLP techniques such as tokenization, which divides the text into discrete words or phrases, and vectorization, which transforms these words into numerical representations. The

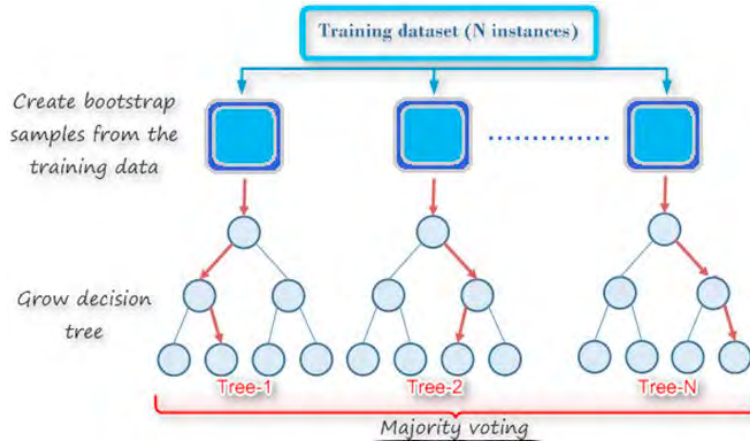


Figure 3.1: Random Forest [3]

retrieved elements can encompass individual words, phrases, or even grammatical patterns that aid in identifying the appropriate emoji to employ. During the training phase, the method produces numerous decision trees, each of which functions on a unique subset of randomly chosen attributes. These trees function autonomously to discern complex patterns in the usage of emojis by Bangla speakers. Once all trees have provided their forecasts, the ultimate projected emoji is determined through a majority vote. This ensemble technique enhances the precision of the model while mitigating the risk of excessive accuracy.

### 3.1.2 BERT-based Deep Learning

The innovative application of the Bidirectional Encoder Representations from Transformers (BERT) model is explored for emotion detection in the Bangla language, a pivotal contribution to the domain of Natural Language Processing (NLP). This research focuses on leveraging the advanced linguistic comprehension capabilities of a pre-trained Bangla BERT model, meticulously fine-tuning it to the specialized task of identifying emotional nuances in textual data.

#### Working Process

The utilization of BERT commences with a fundamental yet crucial tokenization process using 'BertTokenizer.' This initial phase involves transforming the Bangla text into a structured format compatible with the BERT model, encompassing tokenization and the generation of attention masks. These masks are instrumental in directing the model's concentration toward pertinent text segments. Subsequently, 'BertForSequenceClassification,' initially pre-trained on an extensive corpus of Bangla text, undergoes a precise fine-tuning process aimed at emotion detection. A key aspect of this fine-tuning involves hyperparameter optimization, a critical step for tailoring the model to this specific task. Integral hyperparameters such as the learning rate, batch size, and training epochs are strategically selected and fine-tuned.

The learning rate, typically set around  $2e-5$ , balances the optimization speed and

effectiveness. Concurrently, the batch size and number of epochs are determined by considering the dataset’s intricacies and size, ensuring a harmonious balance between computational expediency and exhaustive training. Employing the AdamW optimizer, renowned for its synergy with BERT models, the fine-tuning process is optimized, ensuring the model is aptly aligned with the subtleties of emotion detection in Bangla text. During the training phase, the model’s performance undergoes rigorous evaluation, with metrics such as loss and accuracy providing insights into its learning efficacy. This continuous assessment is pivotal in ascertaining that the model adeptly learns and categorizes the diverse emotional expressions inherent in the text. In the final evaluation phase, the model is subjected to a comprehensive analysis of a test dataset. This evaluation is critical in scrutinizing the model’s ability to accurately discern and classify a range of emotions within Bangla text, highlighting its proficiency and adaptability in this specialized NLP task.

### 3.1.3 Linear SVC(SVM)

A linear vector classifier, often known as SVC, is a type of support vector machine (SVM) that is specifically designed for classification tasks. The operation involves identifying a hyperplane in the feature space that categorizes the data points into separate classes while maximizing the distance between the classes.

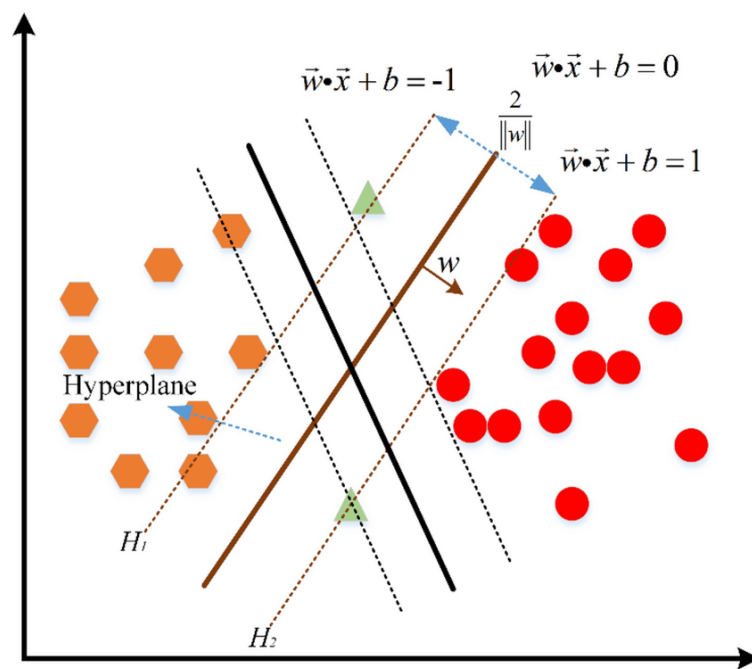


Figure 3.2: SVC  
[20]

#### Working Process

The linear support vector classifier (SVC) classifies data points into separate classes by finding a hyperplane in the feature space that maximizes the distance between

each class. At first, the Bangla language data is treated beforehand to eliminate irrelevant information and identify crucial characteristics using methods like tokenization and vectorization. During the training phase, the data is inputted into the linear Support Vector Classifier (SVC) model. The model then learns to identify an optimal hyperplane that maximizes the distance between different classes, resulting in improved accuracy in classification.

