Name Gender Recognition System

BY

Labannya Saha
ID: 172-15-10181

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Fizar Ahmed

Assistant Professor

Department of CSE

Daffodil International University

Aniruddha Rakshit
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

May 31 2021

APPROVAL

This Project titled "**Name Gender Recognition System**", submitted by **Labannya Saha** 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 31st May 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

Azmoon

Internal Examiner

Nazmun Nessa Moon
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Aniruddha Rakshit

Internal Examiner

Aniruddha Rakshit
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

____

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Fizar Ahmed, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

____

Fizar Ahmed
Assistant Professor
Department of CSE
Daffodil International University

Aniruddha Rakshit

____

Aniruddha Rakshit
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

Labannya Saha

____

Labannya Saha
ID: 172-15-10181
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Fizar Ahmed**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Natural Language and Processing” and “Machine Learning” to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to**, Dr. Touhid Bhuiyan,** Head**,** Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

Person names are extremely important in various types of computer applications. The majority of people's names have a possible refinement between sexual orientations. Recognizing sexual orientations from English character-based Bangladeshi names with greater accuracy can be particularly difficult. In this paper, we present a characterization system focused on machine learning and deep learning that is capable of recognizing sexual introductions from Bangladeshi people's names. With an accuracy of 88 percent, the English character-based name was developed.

We have compared various machine learning and deep learning classifiers, such as Logistic Regression, Random Forest, SVM, and others, to see which calculations provide the best results. Aside from that, a pre-trained python demonstrate on sexual orientation identifiable proof by Bangla's title was discovered.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

LIST OF FIGURES

FIGURES **PAGE NO.**

LIST OF TABLES

TABLE NO. **PAGE NO.**

CHAPTER 1

INTRODUCTION

1.1 Introduction

The names of Bangladeshi people are associated with ethnic and multilingual characteristics. It has both long-term and short-term consequences. Some non-neighborhood areas, such as Arabic, Persian, Vedic, and so on, have a noticeable impact on having a good selection of different strict gatherings terms. Since a vast number of people are Muslims, Arabic prospectuses have a major influence on titles. A few of the names come from the Bengali and Sanskrit languages. Different cultures have different examples of stimulating males and females. Gender is frequently a factor in the naming of persons. As a result, formulating a mechanism for inferring a general trend that can distinguish genders from a large number of names is difficult.

In today's world, Artificial Intelligence (AI) can be a helpful procedure for addressing phonetic problems. We are gradually approaching a time in which specially prepared insights systems can be used. Machines nowadays provide us with a plethora of options for dealing with everyday problems. AI could be the most effective tool for extracting a far-fetched plan from a massive amount of data. If a situation arises in which male and female names are separated by means of strategies. Machine learning (ML) and deep learning (DL) are capable of performing critical tasks. Natural Language Processing (NLP). And the subfields of Artificial Intelligence are Machine Learning and Deep Learning.

In this case, our primary contribution would be to unite the n-grams set of natural language processing before pre-processing the data in order to understand the gender dependent on an individual identity. For amplifying the information accuracy, we used the 2-grams and 3-grams strategies. At that stage, all of the attested data was transferred to the feature extraction. In our case, we followed the counter vector, whose primary task is to convert content into vectors, ensuring that the input is compatible with the algorithms. We used 13 different types of machine learning and deep learning classifiers to prepare the machine learning procedures, including Random Forest Classifier, Support Vector Machine and the shadow of execution was compared using a Assist, we included several focus points and impediments in our process, as well as how it can be produced for better implementation and use.

The obligations are summed up as follows:

• Used Bangladeshi gender-discriminating evidence mechanisms to present the English character.

• A number of deep learning and machine learning classifiers are used, and their performance is compared using a number of quantative performance investigations.

1.2 Motivation

Gender and age remains an important for predicting of demographic information, especially in intervention of unintentional gender/age bias in recommender systems. As a result, the gender of those users who did not give this information during registration must be inferred. We look at the issue of predicting a registered user's gender based on their given name. And we know that Bengali names are related with multicultural and multi linguistic qualities. It has both chronicled and strict impacts. Having a decent variety of various strict gatherings names are conspicuously affected by some non-neighborhood societies like Arabic, Persian, Vedic and so forth. Larger part of individuals are Muslims and accordingly names are especially impacted by Arabic prospectuses. A few names are from Bengali and Sanskrit dialects. Distinctive culture has different examples of syllabi of inducing male and females. Consequently, it is hard to formulate a framework for inferring a common pattern that can differentiate genders form a great diversity of names.

1.3 Problem Definition

For many research questions, demographic information about individuals (such as age, gender or ethnic background) is highly beneficial but often particularly difficult to obtain. To tackle this problem, researchers often depend on automated methods to infer gender from name information provided on the web. Very little work has been done in Bengali language.

1.4 Research Questions

Here are the main questions those are focuses in this thesis are given below:

- What is Gender name detection?
- How does AI based gender identification from Bangladeshi people name work?
- What are the advantages of this work?
- What is the dissimilation among other country's work?
- How to preprocess name text data in Machine learning and NLP?
- How to training machine learning and deep learning model on name dataset?
- What are the future works of gender identification?
- What is the best model for name gender recognition for Bangla language?
- How to solve the limitations for the name gender recognition?

