

BANGLADESH RELATED FACTOID QUESTION CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

BY

Md. Jahid Hossain
ID: 191-15-12250

Md. Nahid Hossain
ID: 191-15-12251

Sadiha Nowmi
ID: 191-15-12341

This report is being submitted in partial fulfillment of the requirements for the Bachelor of Science in Computer Science and Engineering degree.

Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Narayan Chakraborty
Associate Professor
Department of CSE
Daffodil International University




DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY, 2023

APPROVAL


This Project titled “Bangladesh Related Factoid Question Classification Using Machine Learning Techniques”, submitted by Md. Nahid Hossain (191-15-12251), Md. Jahid Hossain (191-15-12250) and Sadiha Nowmi (191-15-12341) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on 26th January, 2023.

BOARD OF EXAMINERS


Chairman


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

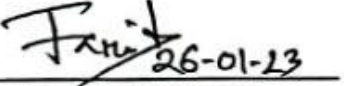
Internal Examiner


Subhenur Latif
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner


Mohammad Monirul Islam
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner


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

DECLARATION


We hereby declare that under the guidance of **Md. Sadekur Rahman, Assistant Professor, Department of CSE Daffodil International University**, we completed this thesis. We also declare that no portion of this thesis or any other portion has ever been submitted anywhere for consideration of a degree or diploma.

Supervised By:



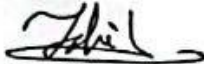
Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By:

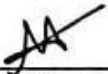


Narayan Chakraborty
Associate Professor
Department of CSE
Daffodil International University

Submitted By:



Md. Jahid Hossain
ID: 191-15-12250
Department of CSE
Daffodil International University



Md. Nahid Hossain
ID: 191-15-12251
Department of CSE
Daffodil International University



Sadiha Nowmi
ID: 191-15-12341
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We begin by expressing our sincere gratitude to Almighty Allah for giving us the ability to successfully complete the final thesis through His wonderful grace.

We sincerely appreciate your help and want to express our gratitude to **Md. Sadekur Rahman**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Our supervisor's extensive knowledge and keen interest in the topic of "Natural Language Processing and Machine Learning" to carry out this thesis successfully. His academic supervision, persistence, insightful criticism, inspiration, consistent and energetic inspection, constant encouragement, insightful counsel, reviewing numerous substandard versions, and revising them at every level have enabled the completion of this thesis. We also want to thank our co-supervisor **Narayan Chakraborty**, Associate Professor, Department of CSE Daffodil International University, Dhaka. He provided us with helpful thoughts and solutions when we faced problems. He inspired and assisted us in completing our task.

For his inspiration and appreciation, we would like to express our sincere gratitude to Professor **Dr. Touhid Bhuiyan, Professor & Head**, Department of CSE. We also like to express our gratitude to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank every one of our classmates at Daffodil International University who participated in this discuss while completing the course work.

Finally, we want to express our gratitude to our parents and friends for constantly encouraging and constructively criticizing our work. At the very least, we thank all of them from the core of our heart.

ABSTRACT

Question answering (QA) is a branch of natural language processing research that is aimed to give human users a simple and practical information retrieval application. Despite being one of the most widely spoken languages in the world, Bengali still has issues with computational linguistics. Question classification is the very first step before developing a factoid question answering system. Appropriate classification for questions is essential because the performance of the whole system depends on it. This paper demonstrates question classification for Bengali language question for developing a Bengali factoid question answering system. We collected data from different Bangla books, newspapers, nobles and so on. Then we have made our dataset with four categories like HUM, NUM, LOC and ENTY. At last, we have collected 1400 questions for our dataset. Before applying any Natural Language Processing (NLP) model, we have to preprocess our dataset. Machine learning is frequently used in classification or prediction. So, we present ten different machine learning algorithms and they are Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid) and Support Vector Machine (Kernel = poly). Among these models, we have got the best performance measures by Logistic Model and obtained 88.57% accuracy.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Question	3
1.5 Expected Output	3
1.6 Report Layout	4
CHAPTER 2: BACKGROUND	5-9
2.1 Terminologies	5
2.2 Related Works	5

2.3 Comparative Analysis and Summary	7
2.4 Scope of the Problem	8
2.5 Challenges	9
CHAPTER 3: RESEARCH METHODOLOGY	10-24
3.1 Research Subject and Instrumentation	10
3.2 Data Collection Procedure	10
3.3 Statistical Analysis	12
3.4 Proposed Methodology	13
3.5 Implementation Requirements	23
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	25-34
4.1 Experimental Setup	25
4.2 Experimental Results & Analysis	26
4.3 Discussion	33
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	35-36
5.1 Impact on Society	35
5.2 Impact on Environment	35

5.3 Ethical Aspects	35
5.4 Sustainability Plan	36
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	37-39
6.1 Summary of the Study	37
6.2 Conclusions	38
6.3 Recommendation	38
6.4 Implication for Further Study	39
REFERENCES	40-43

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Proposed Methodology	13
Figure 3.2: Count vectorizer example	15
Figure 3.3: Gaussian Naïve Bayes working illustration	16
Figure 3.4: Logistic model curve	17
Figure 3.5: KNN model classification	18
Figure 3.6: Random Forest prediction	18
Figure 3.7: Decision Tree structure	19
Figure 3.8: Linear SVM example	20
Figure 3.9: RBF SVM example	20
Figure 3.10: Sigmoid SVM Example	21
Figure 3.11: Poly SVM Example	21
Figure 3.12: Confusion Matrix Example	22
Figure 4.1: Confusion Matrix of Multinomial Naive Bayes algorithm	27
Figure 4.2: Confusion Matrix of Gaussian Naive Bayes algorithm	27
Figure 4.3: Confusion Matrix of Logistic Regression algorithm	28
Figure 4.4: Confusion Matrix of K-Nearest Neighbor algorithm	28
Figure 4.5: Confusion Matrix of Random Forest algorithm	29
Figure 4.6: Confusion Matrix of Decision Tree algorithm	29
Figure 4.7: Confusion Matrix of Support Vector Machine (Kernel = Linear) algorithm	30
Figure 4.8: Confusion Matrix of Support Vector Machine (Kernel = rbf) algorithm	30
Figure 4.9: Confusion Matrix of Support Vector Machine (Kernel = Sigmoid) algorithm	31
Figure 4.10: Confusion Matrix of Support Vector Machine (Kernel = Poly) algorithm	31
Figure 4.11: Algorithms Accuracy Bar Chart	34

LIST OF TABLES

TABLE NO.	PAGENO
Table 2.1 Comparison Between Previous Factoid Question Classification Algorithms Paperwork	7-8
Table 3.1: Dataset of Factoid Question Classification	11
Table 3.2: Dataset detailed information	12
Table 3.3: Sample Dataset with Length	12
Table 4.1: Model Accuracy table	25
Table 4.2: Classification Report all models	32-33

CHAPTER 1

INTRODUCTION

1.1 Introduction

As the volume and variety of online text documents are increasing rapidly nowadays a vast amount of material is available online, users facing difficulties finding their exact desired answers. Users find the program-generated replies which is very inconvenient because they only offer sorted lists of texts that require manual browsing on the part of the user. The majority of the time, while using a natural language processing (NLP) system, customers want precise answers to their queries. Here the necessity of the existence of Question Answering systems becomes evident because it provides the precise answer to a given question [1]. However, the first prerequisite for creating a top-notch question answering system is accurate question classification. So, the categorizing of questions (question classification) is a fundamental task [2]. Finding the correct question type improves the retrieval of more precise answers.

A significant amount of research has already been accomplished on several languages in this particular topic, but comparatively very few has been done on the Bengali language. Bengali is the seventh most spoken languages in the world. Moreover, there are two categories into which the general question answering system domain can be subdivided: open-domain and close/restricted domain. When a question answering system can respond to a diverse query, it can be distinguished as an open-domain. Apple's Siri and IBM's Watson are two well-known examples of this field, where the question answering system may respond to user-provided queries that are not specific to any particular domain.

Our paper is based on close domain question answering system. We have made our dataset with the questions only about Bangladesh.

For example: What is the name of national fruit of Bangladesh?

Answer: Jackfruit

To get more accuracy in our system, a huge dataset is required. The advantage of having big dataset is that we can have enough amount of training and test data. So, dataset play a vital role in the accuracy of desired outcome and performance of the system.

In this paper work, we have only classified the category of factoid questions about Bangladesh. We have used ten types of classifiers and they are Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid) and Support Vector Machine (Kernel = poly). We discuss elaborately the whole process in Chapter-3: Research Methodology.

1.2 Motivation

Natural language processing (NLP) is the ability of a computer program to understand both spoken and written human language. That means it is a way to build the communication between human and a computer. Using natural language processing techniques, lots of factoid question answering system has been developed in different languages. But there are very few factoids question answering systems about Bangladesh in comparison to other countries. A question-answering system's goal is to automatically identify concise responses to every given query that has been given a natural language phrase [3]. As these types of research are very few to be found in our country, so sufficient resources are unavailable. So, we have decided to develop a system that will provide a vast amount of information about our country. So, people can easily get answers to their queries through our system. Moreover, in the future, if anyone wants to work on this topic, our research paper will assist them as a resource. From all these, we were motivated to build a model which can provide the exact answers of the questions about Bangladesh. So, we have classified the questions dataset till now.

1.3 Rational of the Study

Though there are a lot of question answering systems available, the majority of them are in the English language. Very few factoids question answering systems in the Bangla language are to be found. People are not very much interested to work in this sector, so the availability of resources is very limited. Bengali is our mother language and it's our responsibility to enhance our language in all possible best ways.

Moreover, if we search specific questions on Google, sometimes it doesn't provide the exact answer. We have to read an article or a paragraph to find the desired answer. It's very time consuming and boring work to do. Furthermore, modern tools are updated frequently in this era of

advanced technology. So, we want to represent our language to the next generation in a modern way so that they can get all the answers to their queries in a second. That's why we created a dataset about Bangladesh and classified the questions category to develop a factoid question answering system about Bangladesh.

1.4 Research Questions

Some questions regarding this work come up during the research study. The following list of key queries informs our work:

1. How to collect factoid questions without duplicating the data?
2. How to preprocess and extract features from the dataset?
3. Which classifiers perform better to classify factoid questions?

1.5 Expected Output

Although this is an experimental initiative, our main objective is to write a paper about this project. We discovered a good number of papers that are relevant to the factoid question answering system. Most of them are of different languages like English, Arabic, and Japanese. We have also found some Bengali language question answering system-related research papers. As the number of Bengali factoid question answering systems is very few and there's no work has been done previously on the Bangladesh domain, so we decided to conduct some research on this topic and intend to develop a model using machine learning algorithms.

At the end of this research work, our expected outcomes are:

1. Create a dataset of factoid questions and answers.
2. Preprocessing of the dataset.
3. Classify the dataset with the best accuracy.
4. Develop a model for a factoid question answering system about Bangladesh.
5. Build a more accurate factoid system.
6. Publish one or more papers on International Conference.
7. Our main expected outcome is our system can provide an exact answer to the asked query of the user.