### **3.1.4 Naive Bayes Classifier**

A Statistical classifier that assumes strong independence between features and uses Bayes' theorem for probabilistic classification. Its ability to effectively handle data with a large number of dimensions makes it well-suited for jobs involving the classification of text.

#### **Working Process**

The multinomial naive Bayes classifier is a probabilistic model that utilizes Bayes' theorem and assumes feature independence. Given its ability to effectively handle data with many dimensions, this classifier is well-suited for problems involving text classification. The technique commences by preprocessing the Bangla language dataset to remove any unwanted disturbances and identify essential characteristics. Feature extraction is converting text input into a numerical form suitable for machine learning techniques. During the training phase, the model calculates the probability of each class (emojis) occurring and the odds of each feature (words or phrases) being associated with each class. These probabilities are estimated based on the observed frequencies in the training data.

### **3.1.5 Enhanced Ensemble Method: Integrating Random-Forest and SVM**

RandomForest and Support Vector Machine (SVM) is a composite classifier that amalgamates the robust decision-making ability of RandomForest with the dimensional dexterity of SVM. This ensemble model excels in text classification within the NLP realm, with RandomForest adeptly reducing overfitting and deciphering intricate data patterns. SVM efficiently navigates through high-dimensional textual feature spaces.

#### **Working Process**

The ensemble model begins with the crucial phase of TF-IDF vectorization, where textual data transforms into a numerically interpretable format while retaining essential linguistic features. This step is fundamental for preparing the data for machine learning algorithms. In the ensemble architecture, the RandomForest classifier, comprising numerous decision trees, specializes in identifying non-linear relationships within the data, thereby enhancing the model's generalizability and reducing the risk of overfitting. Simultaneously, employing a linear kernel, the SVM adeptly



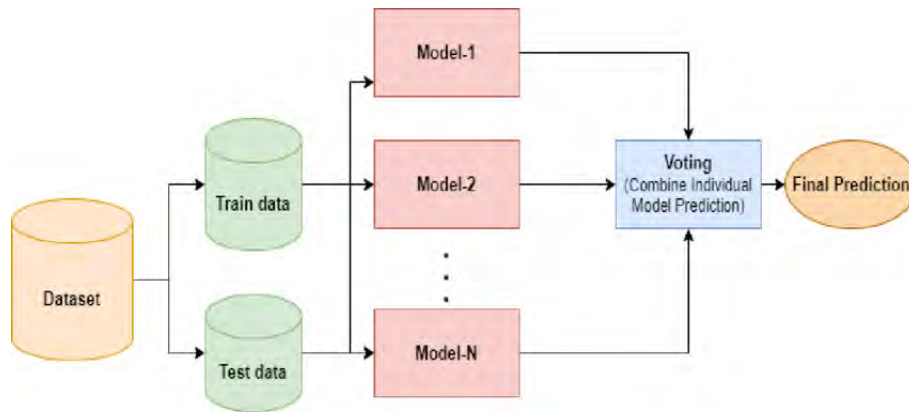


Figure 3.3: Ensemble Model  
[19]

manages the high-dimensional spaces typical of text data, which often include a vast array of unique words and phrases.

The integration of RandomForest and SVM is executed via sci-kit-learn's VotingClassifier using a 'soft' voting scheme. This approach combines the probabilistic outputs of both classifiers, ensuring that the final decision reflects a nuanced understanding based on the combined strengths of both models. During the training phase, the ensemble model capitalizes on the complementary capabilities of RandomForest and SVM. This synergistic integration allows the ensemble to capture complex patterns through RandomForest's analysis and efficiently navigate the dimensional complexities of the text data via SVM. By uniting the distinct advantages of RandomForest and SVM, this ensemble model provides a robust and comprehensive solution to the challenges of text classification in NLP. It stands as a compelling example of the effectiveness of combining diverse machine learning techniques, enhancing the overall performance and reliability in handling intricate, high-dimensional data scenarios without delving into the specifics of the prediction process.

### **3.1.6 SGD Classifier:**

The stochastic Gradient Descent Classifier is employed for handling extensive and sparsely populated datasets. The Scikit-Learn library includes an implementation of the SGD Classifier.

#### **Working Process**

The SGD Classifier utilizes stochastic gradient descent, which is an exceptionally efficient and direct optimization approach. Stochastic gradient descent optimizes the model's parameters (weights) by iteratively updating them using just one sample at a time, making it very suitable for handling big datasets. It is frequently employed for linear classification problems when the objective is to segregate data points using a hyperplane. This classifier is capable of effectively addressing both binary and multi-class classification issues. Because of its iterative structure and dependence on a single data point at each stage, this method is highly efficient for datasets that exceed the memory capacity, as it does not require loading the entire dataset simultaneously.

## **3.2 Model Implementation**

A function is being developed to streamline the process of training and evaluating machine learning models. This function integrates data preprocessing, model training, and assessment seamlessly. Following the training phase, the models are assessed by their ability to predict the emoticons from the validation set accurately. Furthermore, GridSearchCV is employed to optimize the parameter tuning of the LinearSVC model to achieve the highest level of accuracy. This technique determines the parameters that enhance the forecasting capabilities of the model.

### 3.3 Dataset Description

In this work, the curated datasets Dataset-1 (Test) and Dataset-2 (Train) play an essential role in furthering the study of emotion identification in the form of emoji inside the Bangla language utilizing NLP and machine learning approaches. This dataset provides a solid foundation for model building and represents a rich language environment, capturing a wide range of emotional expressions and circumstances. These datasets demonstrate the study’s commitment to investigating the intricacies of emotion and emoji usage in the Bangla language, which is notably underrepresented in NLP research. Their extensive breadth and particular focus on Bangla text make them ideal for investigating the complex link between written language, sentiment, and emoji use, opening the way for more nuanced and culturally sensitive NLP applications.

After data preprocessing, the model training phase is conducted to ensure it is suitable for machine learning techniques. This modification is based on the TF-IDF (Term Frequency-Inverse Document Frequency) technique. This method turns textual input into numerical values to identify patterns and relationships between words and emojis. The dataset is separated into training and validation sets to ensure a thorough model examination.

