A

Mini Project Report

On

“Redundant Question Detection”

Submitted in partial fulfilment for the award of the degree of

Third Year B-Tech

In

INFORMATION TECHNOLOGY

Submitted by

1.Shubham Sanjay Pawar.(20141236)

2. Vishal Hindurao Kamble.(20141234)

Under the Guidance of

Prof. B. S. Yelure

Government College of Engineering, Karad

(An Autonomous Institute of Government of Maharashtra)

Academic Year 2022-2023

Government College of Engineering, Karad

(An Autonomous Institute of Government of Maharashtra)

Department of Information Technology

CERTIFICATE

This is to certify that the Project entitled **"Redundant Question Detection**" has been carried out by the team:

1. **Shubham Pawar.**
2. **Vishal Kamble**.

under the supervision and guidance of our project guide **Prof. B. S. Yelure** with partial fulfillment for the award of the BACHELOR OF TECHNOLOGY in Department of Information Technology from Government College of Engineering, Karad for the academic Year 2022-23 (Sem – V).

Prof. B. S. Yelure

Project Guide

Dr. S. J. Wagh
Head
Information Technology Department

External Examiner

ACKNOWLEDGEMENT

Apart from individual efforts, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental throughout the project work.

It is our privilege to express our gratitude towards our project guide, **Prof. B. S. Yelure,** for their valuable guidance, encouragement, inspiration, and whole-hearted cooperation throughout the project work. We thank him for being a motivation through all our highs and importantly, our lows.

We deeply express our sincere thanks to our Head of Department **Dr. S. J. Wagh** for encouraging and allowing us to present the project on the topic "Redundant Question Detection" and providing us with the necessary facilities to enable us to fulfill our project requirements as best as possible. We take this opportunity to thank all faculty members and staff of the Department of Information Technology, who have directly or indirectly helped our project.

We pay our respects to honorable Principal **Dr. A. T. Pise** for their encouragement. Our thanks and appreciations also go to our family and friends, who have been a source of encouragement and inspiration throughout the project.

ABSTRACT

Redundant questions in databases can have a significant impact, which has led to the development of detecting redundant questions in databases. Question repetition issue found while using question answering sites like Quora, Stack Overflow, Reddit etc. Redundant questions lead to the loss of a rational searching, statistical segregation, and a scarcity of responses to the questioners. Machine Learning and Natural Language Processing can be used to detect duplicate questions. Tokenization and the deletion of stop words are used to preprocess the dataset of over 40,0000 question pairings obtained from Quora. Feature extraction is performed on this pre-processed dataset. Machine learning techniques of Decision Tree, Random Forest, XGboost, Adaboost are applied on the dataset for detecting duplicate questions.Random Forest outperformed Decision Tree, XGboost and Adaboost classifiers with accuracy 81.69 percent.

List Of figures

ABBREVIATIONS

Acronym	Definition
ML	Machine Learning
AI	Artificial Intelligence
AdaBoost	Adaptive Boosting
XGBoost	Extreme Gradient Boosting

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Background:

- Community question answering platforms: CQA are web-based platforms where peoples across the world post the doubts and questions which they have. When a question is posted, users who know the answer, post it on the platform. Question answering platforms maintain a database of questions and answers. So that users can easily access them. Quora, Stack Overflow, Stack Exchange etc. are examples of community question answering sites[1].

- Redundant questions frequently have been asked by different users, which increases redundancy of the database.

- Redundant questions are questions which have the same semantic meaning but are worded differently, hence have the same answers. Redundant questions if undetected waste memory space, create searching issues, spoil the overall user experience of the platform. So, if redundant questions are detected as soon as they are posted a significant amount of time and money could be saved.

1.2 Motivation

Redundant questions cause major problems in databases by increasing redundancy. Thus detecting redundant questions and taking appropriate measures is necessary for reducing redundancy. Traditional manual identification of redundant questions is no longer feasible due to the huge amount of data. Contemporary techniques of machine learning can be employed to automate the task of identifying redundant questions and thus reducing manpower and time required for doing it. Machine Learning models could be trained on redundant question dataset and then be deployed for the task of identifying redundant questions.

