

**A MACHINE LEARNING APPROACH TO CLASSIFY SINGULAR AND PLURAL
NUMBERS OF BENGALI WORDS**

BY

MOHAMMAD ABU YOUSUF

ID: 172-15-9586 AND

SHARMIN SULTANA SONIA

ID: 172-15-9618

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

Supervised By

Sheikh Abujar

Senior Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Md. Tarek Habib

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JUNE 1, 2021

APPROVAL

This Project titled “**A Machine Learning Approach to Classify Singular and Plural Numbers of Bengali Words**”, submitted by Mohammad Abu Yousuf and Sharmin Sultana Sonia to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of M.sc in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on June 1, 2021.

BOARD OF EXAMINERS

Chairman



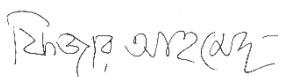
Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Dr. Fizar Ahmed

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology



Md. Azizul Hakim

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Engineering

Jahangirnagar University

External Examiner

DECLARATION

I hereby declare that, this project has been done by us under the supervision of, **Sheikh Abujar, Senior Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



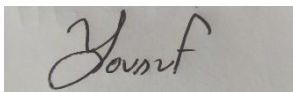
Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Mohammad Abu Yousuf
ID: 172-15-9586
Department of CSE
Daffodil International University



Sharmin Sultana Sonia
ID: 172-15-9618
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

To begin, we would like to express our heartfelt gratitude and gratitude to Almighty God for His celestial gift, which has enabled us to successfully complete the previous year's project.

Mr. Sheikh Abujar Sr. Lecturer, Department of CSE, Daffodil International University, Dhaka, deserves our heartfelt gratitude and our deepest gratitude. To complete this task, our administrator used his extensive knowledge and undeniable interest in the field of "NLP." His consistent perseverance, insightful direction, consistent consolation, consistent and enthusiastic oversight, productive analysis, critical guidance, perusing numerous subpar drafts and adjusting them at all stages made it possible to complete this project. We would like to express our gratitude to Md. Tarek Habib Assistant Professor, our co-supervisor, for his assistance throughout our research. I am really an appreciation to our head, Dr. Touhid Bhuiyan sir for his significant help to do such sorts of research work in the Bengali Language. Likewise, as to thank other employees and the staff of our area of expertise for their backings. We also want to thank our department's other faculty members and volunteers for their unwavering assistance and support. We would like to express our heartfelt gratitude to Almighty Allah and our supervisor for their thoughtful assistance in completing our task, as well as to other employees and the staff of CSE division of Daffodil International University. We might want to thank our entire Daffodil International University course mate who participated in this conversation while finishing the course work.

Finally, we must acknowledge our people's consistent assistance and patience.

ABSTRACT

This paper discusses a method for detecting the singular and plural number of Bengali sentences using context-sensitive grammar rules and NLP that accepts almost all types of Bangla words. The study is likely to result in the development of a system to digitize Bangladesh's vegetable classification system, allowing future generations to learn more about it. The purpose of this project is to classify the singular and plural number from Bengali words using a machine learning classification method. When the noun refers to a single item, the singular number is used, while the plural number is used when the word refers to multiple items. We have now created a model that can recognize singular and plural numbers in Bengali words using a machine learning classification technique. We encountered challenges when collecting Bangla data due to limited availability, but we eventually collected a total dataset of around 1000 words for this project. On 1000 data points from various groups, we achieved a maximum accuracy of 90.6%.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii-iii
Declaration	iv
Acknowledgements	v
Abstract	vi
List of Figure	ix
List of Tables	x
List of Abbreviation	xi
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Rational of the study	2-3
1.4 Research questions	3

1.5 Expected output	3
1.6 Report layout	4
CHAPTER 2: BACKGROUND STUDIES	5-13
2.1 Introduction	5
2.2 Related work	5-9
2.3 Comparative Analysis and Summary	9-12
2.4 Scope of the problem	12
2.5 Challenges	12-13
CHAPTER 3: RESEARCH METHODOLOGY	14-22
3.1 Introduction	14
3.2 Research subject and instrumentation	14-15
3.3 Data Collection	16
3.4 Statistical analysis	16-21
3.5 Implementation requirements	21-22
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	23-27
4.1 Experimental setup	23-25
4.2 Experimental results and analysis	25-27
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	28-29
5.1 Impact on Society	28
5.2 Impact on Environment	28
5.3 Ethical Aspects	29
5.4 Sustainability	29
CHAPTER 6: CONCLUSION AND FUTURE WORK	30-32
6.1 Summary of the study	30
6.2 Conclusion	31
6.3 Implication for further study	31-32
REFERENCES	33-35

LIST OF FIGURES

FIGURES	PAGE NO
Figure 2.1: Tf-Idf vector with dense neural network architecture	6
Figure 2.2 NLP text Classification flowchart [7]	7
Figure 2.3 Proposed model	13
Figure 3.1: Dataset pipeline	16
Figure 3.2: Data preprocessing	17
Figure 3.3: Classification for multinomial Naive Bayes	19
Figure 3.4: Classification for SVM	19
Figure 3.5: Classification for KNN	20
Figure 3.6: Classification for Decision Tree	21
Figure 3.7: Classification for Random Forest	21
Figure 4.1: Models Accuracy comparison	26
Figure 4.2: Accuracy of Train	23
Figure 4.3: Loss of Train	24
Figure 4.4: Accuracy of Validation	24
Figure 4.5: Loss of Validation	25

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: analysis and summary of the related work	9-12
Table 3.1 Bengali Dataset for Singular and Plural form	15

LIST OF ABBREVIATION

LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
NLTK	Natural Language Tools Kit
NLP	Natural Language Processing
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
NMT	Neural Machine Translation

CHAPTER 1

Introduction

1.1 Introduction

NLP is a branch of artificial intelligence (AI) that is an extremely important function in integrating linguistics and computational techniques to translate natural language people, for example. The term "natural language processing" (NLP) has been applied to a variety of scientific fields. Document classification, image classification, and voice classification are examples of technological development. Identification and so on.[1]

Bengali is the fifth most spoken local language in the world and the seventh spoken throughout the language overall. Around 228 million people are speaking Bengali as a first language, and more than 37 million people speak Bangla as a second language. Nouns and pronouns are inflected for case in a number of forms, including nominative, objective, genitive (possessive), and locative. The case marking pattern for each inflected noun is determined by its degree of animacy. Nouns are inflected for numbers when a definite article such as - -ta (singular) or - -gulo (plural) is inserted, as shown in the tables below. Cases are divided into six categories and an additional possessive case in most Bengali grammar books (possessive form is not recognised as a type of case by Bengali grammarians). Events, on the other hand, are usually divided into just four categories in terms of their applications.[3]

1.2 Motivation

One of the most critical aspects of NLP is identification. Speech recognition has been used in a variety of projects. Siri [5], Alexa [4], and other AI assistants have been created.