Train vs Test Dataset Distribution

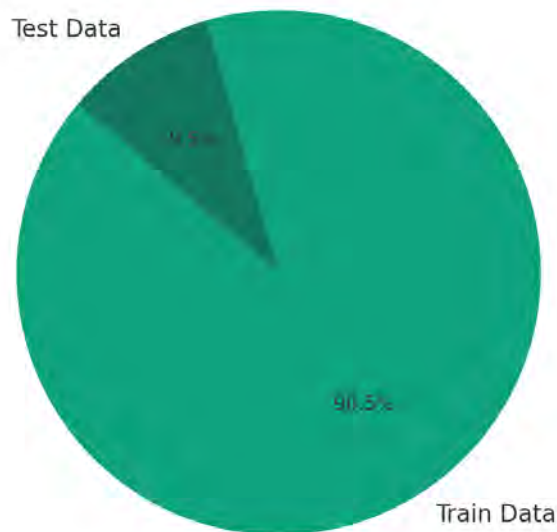


Figure 3.4: Dataset

[21]

- **Dataset-1(Test):** This dataset is meticulously compiled for testing purposes in the domain of Emotion detection in the form of emojis using Natural Language Processing (NLP) and Machine Learning (ML) techniques, specifically focusing on the Bangla language.

**Dataset Content:** The dataset contains various types of information.

- **Text:** Contains the textual content in Bangla, offering a diverse range of subjects and emotional expressions.
- **Sentiment:** Provides sentiment labels associated with each text.
- **Emoji:** Includes emojis related to the text.
- **Emoji Encoded:** Numeric encoding of the emojis for computational convenience.

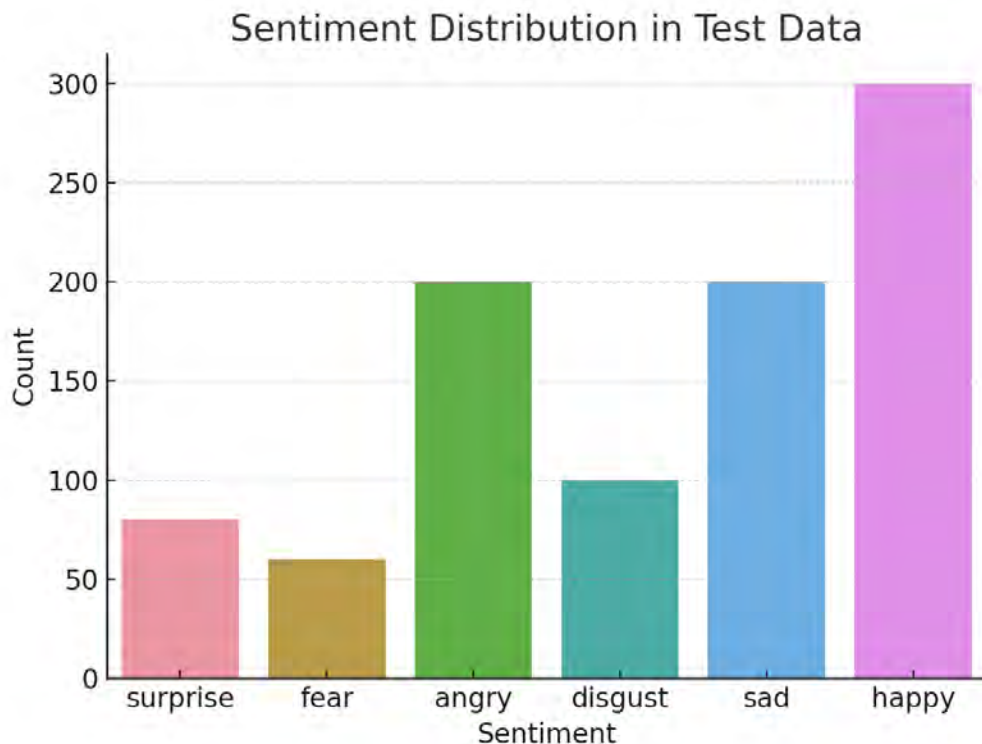


Figure 3.5: DatasetTest

**Objective:** The major goal is to foster testing and experimenting in the advancement of emotion detection using emoji systems. This dataset is critical for assessing the efficacy of various NLP and ML approaches in predicting emojis that match the context and emotion of Bangla texts.

**Dataset Usage:** Training and evaluating ML models for Emoji Prediction.

- Analyzing the interplay between text and emoji usage.
- Assessing the performance of various algorithms in emoji prediction.
- Enhancing understanding of emoji usage in diverse contexts, especially in Bangla language social media texts.

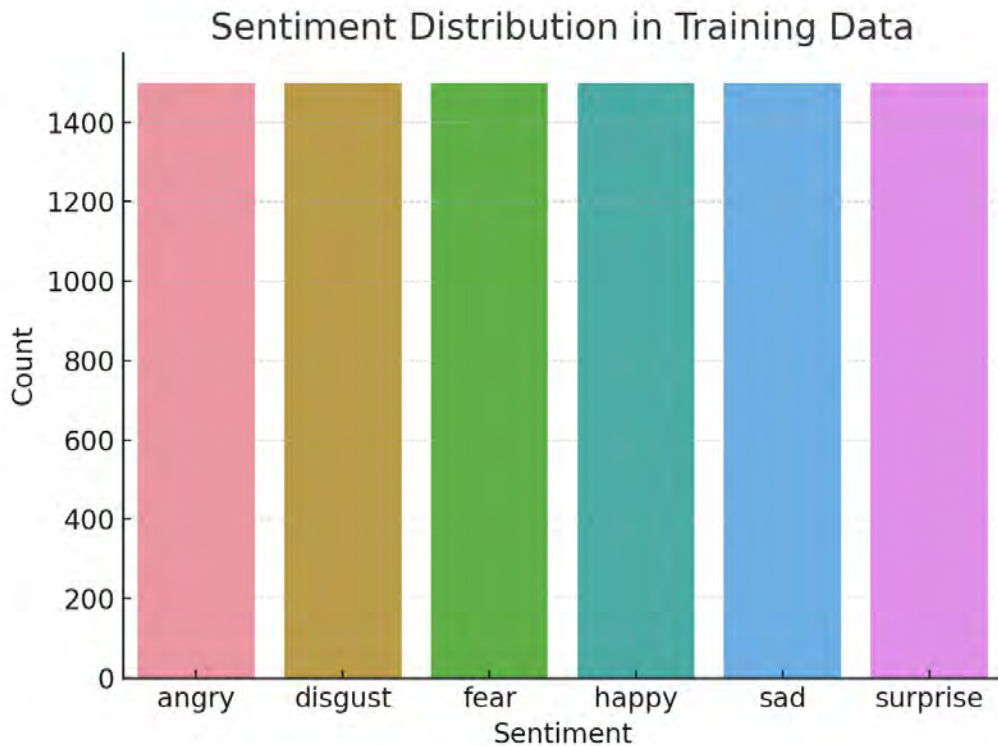


Figure 3.6: DatasetTain

- **Dataset-2(Train):** This dataset was chosen and arranged to train and evaluate machine learning models for Emotion detection in the form of emoji in Bangla-language texts. This dataset includes a diverse variety of tweets, each with sentiment classifications and text content.