1.3 Objective

1) To study Decision Tree, Random Forest, XGboost and Adaboost machine learning classifiers.
2) To build machine learning models using ensemble learning techniques to detect redundant questions.
3) Compare and analyze the performance of Decision Tree, Random Forest, XGboost, Adaboost machine learning algorithms.

1.4 Expected Outcome

1. Redundancy of data will be minimized
2. Significant time savings in searching duplicate question
3. Q&A forums will save money spending on storage and database
4. Consistency across databases will rise
5. After detecting duplicate questions, repeated questions can be removed which will save memory.

1.5 Organization of Report

- In the Chapter 2 - Literature Survey, Research papers on redundant question detection are studied.

- In chapter 3 - problem Definition, Project Problem statement defined with example

- Chapter 4 - Design Methodology includes Dataset Description, Technical Specifications, Block Diagram of Duplicate Question Detection approach.

- Chapter 5 consists of implementation and result.
  Implementation contains descriptions about exploratory data analysis, Data Preprocessing, Feature addition, train test split and model training.
  Confusion matrix and accuracy , recall , precision and F1 score of Decision tree, Random Forest, XGboost, Adaboost algorithms has been discussed in the results.

- Conclusion and Future Scope has been discussed in Chapter 6

Chapter 2

LITERATURE SURVEY

In the past, many researchers have worked on duplicate question detection problem using Machine Learning and Deep Learning approach, here is similar work to our project

Table 2.1. Literature Survey

Research Paper	Objective	Dataset Used	Methodology /Algorithm	Performance Metric	Remark
1	Solve the problem of duplicate questions using Word2vec CNN, RNN	Stack Overflow	DL : WV CNN, WV, RNN, WV LSTM	Recall Rate Top 5 : 82.06 Top 10 : 82.15 Top 20: 82.20	This approach is performed using ML algo. in terms of recall rate.
2	Using semantic matching modeling to capture semantic meaning to detect duplicate questions	CQA upState, MOOC	Semantic Matching Modeling (SMM)	Accuracy / F1 0.897/ 0.781	SMM performance > CNN performance Training Time could be increased for better results
3	Detecting duplicate questions using ML models by considering tf-idf and assigning weight to words	Quora	Logistic Regression, Decision Tree, Decision Tree With Bagging Adaboost	Accuracy Adaboost 81.73%	Building the model by considering word match and tf-idf
4	To Solve Quara question Duplication Problem	Quora	Decision Tree	Accuracy Decision Tree : 73%	SVM gives less accuracy than XGboost

	using Decision Tree, KNN, Naive Bayes algorithm		Naive Bayes KNN	Naive Bayes : 56% KNN : 61%	
5	Identifying Similar Question Pairs Using ML Techniques	Quora	Logistic regression Linear SVM, Gradient Boost algorithm	Log loss function, XGboost 0.357	XGboost improves precision & accuracy
6	Identifying Similar Question Pairs Using ML Techniques	Quora	KNN, XGBoost, Decision Tree, Random forest, Adaboost	Accuracy / F1, KNN : 0.78 / 0.75, XGBoost : 0.82 / 0.80	XGBoost gives higher accuracy
7	Review on Exploring Similarity between Two question using ML	Quora	Naive Bayes Classifier, Logistic regression, SVM	Accuracy : 76%	SVM model has low accuracy
8	Identification of Duplication in question in questions posed on knowledge based platform Quora using ML, DL	Quora	Naive Bayes Classifier, Logistic regression, SVM, XGBoot, ANN, RNN	Accuracy XGBoost : 80.89 % RNN : 79.63	Word2vec can be used instead of BOW models (count & TFIDF vectorizer)
9	Duplicate Question pair Detection with ML	Quora	SVM, Logistic regression, Naive Bayes Classifier, Random Forest	Classification , Accuracy	ML algorithm solve redundant question problem

PROBLEM DEFINITION

2.2 Problem Definition:

Redundant Question Detection Using Machine Learning.

Quora and Stack Exchange are knowledge-sharing platforms where people can ask questions in the hopes of attracting high-quality answers. Often, questions that people submit have previously been asked. Companies like Quora can improve user experience by identifying these duplicate entries. This would enable users to find questions that have already been answered and prevent community members from answering the same question multiple times.

Consider the following pair of questions:

Is talent nurture or nature?

Are people talented by birth or can it be developed?