Bengali is the world's sixth most spoken language, according to Wikipedia. More than 210 million people speak Bengali [Britannica], but surprisingly, no substantial work is being done to address this problem. Some applications, such as spell checkers and grammar checkers, have been created to address these issues, but they are unable to distinguish between singular and plural. They are not enriched in any way. Not only do we need to speak about one and many, but the singular and plural forms of nouns are also significant. They're also important for subject-verb agreement, in which the verb in a sentence is determined by the grammatical number of the noun in the sentence's subject. Since there is never a single thing of anything in this universe, it's critical to learn how to form plural nouns. You'll always find yourself needing to refer to a group of items, and the only way to do so in speech or writing is to use plural nouns. If we don't know what the correct singular and plural words are, we won't be able to use them correctly in a sentence.

In light of this, we decided to develop a system that can distinguish between singular and plural numbers. The aim of this study is to create an efficient framework with a well-defined set of data that can create a machine that is controlled by humans relationship which benefits human beings and provides the best accuracy. As a result, we assume that a digital tool to verify if a term is singular or plural should be available by this project.

1.3 Rationale of the Study

Many studies on Natural Language Processing have been published, but the majority of them are in English. Several automated systems make use of these methods or mechanisms. There seem to be nearly 300 million Bangla speakers worldwide, but only a few scholars work on the language, and the majority of them are inaccurate. In the field of Natural Language Processing (NLP), Bengali number detection systems are scarce.

In recent years, voice-based applications have become increasingly popular. The applications could be used as an AI assistant, census surveys, banking, HR, and marketing, the healthcare industry, mass transit, auto texting, and crime reduction, among other things.

We're excited to work on improving the usage of Bengali accents in speech apps for the benefit of people in rural areas.

1.4 Research Questions

- What is NLP?
- To get the best result, which classification algorithm should be used?
- What methods can be used to extract the features?
- What is a singular plural number?
- What a Bengali number detective works?
- What are the benefits of Bengali number detection?
- How to preprocess Bengali number detection in NLP?
- How Bengali number detection Model work?
- What are the future works of Bengali number detection?

1.5 Expected Outcome

The research is expected to result in the creation of a system to digitize Bangladesh's vegetable classification system so that future generations can learn more about it. This form of technology will collaborate in a variety of ways, including:

- † This technique could help readers identify what they're reading and understand what they're reading by breaking down bangla grammar into several components.
- † This study can be applied to text classification in a number of ways. Classification of languages, words, and so on [15]. Bangladeshi literature could have benefited from it.

1.6 Report Layout

This report's **first chapter** addresses what we're going to do, why we're going to do it, and how we're going to do it. The motive for this project, as well as the expected outcome, is briefly mentioned in this chapter.

This sector's related work has been mentioned in **Chapter 2**. This section also includes a summary of their findings from their work. We set our targets by identifying their limitations and explaining the challenges.

This report's **third chapter** addresses the methods used in this project. This chapter also discusses some theoretical topics that are important to this project. This chapter briefly describes the information gathering process, data preprocessing, feature extraction, and methodologies used in this report.

In **Chapter 4**, the findings from the previous chapters are discussed, as well as a comparison and the best process.

The key subject of **Chapter 5** is the project's summary. Future work, conclusions, limitations, and recommendations are all listed in the report's final chapter.

CHAPTER 2

Background

This chapter will discuss works that are related to our project. Many studies on Bangla language classification, accent classification, spell checker, and word clustering have been conducted. However, our project is to detect singular plural number. In this chapter, we will provide a summary of those studies. The project model, as well as the strengths and weaknesses of those experiments, will also be explored.

2.1 Related Works

Abu Nowshed Chy et al. [2] proposed a technique of classification where they provide users news article classification. They used the Naive Bayes classifier to distinguish between news and news articles. They also used an RSS crawler to collect data before creating a Bangla lexicon and a Bengali stemmer, and finally running a Naive Bayes classifier. Ashwini Thota et al. [6] proposed a technique on how to find fake news. A story made up with the intent to deceive or mislead is known as fake news. Throughout this article, we'll look at they they show up a Deep Learning-based solution to the task of detecting fake news. On test results, their framework experimental results current model architecture by 2.5 percent and achieves a precision of 94.21 percent. They trained their model using three different neural network variations. They present a top standard overview of the network architecture. The following section delves into the details of the projects.

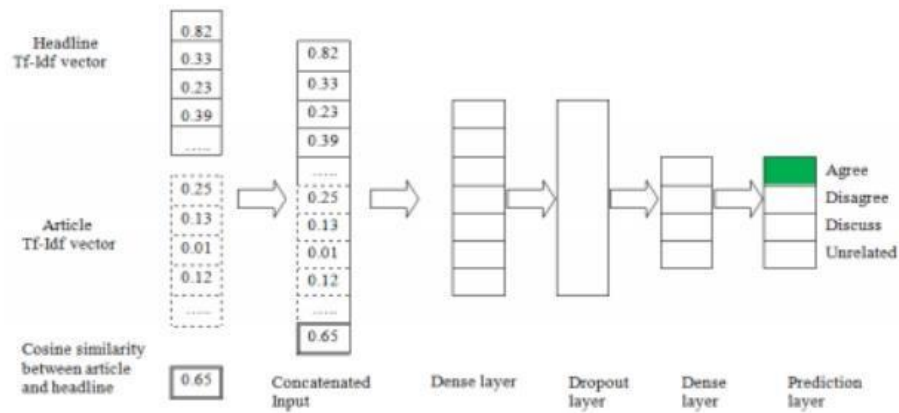


Figure 2.1: Tf-Idf vector with dense neural network architecture depicts the model architecture. To represent text, the BoW with dense network architecture employs a simplified vector space.

Md. Musfique Anwar et al. [7] The paper discusses a framework for analyzing Bangla sentences syntactically using context-sensitive grammar rules. The device works by processing input sentences and transforming them into structural representations (SR) Once the SR for a specific Bangla sentence is formed, the NLP conversion unit converts it to the corresponding English sentence. The usefulness of this approach has been demonstrated by demonstrating various Bangla sentences with 28 decomposition laws and the performance rates in all cases exceed 90%. Their proposed model step like first of all input bangla sentence then tokenizer then Syntax Analyzer/Parser then Grammar Rule Generator then Lexicon etc. Basically This study generates SR for Bangla sentences using parsing methodology in the following ways:

Begin with the input Bangla sentence. Determine the total number of Bangla tokens. Perform the following actions for each Bangla token: Lexicon will help you find the equivalent sections of speech and English sense of the token. Using the grammar rule generator, build the corresponding parse tree/SR. Using the Bangla to English converter, generate the corresponding English sentences.