### **Dataset Details:**

- This has been collected from a variety of sources, including social media sites, to provide a diverse collection of Bengali tweets.
- **Size:** Contains 9000 items, each with a unique identifier, sentiment label, and text content.

**Dataset Columns:** The dataset is structured with the following columns:

- **Text:** The actual text content of the tweet in Bangla.
- **Sentiment:** A label that indicates the sentiment communicated in the tweet.
- **Emoji:** The actual emojis are used in the tweets.
- **Emoji Encoded:** Numeric representation of the emojis.

### **Usage:**

- Training and evaluation of ML models for emoji prediction.
- Developing NLP algorithms for context and sentiment understanding.
- Investigating the relationship between text and emoji used in Bangla tweets.
- Examining ML techniques for automating emoji detection.

These datasets provide a unique chance to improve our understanding of emoji usage to identify emotion in Bangla language tweets and construct predictive models that can forecast emojis based on textual context and mood.

### 3.4 Preliminary Analysis

The datasets have been pre-processed properly to implement in the research work. Multiple phases of Data pre-processing techniques are manipulated to clear and organize the datasets. At the very first stage, the steps are implemented:

Compute basic statistics, including the number of rows and columns, data types, and missing values.

#### Data Preprocessing:

- The test dataset comprises 940 entries, and the training dataset contains 9000 entries, each with four columns: text, sentiment, emoji, and emoji encoded. There are no missing values in either dataset.
- Duplicate Handling: Duplicate entries were thoroughly examined to verify the data's integrity and uniqueness.

Then, gradually, we implemented the following methods:

#### Text Cleaning and Tokenization

- The texts in both datasets were cleaned and tokenized. This includes the elimination of unnecessary characters, normalization to lowercase (where applicable), and preparation for linguistic analysis.

#### Sentiment and Emoji Analysis:

- Sentiment Label Assessment: Each text input is paired with a sentiment label, providing a framework for sentiment analysis.
- Emoji Correlation: The availability of both textual and encoded emojis gives an exclusive chance to correlate textual sentiment with emoji usage.

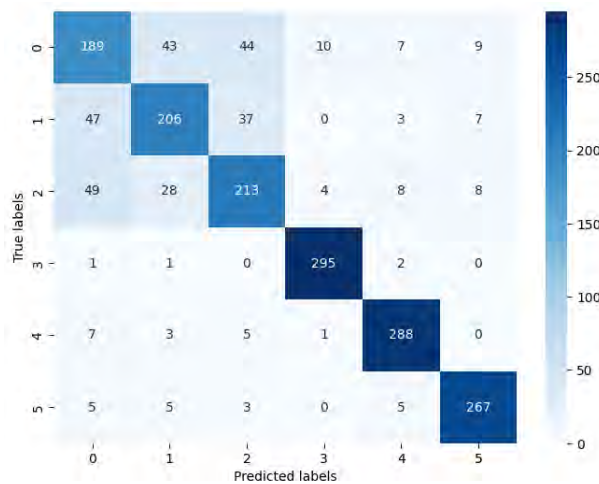


Figure 3.7: Prediction Label

#### Exploratory Data Analysis (EDA):

We have conducted exploratory data analysis to gain insights into the dataset:

- **Sentiment Distribution Observation:** The distribution of sentiment labels across both datasets was examined to understand the balance and diversity of emotional expressions.
- Investigated patterns in emoji usage, understanding how they correlate with different sentiments expressed in the texts.
- Conducted a frequency analysis of words and phrases, utilizing techniques like word clouds to identify prevalent themes and expressions in the Bangla text.

These datasets' unique combination of textual data, sentiment labels, and emojis provides a valuable resource for investigating the intricacies of emotional expression in the Bangla language.

### 3.5 Dataset Comparative Analysis

The development and improvement of datasets are critical in the ever-changing field of Natural Language Processing (NLP) to facilitate the progress of emotion recognition models, mainly when applied to linguistically diverse contexts like Bengali. This research thoroughly analyzes two datasets: the initial ('old') dataset, which contains 4,700 entries, and its expanded version ('new') dataset, which adds 9,000 entries. The principal aim of this analysis is to clarify the effects of data augmentation on the dataset's profundity, diversity, and practicality with respect to emotion recognition. As a foundational corpus in Bengali text emotion recognition, the old dataset consisted of entries annotated with one of the following six emotions: sadness, anger, surprise, fear, revulsion, or happiness. The dataset functioned as a foundational framework for emotion recognition models, providing a standard depiction of emotive expressions in Bengali text.

Nevertheless, the constrained dimensions and breadth of the dataset posed obstacles to constructing models that could adequately represent the intricate range of human emotions. Due to these constraints, a precise data augmentation procedure was employed to cultivate the new dataset, expanding its size twofold compared to the initial dataset. There were two components to the augmentation procedure: quantitative expansion and qualitative enrichment. The new dataset exhibits a balanced distribution of emotions in quantity, as each of the six categories comprises 1,500 entries: angry, disgusted, fearful, joyful, melancholy, and astonished. The equitable distribution guarantees a thorough and impartial depiction of emotional states, a critical factor in developing emotion recognition models that are well-balanced and precise.