These are duplicates; they are worded differently, but they have the same intent. This project focuses on solving the problem of duplicate question identification.

Chapter 3

DESIGN METHODOLOGY

3.1 Block Diagram

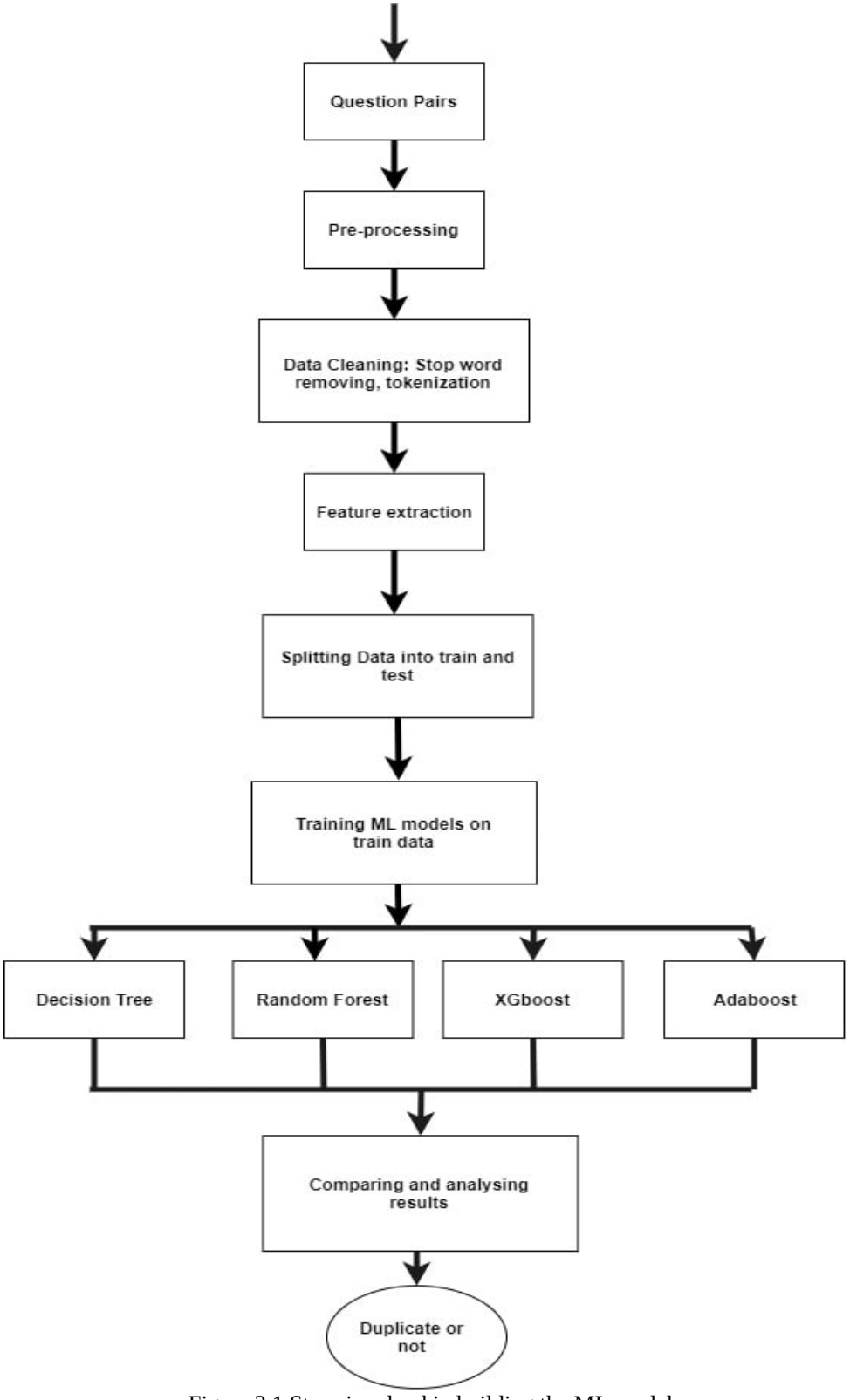

Figure 3.1 Steps involved in building the ML model

3.2 Technical Specifications

Numpy

NumPy (Numerical Python) is an open source Python library that's used in almost every field of science and engineering. It's the universal standard for working with numerical data in Python, and it's at the core of the scientific Python and PyData ecosystems. NumPy users include everyone from beginning coders to experienced researchers doing state-of-the-art scientific and industrial research and development. The NumPy API is used extensively in Pandas, SciPy, Matplotlib, scikit-learn, scikit-image and most other data science and scientific Python packages.

