



FINAL YEAR PROJECT REPORT

**TEXT CLASSIFICATION USING TF-IDF AND
MACHINE LEARNING ALGORITHMS**

**In fulfillment of the requirement
For degree of
BS (COMPUTER SCIENCES)**

By

FARAZ AHMED KHAN

48474 BSCS

M. ZAEEM SHAKIR HUSSAIN

48438 BSCS

ABDUL REHMAN

48414 BSCS

SUPERVISED

BY

MUHAMMAD IQBAL

BAHRIA UNIVERSITY (KARACHI CAMPUS)

FALL-2020

DECLARATION

We hereby declare that this project report is based on our original work except for citations and quotations which have been duly acknowledged. We also declare that it has not been previously and concurrently submitted for any other degree or award at Bahria University or other institutions.

Signature: Faraz


Name: Faraz Ahmed Khan

Reg No.: 48474

Signature : A. Rehman

Name : Abdul Rehman

Reg No. : 48414

Signature : 

Name: Muhammad Zaeem Shakir Hussain

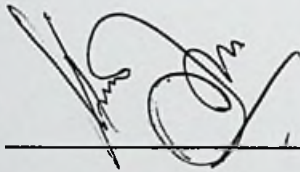
Reg No.: 48438

APPROVAL FOR SUBMISSION

We all guarantee for this project named "TEXT CLASSIFICATION USING TF-IDF AND MACHINE LEARNING ALGORITHMS" was prepared by FARAZ AHMED KHAN, MUHAMMAD ZAEEM SHAKIR HUSSAIN and ABDUL REHMAN has satisfied the necessary guideline for submission in partial fulfilment of the requirement for the award of Bachelor of Computer Science at Bahria University.

Approved by,

Signature :



Supervisor: MUHAMMAD IQBAL

Date : 16-Dec-2020

ACKNOWLEDGEMENTS

We want to thank each and everyone who had helped to the effective completion of this project. We want to offer our thanks to our supervisor, Sir M. Iqbal for his significant advice, guidance and his experience which helped all through the developing stage of this research.

Secondly, we like to offer the special thanks for our caring parents and companions who helped us and gave us some encouragement.

TEXT CLASSIFICATION USING TF-IDF AND MACHINE LEARNING ALGORITHMS

ABSTRACT

The reason for the creation of this system is the need for the classification of the newspaper articles. For the desired purpose we have collected many newspaper articles from many different newspapers. The article will be compared with the articles stored in dataset and if the descriptors and key points for the articles matches with the articles in our dataset, the details of that specific articles will be sent to the system. Some outputs are going to be available which'll show that there is good accuracy of recognition of articles. Starting from Support Vector Machine (SVM) and its variants are gaining momentum among the Machine Learning community. In this paper, we present a quantitative analysis between the established SVM based classifiers on multi-category text classification problem. Here, The dataset is first converted into required format by performing preprocessing activities which involve tokenization and removing irrelevant data. The feature set is constructed as Term Frequency-Inverse Document Frequency matrix, so that representative vectors could be obtained for each document. Experimentally, and we compare the accuracy of different models SVM fits best in accuracy, after making models we ranked those given articles

TABLE OF CONTENTS

DECLARATION		1
APPROVAL FOR SUBMISSION		2
ACKNOWLEDGEMENTS		4
ABSTRACT		5
TABLE OF CONTENTS		6
LIST OF FIGURES		8
LIST OF SYMBOLS / ABBREVIATIONS		9
 CHAPTER		
1	INTRODUCTION	10
	Background	10
	1.2 Problem Statements	10
	1.3 Aims and Objectives	10
	1.4 Scope of Project	10
 2	LITERATURE REVIEW	 12
 3	DESIGN AND METHODOLOGY	 14
	3.1 Data Collection	15
	3.2 Data Pre-processing	15
	3.2.1 Tokenization	15
	3.2.2 Stop Words Removal	15
	3.2.3 Feature Extraction	16
	3.3 News Classification	16
	3.3.1 SVM	16

4	IMPLEMENTATION	18
4.1	Implementing pre-processing	18
4.1.1	Stop word removal	18
4.1.2	Stemming and Lemmatization	18
4.1.3	Label encoding	19
4.1.4	Feature extraction	19
4.2	Data split	19
4.3	Feature reduction	19
4.4	Model Implementation	19
4.5	Results	20
5	RESULTS AND DISCUSSIONS	21
5.1	Training results	21
5.2	Testing results	21
5.2	Confusion matrix	23
5.3	Classification report	23
5.4	Discussion	24
6	CONCLUSION AND RECOMMENDATIONS	25
	REFERENCES	26
	APPENDICES	28