1.5 Research Methodology

In this section of our research paper, we reveal the Experiment Data Set, Data Pre-processing, Architecture of the Model, Learning Rate and Optimizer of the Model and Training the Model. At the end of this chapter performance of the proposed model will be described.

1.6 Research Objectives

There are some benefits for name gender recognition. Some of the technical objectives are given below:

- Develop an efficient model for name gender recognition.
- To inspire the software developers to work with AI using the model.
- Integrate the model in mobile apps and websites.
- To inspire students to work more on Bengali Language.
- Software developers can make an API to predict the gender of Bengali name.

1.7 Research Layout

Chapter 1: Will discuss about introduction, motivation, Problem Definition, Research Question, Research Methodology and the expected outcome of our project.
Chapter 2: Will discuss about background of this research and the related work and current status based on Bengali language perspective.

Chapter 3: Will describe situation of AI in the field of agriculture in Bangladesh.

Chapter 4: Will discuss about the ultimate development in the perspective for NER by using NLP and AI.

Chapter 5: Will discuss about result and benefit of using AI, NLP in Name Gender Recognition.

Chapter 6: The conclusion of this research is described here.

Chapter 7: Here all the references we used for this research are given.

CHAPTER 2

BACKGROUND

2.1 Introduction

In Bangladesh there are very few work or research was done which can detect gender from names. But we will detect gender with higher precision. So the background is the present situation of NLP in Name gender recognition in Bangladesh.

2.2 Related Works

Qianjun Shuai et al. focuses on gender identification in Chinese names in their paper.it employs a number of machine learning algorithms [1] to find a more accurate model. [2] Orestis Giannakopoulos et al. suggested an effective method for determining the gender of Twitter users. To do so, it combined three features like user's name, profile picture and theme color choice as input to infer the gender of the user. And got 87% accuracy by using SVM and PNN. In [3] Fariba Karimi et al. developed a gender inferring system that was both name and image dependent, and it performed effectively.

Anshuman Tripathi et al. proposed an approach that used a Support Vector Machine-based classification method to predict the gender of Indian names.And compared the morphology of Indian and English names[4]. ZHAO Xiao-fan et al. used conditional random fields on 231337 names and acquired 89.30% accuracy. The analysis showed that the last name has a greater impact on gender identification than the first name [5].
Vivek Kumar Verma et al. compared three classifiers [6] SVM, neural network, and adoboost on their face image dataset and showed the best classifier that works for their dataset. [7] Bo WU et al. explored a machine learning image extraction approach using demographic details such as gender, age, and ethnicity.

J. Harris et al. introduced a model [8] for determining people's origins from their names that was more efficient than the conventional method.

John D Burger et al. [9] investigated the feasibility of using word and character level n grams in author profile information to predict gender of authors in tweets and they got around 77% accuracy. In [10] Arjun Mukherjee et al. suggested POS sequence patterns as a new class of features capable of capturing complex stylistic regularities in male and female authors and obtained 77% accuracy.

Using demographic data Ehsan Mohammady et al. [11] used a supervised learning algorithm to infer the characteristics of Twitter users. They acquired 80 % accuracy by using a regression model. [12] Joseph Lmaely used nine different machine learning algorithms to predict gender from

facial image. Sandeep Kumar compared to traditional methods, their proposed state-of-the-art approach yielded better results [13].
Liangliang cao et al. built a system from full body image to infer the gender [14]. Pramit Gupta et al. [15] has used some machine learning algorithms to detect gender from a person's voice. Buyukyilmaz Mucahit et al detect gender from many person's audio files. They were able to predict gender 93.5% accurately [16].

Juan Bekios-Calfa et al. showed the presence of dependencies for instance gender, age, and facial attributes may boost the performance and reliability of gender classifiers [17].

The research can give a spanking framework to support the research. Here all the literature survey paper is talking about gender detection from names and several researchers made a variant solution to recognize the gender. The main goal of this paper is to find out the sexual orientation somehow here all the literature survey paper is related to the goal of this paper.

2.3. Research summary

In this work, we utilized the machine learning and deep learning methods for gender identification of Bangladeshi people names in both in Bengali Form and English character based Bangladeshi names. To implement this model, we have collected a wide range of gender dataset from various sources. Dataset has collected from the school, college, and various institutional resources. At first collect Bengali news articles from different media. At that point make a summary of every piece of information. Our dataset with 1610 Bangla names and 33,000 names in English format. Before applying the machine learning algorithms, we preprocessed our Bangla dataset. In this preprocessing stage, we check all noises and try to remove all of them. We used Count vectorizer then. Which is used to change a given text into a vector-based on the frequency (count) of each word that happens in the whole content.

2.4. Scope of the Problem

For numerous investigate questions, statistic data approximately people (such as age, sexual orientation or ethnic foundation) is exceedingly advantageous but frequently especially troublesome to get. To handle this issue, researchers frequently depend on computerized strategies to induce gender from title data given on the internet. Very little work has been wiped out Bengali language.

2.5. Challenges

First of all, finding a Bangla dataset is not an easy task. All information is available in an unstructured manner. Thusly, information assortment is a test for this research. We couldn’t find a reliable dataset to use. Along these lines, we need another dataset to complete this exploration work. Since the assortment of the dataset, labeling all of the data is another challenging work. Therefore, the preprocessing step needs fresh coding to set up the content as a contribution of a model. We have faced so many problems to run the dataset. Another problem is to run the Bangla dataset as a string. For other languages like English have a well maintain dataset. But for the Bengali language, it’s still in an early stage, so we had to do it on our own. Finding a large amount of exact fake news is another test is in this examination. On the other hand, if the dataset has countless fake information, it would give assists with creating a more precise result.