1.6 Research Layout

There are a total of 6 chapters in our report:

1. In the section of chapter 1, there are several subsections in this chapter that cover all aspects of our research. For instance, we discuss the introduction, motivation, rationale, research question, and expected results.
2. In the section of chapter 2, the topic of factoid question classification, the scope of the selected problem, and the challenges related to this research work were discussed.
3. In the section of chapter 3, describe our factoid question classification procedures, what are the methods we have used and what are the techniques we have used.
4. In the section of chapter 4, discussion and experimental results of the machine learning model we implemented will be discussed.
5. In the section of chapter 5, a description of our work will be given, including the impact of society, environment, ethics, and sustainability plan.
6. In the section of chapter 6, we explained what the work's summary, conclusion, and future research entails.

CHAPTER 2

BACKGROUND

2.1 Terminologies

The term "natural language processing" has been used for decades and is now frequently used in our daily lives. Natural language processing (NLP) is a very hot topic for researchers and a lot of research work has been done using this technique. As we are being dependent on technology and it's important to make sure that machines can understand human languages, so Natural language processing (NLP) is one of the most popular terms in research works. There are so many research works available related to question classification and factoid question answering systems. But there was no available dataset based on factoid questions and answers about Bangladesh. So, we have found a few research papers related to Bengali factoid question classification. In this research paper, we are going to apply question classification algorithms and to do so we have to introduce some new terminologies. When performing NLP, we simply gather raw text input, but in order to apply algorithms, we must convert the data into a numeric value. To convert our dataset into a numerical value, we introduced a new term Count vectorizer. There are a few other terms that we used to evaluate our implemented model performance and they are confusion matrix, precision score, recall score, and F1 score. In the next chapter, we will elaborately discuss these terms. We studied some previous work that is related to factoid question classification in order to implement our work flawlessly and to understand this new concept. We briefly described a few of them below.

2.2 Related Works

A research paper titled 'Multi-Class Classification of Turkish Texts with Machine Learning Algorithms' proposed five different algorithms consisting of Multinomial NB, Bernoulli NB, SVM, KNN, and J48. Their dataset contains 3000 data. With a classification success rate of almost 90%, the multinomial NB method emerged as the top classifier [4].

'Classification of Bengali Question Towards a Factoid Question Answering System' titled paper was published in 2019 and applied a few classification algorithms. They created the dataset with 15355 questions. For their classification process for features, extraction applied both the bi-gram and tri-gram methods and got better results by the bi-gram method. Their used classifier algorithms

are Stochastic Gradient Descent (SGD), Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes (NB). They applied four different kernel Support Vector Machine (SVM) linear, rbf, sigmoid and polynomial. In this paper, Support Vector Machine (SVM) with linear kernel performs the best with an accuracy of 90.60% [5].

In 2018, 'Classification of online toxic comments using the logistic regression and neural networks models' was published, and implemented logistic regression model. They applied this algorithm to the dataset containing 15958 comments [6].

A research paper titled 'Automatic Classification of Academic and Vocational Guidance Questions using Multiclass Neural Network' was published in 2019, proposed five different types of classification such as Support Vector Machine (SVM), Naive Bayes Classifier, K-Nearest Neighbors, Multiclass Logistic Regression, Multiclass Neural Network. Among these classification algorithms Multiclass Neural Network performs best [7].

In 2018, a research paper was released titled 'Classification of factoid questions intent using grammatical features. In their research work, they suggested a system for the classification of factoid questions. They used 3000 questions for their dataset. Two machine learning algorithms were used for question classification in this paper: Support Vector Machine (SVM) and J48. In their datasets, J48 demonstrates high accuracy with 95.8% [2].

In 2021, a research paper titled 'Question Classification for Automatic Question-Answering in Agriculture Domain' was published. For their work, K-NN, Naive Bayes, Decision Tree, Random Forest, SVM, and XGBoost classifier were used. Here 90% data was training data and 10% data was for testing data [8].

In 2018, a paper was published named 'Toward a New Arabic Question Answering System'. They proposed three different types of question classification algorithms e.g.; Support Vector Machine (SVM): kernel function linear, polynomial, radial basis function, and sigmoid, Decision Tree (DT), and Naive Bayes (NB). For their dataset, Support Vector Machine (SVM) classifier shows the most accuracy with 84% percent [9].

Question classification in Persian using word vectors and frequencies' titled paper was published in 2017 and applied a few classification algorithms. They applied word2vec and tf-idf methods for

feature extraction and Support Vector Machine (SVM), Multi Layered Perception (MLP) as question classifiers. Multi Layered Perception (MLP) classification algorithm gave better accuracy with 72.46% [1].

In 2020, a research paper named ‘Bengali Question Answering System Using Seq2Seq Learning Based on General Knowledge Dataset’ was published. In this paperwork, two types of classification algorithms were implemented and they are Naïve Bayes and Decision Tree Classifier. The decision Tree classifier gave the best accuracy and that is 87.63% [10].

2.3 Comparative analysis and summery

We analyzed some previous work that was relevant to our research topic. We are classifying the factoid questions and answers using different types of classification algorithms. That’s why we have collected some research papers relevant to our topic. Basically, we'll analyze which machine learning algorithm work best in factoid questions and answers classification. Table 2.1 compares the results of previous classification algorithms implemented in factoid question classification.

Table 2.1 Comparison Between Previous Factoid Question Classification Algorithms Paperwork

Work Title	Work Type	Best Algorithms Name	Best Accuracy Score
Multi-Class Classification of Turkish Texts with Machine Learning Algorithms	Multiclass	Naive Bayes (Multinomial)	90%
Classification of Bengali Question Towards a Factoid Question Answering System	Multiclass	SVM with linear kernel	90.60%

Work Title	Work Type	Best Algorithms Name	Best Accuracy Score
Classification of factoid questions intent using grammatical features	Multiclass	J48	95.8%
Toward a New Arabic Question Answering System	Multiclass	Support Vector Machine (SVM)	84%
Question classification in Persian using word vectors and frequencies'	Multiclass	Multi Layered Perception (MLP)	72.46%
Bengali Question Answering System Using Seq2Seq Learning Based on General Knowledge Dataset	Multiclass	Decision Tree classifier	87.63%

We studied several previous papers which are related to factoid question classification algorithm and factoid question answering system. But in the above compared table, we only mentioned the best performed algorithm name and the accuracy we have got. Analyzing the above table, we can see that Support Vector Machine (SVM) perform better in NLP based research work. There are some other algorithms like Decision Tree classifier, Logistic Regression Model, Naïve Bayes also perform well.

2.4 Scope of the Problem

Analyzing some previous research papers based on factoid question classification, we have got so many resources about classification algorithms. Among them Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors, Multiclass Logistic Regression, Decision Tree and many

more algorithms that perform well in Natural Language Processing (NLP) based research work. One thing we understood that there is a lot of factoid question answering system exist and majority of them are English. A very few Bengali question answering system paperwork has been conducted yet. Bengali is our mother language and it's our responsibility to enhance our language in all possible best ways. Moreover, we are now living in the world of modern technology and search everything whatever we want to know on Google. So, we want to represent our language to the next generation in a modern way so that they can get all the answers to their queries in a second. It will provide easy access to the information about Bangladesh.

2.5 Challenges

Throughout the whole process of this work, we experienced several challenges. The foremost and most important challenge we face was creating the dataset. At first, we decided to collect data from online sources. But then we faced problems of data duplicity and back-dated data. As the dataset was not available anywhere, so we have to create it manually. Then we started to collect data from books, newspapers, journals, and so on. We also faced difficulties to maintain the ratio of all four categories of data. ENTY type data was less in comparison to the other three categories (HUM, NUM, and LOC). So, we modified some NUM type data to convert it into ENTY type data. As we customized our dataset, so the preprocessing method was a bit tricky.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

We want to develop a factoid question-answering system that can predict the correct answer to factoid questions related to the Bangladesh domain. In this research, we will find a solution for the classification of questions, which is the initial stage in any question-answering system. The user may discover the information they need more quickly with the help of a suitable classification against the short texts [11]. There are two methods for completing the task of question categorization. The first uses manually created rules, whereas the second uses machine learning methods [12]. As our domain is Bangladesh, there are four categories in our work. To classify the question category, we have used many machine learning models. Machine learning models can be approached in one of two ways. The first is supervised machine learning, whereas the second is an unsupervised model. The key difference between Supervised and Unsupervised are, in Supervised learning input and output data are labeled appropriately by human but the input and output data of unsupervised learning is not labeled. Supervised learning is more accurate but, in this approach, we need to train the model first to get the outcome. Whereas in unsupervised learning the model works on its own. It tries to self-learn itself and understand the unlabeled data [4].

In our work, we are training machine learning models with labeled input data and output data. So, all the models we have used are Supervised approaches. Machine learning algorithms we have used in our work are Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly). We are going to discuss all the algorithms elaborately in our proposed methodology portion. Also, the performance of all these models we used will be evaluated by a few benchmarks such as precision, recall score, and f1 score.

3.2 Data Collection Procedure

In supervised learning, Machine learning algorithms can give excellent outcomes but we need to collect our labeled data accurately. So, collecting data for the model properly is very important in this case. The development and training data are included in the data sets used to train classifiers

[13]. We have made our own dataset because there is no ready-made dataset available for our domain. We have faced many problems throughout the data-collecting process. At first, we tried to collect our data from many Bangladeshi websites on the internet but most of the data were backdated and there was a pretty high chance of duplication if we collect our data from the internet. So, we decided to collect our data from Bangla general knowledge books. In this way, we found updated data about Bangladesh, and the chances of duplicate data were very low. It solves some of our problems but still, some data did not meet our requirements so we needed to modify some of them to get our desired data. At first, there were five categories of questions in our dataset (HUM, NUM, LOC, ENTY, ABBR). But then we faced a huge difficulty in our dataset. The amount of data in the ABBR category was very low compared to the other categories and the amount of data in the NUM category was very high (Around 50%) compared to the other categories. Then we eliminated the ABBR category from our dataset and drop some of the data from the NUM category. Finally, we got our dataset which contains 1400 data from four categories (NUM, HUM, LOC, ENTY) with a ratio of 25%.

Table 3.1: Dataset of Factoid Question Classification

Question	Answer	Category
What is the name of Ganga after entering Bangladesh?	Padma	HUM
Who is the architect of Hazrat Shahjalal Airport?	Laros	HUM
Where is Bangladesh Machine Tools Factory situated?	Gazipur	LOC
Where is Panam Bridge situated ?	Sonargaon	LOC
How many Bengali has got Nobel prize?	4	NUM
With how many countries Bangladesh has border?	2	NUM
Which crop in Bangladesh has an improved variety called Eratom?	Rice	ENTY
Which thing called Golden fiber in Bangladesh?	Jute	ENTY

3.3 Statistical Analysis

Collecting data for factoid question classification was pretty challenging but finally, we managed to complete our dataset with 1400 data. Here 350 data belong to the NUM category, 350 data belong to the HUM category, 350 data belong to the LOC category and 350 data belong to the ENTY category. Statistical analysis helps us to understand deeply about the dataset so that we can find the best methods and algorithms to solve our problem. Sometimes it can help us to identify any error or cause of problems or failures. Table 3.2 shows the ratio of data of each class in our dataset. We have also calculated the length of the questions in our dataset. Table 3.3 shows an example of the length calculation. We are providing the statistical analysis of our dataset below.