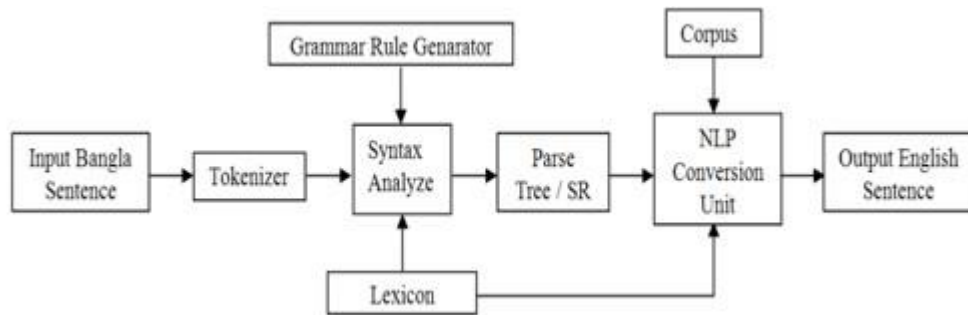


Fig 2.2: NLP text Classification flowchart [7]

Alok Ranjan Pal et al. [8] in their paper based on the meaning definitions of words available in the Bengali WordNet, an attempt is made to automatically classify the sentences into different classes based on their underlying senses. In their experiment, they achieved about 84 percent accuracy on meaning classification across all input sentences. This study's applicability is demonstrated by automated text classification, machine learning, knowledge extraction, and word sense disambiguation. They examined the residual sentences that did not comply with their experiment and had an effect on the results to discover that, in many cases, incorrect syntactic structures and a lack of semantic knowledge are the key barriers to semantic classification. Minhajul Abedin Shafin et al. [9] in this paper because of the COVID-19 pandemic, and because people are under lockdown, online shopping has become the primary outlet for shopping because it is the safest way.

More online product service providers improve things for customers, but it also raises concerns about product quality and service. Their aim is to create a framework that will evaluate consumer feedback from online shopping and provide a ratio of positive to negative feedback. SVM outperformed all other algorithms with the highest accuracy of 88.81 percent. Andrew McCallum et al. [10] in their paper compared different Naive Bayes classifier models in another project. They contrasted the Multivariate Bernoulli Model as well as the Multinomial Model. Each model performs differently depending on the type of data and the size of the data. The Bernoulli Model performed well in a few data sets, especially in small data sets. The Multinomial Model, on the other hand, worked well with

large scale datasets. In 2014, Andronicus A. Akinyelu et al. [11] suggested a new approach for text classification. Since people are being hacked every day via phishing emails, they have implemented a machine learning algorithm called random forest in their system. In this research the outcome was excellent, with a 99.7 percent accuracy score. Baoxun Xu et al. [12] suggested an improved random forest algorithm for text categorizing. Their proposed feature weighting and tree selection methods are an improvement over the random forest algorithm. They effectively minimize subspace size using the modern feature weighting method for subspace sampling and the tree selection method, and increase classification efficiency while keeping the error bound constant. They have collected six datasets, and all of their suggested solutions have been validated. Their improved random forest algorithm achieved 70-90 percent accuracy. Suresh Merugu et al. [13] recently proposed a supervised machine learning method for classifying text messages in 2018. Many supervised algorithms were used, including SVM, Random Forest, K Nearest Neighbor, and BernoulliNB. K Nearest Neighbor performed the worst of all, while Random Forest and BernoulliNB had the highest accuracy (nearly 98 percent). M. Ikonomakis et al. [14] explored a few machine learning strategies for text classification in a previous article. They provided comprehensive information about how an algorithm operates, how to prepare our dataset, and how to preprocess. They also defined the assessment of the results. Timothy P. Jurka et al. [15] addressed RTextTools, a new text classification tool for beginners. RTextTools allows you to identify any text in just ten simple measures. This paper discusses RTextTools from preparation to outcome assessment. "The first work on automated text summarization was done by Luhn [16] in 1958 based on term frequency, and the technique was expanded by Baxendale [17] by using cue terms and sentence location in the document.

These important contributions laid the groundwork for computerized text summarization, and since then, researchers have been willing to contribute in this field of Natural Language Processing. Isonuma et al. [18] also start a document classification task for single document summary. They tested their neural network-based model on documents from two financial news publishers. Because of its efficiency on sentence level classification problems,

Convolutional Neural Network (CNN) is used for sentence embedding rather than word embedding. Another neural network-based architecture, LSTM-RNN, is used to extract document summaries. Das et al. [19] create a topic-based opinion review for a Bengali text. They used an annotation tool to annotate sentences for description by highlighting the root words in order to differentiate the sentiment content. The sentiment terms are defined by the annotator based on their Part of Speech (POS) categories. For combining subject-sentiment, K-means clustering is used. Finally, a theme-based relational graph is used to select the summary sentence, and the page rank algorithm is used to recover information [20].

2.2 Comparative Analysis and Summary

Table 2.1: analysis and summary of the related work

Author	Methodology	Description	Outcome
Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das	naive Bayes classifier	Classifying Bangla news	78%
Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, Nibrat Lohia	Deep Learning- based	Fake News Detection	94.21%

Md. Musfique Anwar, Mohammad Zabed Anwar and Md. Al- Amin Bhuiyan	NLP conversion	Bangla sentences syntactically using context- sensitive grammar rules.	90%
Alok Ranjan Pal, Diganta Saha and Niladri Sekhar Dash	Automated text classification, machine learning, knowledge extraction.	Automatic Classification of Bengali sentences.	84%
Minhajul Abedin Shafin; Md. Mehedi Hasan; Md. Rejaul Alam; Mosaddek Ali	NLP and Machine Learning	Product Review Sentiment	88.81%

Mithu; Arafat Ullah Nur; Md. Omar Faruk			
Andrew McCallum and Kamal Nigam	Multivariate Bernoulli Model and Multinomial Model	Comparison between Multivariate Bernoulli Model and Multinomial Model	Multinomial Model performed 4.8% better Multivariate Bernoulli model

Andronicus A. Akinyelu and Aderemi O. Adewumi	Random Forest	Classifying phishing email from emails.	99.7%
Baoxun Xu, Xiufeng Guo, Yunming Ye and Jiefeng Cheng	Weighting method and tree selection method an improvement for random forest algorithm	They effectively reduce subspace size and increase classification efficiency without raising error bounds by using a new feature weighting method for subspace sampling and a tree selection method.	70-90% for six different datasets
Suresh Merugu, M. Chandra Shekhar Reddy, Ekansh Goyal and Lakshay Piplani	SVM, Random Forest, K Nearest Neighbor and BernoulliNB	I took a dataset of 5000 messages and used 90% of them for training and the rest for research. For classifying, various supervised Machine Learning algorithms were used.	SVM and Random Forest got accuracy 98% and BernoulliNB 97.6%
M. Ikonomakis, S. Kotsiantis and V. Tampaka	Different Machine learning algorithms	Different machine learning algorithms and techniques were discussed.	Discussed about algorithms

Timothy P. Jurka, Loren Collingwood, Amber E. Boydston, Emiliano Grossman, and Wouter van	RTextTools	They spoke about RTextTools, which allows you to quickly identify text in ten steps.	Discussed about RTextTools
---	------------	---	-------------------------------

2.3 Scope of the Problem

We try to do our best to detect Bengali singular and plural numbers through nlp. As far as we know, no NLP-based method has been used for Bengali number detection.