2.6. Bangladesh Perspective

In Bangladesh's perspective, there are a few work has been done on gender recognition. Demographic data like sexual orientation about an individual is vital for some research question. An individual's name has a great deal of data. The naming interaction of an individual relies upon religion, convictions, custom and locale of a country.

In Bangladesh, a few names have significant sense. some are started from their past ancestor. Furthermore, a few names really have no significance. A few names are both used for female and male. In Bangladesh, some names are taken from father or mother’s name.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

To decide sex utilizing machine learning, we gotten five critical stages as data collection, data preprocessing, preparing data joining ML algorithms(Models). performance. The tall number of tests and fittingly orchestrated combination is guaranteeing the quality of information. Other than, information preprocessing has made the information ceaselessly dependable and extraordinary cases free. We have executed the three best machine learning classifiers and significant deep neural systems on the pre-handled data. Our organized expectation show has expected the foremost fitting technique of Sexual orientation discovery with a tall level of precision and dependability.

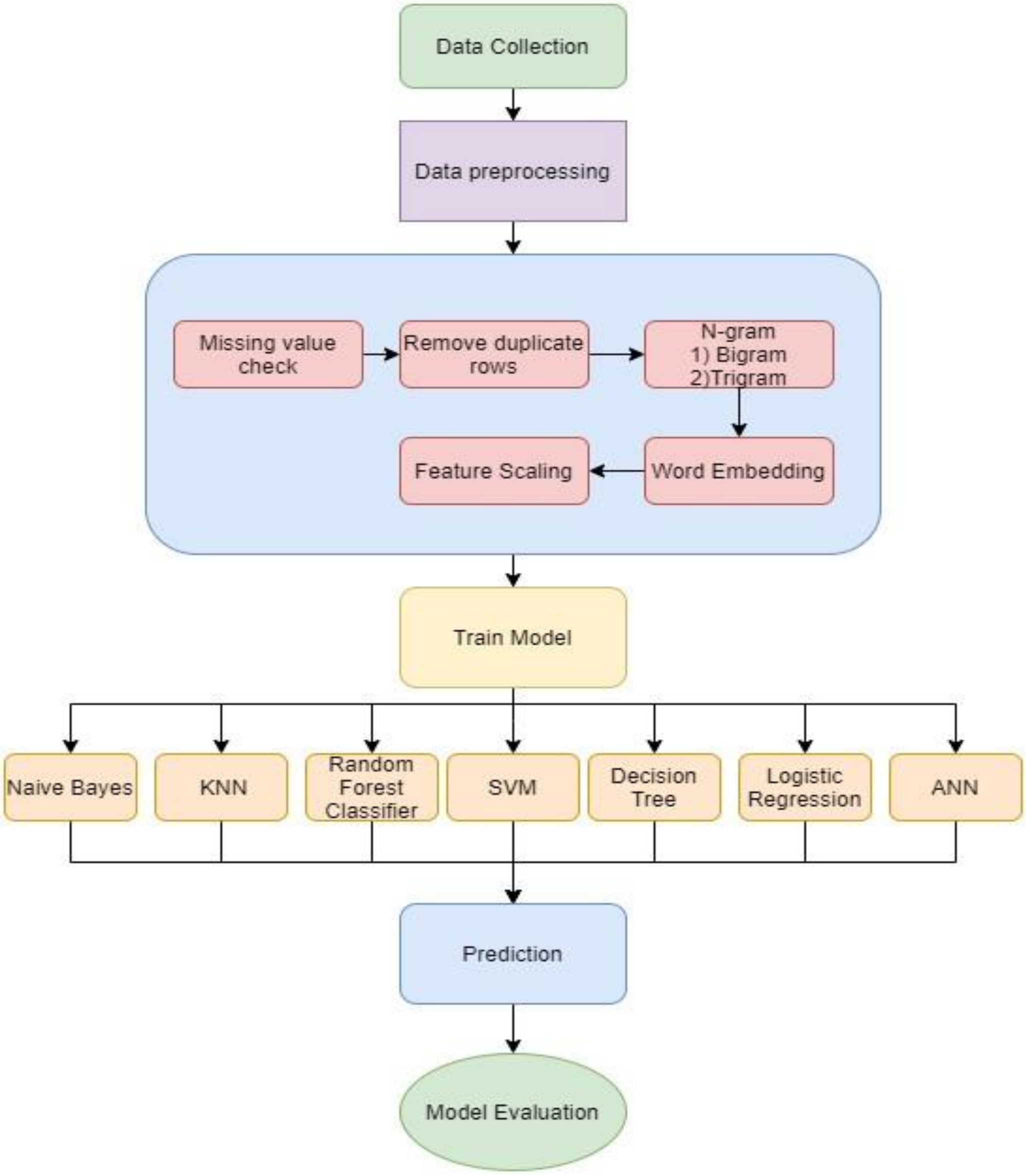


Figure 3.1.1 Methodology of gender detection from Bengali Names

3.2 Experiment Data Set

Most of the names of this dataset were taken from blog, posts, web and others are from list of voter in Bangladesh. There are people living from various convictions and cultures in Bangladesh including Bengalese, local clans and aboriginal minorities. Names from the same or different culture share an equivalent intrigue, that is all the more for the most part female and male names are influenced by certain phonetics. For example, "Arif" is a male name yet "Arifa" is a female name. An additional “a” at the end of the word turns into an improvement of phonology and transforms the name into a female one.