The NumPy library contains multidimensional array and matrix data structures (you'll find more information about this in later sections). It provides ndarray, a homogeneous n-dimensional array object, with methods to efficiently operate on it. NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

Pandas

- Pandas is an open source library in Python. It provides ready to use high-performance data structures and data analysis tools.
- Pandas module runs on top of NumPy and it is popularly used for data science and data analytics.
- NumPy is a low-level data structure that supports multi-dimensional arrays and a wide range of mathematical array operations. Pandas has a higher-level interface. It also provides streamlined alignment of tabular data and powerful time series functionality.
- DataFrame is the key data structure in Pandas. It allows us to store and manipulate tabular data as a 2-D data structure.
- Pandas provides a rich feature-set on the DataFrame. For example, data alignment, data statistics, slicing, grouping, merging, concatenating data, etc

Matplotlib

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing features to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython.Conventionally, the package is imported into the Python script by adding the following statement
–

from matplotlib import pyplot as plt

Scikit-learn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1 Dataset

Dataset for this project is taken from kaggle which has been posted by quora. It consists of 420000 rows and 6 columns. The first feature contains the unique id for each row. The second and third column contains qid1 and qid2 which are the ids for question 1 and 2. The fourth and fifth column contain question 1 and question 2 and the last column contains categorical values 1's and 0's denoting duplicate and non - duplicate questions.

Table 4.1 A snapshot of Dataset

id	qid1	qid2	question1	question2	is_duplicate
339499	665522	665523	Why was Cyrus Mistry removed as the Chairman o...	Why did the Tata Sons sacked Cyrus Mistry?	1
289521	568878	568879	By what age would you think a man should be ma...	When my wrist is extended I feel a shock and b...	0
4665	9325	9326	How would an arbitrageur seek to capitalize gi...	How would an arbitrageur seek to capitalize gi...	0
54203	107861	107862	Why did Quora mark my question as incomplete?	Why does Quora detect my question as an incomp...	1
132566	262554	91499	What is it like working with Pivotal Labs as a...	What's it like to work at Pivotal Labs?	0

4.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a procedure to understand the data and recognize underlying patterns in the data. In this work EDA was performed on the dataset. It gave information about total number of duplicate questions and non-duplicate questions, number of repeated questions. The pie chart given in fig.3. Shows that about 63 percent of the questions in the dataset are non-duplicate while 37 percent are duplicate.

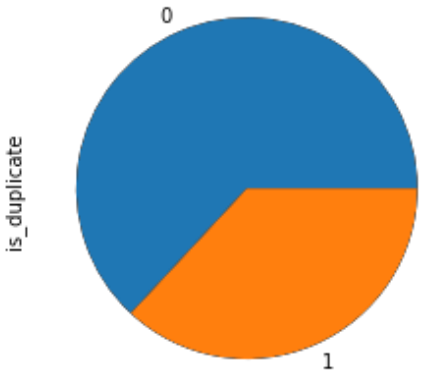

Fig.4.1. Percentage of duplicate (1) and non-duplicate (0) questions in dataset

4.3 Data Cleaning and Preprocessing

The analysis results obtained from the EDA provide the data that are repeated. Thus those need to be removed from the dataset. Removing repeated data will reduce data size. If not removed, duplicate data unnecessarily consumes memory space and machine learning models take more time to train.

In Pre-processing step stop words were removed from the dataset since stop words do not contribute to the semantic meaning of the sentence. Stop words are basically the set of most commonly used words in any language. In English language - the, a, for, to, but, yet, so, are some of the examples of the stop words. Removing stop words further reduces the size of the dataset. The nltk (natural language toolkit) library provides a collection of stop words in the English language; it was used for the task of removing stop words. Conversion of both the questions to lowercase was done. Further in the pre-processing step acronyms were transformed to their full form so the ML model could better capture the insights. Special symbols like %, $, @ were transformed into words like percent, dollar, at, in the pre-processing step so the dataset comes into standard format usable in ml model training. Numerical figures were also converted to corresponding word representation for ex. 000,000 were written ‘m’ and 000 were written as ‘k’ and so on.