2.4 Challenges

- **Data Collection:** Data gathering is among the most major issues. The main challenge in Bengali Language is data collection. Since there are no coherent datasets for Bangla. This was also a significant challenge because we were unable to collect any type of data. As a result, we had to rely on numerous YouTube videos, Google, and other resources.
- **Collect the exact data:** As previously stated, we needed a unique type of data for this project, which was also the main challenge.
- **Model selection:** Despite the fact that a great deal of research has been done on NLP. Bengali Language research is still difficult. Because the majority of research has been conducted in English, many models have already been introduced. It was strenuous to choose the best layout for Bengali Language that would furnish the highest rate of exactness.

- **Proposed Workflow:** We had to test with a number of processes in order to determine the perfect procedure for this project. This project's workflow is as follows:

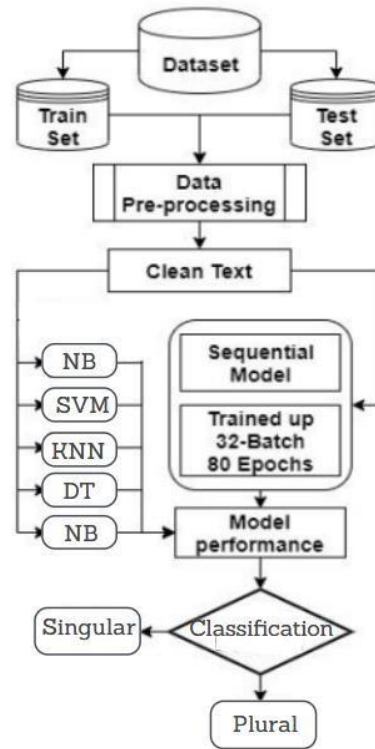


Fig 2.3: Proposed model

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

At around this stage, we'll go over the tools and methods we were using to collect data for our analysis. Natural language processing using machine learning necessitates a vast quantity of data, which is the most essential element of any type of research.

- 1) **Manually:** In the first approach, we collected data from Bangla grammar books (NCTB). But the problem was that it took too long and was difficult to find the information we needed.
- 2) **Online resources:** In order to collect better and more data, we began collecting data from YouTube videos in which we obtained various data.

3.2 Dataset Utilized

The most popular topic for determining their classification approach is the Bengali text of any sentence among singular and plural. The following table summarizes the output of the six classifiers listed above.

When the data collection is more robust to feed the reinforcement, machine learning pays close attention. The efficiency of classifiers is requested in our Bengali language dataset in singular and plural form. We prepared a dataset on behalf of using two forms of Bengali from using several online Bengali sources and NCTB approved textbooks. The dataset collection is pretty good for taking different types of words in both singular and plural forms, such as different news, fiction, historical, political, and so on. There are over 848 words in the learning dataset.

With machine learning classifiers, which perception of level is most convenient. Bengali researchers are looking forward to being able to process language in the future. The method of collecting data for this research project, as well as the methods and algorithms we used to make it simpler, will be demonstrated.

We used the preceding texts for this construction process:

- Singular
- Plural

Table 3.1: Bengali Dataset for Singular and Plural form

Text	Type	Text	Type
অধ্যাপক	একবচন	অধ্যাপক মন্ডলী	বহুবচন
অনুচর	একবচন	অনুচরবর্গ	বহুবচন
অনুষ্ঠান	একবচন	অনুষ্ঠান মালা	বহুবচন
কববতা	একবচন	কববতাগুচ্ছ	বহুবচন
গ্রন্থ	একবচন	গ্রন্থাবল	বহুবচন

3.3 Statistical Analysis

This section will go over the specifics of the dataset. We've already spoken about the volumes of information and the groups in our functioning repository.

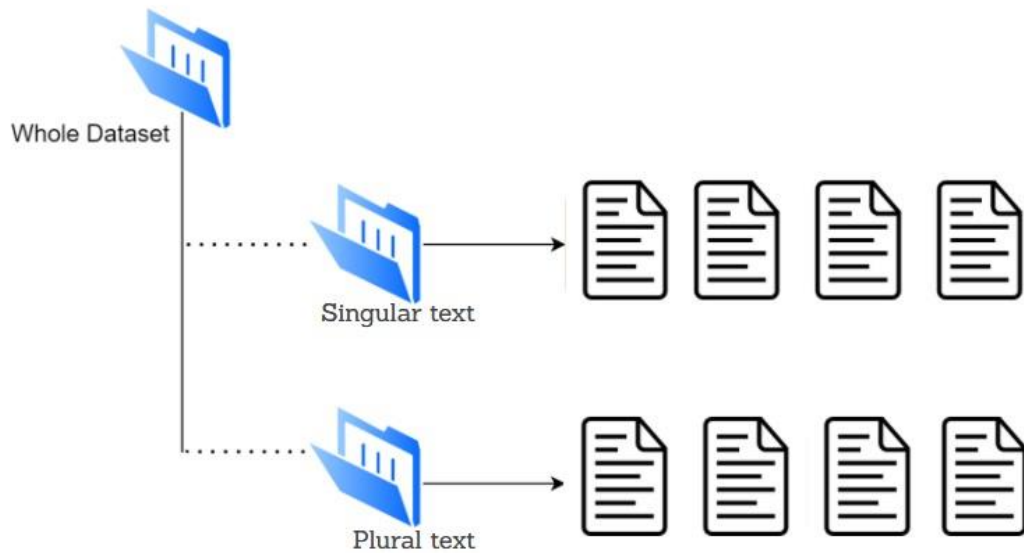


Fig 3.1: Dataset pipeline

3.4 Proposed Methodology

Simple Bayesian classifiers have recently gained popularity, and they have been shown to function admirably (Friedman 1997; Friedman et al. 1997; Sahami 1996; Langley et al. 1992). These probabilistic techniques construct a probabilistic model that includes substantial assumptions about how data is generated. They then estimate the parameters of the generative model using a set of labeled training examples. Bayes' rule is used to classify new instances by selecting the class that is most likely to have generated the example. Document classification is an example of a domain with several attributes. Words are the characteristics of the instances to be categorized, and the number of different words can be extremely big. While certain simple document classification tasks can be accomplished

with a vocabulary size of less than a hundred, many sophisticated jobs using real-world data from the Web, UseNet, and newswire articles require vocabulary sizes in the thousands.