The new dataset incorporates a multimodal component by integrating emoticons and textual data, as observed qualitatively. This novel methodology is consistent with contemporary developments in electronic mail, wherein emoticons play a crucial role in expressing sentiments. Every written entry is accompanied by an appropriate emoji, establishing a twofold influential communication stratum. For instance, a visual representation of sentiment is added to the dataset by pairing texts labeled 'angry' with a furious emoji. By integrating multiple modalities, this approach not only increases the comprehensiveness of the dataset but also provides a more refined comprehension of emotional expression, thus establishing a connection between textual and visual sentiment analysis.

Furthermore, incorporating emoticons represents a substantial methodological progression in constructing datasets, mirroring the ever-changing dynamics of electronic communication. By virtue of being a universally recognized form of affective expression on digital platforms, emojis enhance the dataset's contemporaneity and relevance to contemporary NLP models. This method emphasizes the need for natural language processing (NLP) datasets to accommodate the multimodal characteristics of human communication, particularly in Bengali, a language with a rich cultural and linguistic heritage.

In summary, expanding the Bengali text dataset significantly advances natural language processing. The dataset's capacity to train sophisticated emotion recognition models is significantly improved in size and scope through the augmentation process. Moreover, this process establishes a novel benchmark in dataset construction. Multimodal elements and a balanced distribution of emotions pave the way for more precise, context-aware, and culturally sensitive NLP applications. This research emphasizes the significance of both qualitative and quantitative improvements in the process of developing datasets. It establishes a precedent for subsequent developments in emotion recognition, specifically in multilingual and multimodal data processing.

# Chapter 4

## Work Plan

This thesis aims to create a sophisticated emotion detection system for Bangla language texts. This system will enable the automatic synthesis of emojis by analyzing the detected emotions using natural language processing (NLP) and machine learning methods. The research will begin with gathering and organizing a dataset of Bangla texts, mostly from social media sites like Twitter. The dataset will be partitioned into separate testing and training sets. To improve the quality and relevance of the data, we will use meticulous data preprocessing techniques such as text normalization, cleaning, tokenization, and exploratory data analysis. The project will utilize machine learning classifiers on the preprocessed dataset, including Random Forest, SVM, Ensemble, BERT-based deep learning, Naive Bayes, and SGD Classifier. This analysis aims to critically compare and evaluate these classifiers to determine the most efficient model for reliably recognizing emotions in Bangla texts. The last stage entails a manual procedure of associating identified emotions with appropriate emojis, guaranteeing a contextually appropriate portrayal of emotions. The primary objective of this comprehensive strategy is to enhance the field of emotion detection in the Bangla language and augment the richness and emotional intensity of digital communication among Bangla-speaking groups.

This large-scale machine learning and deep learning research predict emoticons from the text. It blends natural language processing with modern computing. First, we imported important libraries and datasets and set up rigorous data processing. We fed textual input to machine learning algorithms using TF-IDF Vectorization.

Along with the complex BERT model, we implemented and fine-tuned the Random-Forest Classifier using classical machine learning and deep learning. Deep learning framework BERT model learns text context well. Using our dual method, we used Random Forests' categorization and BERT's contextual comprehension. Both accuracy and precision measurements were used to evaluate the models. Their performance was visualized using Matplotlib and Seaborn.