Table 3.2: Dataset detailed information

Category of question	Total data count	Percentage of total data
NUM	350	25%
HUM	350	25%
ENTY	350	25%
LOC	350	25%

Table 3.3: Sample Dataset with Length

Question	Category	Length
What is the name of the player who achieved man of the match on Test and ONE-day debut in the history of cricket?	HUM	23
What is the lowest stage of Bangladesh administrative formation?	HUM	10
How many people were awarded the Bangla Academy Literary Award 2018?	NUM	11
When was at first Bangladesh participated in the World Cup selection part?	NUM	12
What is the name of the lowest poverty district in Bangladesh in 2022?	LOC	13
Which is the largest division of Bangladesh in respect of area?	LOC	12
Which crops are cultivated in Bangladesh during the winter season?	ENTY	10
Which plant is suitable for cultivation in Jashor area?	ENTY	9

1. Our dataset has 1400 data.
2. Our dataset contains 4 categories of data with the rate of 25%
3. We got 3 columns of data in our dataset
4. Among all the questions, the highest length of question in our dataset is 25.
5. 2177 unique words available in our dataset

3.4 Proposed Methodology

The purpose of this section is to discuss the methodology we used to conduct our research. We have implemented ten supervised machine learning classifiers Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly) to classify the factoid question category.

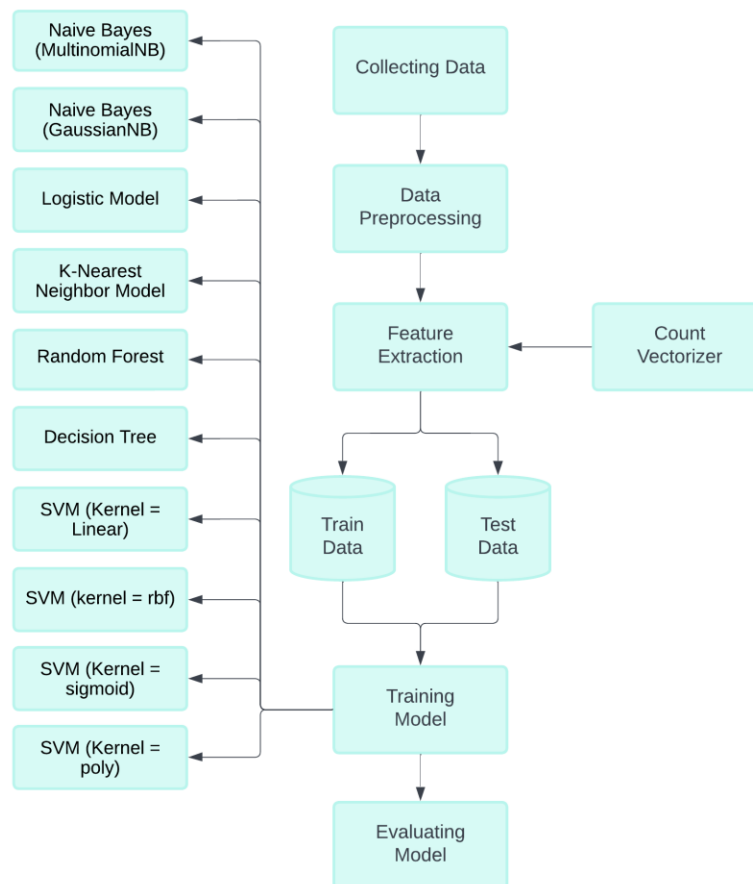


Figure 3.1: Proposed Methodology

As we have used supervised machine learning models so, we need a database to train the model. We have made our database by ourselves. In an effort to get all the data we needed, we have tried to find it and modify it as necessary. A few steps have been broken down to explain the whole process. Each step of our methodology is depicted in Figure 3.1. Below is a description of the remaining methodology phases.

3.4.1 Data Preprocessing

Data Pre-processing is a technique for getting rid of undesired phrases including outliers, stop words, and hyperlinks that don't improve text classification [14]. One of the best ways to ensure the accuracy of the data is through pre-processing. Before to implementing the algorithms for analysis [15]. Finding a rule that divides the data into distinct categories is the process of classification [16]. The factoid questions were categorized using ten machine learning algorithms. But we cannot just feed our raw text data to our models. A random dataset contains a lot of error factors such as duplicate data, null values, data category imbalance, unacceptable values, etc. If we train our models with this kind of dataset then it will lead us to very poor performance and accuracy. To make the process faster and more accurate we do preprocess on our dataset. As we have made our own custom dataset so, the dataset was pretty clean and noise-free by default. To double check if any null value is present in our dataset or not, we have checked whether there is any null value present or not. If any null value is present then we have dropped that particular data. we have also checked if there is any duplicate data present in our dataset. If it finds any duplicate data in our dataset, it drops that particular data. Our data has four categories such as NUM, HUM, LOC, and ENTY. Each category has a 25% ratio of data. So, our dataset does not contain any imbalanced data categories. Then the text data was converted into numerical values using Count vectorizer so we could train our models.

3.4.2 Feature extraction

The majority of current document classification algorithms use a vector space model to describe documents, which interprets a document as a "bag of terms" [17]. After pre-processing, the questions are now prepared for the features extraction step. Working with a dataset that contains hundreds even thousands of characteristics is getting typical these days. Here feature extraction comes into the picture. Basically, feature extraction reduces the number of features to work with. To get high performance in any classification system we definitely need to use feature extraction [1]. In our work, we have used the Count vectorizer for feature extraction. Vectorization is the process of changing an algorithm from

working with one variable at a time to working with many values simultaneously [18]. It will first convert our questions into vectors which are in numerical values. Based on the frequency of a word occurring in a question, count vectorizers create a vector from the given question, which makes this method extremely useful for NLP applications. By counting vectors with unique values, a vector matrix is generated. Afterward, when it reads any question, it tries to match the words with that matrix and increase the counter if it finds any. Figure 3.2 helps us to understand the count vectorizer graphically.

UNIQUE WORDS

	CAT	DOG	EAT	FOOD
NO. OF SENTENCES	0	2	0	0
1	0	0	1	0
3	0	5	0	0
4	2	0	0	0

Figure 3.2: Count vectorizer example

Count vectorizer is a very powerful way to extract features from any given text. As our dataset got a lot of text data in the question column so, count vectorizer is the best way to extract features from our dataset. Using this method, we got good accuracy from the implemented models. We will discuss more about the results in chapter 4.

3.4.3 Model Selection

Selecting the best model is crucial to achieving the best classification result. In machine learning, there are two approaches we can use. one is supervised and another one is unsupervised. As, we have made labeled train data and test data so, in our case we are using supervised technique. In our work we have used Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel =

Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly) models. A training dataset and a testing dataset were created after applying the Count vectorizer. 70% of the data is used for training and 30% for testing. Among all the models Logistic Model has the best performance. Now we will briefly discuss about the algorithms we used on our dataset.

3.4.3.1 Naive Bayes (Multinomial)

The Multinomial naive Bayes classification method is a great way to analyze text data and solve problems with multiple classes. When dealing with problems with multiple classes and analyze text data, multinomial naive Bayes classification is very useful. Because of its simple implementation and low complexity, the Nave Bayes algorithm is frequently employed in text classification. For categorizing texts, Naive Bayes mostly uses multivariate Bernoulli and multinomial models as event models [19]. This algorithm can be implemented on both continuous and discrete data. This machine learning algorithm performs very well for text data. As most of our data in the dataset are in text form so, this is a perfect model for our work. The formula of Naive Bayes (Multinomial) is:

$$P(H | E) = \frac{P(E | H) * P(H)}{P(E)} \quad [20]$$

3.4.3.2 Naive Bayes (Gaussian)

The Gaussian Naive Bayes model supports continuous-valued features. In continuous data analysis,

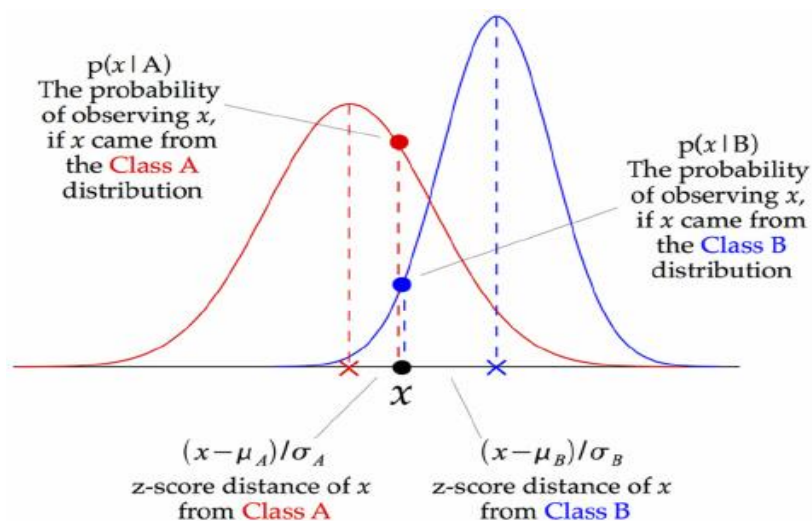


Figure 3.3: Gaussian Naive Bayes working illustration [21]

it is common to assume that the continuous values correspond to each class are Gaussian distributed. The Gaussian Naive Bayes machine learning classifier is demonstrated in the figure 3.3.

3.4.3.3 Logistic Model

For text-valued features, logistic models are excellent machine learning algorithms. Generalized linear models called logistic regressions are used to analyze the association between characteristics (independent variables) and a binary outcome [22]. The probability of an outcome that can only have two values is most likely predicted via logistic regression. The prediction of this model is based on the use of multiple numerical or categorical predictors. The logistic model also creates a logistic curve which we can understand better from figure 3.4 which represents a graphical view of it. Performance of this algorithm was best for our dataset.