This thesis adopted a protocol or technique to obtain the desired result. Classification of test documents can be accomplished using estimations of these parameters determined from training documents by calculating the posterior probability of each class given the evidence of the test document and picking the class with the highest probability. This is expressed using the following rule:

$$P(c_j|d_i; \hat{\theta}) = \frac{P(c_j|\hat{\theta})P(d_i|c_j; \hat{\theta}_j)}{P(d_i|\hat{\theta})}$$

By first substituting Equations 1 and 4 into the right-hand side, the right-hand side can be enlarged. Then, depending on the event model chosen, the expansion of individual terms for this equation is determined. For the multivariate Bernoulli event model, use Equations 2 and 3. For the multinomial, use Equations 5 and 6. Figure will provide a brief overview of our methodology.

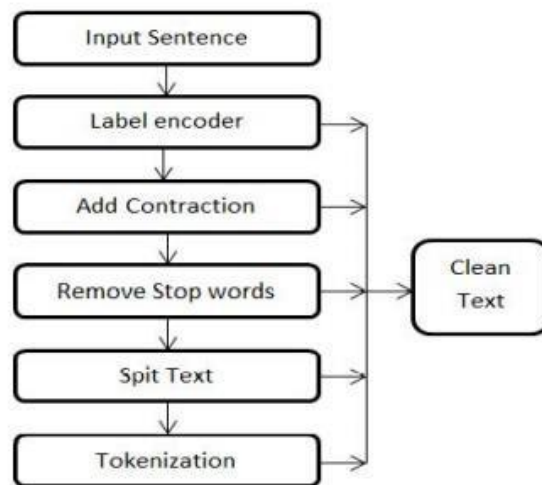


Fig 3.2: Data preprocessing

In a Bayesian learning framework, consider the task of text classification. This method assumes that the text data was generated by a parametric model and calculates Bayes-optimal estimates of the model parameters using training data. Then, armed with these estimates, it uses Bayes' rule to classify fresh test documents, reversing the generative model and calculating the posterior probabilities. Selecting the most likely class becomes the easiest part of the classification process. In all instances, a mixture model parameterized by θ is used to generate text documents. Mixture components c_j , $j = 1, \dots, |C|$ make up the mixture model.

A disjoint subset of θ is used to parameterize each component. As a result, a document, d_i , is formed by first picking a component based on priors, $P(c_j | \theta)$, and then having the mixture component build a document based on its own parameters, $P(d_i | c_j; \theta)$. A sum of total probability over all mixture components can be used to characterize the likelihood of a document:

$$P(d_i | \theta) = \sum_{j=1}^{|C|} P(c_j | \theta) P(d_i | c_j; \theta).$$

A class label is assigned to each document. We assume that classes and mixture model components have a one-to-one relationship, thus we use c_j to denote both the j th mixture component and the j th class. The traditionally "hidden" indicator variables for a mixture model are presented as these class labels in this situation (supervised learning from labeled training instances).

```

Score: 60.68 %
Classification Report:
              precision    recall  f1-score   support

     0           0.48       0.45       0.47         44
     1           0.68       0.70       0.69         73

 accuracy          0.61         117
 macro avg          0.58         117
 weighted avg       0.60         117

```

Fig 3.3: Classification for multinomial Naive Bayes

Support-vector machines (SVMs, also known as support-vector networks) are supervised learning models that examine data for classification and regression analysis in machine learning. Vladimir Vapnik and colleagues at AT&T Bell Laboratories developed it (Boser et al., 1992, Guyon et al., 1993, Vapnik et al., 1997). SVM maps training examples to points in space in order to widen the distance between the two categories as much as possible. New examples are then mapped into the same space and classified according to which side of the gap they fall on.

```

Score: 88.03 %
Classification Report:
              precision    recall  f1-score   support

     0           0.79       0.93       0.85         44
     1           0.95       0.85       0.90         73

 accuracy          0.88         117
 macro avg          0.87         117
 weighted avg       0.89         117

```

Fig 3.4: Classification for SVM

K Nearest Neighbor (KNN) is a machine learning technique that is basic, easy to grasp, and versatile. Finance, healthcare, political science, handwriting detection, picture recognition, and video recognition are just some of the uses for KNN. Financial institutions will forecast a customer's credit rating through credit ratings. Financial institutions will predict whether a loan is safe or dangerous when it is disbursed. In political science, potential voters are

divided into two groups: those who will vote and those who will not vote. Both classification and regression issues are solved using the KNN algorithm. The KNN algorithm is based on feature similarity. KNN is a slow and non-parametric learning method. The term "non-parametric" refers to the absence of any assumptions about the underlying data distribution. To put it another way, the model structure is based on the dataset.

```

Score: 86.32 %
Classification Report:
              precision    recall  f1-score   support

     0           0.78       0.89       0.83         44
     1           0.93       0.85       0.89         73

 accuracy              0.86         117
 macro avg           0.85       0.87       0.86         117
 weighted avg        0.87       0.86       0.86         117

```

Fig 3.5: Classification for KNN

Decision Tree is a supervised learning technique that may be used to solve both classification and regression problems, however it is most commonly employed to solve classification issues. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The Decision Node and the Leaf Node are the two nodes of a Decision tree. Leaf nodes are the output of those decisions and do not contain any more branches, whereas Decision nodes are used to make any decision and have several branches. The decisions or tests are made based on the characteristics of the given dataset. It's a graphical depiction for obtaining all feasible solutions to a problem/decision depending on certain parameters. It's termed a decision tree because, like a tree, it starts with the root node and grows into a tree-like structure with additional branches.

```

Score: 84.62 %
Classification Report:
      precision    recall  f1-score   support

     0       0.73      0.93      0.82         44
     1       0.95      0.79      0.87         73

 accuracy          0.85         117
 macro avg          0.84      0.86      0.84         117
 weighted avg          0.87      0.85      0.85         117

```

Fig 3.6: Classification for Decision Tree

A random forest is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting a number of decision tree classifiers on various sub-samples of the dataset. If `bootstrap=True` (default), the sub-sample size is regulated by the `max samples'` argument; otherwise, the entire dataset is utilized to create each tree. In the form `class label: weight`, weights are associated with classes. All classes are intended to have weight one if it isn't stated. A list of dicts in the same order as the columns of `y` can be provided for multi-output issues.

```

Score: 90.6 %
Classification Report:
      precision    recall  f1-score   support

     0       0.85      0.91      0.88         44
     1       0.94      0.90      0.92         73

 accuracy          0.91         117
 macro avg          0.90      0.91      0.90         117
 weighted avg          0.91      0.91      0.91         117

```

Fig 3.7: Classification for Random Forest

3.5 Implementation Requirements

Python 3.8:

The Python programming language is a high-level programming language [16]. It has recently been popular for Data Science and AI implementation, and it may also be utilized for online and mobile application development. It is accepted by all types of developers due to its simple structure and grammar. These benefits encourage us to use this language in our job. And, of course, its sizable community is constantly behind it.