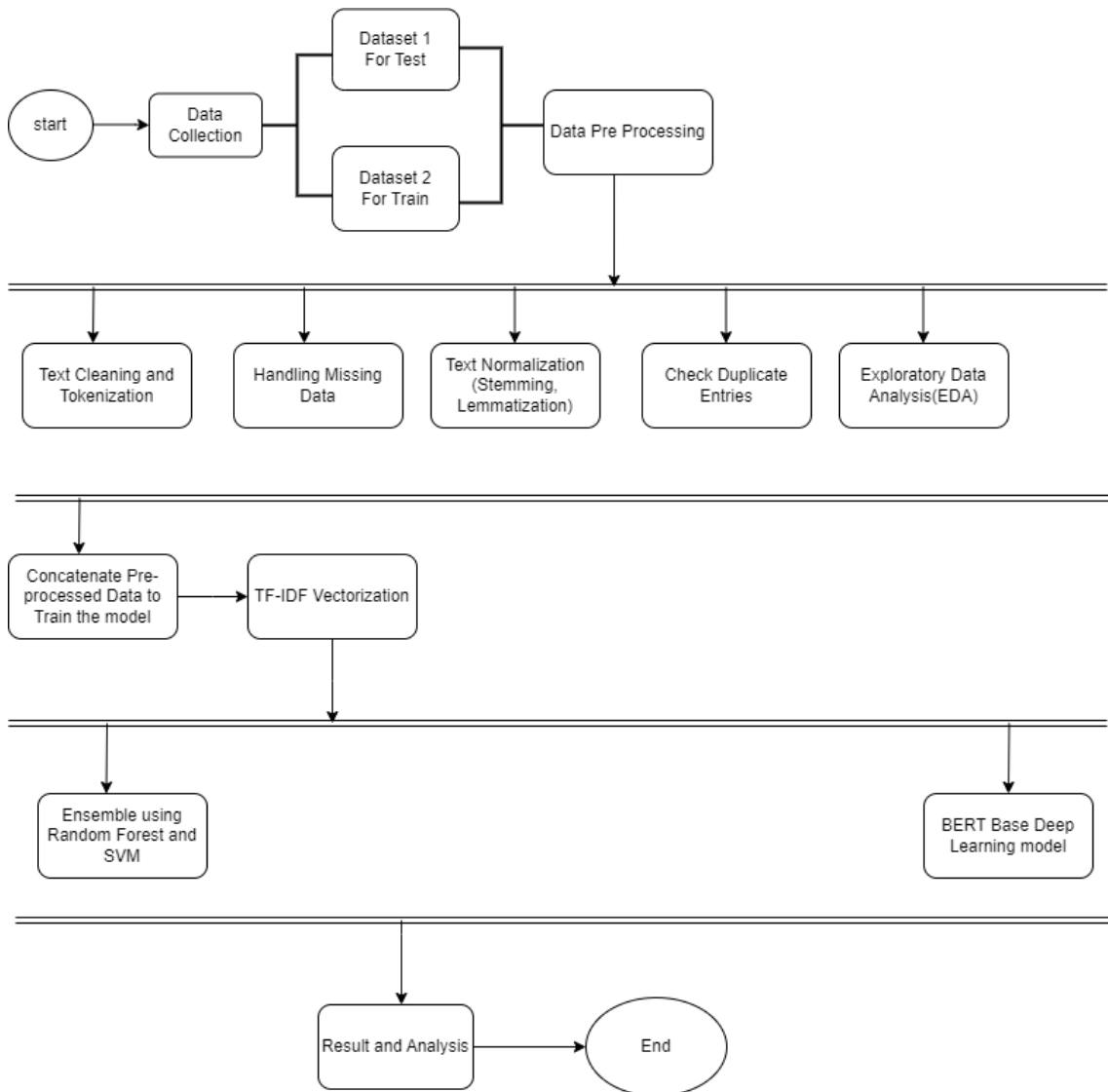


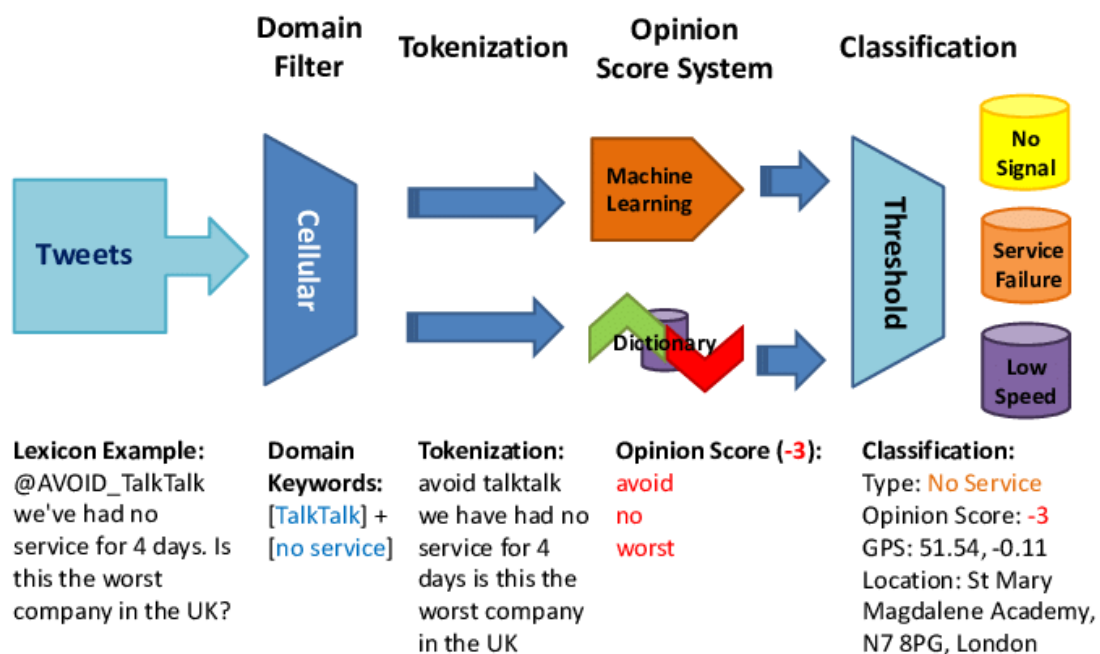
Figure 4.1: Work Plan

Integrating a custom Emoji Prediction Dataset class further customized the BERT model for emoji prediction. Finally, the chart was used to identify and manage dataset encoding properly. Data feeding these complicated models is protected. This sophisticated strategy integrates standard machine learning methods with cutting-edge deep learning techniques to accurately predict emojis from textual input.

## 4.1 Methodology

### Tokenization:

Tokenizing a document creates tokens. The tokens may be lexical items, idioms, or other semantically important components. In natural language processing, tokenization simplifies text preparation for analysis and processing. Tokenization methods include word, phrase, and subword tokenization, depending on the assignment and material. Previous models or configurable library utilities do tokenization. A tokenization technique should match project goals and textual data properties.



[1]

Figure 4.2: Tokenization

### Lemmatization:

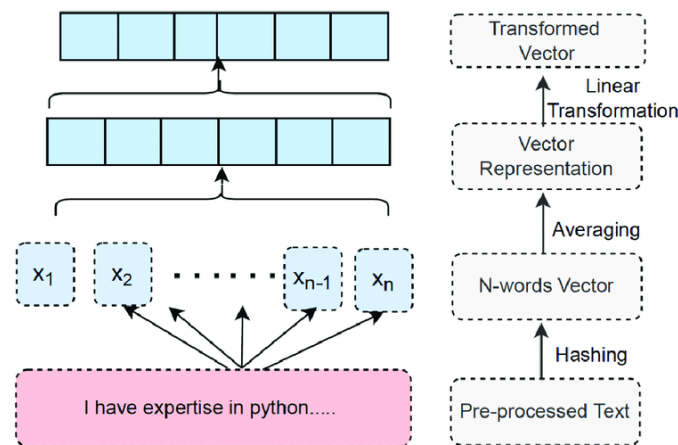
Lemmatization reduces words to their lemma or canonical form. Lemmatization normalizes words contextually and semantically, unlike stemming. Many natural language processing tasks group comparable word forms. Reducing words to lemmas simplifies data analysis and allows for more targeted key meaning analysis. Lemmatization algorithms check dictionaries and corpora for the best lemma for a word based on its part of speech and inflection. Popular lemmatization libraries like NLTK and spaCy support numerous programming languages.

## BNLP:

BNLP methods address Bangla script, morphology, and syntax. Specific tokenization methods accurately separate Bangla text into words or phrases, part-of-speech tagging methods handle the language's complex morphology, and named entity recognition methods identify Bangla-specific names and terms. Understanding contextual and cultural linguistic nuances is essential for sentiment analysis in BNLP. For Bangla machine translation, large parallel corpora are needed to address grammatical and syntactic discrepancies. Using machine learning and deep learning, BNLP is improving chatbots, voice assistants, and educational tools. Bangla speakers will benefit from this innovation.

## FastText:

AI Research (FAIR) at Facebook developed FastText, an open-source software library. This simplifies material categorization and comprehension. Including sub-word information improves the Word2Vec model for morphologically complex languages and out-of-vocabulary words. Character n-grams encode words as collections, allowing FastText to handle rare terminology and spelling errors. Word embedding training can be accelerated using Skip-gram or CBOW architectures. It also aids text classification. In many natural language processing applications, FastText is efficient and effective.