3.3 Data Pre Processing

Information preprocessing might be an uncommonly basic step inside the information mining handle. It incorporates changing over the crude information from different sources into a recognizable organization. Legitimately preprocessed information guarantees the foremost great preparing of calculations. To form our system stronger, we had multi-stage preprocessing of the dataset indicated inside a extend of procedures.

3.3.1. Missing value check

There could be 2 sorts of missing values. The very first one is the missing names. In this specific case we erased the column. Since we cannot presume the name. Another is the missing label. On the off chance that the regular name is female, we place the name is 'f' in any case 'm'. In the event that we can't get whether the name is a female or a male one, we erased the line.

3.3.2. N-Gram

The N-gram system divides a string into N length substrings, where N is the length of the string. In a broad sense, an N-gram is any concatenation of relative characters of length N that can be found in the source text material. We used infectious bi-grams and trigrams with N=2 as a human respect in our work.

3.3.3. Feature embedding

Word embedding is a strategy where words are represented in a way that comparative words are represented essentially. Count vector is an exceptionally effective system in conventional Natural language processing. This strategy produces an N×M grid where N is the quantity of data components and M is the quantity of unit components display within the data components.

Each data component is represented by the recurrence of unit components displayed within the data component. In our investigation, the information components are English names like "Prottasha" etc. The unit components of these names are bigrams and trigrams.

3.3.4. Feature scaling

Feature scaling is used to standardize the autonomous functions over a data set within a given range. Unit vector, standardization, mean normalization, and min-max scaling are some of the function scaling methods available. Mean normalization was used as the highlight scaling tool in our study.

3.3.5. Dimensionality reduction

Reducing dimension is a methodology for reducing feature of the dataset without hampering the ML model accuracy. Basically, the NLP problems deal with a huge dimension. English character-based Bangladeshi names in our datasets after performing word embedding. We reduced that down to 400 columns only. In this work, we have utilized PCA.

CHAPTER 4

ALGORITHM DESCRIPTION

4.1. Introduction

The work imminent here utilized seven classification calculations to anticipate the sex title of people of Bangladesh. The applying classifiers are Logistic regression, Naive Bayes, Decision Tree, Random forest Classifier, SVM, and K-NN, and ANN. The information set for for this work was accumulated and connected on each classifier to predict or target the gender from title and the accomplishment of the classification calculations is assessed based a few statistical techniques and execution analysis.

4.2. Algorithms

4.2.1. Logistic Regression

Calculated logistic regression might be a quantifiable and numerical procedure in orchestrate to examine a data set and it has one or more than one free variety that holds a result. The result is thought with a diploid variety. The foremost point of calculated relapse is to capture the driving likely suitable procedure to gather the relationship between the diploid way of a fragment and a set of free varieties. Calculated relapse actuates the coefficients of a plan to recognize a logit formalization of the chance of the nearness of the specific of interested:

$$\theta_i = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)\right]}$$

4.2.2 K-Nearest Neighbor Classification (K-NN)

The K-NN isn't a parametric appear that's imperative to disconnected data and prescient Backslide. In each coordinate case, the entered data are formed of the K best closest planning event inside the components put. K-NN is one kind of event set up learning. In K-NN classification, the result contains a bunch of people. The division of a bunch is chosen by the assortment choice by the relate of data. On the off chance that K = 1, at that point the course has an along closest neighbor. Within the occasion that k = 2, at that point the course has bi or twofold closet and so on. In common, the weighting upgrades, the hail neighbor is pronounced to whole of 1/x where x is the length to the neighbor to the point which is confined. The slightest length between any two relate is persistently a solid line and the length is called Euclidean partitioned. Numerically, the term of Euclidean distance is

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

Here, d (p, q) means Euclidean length between b and a. In our method, to go the data with the help KNN, we have taken k = 5 with neighbors.

4.2.3. Decision Tree

The choice tree may be a strategy which livelihoods the structure of a tree and diagram the system conceivable comes about concurring to result of event, capacity to classify from their costs, and advantage. In this strategy, it displays a system that as it were holds on conditional control [11]. It can be a strategy in organize to seem a system that because it was holds on conditional control statements. Choices trees are for the foremost portion held in a couple of regions such as operations ask almost, choice examination, to assist disconnected a technique most likely to reach a point. We utilized the entropy mishap to actualize the choice tree.
Here is the condition.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

4.2.4. Naive Bayes Classifier

The Naive Bayes Classifier framework is set up on probably is known as Bayesian framework and is uncommonly set up when the numbers of the inputs are as well enormous. In show disdain toward of its least demanding, Naive Bayes is the foremost usable calculation within the field of arithmetic or statistics. A Naive Bayesian demonstrate can be created effectively without complexity and has parametric estimation which makes intrinsically utility for huge datasets. Here, P(B|A) is the predator chance of a lesson or gather, P(c) is the framework chance of lesson, and P(B|A) is the likelihood.

$$p(B|A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} \frac{p(A_i|B)}{p(A_i)}\ p(B)$$

4.2.5. Support Vector Machine (SVM)

Support Vector Machine (SVM) is known as well-established framework to partitioned data among themselves. It could be a state-of-the-art advancements and division classifier intrinsically construct by a disconnected hyper line or plane. In SVM all given labeled training data (directed learning), the algorithm comes about an ideal hyperplane which categorizes unused cases. In two-dimensional space, this hyper line or plane may be a line to isolated a plane in two bunches or a portion wherein each bunch lay in either side or another . SVMs can have exactness rate to execute a nonlinear classification holding what is said the bit trap, so also, SVM is utilized for mapping their huge information into high-dimensional component spaces.