Anaconda 4.5.12:

This is the free and open-source Python distribution [17]. This is also possible with the R programming language. This is a pack installer, not a standalone program. By introducing anything, it introduces a slew of crucial information science instruments. It even goes with the concept of a virtual climate. We can disconnect several tasks from one another so that we can use different requirements for each of them. We used the 4.5.12 version of boa constrictor, which was the most recent version at the time.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

In the earlier section., we finished data collection, preprocessing, feature extraction, and established a procedure to reach our vision. We completed our data arrangement, preprocessing, included extraction, and established a method to deal with show up at the target in the previous segment. According to the system, we also need to use RF to get better results, so we'll look at these three methods in this section. In our dataset, we focused on the subjective forest area. We used 80% of the data to generate a ready-to-use explanation and 20% for testing.

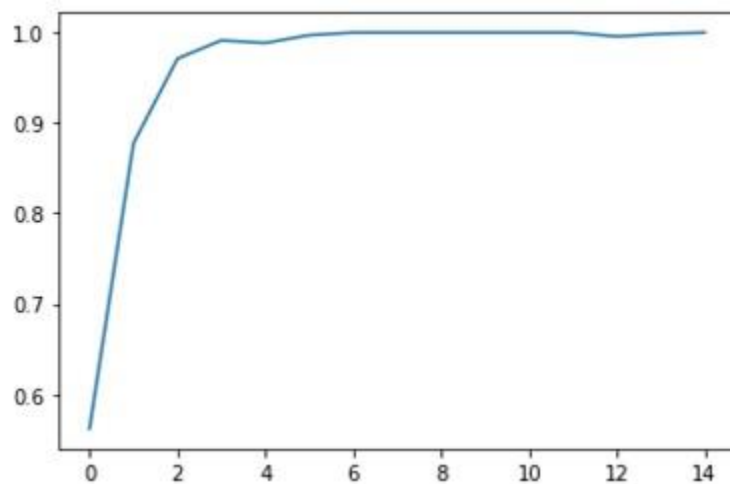


Figure 4.2: Accuracy of Train

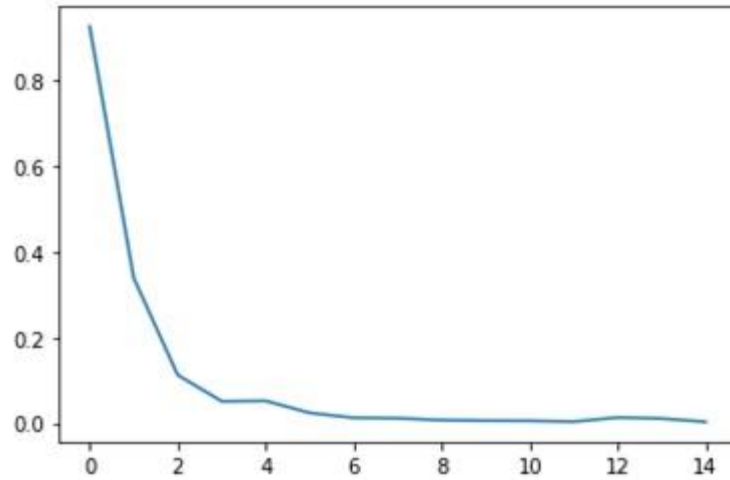


Figure 4.3: Loss of Train

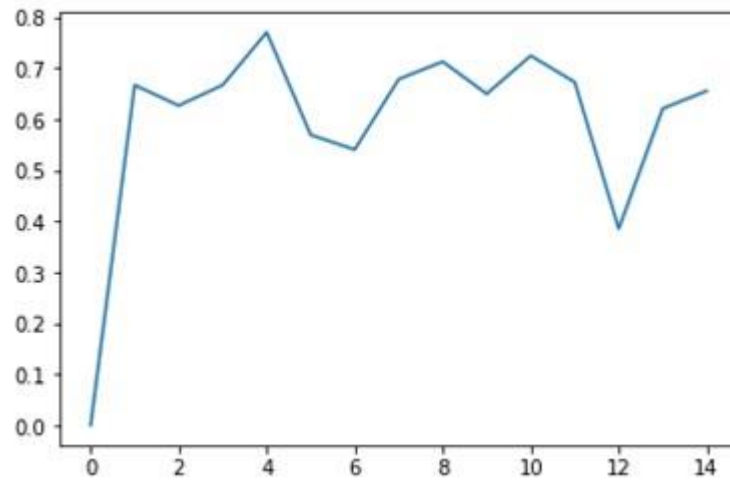


Figure 4.4: Accuracy of Validation

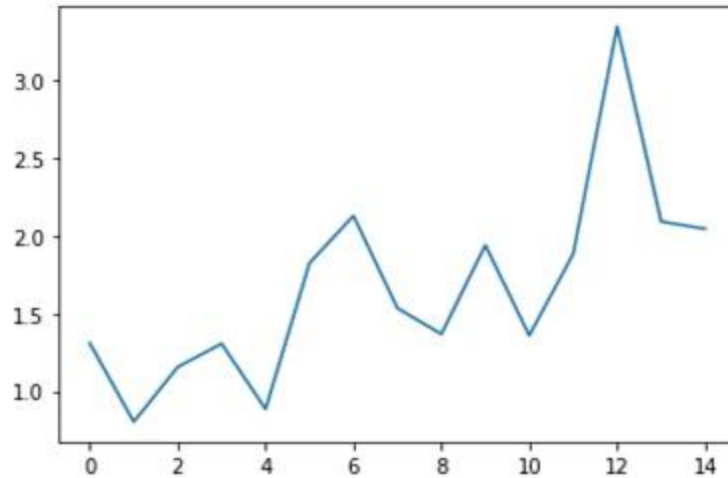


Figure 4.5: Loss of Validation

So, from above figures we got train accuracy 91% and validation accuracy near 73%