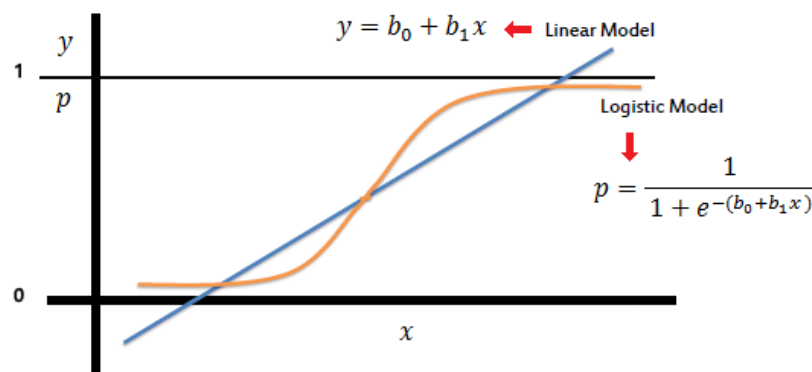


Figure 3.4: Logistic model curve [23]

3.4.3.4 K-Nearest Neighbor Model

K-Nearest Neighbors Algorithm is an algorithm for supervised learning. This algorithm is also known as the KNN algorithm. This algorithm focuses on keeping things that are similar to one another close together [24]. This algorithm uses proximity to classify or predict the grouping of a dataset. A document's relevance to a query is determined in the KNN method based on its Euclidean distance to the query vector. This statistic has only marginal success [25]. A class label is selected for classification based on plurality voting or majority voting which means that the label is mostly expressed around that particular data point. Majority voting means more than 50% voting goes for one label of data which is used for only two classes but for multiclass classification we don't need

more than 50% vote to draw out a conclusion [26]. An example of a KNN model classification process can be seen in Figure 3.5.

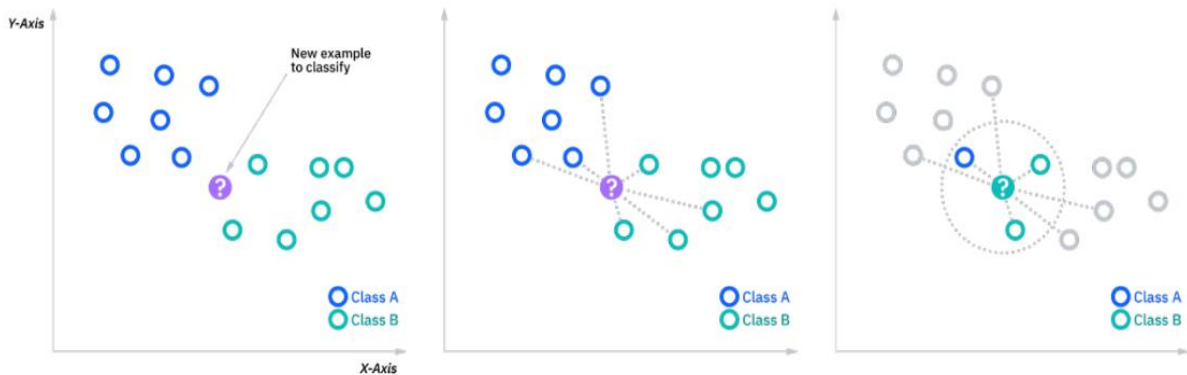


Figure 3.5: KNN model classification [26]

3.4.3.5 Random Forest Model

Machine learning algorithms such as random forests are also supervised. A simple, accurate, and flexible algorithm like this is widely used. It is more applicable to a variety of datasets because it can be used with both linear and non-linear data. This model is called a forest because it creates a forest of multiple decision trees. Among those, it calculates its predictions.

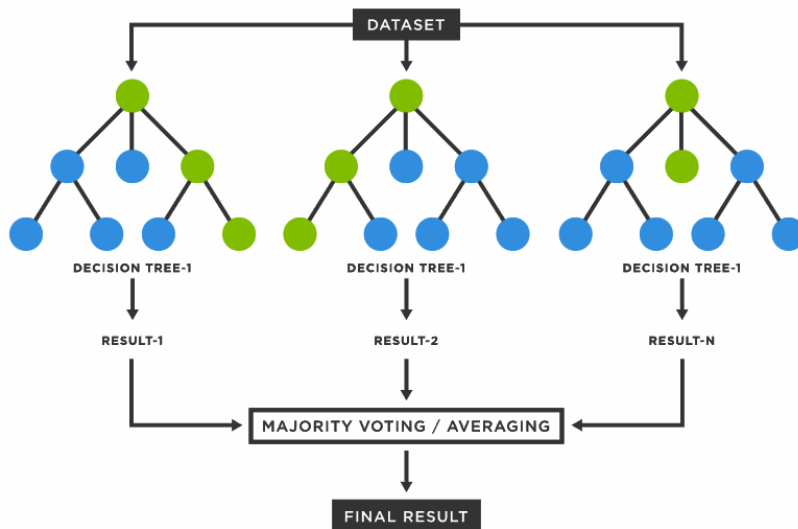


Figure 3.6: Random Forest prediction [27]

A random forest needs basic factors in order to function well:

1. An identifiable signal to prevent models from merely guessing.
2. A low correlation between the predictions provided by the trees and those of the other trees is required.
3. Features with some degree of predictability include: GI=GO [27].

3.4.3.6 Decision Tree Model

Machine learning algorithms such as decision trees are supervised. This algorithm is used for both regression and classification purposes. In a decision tree structure, the edges indicate a collection of features that lead to additional unique features in the groups while the leaves represent the various groups [28]. In our case, we are using this algorithm to classify our question categories. The decision tree starts with a root node and does not have any incoming nodes. Nodes that are outgoing from the root node are called decision nodes. In decision tree learning, the split points in a tree are identified through a greedy search [29]. We can understand this properly from the visual representation of Decision Tree in figure 3.7.

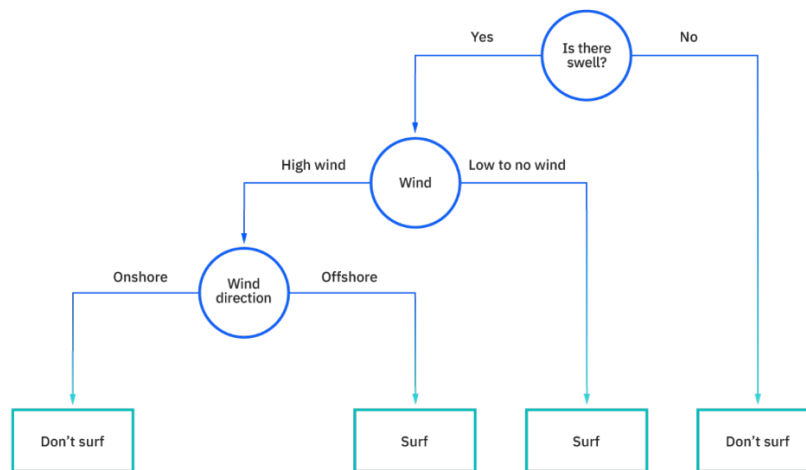


Figure 3.7: Decision Tree structure [31]

3.4.3.7 Support Vector Machine (Kernel = Linear)

Linear Kernel SVM is an algorithm used when there is a lot of features available in the dataset. As we are doing question classification here it has a lot of features. So, we are also using this algorithm to classify our questions. Figure 3.8 shows an example of Linear support vector machine algorithm.

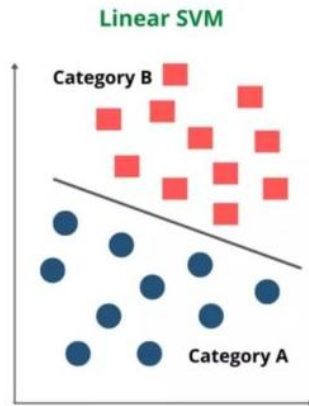


Figure 3.8: Linear SVM example [30]

3.4.3.8 Support Vector Machine (Kernel = rbf)

A fascinating algorithm for machine learning is the RBF kernel SVM. RBF means Radial Basis Function. This algorithm can classify data points divided by radial-based geometries like figure 3.9. The notable point about this algorithm is its capacity to precisely separate data points by hugging them together.

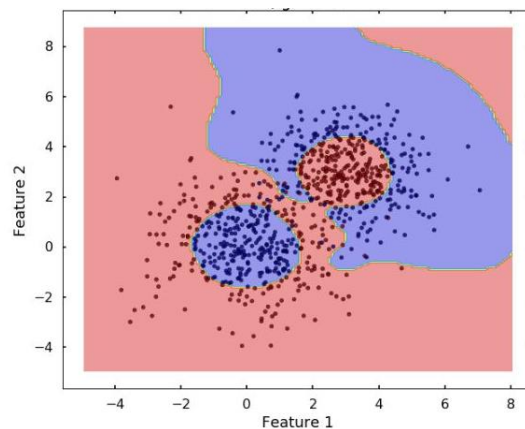


Figure 3.9: RBF SVM example [31]

3.4.3.9 Support Vector Machine (Kernel = Sigmoid)

Sigmoid kernel $K(x_i, x_j) = \tanh(ax_i^T * x_j + r)$, this requires two variables a and r . Here, a is the scaling parameter of the input data, while r is the shifting parameter that determines the threshold for mapping. The input data's dot-product is scaled and then inverted for a value of $a < 0$. The behavior

for various parameter combinations, which will be covered in the remaining paragraphs of this section, is summarized in the table below. The figure 3.10 show an example of Sigmoid kernel SVM example.

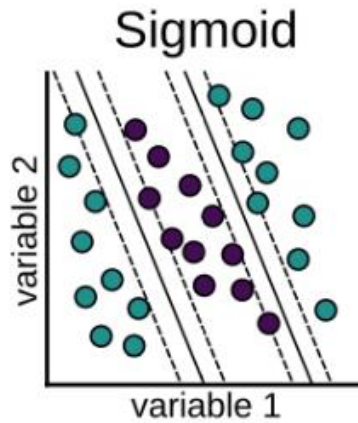


Figure 3.10: Sigmoid SVM Example [32]

3.4.3.10 Support Vector Machine (Kernel = Poly)

A polynomial kernel of SVM maps data into a higher-dimensional space using a polynomial formula. This is accomplished by taking the dot product of the polynomial function in the new space with the original space's data points. SVM kernels are based on polynomial functions that transform data into a higher-dimensional space. The polynomial function in the new space is then taken as the dot product of the data points in the original space. In svm classification applications when the data cannot be separated linearly, the polynomial kernel is frequently utilized [33]. The figure 3.11 represents the graphical representation of an example of Polynomial kernel SVM.

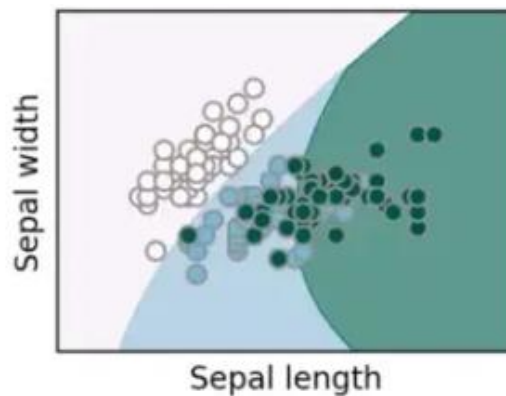


Figure 3.11: Poly SVM Example [34]

3.4.4 Model Evaluation

We cannot evaluate our implemented methods based on training and testing accuracy. To evaluate the effectiveness and reliability of any specific model, classification performance validation is required [35]. There are a few other parameters to consider to evaluate our implemented method such as the classification report which contains confusion matrix, f1 score, recall score, and precision score. These parameters will be discussed briefly below.