4.2.6. Random Forest Classifier

Random Forest Classifier could be a classification instrument. Random Forest Classifier which is started for calculating a complex's quantitative or unlimited organic work risen on a quantitative portrayal of the compound's atomic structure. It was being moreover roofed in our investigate as one kind of the classifiers for making the T2D medications prescience structure. The outcome of the irregular timberland calculation may be a store of decision trees risen on the unpredictably choice of shape.

4.2.7. ANN

ANN or artificial neural network may be a framework of computation that mirrors human neurons in terms of comprehending and analyzing real-world data and tends to unravel problems that were gathered to fathom cognitively. The essential design of an ANN incorporates an input layer, covered up layer, and yield layer. In our ANN model, there were 50000 inputs and 300 yield measurements inside and out. A basic representation of our ANN architecture.

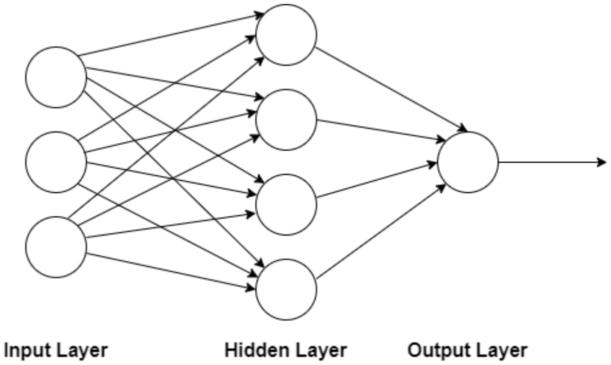


Fig. 4: The architecture of ANN

CHAPTER 5

Experimental Results and Discussion

5.1. Experiment

To explore different avenues regarding our proposed procedures, we have first fabricated the model and trained it. Seven classifiers calculations have been utilized to foresee the most appropriate method of sexual orientation recognizable scheme. The pre-processed training datasets have been utilized to test the presentation of the model. At that point we have tried the exhibition of our model utilizing the test datasets. In a nutshell, we explained a few trials to register the precision of our model.

5.2. Experiment setup

We have executed the entire project in a python programming language having rendition 3.7 in Anaconda distribution. Python library offers different offices to fabricate machine learning and deep learning models. The incredible library for information portrayal is pandas that give immense orders and huge information handling abilities. We have utilized it to peruse and break down information in less code composing. A while later, scikit learn has highlighted for different order, grouping calculations to constructing models. Additionally, Keras joins the upsides of Theano and Tensor Flow to prepare a neural network model. We utilized it to fit and assess capacities to prepare and survey the neural system model individually bypassing similar information and yield, and afterward we applied Matplotlib for graphical perception. Here is the summery of the instruments that we used

Hardware and Software of Local PC:

- Intel Core i7 8GB RAM
- 1 TB HDD
- Google Colab including 35GB TPU

Advancement Tools:

- Windows 10
- Python 3.8
- Sklearn
- Pandas
- TensorFlow Backend Engine
- Keras
- NLTK
- NumPy

5.3. Statistical tools for measurements:

Statistical tools for measurements: We utilized several Statistical tools for measurements, evaluation, and analysis of the performances, and compared all the algorithms. these are:

The F1-score is used to calculate the accuracy of a model. Moreover, the accuracy of the model is calculated. It also estimates binary classification systems Is used, which categorizes the examples as 'positive' or 'negative'. The mathematical definition of the F1score is:
Recall considers the percentage of correct predictions for all the positive categories among which are positive Forecasts can be made. Reduce false-positive errors, although changing limits and maximizing will reduce false-negative errors.

$$\frac{True\ Posiive(TP)}{True\ Positive(TP) + \ \ False\ Negative(FN)} \qquad (3)$$

Accuracy is a metric that evaluates s for the correct prediction rate for the positive class.

$$\frac{True\ Positive(TP)}{True\ Positive(TP) + \ \ False\ Positive(FP)} \qquad (4)$$

The F-beta score is evaluated in the binary classification model based on a configurable single-score for the forecasts made for the positive class.