4.4 Feature Extraction

As the dataset contains only six features which are not enough in deciding duplicity of question pairs , so need to add extra features to the existing ones to get better performance of the model. Basic features like q1 length, q2 length, q1 words, q2 words, word_common, word_share are extracted from the dataset and added thus increasing the number of features [9] .
Basic feature added are as follows

q1 length –number of characters in question1
q2 length – number of characters in question2
q1 words – number of words in question1
q2 words – number of words in question2
word_total – total words in question1 plus question2
word_common –number of words common to both the questions

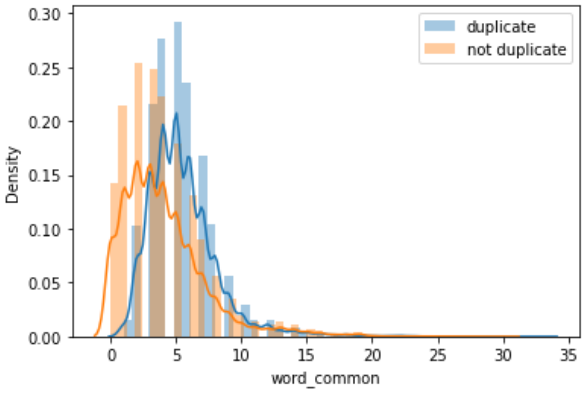

Figure 4.2: plot of word_common in duplicate and non-duplicate questions

The above figure illustrates how important the word similarity that is extracted from the dataset to determine question redundancy. When the word similarity count is greater than 5, there is a higher probability that a question will be redundant. This distinction between duplicate and non-duplicate questions is observed in the Figure 4.2.

word share - ratio of common words to the total words in both the questions[5]

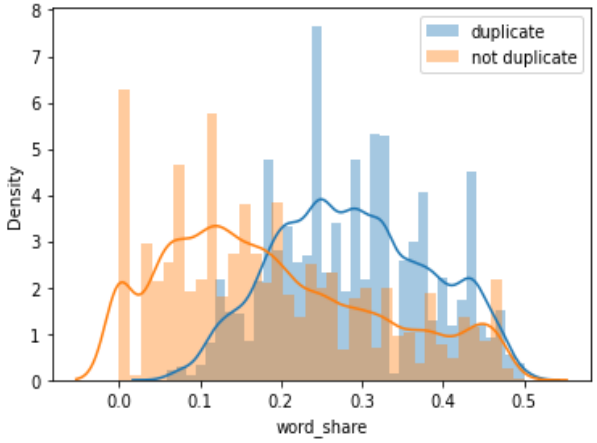

Figure 4.3: Plot of word_share in duplicate and non-duplicate questions

Word share is the ratio of similar words to the total words in both the questions. It is clear from the Figure 4.3 that there is a very minimal chance of questions being redundant when the word share between them is less than 0.2 and a high probability if it is greater than 0.2.

4.5 Advance Features

cwc_min – ratio of number of common words to the minimum number of words among question1 and question2 [9].

cwc_max – ratio of number of common words to the maximum number of words among question1 and question2 [9].

csc_min – ratio of number of common stop words in question1 and question2 to the minimum number of stop words among question1 and question2 [9].

csc_max – ratio of the number of common stop words in question1 and question2 to the maximum number of stop words among question1 and question2

Last_word_eq – numerical value 1 or 0 depending on whether the last word of both the questions is equal or not.

First_word_eq – numerical value 1 or 0 depending on whether the last word of both the questions is equal or not.

Longest_substr_ratio – ratio of the length of common substring to the minimum length question among question1 and question2 [9].

4.6 Vectorization

Bag of words technique is used for vectorization of text data.
Machine Learning algorithms cannot work with raw text data directly, so it's needed to transform the text data into numbers, especially vectors of numbers before a Machine Learning model could be trained as text input. In this work bag of word technique is used for transforming text to vectors. Bag of word counts the occurrence of words in documents and adds each unique word as a new feature to the documents, if a particular word is present in a document then count of the occurrence of that word is added to that feature. For this work 3000 bag of word features for question1 and 3000 bag of word features have been computed and added to the dataset.