3.4.4.1 Classification Report

In a classification report, predictions made by a classification algorithm are evaluated according to their accuracy. In what proportions did the predictions prove correct and in what proportions did they fail to do so. It can be stated that categorization metrics are derived from the True Positives, False Positives, True Negatives, and False Negatives of a categorization report. [36].

3.4.4.1.1 Confusion Matrix

After collecting data, preprocessing it, and implementing them into a model first thing we try to do is see the result in a graphical form or numerical form. We try to evaluate the effectiveness and performance of that model, exactly confusion matrix comes into the limelight there. There can be two or more classes of data when using this method of performance measurement. This function produces a table of four different combinations of values, which include a True positive, a True negative, a False positive, and a False negative.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.12: Confusion Matrix Example [37]

3.4.4.1.2 Precision Score

A precision is calculated using the ratio tp and $(tp + fp)$. Here, tp represents the true positive and fp represents the false positive.

$$\text{Precision} = TP / (TP + FP)$$

3.4.4.2.3 Recall Score

As measured by the model recall score, the model is able to correctly predict positives among real positives. As opposed to precision, which counts the proportion of accurate positive predictions among all positive predictions made by the model, accuracy measures how accurate the prediction was. As an example, if we are predicting the positive review of any product then the recall score will count how many positive reviews is actually positive and how many positive reviews our model has predicted.

3.4.4.1.4 F1 Score

F1 scores are calculated based on precision and recall. A combined statistic based on the F1 score should be calculated for accuracy and recall. Additionally, F1 was designed to function effectively when there is a lack of balance in the data. F1 score formula:

$$F1 \text{ score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.5 Implementation Requirements

In our research, we are Classifying factoid question categories. To classify these categories, we are using a machine learning algorithm. We also used Natural Language Processing techniques to extract features. To run our code or system we can use any IDE or we can easily run the code from any online IDE such as Google Colab, Jupyter Notebook, etc. As these are online IDE so we do not need any high-spec computer to run our codes. from our dataset. All the hardware, software and advanced tools, libraries that we used to complete our work are mentioned below,

Recommended Hardware and Software:

1. Intel Core i3 8th gen processor.
2. 250gb of hard drive.
3. 8gb of ram.

4. Decent internet speed.
5. Windows 10.
6. Google Colab, Jupyter Notebook.

Tools and libraries:

1. Python 3.8.
2. Pandas.
3. Scikit-Learn
4. Seaborn.
5. Matplotlib

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

This section will cover the performance and efficiency of our model. To classify our factoid questions, we have used ten machine learning algorithms. All the algorithms performed pretty close to each other but some of them classified the question very accurately. Preprocessing was a very important part of our work. without preprocessing of dataset our accuracy would be very poor. So, we have preprocessed our data and vectorized the text data to train them through our models. Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly) came up with the accuracy of 87%, 73%, 88%, 82%, 87%, 85%, 87%, 86%, 79%, and 83% respectively. Table 4.1 shows all the model accuracy. We also generated a confusion matrix, and classification report and compare them to finalize our outcome.

Table 4.1: Model Accuracy table

Model Name	Accuracy
Naive Bayes (Multinomial)	87%
Naive Bayes (Gaussian)	73%
Logistic Model	88%
K-Nearest Neighbor Model	82%
Random Forest	87%
Decision Tree	85%
Support Vector Machine (Kernel = Linear)	87%
Support Vector Machine (kernel = rbf)	86%
Support Vector Machine (Kernel = sigmoid)	79%
Support Vector Machine (Kernel = poly)	83%

Among all the models we implemented Logistic model gives us the best performance with the accuracy of 88%.

4.2 Experimental Results and Analysis

This study compares machine learning techniques to see which characteristics and methods may get the best results for identifying the query category [38]. A total of ten machine learning algorithms have been used to categorize factoid questions. Machine learning algorithms can predict question categories for us but it is not 100% accurate. What we can do is try to make out the model as efficient as possible by preprocessing our dataset before feeding it to the model. We can tweak some values or methods to find out the perfect values and methods which will lead us to the highest possible accuracy out of our model. If we feed our models the text data then it will be very unoptimized and not efficient. As the computer works with numerical values so, Data from text forms was vectorized and converted to numerical values. First, we used TF-IDF vectorizer to vectorize our text data. We got a pretty good accuracy score with the TF-IDF vectorizer. Our Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly) came up with the accuracy of 85%, 72%, 87%, 81%, 85%, 80%, 86%, 86%, 79%, and 83% respectively. Then we used Countvectorizer and vectorized our text data and feed the data again to our models. Doing that tweak to our work we noticed some accuracy has slightly increased. Our Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly) came up with the accuracy of 87%, 73%, 88%, 82%, 87%, 85%, 87%, 86%, 79%, and 83% respectively.

We have also evaluated our model outcomes with a few other criteria such as f1 score, precision score, and recall score. We have generated a confusion matrix for all the classification models we implemented. The figure shows a graphical representation of the model performance. Among all the models we tested, Logistic Model had the best accuracy with 88.57% and a decent precision score, f1 score, and recall score. A confusion matrix based on the Multinomial Naive Bayes algorithm is shown in Figure 4.1. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 80, for the HUM category, the value of the true positive = 83, for the LOC category, the value of the true positive = 96, for the NUM category, the value of true positive = 107.

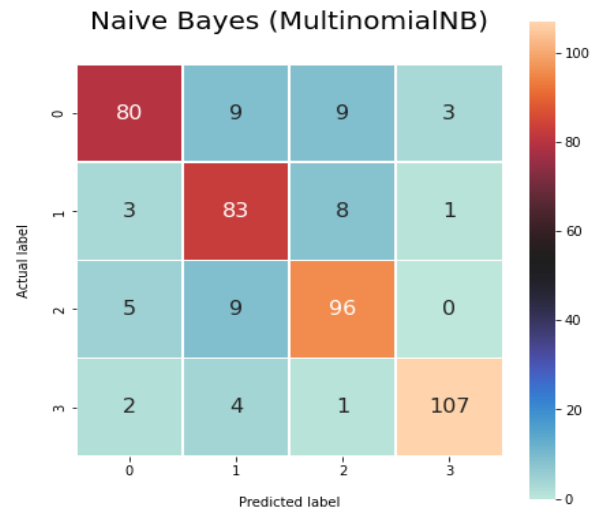


Figure 4.1: Confusion Matrix of Multinomial Naive Bayes algorithm

The confusion matrix of the Gaussian Naive Bayes algorithm we implemented is shown in figure 4.2. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 80, for the HUM category, the value of the true positive = 83, for the LOC category, the value of the true positive = 96, for the NUM category, the value of true positive = 107.

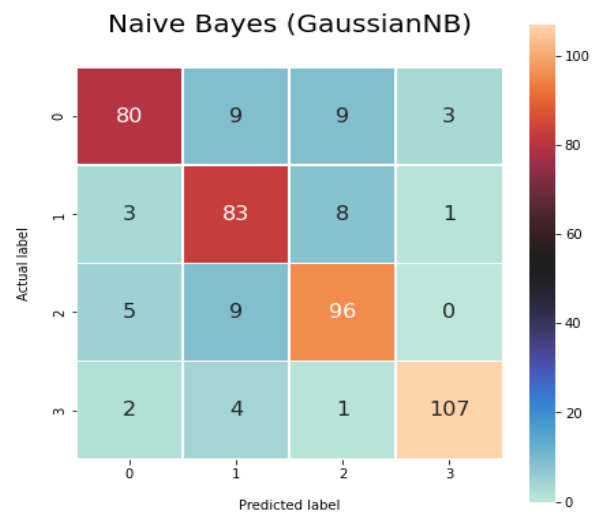


Figure 4.2: Confusion Matrix of Gaussian Naive Bayes algorithm

The confusion matrix of the Logistic Regression algorithm we implemented is shown in figure 4.3. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 86,

for the HUM category, the value of the true positive = 83, for the LOC category, the value of the true positive = 95, for the NUM category, the value of true positive = 108.

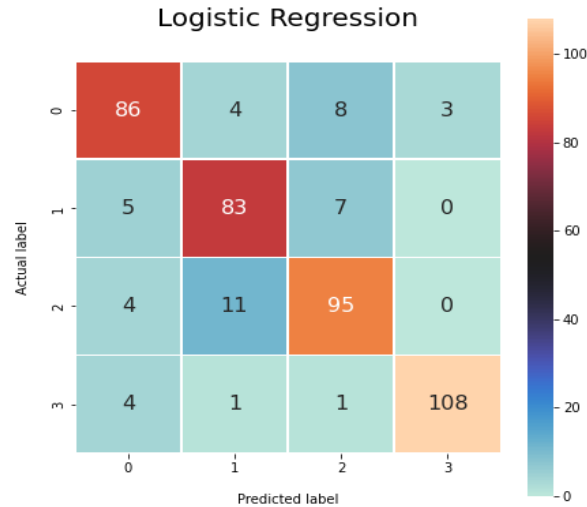


Figure 4.3: Confusion Matrix of Logistic Regression algorithm

The confusion matrix of the K-Nearest Neighbor algorithm we implemented is shown in figure 4.4. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 70, for the HUM category, the value of the true positive = 75, for the LOC category, the value of the true positive = 97, for the NUM category, the value of true positive = 103.

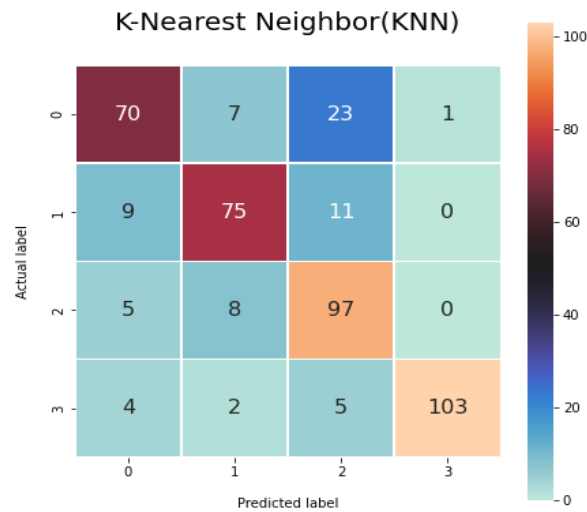


Figure 4.4: Confusion Matrix of K-Nearest Neighbor algorithm

The confusion matrix of the Random Forest algorithm we implemented is shown in figure 4.5. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 79, for the HUM category, the value of the true positive = 84, for the LOC category, the value of the true positive = 94, for the NUM category, the value of true positive = 109.