It's also calculated utilizing precision and recall.

$$\frac{(1 + \ \beta^2)\ .(Precision\ .\ Recall)}{(\beta^2\ .\ Precision\ \ + Recall\ )} \qquad (5)$$

Hamming loss is designed for multi-class while Precision, Recall, F1-beta score represents one clear single-presentation-value for multiple-label cases in contrast to the precision/recall/f1beta score that can be assessed only for independent binary classifiers for each label.

$$\frac{1}{|D|}\sum_{i=1}^{|D|} \frac{|Yi\ \Delta\ Zi|}{|L|} \qquad (6)$$

The Jaccard similarity coefficient: The union of the two label sets is used to compare the set of labels predicted in y_true to mark the separate intersection as a measure by calculation.

$${}_{J}(A,B) = \ \frac{|A \cap B|}{|A \cup B|} \qquad (7)$$

Another is often the loss function utilized in (multinomial) logistic regression and extensions of it, such as neural networks, characterized as the negative log-likelihood of a logistic model that returns y_pred probabilities training data y_true.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\log\big(p(y_i) + (1 - y_i) .\log\log\big(1 - p(y_i)\big)\big) \qquad (8)$$

The Matthews correlation coefficient is used as a standard for binary and multiclass classification in machine learning.

$$\text{MCC} = \frac{TP\ .\ FN - FP\ .FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \qquad (9)$$

Average accuracy (MAP or sometimes simply AP) is a popular metric used to measure the effectiveness of models used to perform document retrieval and object detection tasks.

$$\text{A}=\textstyle\sum_{k=0}^{k=n-1}\ [Recalls(k) - Recalls\ (k+1) * Precisions(k)] \qquad (10)$$

The balanced accuracy: The balanced accuracy in binary and multiclass classification issues to deal with imbalanced datasets.

$$\frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \qquad (11)$$

Cohen's kappa is a statistic that measures inter-annotator agreement.

$$\text{K} = \frac{P_{0-}\ P_E}{1-\ P_E} \qquad (12)$$

5.4. Result and Performance Analysis

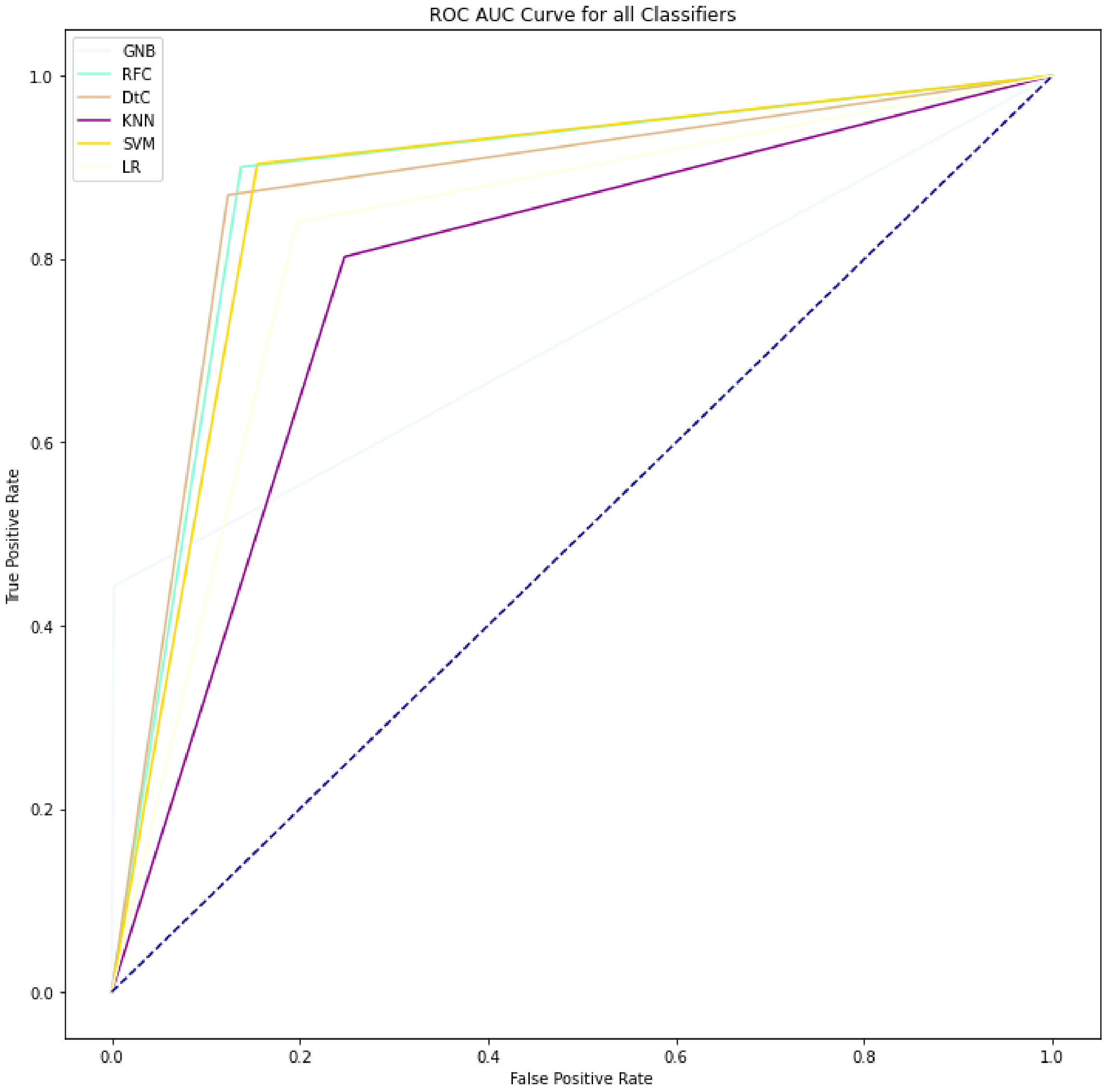
In our experiment, we improved the methods and tried to utilize their thirty-two classifier parameters to gain way better performance. We can analyze the 11 statistical measurements.

