A

PROJECT REPORT

ON

NEWS CLASSIFICATION BASE ON TWENTY NEWS GROUPS

Submitted in Partial fulfilment of the requirements for the award of degree of Master of Computer Applications (MCA) 3rd Semester Examination, December 2023

Submitted by:

Potshangbam Kirankumar Singh

MCA 3rd 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

Department of Computer Science
Manipur University, Canchipur
Imphal-795003
Manipur, India

CERTIFICATE

This is to certify that the project entitled "**News Classification Base On Twenty News Groups"** 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:

(**Sir Tejmani Singh)**
Head of Department
Department of Computer Science
Manipur University

Department of Computer Science
Manipur University, Canchipur
Imphal-795003
Manipur, India

CERTIFICATE

This is to certify that the project entitled "**News Classification Base on Twenty News Groups"** 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

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 3rd Semester
Roll No.: 2250026
Regd. No.: 17150392
Department of Computer Science
Manipur University, Canchipur
Imphal-795003

ABSTRACT

The text 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 the feature extraction to convert the text data to numerical representation. CountVectorizer and TFIDF are used. The conversion is a must in order to make machines understand.

There are six algorithms used in this. Among them, SGDClassifier gives the highest evaluation factors. Using this algorithm, a simple model to classify a text data to news groups is built.

TABLE OF CONTENTS

Title	Page No.
ACKNOWLEDGEMENT	i
ABSTRACT	ii
TABLE OF CONTENTS	iii-iv
CHAPTER 1 INTRODUCTION	1
1.1 Data Growth	2-3
1.2 Text Classification	4
1.3 Problem Statement	4
1.4 Motivation	5
1.5 Contribution	5
CHAPTER 2 RELATED WORK	6
CHAPTER 3 METHODOLOGY	7
3.1 Dataset	8
3.2 Preprocessing	8-9
3.3 Feature Extraction	10
3.3.1 TF-IDF	10-11
3.3.2 CountVectorizer	11
3.4 Classification	12
3.4.1 Classifiers	13-15
3.5 Evaluation Matric	16
CHAPTER 4 TOOLS	17
4.1 Python Programming Language	17
4.1.1 Python Over Other Programming Language	18
4.1.2 Python in Machine Learning	19
4.1.3 Python in Natural Language Processing	19
4.2 Libraries	20

CHAPTER 5 RESULT, DISCUSSION AND IMPLEMENTATION
5.1 Result 21-30
5.2 Discussion 31
5.3 Implementation 32-35
CHAPTER 6 CONCLUSION 36
6.1 Scope 37
CHAPTER 7 REFERENCES 38