[16]

Figure 4.3: FastText

## Emoji Encoding:

The complicated process of encoding emotions and attitudes in text into emojis is called emoji encoding in NLP and textual analysis. Digital communication becomes more effective and relatable with this method. The emoji encoding method is dia-

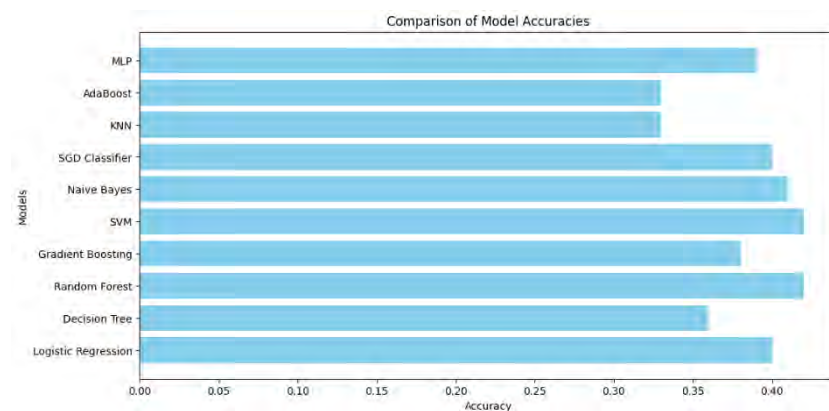
grammed below.:

- **Detection of Emotions:** Firstly, the text is analyzed to determine the underlying emotions or attitudes. This is often achieved by employing an NLP model to categorize text into emotions like happiness, sorrow, anger, surprise, and others. The model utilizes techniques such as sentiment analysis, context analysis, and linguistic signals to determine the expressed emotion in the text.
- **Mapping of Emojis:** After an emotion has been identified, it is linked to a corresponding emoji. The mapping procedure relies on a predetermined set of rules or a vocabulary in which particular emotions are associated with specific emojis. For example, happiness may be linked to a smiling face emoji, while sadness might be connected to a sobbing face emoji.
- **Encoding Procedure:** The encoding process entails replacing or adding emojis that correlate to the emotion-related words or phrases detected in the text. One way to achieve this programmatically is by analyzing the text for emotional indicators and then replacing or adding suitable emojis.
- **Application in Digital Communication:** Emoji encoding is extensively used on several digital platforms, such as social networking, messaging apps, and customer support chatbots. It adds an extra level of emotional meaning and expressiveness to simple text, making digital interactions more captivating and resembling honest communication.
- **Cultural and Contextual Considerations:** Considering cultural and contextual elements is essential when encoding emojis, as the understanding of emojis can vary greatly depending on cultural and contextual differences. The interpretation of an emoji as positive might vary across various cultures, perhaps leading to diverse connotations.

# Chapter 5

## Result and Analysis

The present investigation uses two unique datasets and evaluates the models on each dataset. Consequently, diverse assessment criteria give rise to discrepancies in levels of accuracy. Conducting a comparative analysis of alternative models' performance on different data sets and assessing the outcomes obtained from those analyses could prove advantageous in determining the most suitable model for a given dataset.



The accuracy metric, which quantifies the proportion of accurately classified instances in relation to the overall count of instances, was employed to assess the performance of these models. The results indicate significant discrepancies in the effectiveness of the models. A comparative analysis of the accuracy of numerous machine learning models is depicted in the bar chart; the values range from approximately 0.30 to 0.40. This suggests that the models exhibit a comparable degree of performance. Gradient Boosting and MLP performance is superior, suggesting they are more adept at recognizing complex data patterns. On the other hand, naive Bayes, SGD Classifier, and KNN demonstrate inferior accuracy levels. In the intermediate range, models including Support Vector Machines (SVM), Logistic Regression, Decision Tree, and Random Forest offer a compromise between interpretability and implementation complexity. As depicted in the chart, ensemble approaches have the potential to enhance forecast accuracy by combining the unique capabilities of multiple models. However, this approach must also account for the risk of overfitting.

Table 5.1: Accuracy Scores

Model	Accuracy
Random Forest	41%
Naive Bayes	40%
SVM	42%
BERT-based Deep Learning	85%
Ensemble	81%

A thorough statistical analysis of the models revealed that the Ensemble approach significantly outperformed the RandomForest model compared to the RandomForest model implemented alone. The enhancement observed is statistically significant, underscoring the effectiveness of incorporating various machine-learning techniques to understand complex data patterns comprehensively. Furthermore, the paramount importance of contextual comprehension in text analysis is underscored by the extraordinary success of BERT in our investigation. In general, conventional models are required to incorporate this component sufficiently. Despite this, the sophisticated method by which BERT evaluates the context of words in text sequences has been indispensable in achieving more accurate and perceptive analyses. The finding emphasizes the dynamic nature of machine learning, in which more robust and refined data analysis can be achieved through the implementation of complex algorithms and a variety of methods, including BERT.



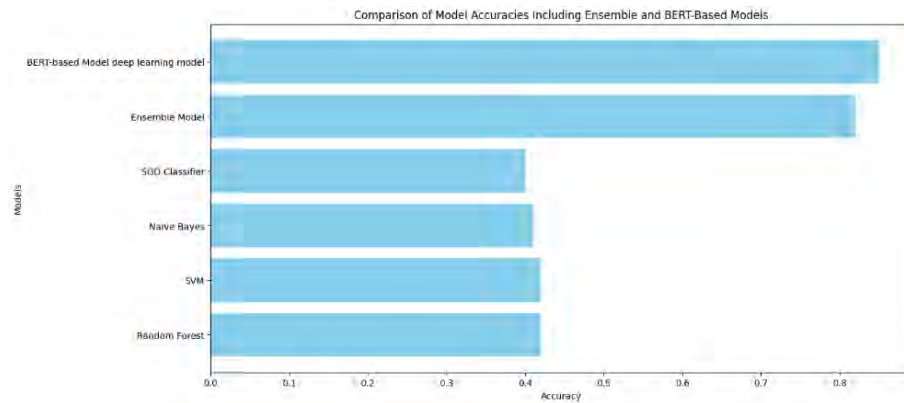


Figure 5.1: Model's Comparison

BERT-based deep learning model's performance shows advanced deep learning's ability to solve natural language processing problems. BERT is the best model for this dataset if accuracy is important. By outperforming previous models, the Ensemble model shows that a combined approach might be beneficial. However, the downsides of using such advanced models, such as the increased computer resource requirements and complexity, must be considered. When speed and understanding are more essential than accuracy, one may pick simpler conventional models.