At the point when the models were established, we prepared it utilizing dataset for the training. While setting up the model, we tuned the significant parameter to build the algorithms increasingly precise. After the model was arranged appropriately, we ran our test informational collection to figure out the methodology of gender detection. For per unique calculation, the value was classified, and we discovered a trustworthy accuracy score, precision, recall, Cohen's kappa, F1 score and ROC AUC. For boosting execution, it is constantly a superior plan to expand information size as opposed to relying upon predictions and frail correlations. Additionally, including a hidden layer may speed up in light of its propensity to make a training dataset overfit. In any case, halfway, it relies upon the multifaceted nature of the model. Conflictingly, expanding the epochs number improves execution however it some of the time over fits training information. It 10 functions admirably for the deep network than the shallow framework while pondering the regulation factor.

Table.2. Performance and Result Analysis

Name	Naïve Bayes	RFC	Decision Tree	KNN	SVM	Logistic Regression	ANN
Accuracy	0.721947	0.881376	0.873215	0.777324	0.874381	0.821626	0.863888
F1S	0.697954	0.881354	0.873205	0.777234	0.874309	0.821596	0.863627
Recall	0.720249	0.881491	0.873194	0.777477	0.87456	0.821737	0.864188
Precision score	0.819595	0.881891	0.873228	0.778063	0.875638	0.822062	0.86739
F-Beta score	0.744909	0.881595	0.873218	0.777603	0.874916	0.821803	0.86536
HL	0.278053	0.118624	0.126785	0.222676	0.125619	0.178374	0.136112
Jaccard similarity	0.542638	0.78788	0.774948	0.635658	0.776699	0.697217	0.760035
Matthews correlation	0.530624	0.763381	0.746422	0.55554	0.750197	0.6438	0.731571
AUC	0.720249	0.881491	0.873194	0.777477	0.87456	0.821737	0.864188
Balanced Accuracy	0.720249	0.881491	0.873194	0.777477	0.87456	0.821737	0.864188
Cohen's kappa	0.44199	0.762799	0.746412	0.554775	0.748844	0.643323	0.727932

Here from the table we can depict that the random forest tree showed the best accuracy (88%) and f1 score (88%). Here it also provided the lowest hamming loss which is around 0.11 where the most hamming loss is 0.27 of Naive Bayes. The Jaccard similarity and Matthews correlation coefficient of random forest tree are 78% and 76% with AUC curve of 88%. Form this table we can also analysis that decision tree is the second place from the perspective of all accuracy. Besides, in deep neural network, ANN took the third positon with 86% of accuracy algorithm with f1 score. Nevertheless, the Gaussian Naive Bayes took the lowest accuracy among all trained models which is 72% accuracy and 69% f1 score.



CHAPTER 6

Conclusion and Future Work

6.1. Summary of the Study

Our entire research work is identified with the NLP and machine learning. In this task, we have utilized machine learning models for detecting gender from name. We have finished this research in more than half year. The entire cycle of work is partitioned into certain parts. The entire synopsis of the research is given beneath bit by bit.

Step 1: Planning

Step 2: Problem Analysis

Step 3: Model design

Step 4: Data collection form various institution

Step 5: Summarize the collected data

Step 6: Labeling all the missing data

Step 7: Data preprocessing

Step 8: Data transformation

Step 9: Feature extraction

Step 10: Training the models

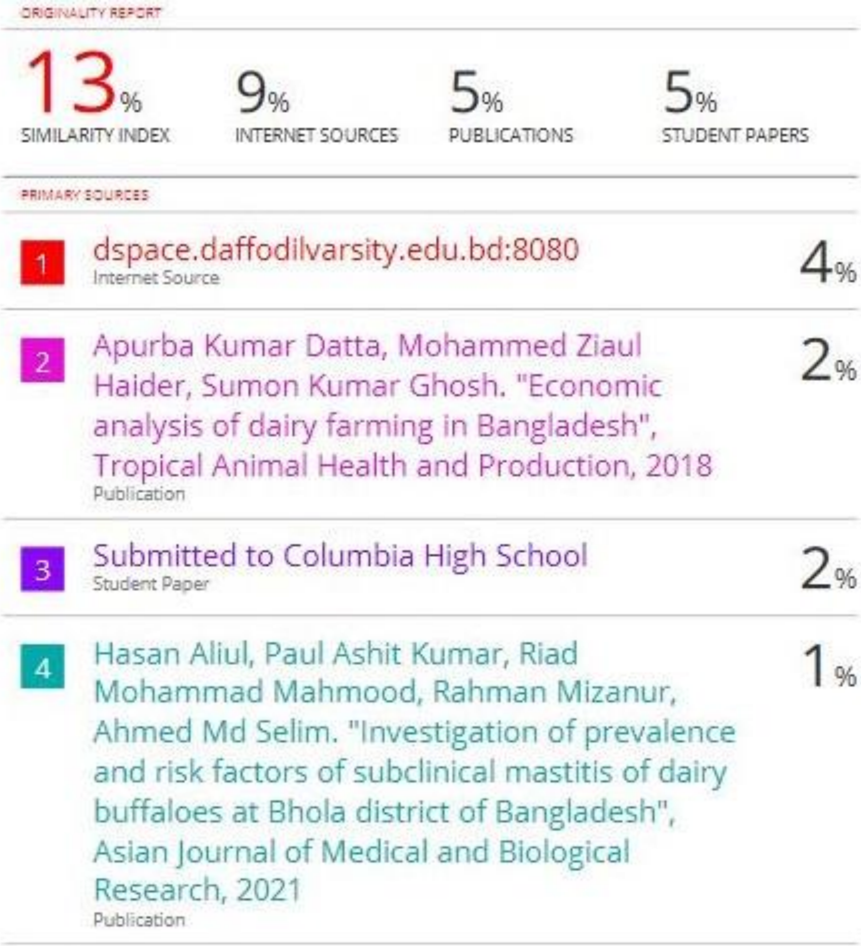
Step 11: Check the outcome and examination of the reaction of the machine