4.2 Experimental Results & Analysis

Feature selection is done when lowering the vocabulary size by choosing words with the highest average mutual information with the class variable (Cover and Thomas). This strategy has been utilized frequently with text (Yang and Pederson; Joachims; Craven et al.). This has been done in all previous work that we are aware of by calculating the average mutual information between the (1) document class and (2) the lack or presence of a word in the document, i.e. utilizing a document event model, the multivariate model. Bernoulli's theorem. We write C for a random variable spanning all classes, and W_t for a random variable spanning the lack or presence of the word w_t in a document, where W_t takes on the values 0 and 1, with $w_t = 0$ indicating the absence of w_t and $w_t = 1$ indicating the presence of w_t . The difference between the entropy of the class variable, $H(C)$, and the entropy of the class variable conditioned on the absence or presence of the word, $H(C|W_t)$, is the average mutual information (Cover and Thomas). We tested this strategy as well as an event model that corresponds to the multinomial: computing the mutual information between (1) the document class from which a word occurrence is taken and (2) a random

variable across all word occurrences. The incidents are referred to as occurrences. This section presents empirical evidence that the multinomial event model outperforms the multivariate Bernoulli model in most cases. Five separate data sets were used to generate the results. With only 800 words, the multivariate Bernoulli event model achieves a maximum accuracy of 91%. It's worth noting that the multivariate Bernoulli performs best with a smaller vocabulary than the multinomial, while the multinomial performs better with a bigger vocabulary.

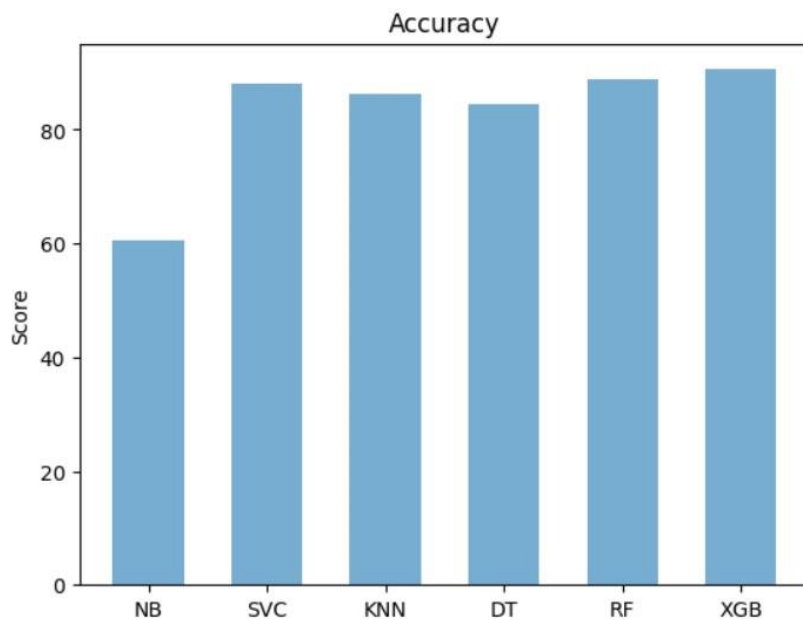


Fig 4.1: Models Accuracy comparison

A minimal vocabulary is sufficient for great performance on simple categorization problems. The Reuters categorization jobs are an example of this—it is well-known that high accuracy can be achieved in several of the categories with only a few words, occasionally even the single word that represents the category title. Our findings support this, indicating that modest vocabulary sizes often yield the best results. Many real-world

classification jobs lack this characteristic (i.e., all papers in a category are about a single narrow subject with limited vocabulary), instead consisting of a variety of topic areas with overlapping vocabularies. For appropriate classification accuracy in such jobs, vast vocabulary is required. Because our findings reveal that the multivariate Bernoulli model struggles with huge vocabularies, the multinomial event model is a better fit for these difficult classification tasks.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Language is vital in every civilization, nation, and to the people of a country. Language may be expressed in two ways: by speaking and by writing. Writing is the component of a language that has been documented. When we write something, it is a means for us to reflect our language, and therefore spelling of every word, phrase is also a very significant element of our language. Bangla is our national language. As a result, it is critical to be well-versed in Bangla grammar. We always utilize Bangla for different reasons, and we get confused about which is singular and which is plural. And we are unable to use the right form in a variety of situations. As a result, our project "classify singular and plural number" will have a significant influence on our society.

5.2 Impact on Environment

Our research will have no direct influence on the environment. We cannot argue that we have a direct influence on the environment, but as we all know, a nation's culture is frequently linked to its environment, therefore we have a positive influence on the environment as well. Culture, on the other hand, is a component of our surroundings. And our language, Bangla, is a vital and old part of our culture. It is the language that we have been using for hundreds of years. Our task is to categorize Bengali words into singular and plural forms. In this epidemic condition, everybody who can learn from their own home will benefit our society and the environment. People will be able to maintain social distance as a result. Furthermore, people can save time. This will be beneficial to the environment. We did not utilize any equipment or gear that could hurt the environment. Our project will help the entire human race.

5.3 Ethical Aspects

This proposed model is not morally repugnant and does not violate basic liberties in this way. This model does not collect personal information such as name or IP address. As a result, there is no need for protection. People can check easily. In that circumstance, there is a possibility of data purification. People may be concerned about the security of their personal information. In this regard, we shall not make our data public. People will only be able to see the information they have supplied. They will not be able to see the info of others. The ethical part of our initiative is to educate individuals on the difference between single and plural numbers. To learn and apply the proper number in everyday situations. We want individuals to be able to accurately utilize their preferred term while writing essential documents such as government documents, property related paperwork, and so on.

5.4 Sustainability

In this context, we attempted to classify Bengali word singular and plural numbers using several methods. Our project's work will not end here, and we have a future strategy for it. This model has several limitations, such as operating with a tiny dataset and a restricted amount of words. The model, on the other hand, is intended for future growth. Because research is a never-ending process. As a consequence, this model will be built on a daily basis for the Bengali language. Any job need more investigation in order to discover an appropriate answer. Then, after much investigation, an appropriate solution to a particular problem is uncovered. As a result, new research effort must be implemented or developed. The previous work's restrictions have an impact on the future implementation. Resolving the shortcomings of prior work adds to the development of an efficient system. The next phase in this research will be to extend the Bengali word dataset. We must always try to make things better. Furthermore, there are other aspects in this project that should be improved.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

Our entire project revolves around Bengali NLP. We created a deep learning model for Bengali number detection in this project. That is extremely useful. We completed this project in a timely manner. The entire project is divided into sections. The project's overall summary is provided below, with step-by-step instructions.

Step 1: Data collection form NCTB

Step 2: Summarize the collected data

Step 3: Data preprocessing

Step 4: Vocabulary count

Step 5: Add special token

Step 6: Define Encoder and Decoder with LSTM

Step 7: Build sequence to sequence model

Step 8: Train model

Step 9: Check the result analysis the response of the machine

This model will aid our Bengali NLP research community in developing a fully dependent automatic detective number as well as furthering our understanding of Bengali singular plural detection. Now we will talk about the future work and the conclusion of this research project.