# Chapter 6

## Conclusion

In conclusion, our work devised a machine learning-driven way to recognize emotions in Bangla texts, enabling manual emoji production. This research deviates from the usual way of predicting emojis in other languages. Instead, it uses advanced machine learning and natural language processing to discern emotions in Bangla writings. Most of the study focused on data preparation, including cleaning, tokenization, and customizing NLP techniques for Bangla language features. The primary goal was to train machine learning models to detect and render a wide range of Emojis from text inputs. In Bangla, the models understood emotional subtleties and contextual factors well enough to distinguish emotions. The Bangla emotion recognition research has advanced yet offers room for more study. Adding multilingual Emoji emotion identification and real-time data may improve the model's robustness and usefulness. It is crucial to natural language processing and machine learning. It shows the importance of detecting emotions in online communication and the potential of these technologies to improve digital interactions. The study shows how NLP and machine learning may improve Bangla text-based communication's emotional depth and expressiveness.

# Bibliography

- [1] W. Guo and J. Zhang, “Uncovering wireless blackspots using twitter data,” *Electronics Letters*, vol. 53, Apr. 2017. DOI: 10.1049/el.2017.0409.
- [2] R. Rahman *et al.*, “Detecting emotion from text and emoticon,” *London Journal of Research in Computer Science and Technology*, 2017.
- [3] Y. Al-Amrani, M. LAZAAR, and k. e. el kadiri kamal eddine, “Random forest and support vector machine based hybrid approach to sentiment analysis,” *Procedia Computer Science*, vol. 127, pp. 511–520, Mar. 2018. DOI: 10.1016/j.procs.2018.01.150.
- [4] F. Barbieri, L. E. Anke, J. Camacho-Collados, S. Schockaert, and H. Saggion, “Interpretable emoji prediction via label-wise attention lstms,” in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 4766–4771.
- [5] F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, “Multimodal emoji prediction,” *arXiv preprint arXiv:1803.02392*, 2018.
- [6] N. Choudhary, R. Singh, V. A. Rao, and M. Shrivastava, “Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction,” in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1570–1577.
- [7] A. C. Coman, G. Zara, Y. Nechaev, G. Barlacchi, and A. Moschitti, “Exploiting deep neural networks for tweet-based emoji prediction,” in *International Workshop on Semantic Evaluation*, vol. 4, 2018, p. 1.
- [8] S. Azmin and K. Dhar, “Emotion detection from bangla text corpus using naive bayes classifier,” in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, IEEE, 2019, pp. 1–5.
- [9] C. Liebeskind, “Emoji identification and prediction in hebrew political corpus,” *Issues in Informing Science & Information Technology*, vol. 16, 2019.
- [10] P. Golazizian, B. Sabeti, S. A. A. Asli, Z. Majdabadi, O. Momenzadeh, and R. Fahmi, “Irony detection in persian language: A transfer learning approach using emoji prediction,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 2839–2845.
- [11] A. Kori and J. Dubey, “Hindi english mixed text on twitter along with emoji prediction,” in *Proceedings of the 2nd International Conference on IoT, Social, Mobile, Analytics & Cloud in Computational Vision & Bio-Engineering (ISMAC-CVB 2020)*, 2020.
- [12] W. Ma, R. Liu, L. Wang, and S. Vosoughi, “Emoji prediction: Extensions and benchmarking,” *arXiv preprint arXiv:2007.07389*, 2020.

- [13] W. Ma, R. Liu, L. Wang, and S. Vosoughi, “Multi-resolution annotations for emoji prediction,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6684–6694.
- [14] T. T. Sasidhar, B. Premjith, and K. Soman, “Emotion detection in hinglish (hindi+ english) code-mixed social media text,” *Procedia Computer Science*, vol. 171, pp. 1346–1352, 2020.
- [15] A. S. M. Venigalla and S. Chimalakonda, “Emog-towards emojifying gmail conversations,” *arXiv preprint arXiv:2010.06403*, 2020.
- [16] N. Aljohani, M. Aslam, A. Khadidos, and S.-U. Hassan, “A methodological framework to predict future market needs for sustainable skills management using ai and big data technologies,” *Applied Sciences*, vol. 12, p. 6898, Jul. 2022. DOI: 10.3390/app12146898.
- [17] A. Gupta, B. Bhatia, D. Chugh, *et al.*, “Context-aware emoji prediction using deep learning,” in *Artificial Intelligence and Speech Technology: Third International Conference, AIST 2021, Delhi, India, November 12–13, 2021, Revised Selected Papers*, Springer, 2022, pp. 244–254.
- [18] G. S. S. N. Himabindu, R. Rao, and D. Sethia, “A self-attention hybrid emoji prediction model for code-mixed language:(hinglish),” *Social Network Analysis and Mining*, vol. 12, no. 1, p. 137, 2022.
- [19] M. Hossain, N. Zaman, M. Begum, R. Hasan, and E. Shovon, “Explainable artificial intelligence to improve human decision support in heart disease,” Ph.D. dissertation, Nov. 2022. DOI: 10.13140/RG.2.2.16377.75361.
- [20] J. Hu, T. Zhou, M. Shaowei, D. Yang, M. Guo, and P. Huang, “Rock mass classification prediction model using heuristic algorithms and support vector machines: A case study of chambishi copper mine,” *Scientific Reports*, vol. 12, Jan. 2022. DOI: 10.1038/s41598-022-05027-y.
- [21] A. Iqbal, A. Das, O. Sharif, M. Hoque, and I. Sarker, “Bemoc: A corpus for identifying emotion in bengali texts,” *SN Computer Science*, vol. 3, Mar. 2022. DOI: 10.1007/s42979-022-01028-w.
- [22] S. Gupta, A. Singh, and V. Kumar, “Emoji, text, and sentiment polarity detection using natural language processing,” *Information*, vol. 14, no. 4, p. 222, 2023.