6.2. Conclusions & Future Works

Identifying gender from the Bangla names is a critical issue. To achieve this goal n-grams techniques such as 2-grams has been used for extending the data quality which is used for feature extraction. Besides, the counter vector makes the data compatible with the algorithms. In this work ten machine learning classifiers such as Calculated Relapse, SVM, Credulous Bayes Classifier, k-Nearest Neighbors, Choice Tree, Arbitrary Timberland, Affect Learning and profound manufactured neural systems have been connected and among these Affect Learning gives the most excellent yield. In the long run we would like to create API that would be utilized by other applications. Other than, we'll attempt to utilize more information and other progressed calculations and advances to form the show more grounded. Additionally, we have plan to utilize arrangement implanting with the checking name's portion, since each title take after a grouping. In future we have to arrange to execute more profound learning calculation to figure out the finest show, and utilize this framework in genuine life application.

REFERENCE

1. Shuai, Qianjun, et al. "Research on gender recognition of names based on machine learning algorithm." *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. Vol. 2. IEEE, 2018.
2. Giannakopoulos, Orestis, et al. "Gender recognition based on social networks for multimedia production." *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2018.
3. Karimi, Fariba, et al. "Inferring gender from names on the web: A comparative evaluation of gender detection methods." *Proceedings of the 25th International conference companion on World Wide Web*. 2016.
4. Tripathi, Anshuman, and Manaal Faruqui. "Gender prediction of Indian names." *IEEE Technology Students' Symposium*. IEEE, 2011.
5. Zhao, X. F., Dan Zhao, and Y. G. Liu. "The automatic gender recognition of Chinese name using conditional random fields." *Microelectronics & Computer* 28.10 (2011): 122-124.
6. Verma, Vivek Kumar, et al. "Local invariant feature-based gender recognition from facial images." *Soft computing for problem solving*. Springer, Singapore, 2019. 869-878.
7. Wu, Bo, Haizhou Ai, and Chang Huang. "Facial image retrieval based on demographic classification." *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*. Vol. 3. IEEE, 2004.
8. Harris, J. Andrew. "What's in a name? A method for extracting information about ethnicity from names." *Political Analysis* 23.2 (2015): 212-224.
9. Burger, John D., et al. *Discriminating gender on Twitter*. MITRE CORP BEDFORD MA BEDFORD United States, 2011.
10. Mukherjee, Arjun, and Bing Liu. "Improving gender classification of blog authors." *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*. 2010.
11. Ardehaly, Ehsan Mohammady, and Aron Culotta. "Using county demographics to infer attributes of twitter users." *Proceedings of the joint workshop on social dynamics and personal attributes in social media*. 2014.
12. Lemley, Joseph, et al. "Comparison of Recent Machine Learning Techniques for Gender Recognition from Facial Images." *MAICS* 10 (2016): 97-102.
13. Kumar, Sandeep, Sukhwinder Singh, and Jagdish Kumar. "Gender classification using machine learning with multi-feature method." *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2019.
14. Cao, Liangliang, et al. "Gender recognition from body." *Proceedings of the 16th ACM international conference on Multimedia*. 2008.
15. Gupta, Pramit, Somya Goel, and Archana Purwar. "A stacked technique for gender recognition through voice." *2018 Eleventh International Conference on Contemporary Computing (IC3)*. IEEE, 2018.
16. Buyukyilmaz, Mucahit, and Ali Osman Cibikdiken. "Voice gender recognition using deep learning." *2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016)*. Atlantis Press, 2016.
17. Bekios-Calfa, Juan, José M. Buenaposada, and Luis Baumela. "Robust gender recognition by exploiting facial attributes dependencies." *Pattern recognition letters* 36 (2014): 228-234.



ORIGINALITY REPORT

13%	9%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Apurba Kumar Datta, Mohammed Ziaul Haider, Sumon Kumar Ghosh. "Economic analysis of dairy farming in Bangladesh", Tropical Animal Health and Production, 2018 Publication	2%
3	Submitted to Columbia High School Student Paper	2%
4	Hasan Aliul, Paul Ashit Kumar, Riad Mohammad Mahmood, Rahman Mizanur, Ahmed Md Selim. "Investigation of prevalence and risk factors of subclinical mastitis of dairy buffaloes at Bhola district of Bangladesh", Asian Journal of Medical and Biological Research, 2021 Publication	1%

Fig: Plagiarism Report