A

PROJECT REPORT

ON

LABEL THE UNLABELED DATA USING

SUPERVISED LEARNING

Submitted in Partial fulfilment of the requirements for the award of degree of Master of Computer Applications (MCA) 4th Semester Examination, June 2024

Submitted by:

Potshangbam Kirankumar Singh

MCA 4th Semester

Roll No.: 2250026

Regd. No.: 17150392, 2017

Under the guidance of:

Dr. Ksh. Nareshkumar Singh

Guest Faculty

Department of Computer Science

Manipur University

DEPARTMENT OF COMPUTER SCIENCE

MANIPUR UNIVERSITY, CANCHIPUR

IMPHAL- 795003

CERTIFICATE

This is to certify that the project entitled **"LABEL THE UNLABELED DATA USING SUPERVISED LEARNING"** is a comprehensive work carried out by **Potshangbam Kirankumar Singh**, a student of **MCA** in the **Department of Computer Science, Manipur University**, bearing roll number **2250026**, under the supervision of **Dr. Ksh. Nareshkumar Singh**, Guest Faculty, Department of Computer Science, Manipur University for the partial fulfilment of the requirement for the degree of **Master of Computer Applications** under Manipur University. To the best of my knowledge, the work has not been submitted to any University or Institution for award of any degree or diploma. I am pleased to recommend the project to the **Department of Computer Science**, Manipur University for necessary evaluation.

Date:

(**Tejamani Sinam**)
Head of Department
Department of Computer Science
Manipur University

CERTIFICATE

This is to certify that the project entitled **"LABEL THE UNLABELED DATA USING SUPERVISED LEARNING"** is a comprehensive work carried out by **Potshangbam Kirankumar Singh**, a student of **MCA** in the **Department of Computer Science, Manipur University**, bearing roll number **2250026**, under the supervision of **Dr. Ksh. Nareshkumar Singh**, Guest Faculty, Department of Computer Science, Manipur University for the partial fulfilment of the requirement for the degree of **Master of Computer Applications** under Manipur University. To the best of my knowledge, the work has not been submitted to any University or Institution for award of any degree or diploma. I am pleased to recommend the project to the **Department of Computer Science**, Manipur University for necessary evaluation.

Date:

(**Dr. Ksh. Nareshkumar Singh**)
(Guide)
Guest Faculty
Department of Computer Science
Manipur University

DECLARATION

I hereby declare that the project work entitled as **"LABEL THE UNLABELED DATA USING SUPERVISED LEARNING"** under the guidance and supervision of respected Sir, **Dr. Ksh. Nareshkumar Singh**, Guest Faculty, Department of Computer Science, Manipur University, submitted in partial fulfilment of the requirements for the award of Master of Computer Application (MCA) 4th Semester under Manipur University.

I also declare that all the contents or works incorporated with this documentation of the proposed project have not been submitted to any university or institutions for award of any degree nor diploma.

Date:

Potshangbam Kirankumar Singh
MCA 4th Semester
Roll No.: 2250026
Regd. No.: 17150392
Department of Computer Science
Manipur University, Canchipur
Imphal-795003

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and thankfulness to my project guide, **Dr. Ksh. Nareshkumar Singh**, Department of Computer Science, Manipur University for his priceless guidance and knowledge given to me.

I would also like to express my sincere thanks to **Sir Tejmani Sinam**, Head of Department, Department of Computer Science, Manipur University and other faculties of the department in helping me to overcome the problem I had and directed me to the rightful direction.

Date: **Potshangbam Kirankumar Singh**

MCA 4th Semester
Roll No.: 2250026
Regd. No.: 17150392
Department of Computer Science
Manipur University, Canchipur
Imphal-795003

ABSTRACT

The text classification (document classification) is a process of classifying or categorizing a text to a certain category or class using machines or computer. However, machines cannot understand the human textual languages. The text data should be converted into machines understandable form.

The methodology used here is using of an algorithm that allows us to label the documents in the unlabeled dataset. The said algorithm is tried in the labeled dataset first with train and test different sizes and then later applied to unlabeled documents, and also the feature extraction to convert the text data to numerical representation used in this project are CountVectorizer, CountVectorizer with n-grams, TFIDF, TFIDF with n-grams and Doc2Vec. The conversion is a must in order to make machines understand.

There are four algorithms (classifiers) used in this. Among them, SGDClassifier, LogisticRegression and MultinomialNB give the highest evaluation factors.