4.7 Splitting Dataset

Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. Data splitting is an important aspect of Machine Learning, particularly for creating models based on data.
Prior to the training ML models on available data, the dataset is divided as 70 percent for training and 30 percent for testing.

4.8 Machine Learning Models

Following five machine learning classifiers are used for this work.

Decision Tree: Decision tree is the most powerful tool for classification problems. Decision tree is a tree based classification algorithm in which intermediate nodes represent decision nodes and leaf nodes represent the predicted output. Decisions are taken by splitting the nodes based on attribute selection measures. [5]

Random Forest: Random Forest is the ensemble learning technique which combines a number of decision trees. Decision trees are the building blocks of random forest. Number of decision trees operate as ensemble independent of each other and take decisions by majority voting.[5]

XGBoost: XGBoost is short for extreme gradient boosting. It is an ensemble, decision tree based machine learning algorithm that uses gradient boosting framework. Gradient boosting is a machine learning technique which ensemble weak prediction models to make a strong prediction model.

Adaboost: Adaboost is a boosting technique which ensembles multiple weak classifiers into strong classifiers. A poorly performing classifier is termed as weak classifier. Adaboost combines such weak classifiers to form the strong classifier thus improving performance. Adaboost is short for Adaptive Boosting.[5]

4.9 RESULT

Confusion Matrix

The confusion matrix was utilized for the performance evaluations of the methods used after the classification. It's a two dimensional matrix one representing actual values and others representing predicted values. Following are confusion matrix for Decision Tree, Random Forest, XGboost and Adaboost.