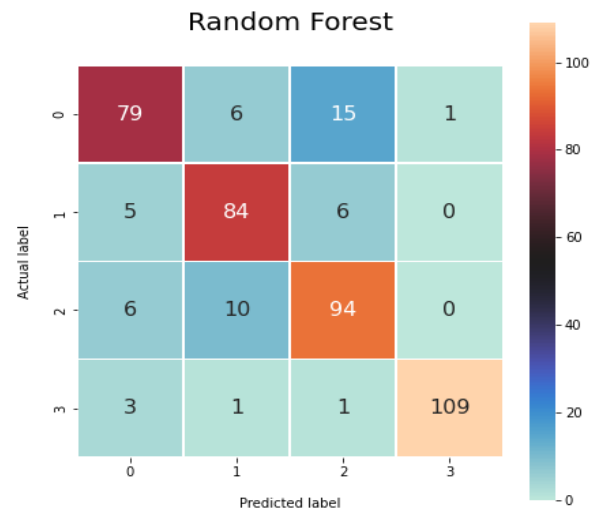


Figure 4.5: Confusion Matrix of Random Forest algorithm

The confusion matrix of the Decision Tree algorithm we implemented is shown in figure 4.6. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 76, for the HUM category, the value of the true positive = 83, for the LOC category, the value of the true positive = 87, for the NUM category, the value of true positive = 109.

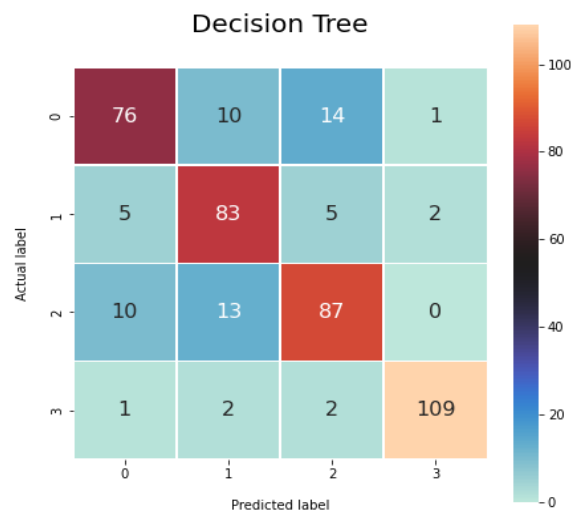


Figure 4.6: Confusion Matrix of Decision Tree algorithm

The confusion matrix of the Support Vector Machine (Kernel = Linear) algorithm we implemented is shown in figure 4.7. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 85, for the HUM category, the value of the true positive = 84, for the LOC category, the value of the true positive = 94, for the NUM category, the value of true positive = 104.

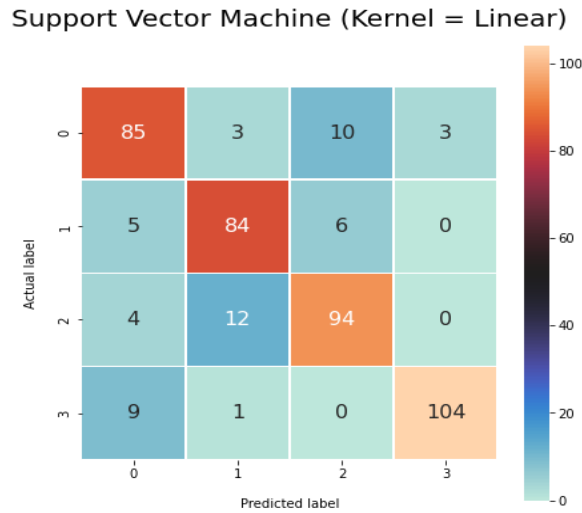


Figure 4.7: Confusion Matrix of Support Vector Machine (Kernel = Linear) algorithm

The confusion matrix of the Support Vector Machine (Kernel = rbf) algorithm we implemented is shown in figure 4.8. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 79, for the HUM category, the value of the true positive = 81, for the LOC category, the value of the true positive = 95, for the NUM category, the value of true positive = 108.

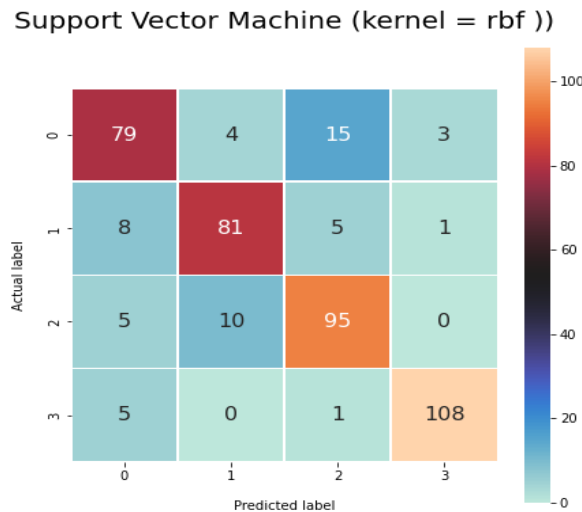


Figure 4.8: Confusion Matrix of Support Vector Machine (Kernel = rbf) algorithm

The confusion matrix of the Support Vector Machine (Kernel = Sigmoid) algorithm we implemented is shown in figure 4.9. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 86, for the HUM category, the value of the true positive = 71, for the LOC category, the value of the true positive = 79, for the NUM category, the value of true positive = 99.

Support Vector Machine (kernel = sigmoid)

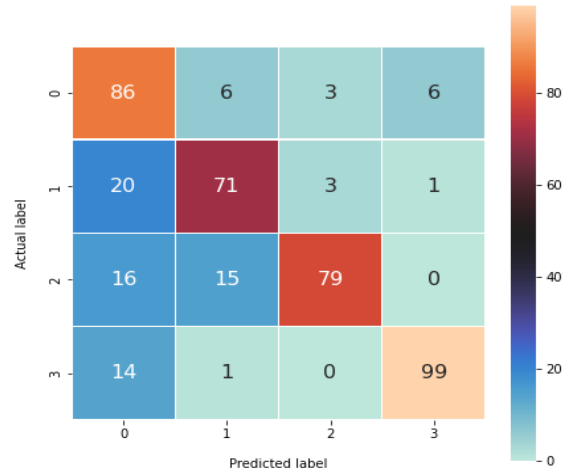


Figure 4.9: Confusion Matrix of Support Vector Machine (Kernel = Sigmoid) algorithm

The confusion matrix of the Support Vector Machine (Kernel = Poly) algorithm we implemented is shown in figure 4.10. From the confusion matrix, we can see that for ENTY category, the value of the true positive = 74, for the HUM category, the value of the true positive = 67, for the LOC category, the value of the true positive = 102, for the NUM category, the value of true positive = 108.

Support Vector Machine (kernel = poly)

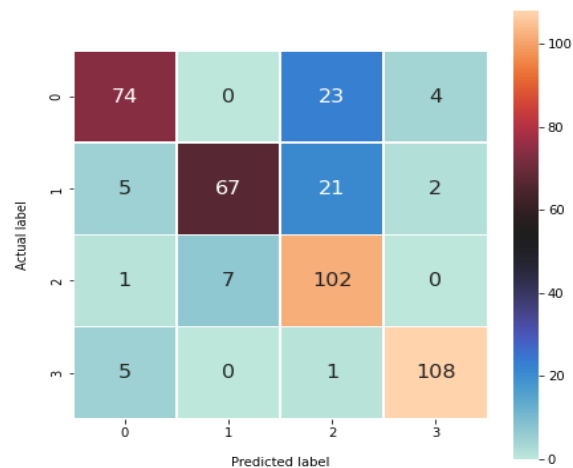


Figure 4.10: Confusion Matrix of Support Vector Machine (Kernel = Poly) algorithm

Table 4.2 shows the classification report for all the models we implemented.

Table 4.2: Classification Report all models

Algorithm Name	Class	Precision Score	Recall Score	F1 Score	Accuracy
Naïve Bayes (Multinomial)	ENTY	0.89	0.79	0.84	0.87
	HUM	0.79	0.87	0.83	
	LOC	0.84	0.87	0.86	
	NUM	0.96	0.94	0.95	
Naïve Bayes (Gaussian)	ENTY	0.88	0.77	0.86	0.73
	HUM	0.78	0.88	0.81	
	LOC	0.84	0.87	0.86	
	NUM	0.98	0.93	0.94	
Logistic Model	ENTY	0.87	0.85	0.86	0.88
	HUM	0.84	0.87	0.8	
	LOC	0.86	0.86	0.86	
	NUM	0.97	0.95	0.96	
K-Nearest Neighbors Algorithm	ENTY	0.80	0.69	0.74	0.82
	HUM	0.82	0.79	0.80	
	LOC	0.71	0.88	0.79	
	NUM	0.99	0.90	0.94	
Random Forest	ENTY	0.83	0.83	0.83	0.87
	HUM	0.87	0.84	0.86	
	LOC	0.83	0.87	0.85	
	NUM	0.98	0.96	0.97	
Decision Tree	ENTY	0.78	0.79	0.78	0.85
	HUM	0.82	0.88	0.85	

Algorithm Name	Class	Precision Score	Recall Score	F1 Score	Accuracy
	LOC	0.84	0.79	0.82	
	NUM	0.97	0.95	0.96	
SVM (Kernel - linear)	ENTY	0.83	0.84	0.83	0.87
	HUM	0.84	0.88	0.86	
	LOC	0.85	0.85	0.85	
	NUM	0.97	0.91	0.94	
SVM (Kernel - rbf)	ENTY	0.81	0.78	0.80	0.86
	HUM	0.85	0.85	0.85	
	LOC	0.82	0.86	0.84	
	NUM	0.96	0.95	0.96	
SVM (Kernel - sigmoid)	ENTY	0.63	0.85	0.73	0.79
	HUM	0.76	0.75	0.76	
	LOC	0.93	0.72	0.81	
	NUM	0.93	0.87	0.90	
SVM (Kernel - poly)	ENTY	0.87	0.73	0.80	0.83
	HUM	0.91	0.71	0.79	
	LOC	0.69	0.93	0.79	
	NUM	0.95	0.95	0.95	

4.3 Discussion

Our objective is to contribute to the classification of the question of the factoid questions about Bangladesh's research domain. Various types of languages have been the subject of much research worldwide. This is our small effort to add Bangladesh to that list of research work. We dreamed of developing a question-answering system about Bangladesh and this research work is half of our final work. So, we are very happy that we are almost ahead of our dream and that we have completed this research work. The purpose of this study is to develop a machine-learning model

capable of classifying questions about Bangladesh. We used ten different machine learning-based algorithms in our model and they are- Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (Kernel = Linear), Support Vector Machine (kernel = RBF), Support Vector Machine (Kernel = sigmoid), Support Vector Machine (Kernel = poly). The accuracies of all models are shown in Figure 4.2. Among these ten algorithms, the logistic regression model is able to classify the question of Bangladesh with the highest accuracy of 88%.