6.2 Conclusions

The primary goal of this research is to expand and develop the Bengali NLP research area. We used the Bengali word as the model's input and generated a detective singular or plural Bengali word as the model's output. Normally, encoders and decoders function in the same way for detection. For Bengali text, our dataset is not large. However, the machine provides excellent results for this dataset. This model was created to classify Bengali singular plurals. We entered our data, and our machine told us whether the word was a singular or plural number. This is our model's main limitation. This is an additional research area for Bengali singular plural classification. As a result, a preprocessing library for Bengali text must be created. After all, no machine produces completely accurate results. Every machine has some constraints in its working field. Likewise, our detection model has some limitations. The important thing is that the model can generate a detector for the Bengali language. This is a significant accomplishment for our Bengali NLP field, and it will be useful for future research.

6.3 Implication for Further Study

This model has some limitations, such as working for a limited number of words and a small dataset. However, the model is designed for future expansion. Because research is a continuous process. As a result, this model will be developed for the Bengali language on a daily basis. Any work requires additional research to find a suitable solution. Then, after conducting extensive research, a suitable solution to a specific problem is discovered. As a result, future implementation or development of research work is required. The limitations of the previous work influence the future implementation. Solving the limitations of previous work contributes to the creation of an efficient system. The next step in this project will be to expand the dataset of Bengali words. We must always strive to improve the situation. In addition, in this project, there are several items that could be enhanced. The following are a few ideas for future projects:

- Increasing the number of data

- Attempting additional classifiers and comparing them.
- Besides enriching Bangla grammar.

Reference:

[1]. Salau A.O., Olowoyo T.D., Akinola S.O. (2020) Accent Classification of the Three Major Nigerian Indigenous Languages Using 1D CNN LSTM Network Model. In: Jain S., Sood M., Paul S. (eds) Advances in Computational Intelligence Techniques. Algorithms for Intelligent Systems. Springer, Singapore.

[2]. Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das, “Bangla News Classification using Naive Bayes classifier”, 16th Int'l Conf. Computer and Information Technology, 810 March 2014, Khulna, Bangladesh.

[3]. “Bengali language, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Bengali_language

[Last accessed: 1- May- 2020]

[4]. “Alexa” [Online]. Available: <https://www.alexa.com/> [Last accessed: 1- May- 2020]

[5]. “Siri” [Online]. Available: <https://www.apple.com/siri/> [Last accessed: 1- May- 2020]

[6]. Aswini Thota , Priyanka Tilak , Simrat Ahluwalia , Nibrat Lohia “Fake News Detection: A Deep Learning Approach”, SMU Data Science Review, Vol. 1 [2018], No. 3, Art. 10.

[7]. Md. Musfique Anwar, Mohammad Zabed Anwar and Md. Al-Amin Bhuiyan “Syntax Analysis and Machine Translation of Bangla Sentences”, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009.

[8]. Alok Ranjan Pal , Diganta Saha and Niladri Sekhar Dash “AUTOMATIC CLASSIFICATION OF BENGALI SENTENCES BASED ON SENSE DEFINITIONS PRESENT IN BENGALI WORDNET” ,International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.1, January 2015 .

© Daffodil International University

[9]. Minhajul Abedin Shafin; Md. Mehedi Hasan; Md. Rejaul Alam; Mosaddek Ali Mithu; Arafat Ullah Nur; Md. Omar Faruk "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language", 2020 23rd International Conference on Computer and Information Technology (ICCIT).

[10]. Andrew McCallum and Kamal Nigam," A Comparison of Event Models for Naive Bayes Text Classification," Published 1998.

[11]. Andronicus A. Akinyelu and Aderemi O. Adewumi, " Classification of Phishing Email Using Random Forest Machine Learning Technique," Journal of Applied Mathematics Volume 2014, Article ID 425731, 6 pages.

[12]. Baoxun Xu, Xiufeng Guo, Yunming Ye and Jiefeng Cheng " An Improved Random Forest Classifier for Text Categorization," journal of computers, vol. 7, no. 12, December 2012.

[13].Suresh Merugu, M. Chandra Shekhar Reddy, Ekansh Goyal and Lakshay Piplani, "Text Message Classification Using Supervised Machine Learning Algorithms," International Conference on Communications and Cyber Physical Engineering 2018, ICCCE 2018: ICCCE 2018 pp 141-150.

[14]. M. Ikonomakis, S. Kotsiantis and V. Tampaka ," Text Classification Using Machine Learning Techniques," WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966- 974.

[15].Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt," RTextTools: A Supervised Learning Package for Text Classification" The R Journal Vol. 5/1, June ISSN 2073-4859.

[16]. Luhn, Hans Peter. "The automatic creation of literature abstracts." IBM Journal of research and development 2.2 (1958): 159-165.

- [17]. Baxendale, Phyllis B. "Machine-made index for technical literature—an experiment." IBM Journal of research and development 2.4 (1958): 354-361.
- [18]. Isonuma, Masaru, et al. "Extractive summarization using multi-task learning with document classification." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.
- [19]. Das, Amitava, and Sivaji Bandyopadhyay. "Topic-based Bengali opinion summarization." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010. Their method is effective for detecting themes.
- [20]. Hahn, Udo, and Inderjeet Mani. "The challenges of automatic summarization." Computer 33.11 (2000): 29-36.

A Machine Learning Approach to Classify Singular and Plural Numbers of Bengali Words

ORIGINALITY REPORT

19%

SIMILARITY INDEX

16%

INTERNET SOURCES

5%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	www.cs.cmu.edu Internet Source	6%
3	en.wikipedia.org Internet Source	2%
4	Abdullah Al Munzir, Md. Lutfor Rahman, Sheikh Abujar, Ohidujjaman, Syed Akhter Hossain. "Text analysis for Bengali Text Summarization using Deep Learning", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019 Publication	1%
5	Minhajul Abedin Shafin, Md. Mehedi Hasan, Md. Rejaul Alam, Mosaddek Ali Mithu, Arafat Ullah Nur, Md. Omar Faruk. "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language", 2020	1%

23rd International Conference on Computer and Information Technology (ICIT), 2020

Publication

6 citeseerx.ist.psu.edu 1%

Internet Source

7 mdp-toolkit.sourceforge.net 1%

Internet Source

8 "Enhancing Reusability and Measuring Performance Merits of Software Component using Data Mining", International Journal of Innovative Technology and Exploring Engineering, 2019 1%

Publication

9 Xu, Baoxun, Xiufeng Guo, Yunming Ye, and Jiefeng Cheng. "An Improved Random Forest Classifier for Text Categorization", Journal of Computers, 2012. 1%

Publication

Exclude quotes On

Exclude matches < 1%

Exclude bibliography On