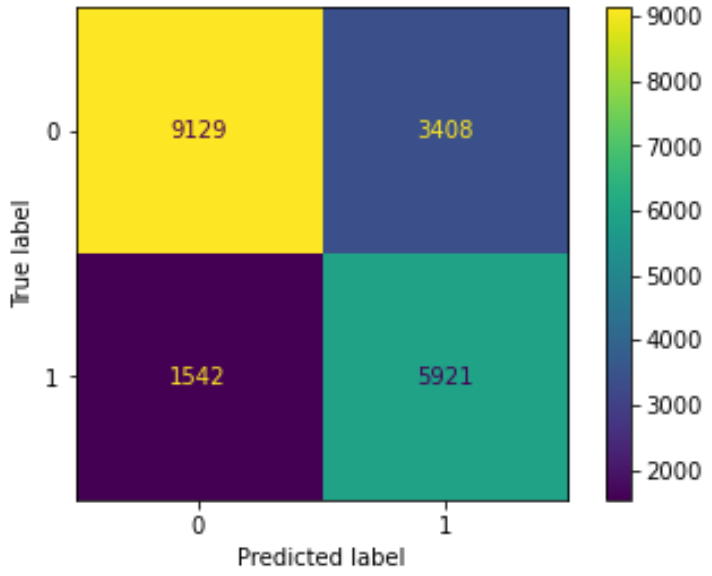

Fig 4.4 Decision Tree

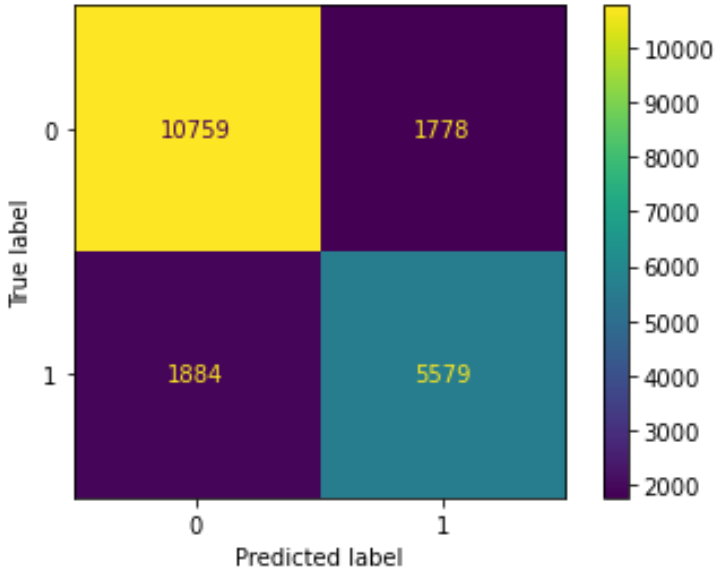

Fig. 4.5 Random Forest

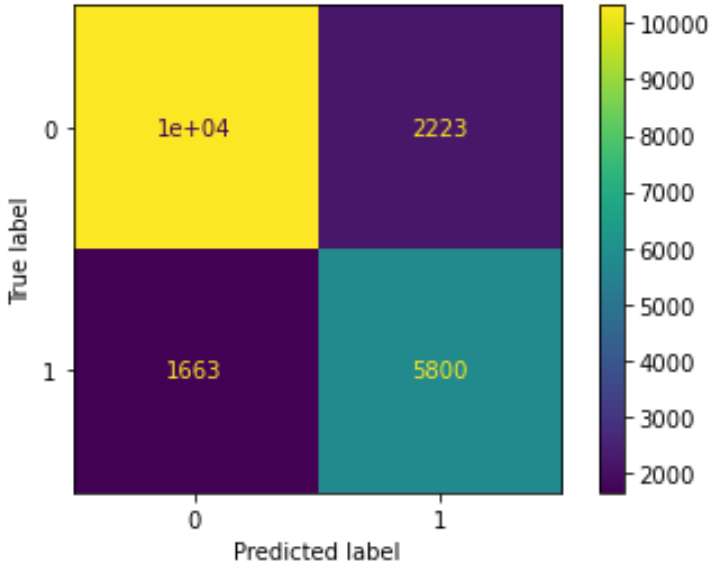

Fig. 4.6 XGboost

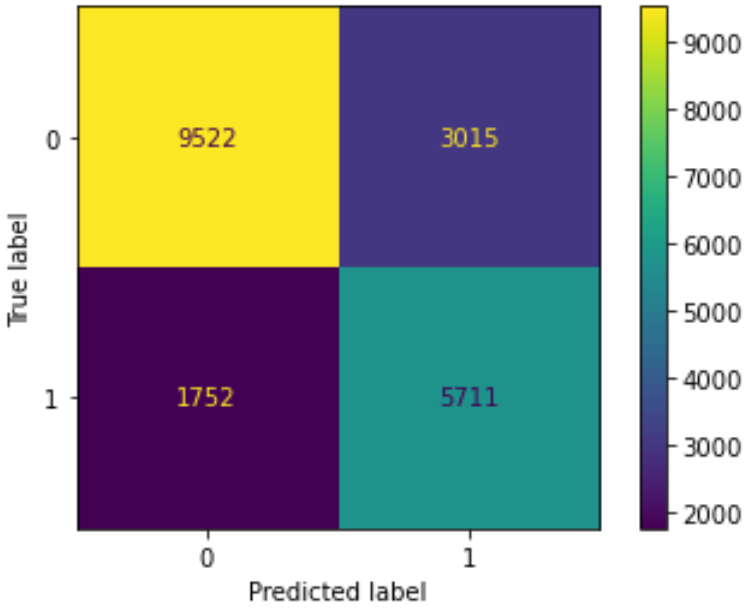

Fig 4.7 Adaboost

True Positive (TP) = the ratio of total number of observation is positive and is predicted to be positive to the total number of predictions made.

True Negative (TN) = the ratio of total number of observation is negative and is predicted to be negative to the total number of predictions made.

False Positive (FP) = the ratio of total number of observation is negative and is predicted to be positive to the total number of predictions made.

False Negative (FN) = the ratio of total number of observation is positive and is predicted to be negative to the total number of predictions made.

Accuracy (%) = the percentage of ratio of correct predictions to the total predictions done.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 score} = \frac{TP}{TP*\frac{1}{2}*(FP+FN)}$$

Table 4.1 Model performance comparison

	Decision Tree	Random Forest	XGBoost	AdaBoost
Accuracy	0.7525	0.8169	0.8057	0.7498
Precision	0.77	0.82	0.81	0.74
Recall	0.75	0.82	0.81	0.76
F1	0.76	0.82	0.81	0.74