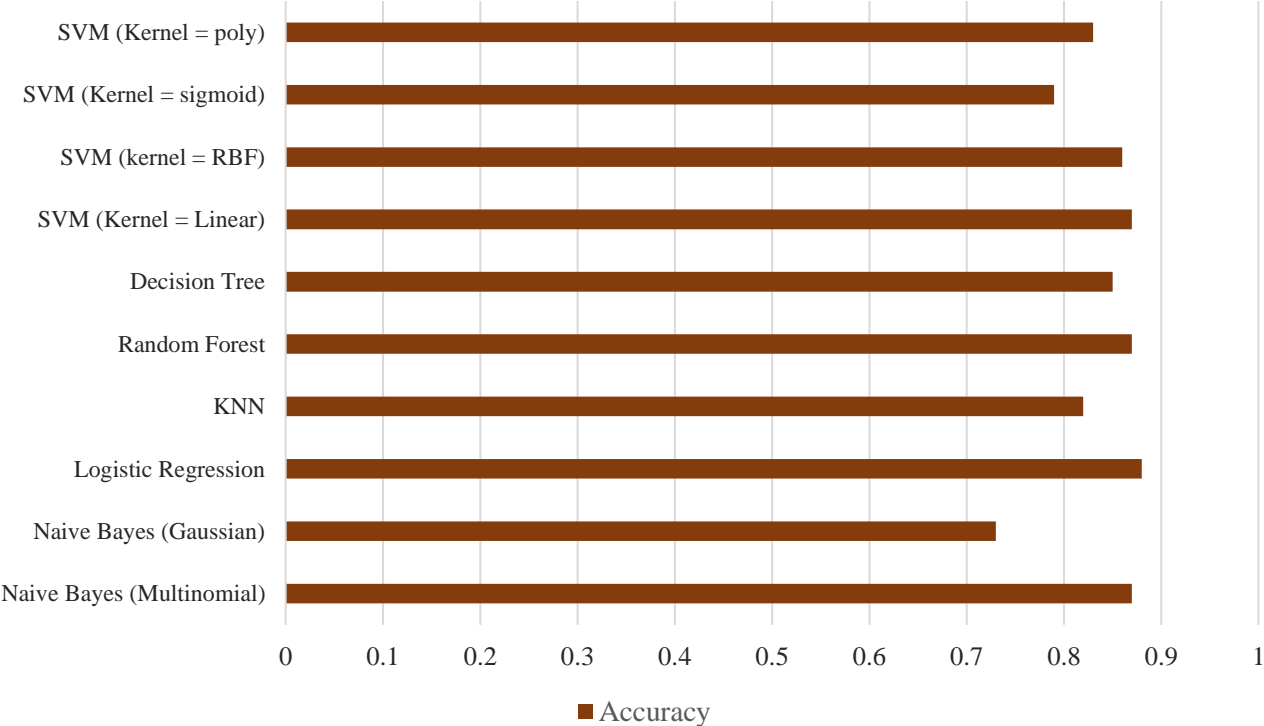


Figure 4.11: Algorithms Accuracy Bar Chart

Chapter: 05

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

It's very important to know about your country. We have worked on a model that is about the classification of factoid questions in Bangladesh. We have about 1400 questions about Bangladesh. People will know about our country. It will have a big impact on our society. People can learn a lot more about our country about our society such as our historical events in the time of kings, our historical liberation war, the progress of our country, our developing projects, our financial progress, our historical places, our own Bengali culture, our Bengali heritage, our Bengali celebration, our national affairs, the geographical knowledge of our country, etc. This thing will also attract foreigners as they will get to know about our dedication to doing such work for Bangladesh. Our society will gain a greater understanding of the country and a greater respect for it.

5.2 Impact on Environment

We believe that our project has the potential to have a significant meaningful impact on society. Then its effect on the environment will also be seen. We know that society is always made up of people around society and they all live in an environment. Since we did this project to categorize factoid questions about Bangladesh. Basically, it is a new job for people in society. They will know details about the country. Which will help in developing a deeper piece of knowledge about the country. For example, do we know when we fought for our liberation? Basically, we all know this answer but when someone asks us do we know, who was the commander of Sector 01 during the Bangladesh Liberation War? Most of us will be confused or not able to give the correct answer but the answer will be found in our project and answered by Major Ziaur Rahman. In the context of the environment, we can say that the work we do is significant.

5.3 Ethical Aspects

From our point of view, this work is our duty, love, and dedication to doing something for our country. As citizens of the country, we all have a duty to do something for the country. Which will improve the knowledge about our country. We have some ethical aspects to doing this work and they are-

1. Finding information about our subject is very difficult. Through this research work, freshers who actually want to work in this field will get some guidance and a good amount of resources from our research work.
2. Our country-based factoid question-answer-related work is very less. For this reason, we are very happy to do such research work. It will be useful to develop a factoid question answering system for Bangladesh in the future based on this work.
3. This research work will help to increase the knowledge about our country.

5.4 Sustainability Plan

The purpose of this research work is to spread knowledge of our country and encourage people to get involved. Due to a very busy lifestyle, people don't have much time to study about their country in depth. Therefore, we believe this research work will have a lasting impact on the people of our society. Day by day we will add more data to our dataset and this thing will create a huge data resource about our country. Our research is primarily concerned with gathering raw data about our country and transforming it into realistic questions. After generating the factoid question-answering dataset it needs to be preprocessed. We have the plan to turn this study into a factoid question-answering system for Bangladesh. This is one of the main reasons why we create models and categorize questions. This is why we want to sustain this model properly. We also need a step-by-step approach to how to sustain this model. We are working on it and we are pretty sure that our research work will survive in the environment.

Chapter: 06

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

ML is the focus of our research. We get natural language processing (NLP) if we delve deeper into this topic of research work. Dataset creation is a fundamental and mandatory step in the research process that allows the research work to continue. Using a factoid question dataset about Bangladesh, our research work is developing a machine learning-based model for classifying questions. Maintaining our workflow and getting to this stage was not easy because of the data collection part. We searched a lot but our domain-related dataset was nowhere to be found. In order to create our own customized dataset, we have created our own dataset. Where we will create factoid question answers with question types on Bangladesh subjects. All steps and job summaries are given below -

Step 01: Defining the problem.

Step 02: Books and websites pertaining to GK are used to collect data.

Step 03: From custom made dataset, preprocess the data.

Step 04: Feature extraction.

Step 05: Splitting the dataset into train and test data.

Step 06: Implementing models (Naive Bayes (Multinomial), Naive Bayes (Gaussian), Logistic Model, K-Nearest Neighbor Model, Random Forest, Decision Tree, Support Vector Machine (kernel = Linear), Support Vector Machine (kernel = rbf), Support Vector Machine (kernel = sigmoid), Support Vector Machine (kernel = poly)).

Step 07: Finding model accuracy.

Step 08: Generating the confusion matrix and other performance factors.

Step 09: Evaluating our final model.

After performing the steps, we were able to get our expected model which is able to classify our Bangladesh-related query. As a result of our research, we are able to classify questions regarding

the Bangladesh domain for factoid questions. Working on our country-related topic is our strength to complete this task. It is our pride to work on this subject and we have easily recognized our country in the global domain.

6.2 Conclusion

We found that there are many works on text classification in different domains. We will be able to build a factoid question answering system in the future by classification of questions related to Bangladesh in this machine learning domain. Among the algorithms we've used in our research, there are some which are particularly effective on text data from supervised learning. We chose to use ten different algorithms to build a model to classify our factoid questions about Bangladesh, such as naive bayes (multinomial), naive bayes (gaussian), logistic model, k-nearest neighbor model, random forest, decision tree, support vector machine (kernel = linear), support vector machine (kernel = rbf), support vector machine (kernel = sigmoid), support vector machine (kernel = poly). Since our research work requires data that is not previously available, creating a dataset is very difficult. Thereafter manually creating the dataset we needed, we created a dataset of 1800 records. After creating the dataset, we faced some problems and we chose 1400 data out of 1800 data to get better performance from the model used. Almost all classification algorithms perform very well but the logistic regression model performs very well in our case. It came with the highest accuracy and performance scores. We obtained 88% accuracy using this logistic regression model on our custom-made dataset. But our work has some limitations. Due to time and resource constraints, we could not generate a large dataset. If we are able to collect a large amount of data our model will give us more accuracy than we have now. All of us who work with ML know that the larger the dataset, the higher the classifier's accuracy.

6.3 Recommendation

Based on the research we have done; we have made some recommendations. To improve the accuracy of the model, we will increase the number of data points in our dataset. Some machine learning algorithms were used in our research. We used only a vectorization technique to convert the text data. There are other techniques and algorithms available for a large dataset as a result, the models and techniques will more accurately predict or classify questions from the factoid question-answering for the Bangladesh dataset. Below are some recommendations for our work-

1. Create a huge dataset for question classification for factoid questions and answers about Bangladesh.
2. Try to understand the factoid question answering patterns more precisely.
3. Understand other text conversion techniques and improve techniques on the datasets we use.
4. Attempt to implement a better classification model.
5. Enhance the accuracy of the performance.

6.4 Implication for Further Study

As a result of our research work, we have also encountered some limitations. For instance, we have utilized multinomial and Gaussian Nave Bayes algorithms, logistic models, K-nearest neighbor models, random forests, decision trees, support vector machines, etc. In addition to implementing all these algorithms, we have used count vectorized text conversion techniques. The dataset has a limitation in that we want to make our dataset bigger than the current dataset for more accuracy. Because we want to apply the BERT model and other necessary models to turn it into a factoid question-answering system. This task is not easy but we will use NLP to complete this task. But for this time, we are doing question classification which is a very important and mandatory part of the factoid question-answering method. In the classification part of the dataset questions, we got 88% accuracy from the logistic regression model and we will try our best to make this research work more accurate and make it a realistic question-answering system about Bangladesh. We would also like to add, we will make our dataset an open resource for NLP researchers. Using this dataset, they can try and we want them to get more accuracy. We have a future plan like when we develop our factoid question and answer system about Bangladesh and we will put this system on the website. People can know more about our country by using this website. This will be our achievement and dream.

REFERENCES

- [1] M. Razzaghnoori, H. Sajedi, and I. K. Jazani, "Question classification in Persian using word vectors and frequencies," *Cognitive Systems Research*, vol. 47, pp. 16–27, Jul. 2017, doi: 10.1016/j.cogsys.2017.07.002. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S1389041717300013>. [Accessed: Dec. 23, 2022]
- [2] A. Mohasseb, M. Bader-El-Den, and M. Cocea, "Classification of factoid questions intent using grammatical features," *ICT Express*, vol. 4, no. 4, pp. 239–242, Dec. 2018, doi: 10.1016/j.ict.2018.10.004. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405959518304491>. [Accessed: Dec. 17, 2022]
- [3] M. Mishra, V. K. Mishra, and S. H.R., "Question Classification using Semantic, Syntactic and Lexical features," *International journal of Web & Semantic Technology*, vol. 4, no. 3, pp. 39–47, Jul. 2013, doi: 10.5121/ijwest.2013.4304. [Online]. Available: <https://bit.ly/3Z25zAM>. [Accessed: Dec. 05, 2022]
- [4] F. Gürcan, "Multi-Class Classification of Turkish Texts with Machine Learning Algorithms," *IEEE Xplore*, Oct. 01, 2018. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8567307>. [Accessed: Dec. 05, 2022]
- [5] S. T. Alam Monisha, S. Sarker, and M. M. Hasan Nahid, "Classification of Bengali Questions Towards a Factoid Question Answering System," *IEEE Xplore*, May 01, 2019. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8934567>. [Accessed: Dec. 07, 2022]
- [6] M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, "Classification of online toxic comments using the logistic regression and neural networks models," *AIP Publishing*, vol. 2048, no. 1, Dec. 2018, doi: 10.1063/1.5082126. [Online]. Available: <https://aip.scitation.org/doi/abs/10.1063/1.5082126>. [Accessed: Dec. 07, 2022]
- [7] O. Zahour, E. Habib, A. Eddaoui, and O. Hourrane, "Automatic Classification of Academic and Vocational Guidance Questions using Multiclass Neural Network," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, pp. 550–556, 2019, doi: 10.14569/ijacsa.2019.0101072. [Online]. Available: <https://bit.ly/3IgmLN5>. [Accessed: Dec. 23, 2022]
- [8] Z. Cho, Y. Oo, T. Kyaw, H. Nwe, and H. Thant, "Question Classification for Automatic Question-Answering in Agriculture Domain," *JOURNAL OF INTELLIGENT INFORMATICS AND SMART TECHNOLOGY*, vol. 6, pp. 10–18, 2021. [Online]. Available: https://jiist.aiat.or.th/assets/uploads/16358531678571fGUM1635597555523j34iwques_ans.pdf. [Accessed: Dec. 05, 2022]
- [9] I. Lahbari, E. Alaoui, and K. Zidani, "Toward a New Arabic Question Answering System," *The International Arab Journal of Information Technology*, vol. 15, no. 3A, pp. 610–619, Apr. 2018. [Online]. Available: <http://iajit.org/PDF/Special%20Issue%202018,%20No.%203A/17416.pdf>. [Accessed: Dec. 23, 2022]

- [10] M. Keya, A. K. M. Masum, B. Majumdar, S. A. Hossain, and S. Abujar, "Bengali Question Answering System Using Seq2Seq Learning Based on General Knowledge Dataset," IEEE Xplore, Jul. 01, 2020. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9225605>. [Accessed: Dec. 11, 2022]
- [11] W. Meng, L. Lanfen, W. Jing, Y. Penghua, L. Jiaolong, and X. Fei, "Improving Short Text Classification Using Public Search Engines," Lecture Notes in Computer Science, vol. 8032, pp. 157–166, 2013, doi: 10.1007/978-3-642-39515-4_14. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-39515-4_14. [Accessed: Dec. 07, 2022]
- [12] S. M. H. Nirob, Md. K. Nayeem, and Md. S. Islam, "Question classification using support vector machine with hybrid feature extraction method," IEEE Xplore, Dec. 01, 2017. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8281790>. [Accessed: Dec. 11, 2022]
- [13] Y. Hou, C. Tan, X. Wang, Y. Zhang, J. Xu, and Q. Chen, "HITSZ-ICRC: Exploiting Classification Approach for Answer Selection in Community Question Answering," 2015 [Online]. Available: <https://aclanthology.org/S15-2035.pdf>. [Accessed: Dec. 07, 2022]
- [14] O. Aborisade and M. Anwar, "Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Jul. 2018, doi: 10.1109/iri.2018.00049. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8424720>. [Accessed: Dec. 11, 2022]
- [15] M. B. Rissan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," Indonesian Journal of Electrical Engineering and Computer Science, vol. 28, no. 1, p. 375, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383. [Online]. Available: <https://ijeecs.iaescore.com/index.php/IJEECS/article/view/26843>. [Accessed: Dec. 07, 2022]
- [16] S. W. Mohod, Dr. C. A. Dhote, and Dr. V. M. Thakare, "Modified Approach of Multinomial Naïve Bayes for Text Document Classification," csjournals, Apr. 02, 2015. [Online]. Available: <http://www.csjournals.com/IJCSC/PDF6-2/30.%20Mohod.pdf>. [Accessed: Dec. 05, 2022]
- [17] M. J. Meena and K. R. Chandran, "Naïve Bayes text classification with positive features selected by statistical method," IEEE Xplore, Dec. 01, 2009. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5378273>. [Accessed: Dec. 10, 2022]
- [18] B. Vimal and Dr. S. A. Kumar, "Application of Logistic Regression in Natural Language Processing," International Journal of Engineering Research and, vol. V9, no. 06, Jun. 2020, doi: 10.17577/ijertv9is060095. [Online]. Available: <https://www.ijert.org/application-of-logistic-regression-in-natural-language-processing>. [Accessed: Dec. 6, 2022]
- [19] Z. H. Kilimci and M. C. Ganiz, "Evaluation of classification models for language processing," IEEE Xplore, Sep. 01, 2015. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7276787>. [Accessed: Dec. 10, 2022]

- [20] “Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022,” upGrad blog, Jan. 03, 2021. [Online]. Available: <https://www.upgrad.com/blog/multinomial-naive-bayes-explained/#:~:text=The%20Multinomial%20Naive%20Bayes%20algorithm%20is%20a%20Bayesian%20learning%20approach.> [Accessed: Dec. 17, 2022]
- [21] “Gaussian Naive Bayes,” OpenGenus IQ: Learn Computer Science, Feb. 23, 2020. [Online]. Available: <https://iq.opengenus.org/gaussian-naive-bayes/>. [Accessed: Dec. 22, 2022]
- [22] M. Kirschner, R. Bernardi, M. Baroni, and L. T. Dinh, “Analyzing Interactive QA Dialogues Using Logistic Regression Models,” *AI*IA 2009: Emergent Perspectives in Artificial Intelligence*, vol. 5883, pp. 334–344, 2009, doi: 10.1007/978-3-642-10291-2_34. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-10291-2_34. [Accessed: Dec. 6, 2022]
- [23] “Logistic Regression,” Saedsayad.com, 2019. [Online]. Available: https://www.saedsayad.com/logistic_regression.htm. [Accessed: Dec. 19, 2022]
- [24] K. Shah, H. Patel, D. Sanghvi, and M. Shah, “A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification,” *Augmented Human Research*, vol. 5, no. 1, Mar. 2020, doi: 10.1007/s41133-020-00032-0. [Online]. Available: <https://link.springer.com/article/10.1007/s41133-020-00032-0>. [Accessed: Dec. 6, 2022]
- [25] A. Moldagulova and R. Bte. Sulaiman, “Using KNN algorithm for classification of textual documents,” *IEEE Xplore*, May 01, 2017. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8079924>. [Accessed: Dec. 11, 2022]
- [26] “What is the k-nearest neighbors’ algorithm? | IBM,” www.ibm.com. [Online]. Available: <https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%2C%20also%20know%20as%20KNN%20or.> [Accessed: Dec. 17, 2022]
- [27] “What is a Random Forest?” TIBCO Software. [Online]. Available: <https://www.tibco.com/reference-center/what-is-a-random-forest>. [Accessed: Dec. 22, 2022]
- [28] F. S. Gharehchopogh and Y. Lotfi, “Machine Learning based Question Classification Methods in the Question Answering Systems,” *International Journal of Innovation and Applied Studies*, vol. 4, no. 2, pp. 264–273, 2013 [Online]. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=71a78f3b9e9a84eddf3abf6ad78958f8b49422c1>. [Accessed: Dec. 11, 2022]
- [29] IBM, “What is a Decision Tree | IBM,” www.ibm.com. [Online]. Available: <https://www.ibm.com/topics/decision-trees>. [Accessed: Dec. 20, 2022]

- [30] K. Buvaneshwaran, "Support Vector Machine (SVM) In Machine Learning - CopyAssignment," copyassignment, Aug. 14, 2022. [Online]. Available: <https://copyassignment.com/support-vector-machine-svm-in-machine-learning/>. [Accessed: Dec. 20, 2022]
- [31] Data Warehousing and Data Science, "SVM with RBF Kernel," Data Warehousing and Data Science, May 24, 2021. [Online]. Available: <https://dwbi1.wordpress.com/2021/05/24/svm-with-rbf-kernel/>. [Accessed: Dec. 17, 2022]
- [32] A. Gupta, "Kernel Tricks in Support Vector Machines," Geek Culture, Jun. 01, 2021. [Online]. Available: <https://medium.com/geekculture/kernel-methods-in-support-vector-machines-bb9409342c49>. [Accessed: Dec. 19, 2022]
- [33] PyCodeMates, "SVM Kernels: Polynomial Kernel - From Scratch Using Python.," PyCodeMates. [Online]. Available: <https://www.pycodemates.com/2022/10/svm-kernels-polynomial-kernel.html>. [Accessed: Dec. 19, 2022]
- [34] H. Marius, "Multiclass Classification with Support Vector Machines (SVM), Kernel Trick & Kernel Functions," Medium, Sep. 09, 2020. [Online]. Available: <https://towardsdatascience.com/multiclass-classification-with-support-vector-machines-svm-kernel-trick-kernel-functions-f9d5377d6f02>. [Accessed: Dec. 19, 2022]
- [35] S. Fong, Y. Zhuang, K. Liu, and S. Zhou, "Classifying Forum Questions Using PCA and Machine Learning for Improving Online CQA," Communications in Computer and Information Science, vol. 545, pp. 13–22, Nov. 2015, doi: 10.1007/978-981-287-936-3_2. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-287-936-3_2. [Accessed: Dec. 22, 2022]
- [36] M. Krishnan, "Understanding the Classification report through sklearn," Muthukrishnan, Jul. 07, 2018. [Online]. Available: <https://muthu.co/understanding-the-classification-report-in-sklearn/>. [Accessed: Dec. 17, 2022]
- [37] Sarang Narkhede, "Understanding Confusion Matrix," Medium, May 09, 2018. [Online]. Available: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>. [Accessed: Dec. 11, 2022]
- [38] H. Priyambowo and M. Adriani, "Insincere Question Classification on Question Answering Forum," IEEE Xplore, Jul. 01, 2019. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8988798>. [Accessed: Dec. 05, 2022]

PLAGIARISM REPORT

FACTOID QUESTION CLASSIFICATION

ORIGINALITY REPORT

10%

SIMILARITY INDEX

9%

INTERNET SOURCES

6%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	mdpi-res.com Internet Source	2%
3	academic-accelerator.com Internet Source	1%
4	researchportal.port.ac.uk Internet Source	1%
5	Sourav Sarker, Syeda Tamanna Alam Monisha, Md Mahadi Hasan Nahid. "Bengali Question Answering System for Factoid Questions: A statistical approach", 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019 Publication	1%

Exclude quotes Off

Exclude matches < 1%

Exclude bibliography On