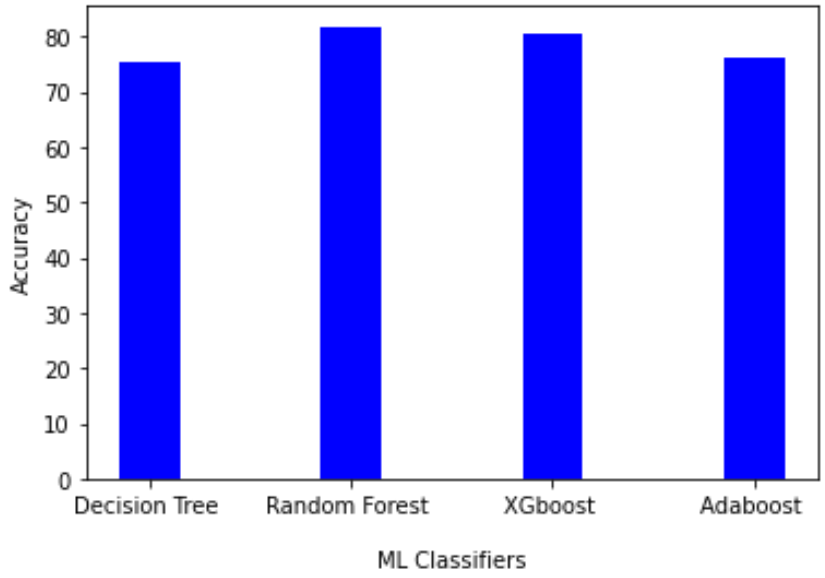

Fig. 4.8 Comparison of Classifiers

As represented in Figure 4.8, Random Forest outperforms Decision Tree by a margin of accuracy of 6.44 percent. Similar to XGboost and Adaboost, the Random Forest classifier outperforms them with accuracy rates of 1.12 and 6.71 percent. In light of this, it can be said that Random Forest is the best machine-learning classifier for redundancy detection.

Chapter 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This study uses Machine Learning and Natural Language Processing to classify whether question pairings are duplicates or not in Q&A forums. The use of minimal cost architecture and the selection of highly dominating elements from the questions make it an effective template for detecting duplicate inquiries and subsequently finding high-quality answers. Among the proposed models Random Forest beats Decision Tree, XGboost and Adaboost ML models. Random Forest also reduces false positive rate which is of utmost importance while detecting redundant questions.

5.2 Future Scope

This research work provides good results and can be used in predicting duplicate questions for study purposes. However, few complications like extraction of features and vectors, heavy use of memory by .csv file or any other file has to be taken care of in future work. Due to memory issues it is difficult to load and save any changes every single time. Deep Learning techniques could be used for better capturing semantic meaning of questions. An application of this work could be in detecting similarity of questions in question papers.

REFERENCES

[1] https://www.kaggle.com/c/quora-question-pairs- 25 July 20222

[2]https://www.quora.com/How-does-the-Quora-Digest-work -26 July 2022

[3] JING JIANG, LI ZHANG AND LITING WANG "Duplicate Question Detection With Deep Learning in Stack Overflow"(2020).[DOI:10.1109/ACCESS.2020.2968391]

[4] ZHUOJIA XU AND HUA YUAN "Forum Duplicate Question Detection by Domain Adaptive Semantic Matching"(2020).[DOI:10.1109/ACCESS.2020.2982268]

[5] BASAVESHA D, Dr. Y S NIJAGUNARYA "Detecting Duplicate Questions in Community Based Websites Using Machine Learning".
[https://ssrn.com/abstract=3835083]

[6]Ms. Vishwaja M. Tambakhe, Dr. Kishor P.Wag "Review on Exploring Similarity between two Questions using Machine Learning".
[DOI: https://doi.org/10.32628/CSEIT217360]

[7]R. Rishickesh, R.P. Ram Kumar, A.Shahina, A. Nayeemullah Khan "Identification of Duplication in Questions Posed on Knowledge Sharing Platform Quora using Machine Learning Techniques"'. [DOI:10.35940/ijitee.L3017.1081219]

[8]Chakaveh Saedi, Joao Rodrigues, Jo ˜ ao Silva, Ant ˜ onio Branco, Vladislav Maraev "Learning Profiles in Duplicate Question Detection".[DOI 10.1109/IRI.2017.39]

[9]Vivek Bhalerao,Sathya Ar,Sandeep Kumar Panda "A Machine Learning Model to identify Duplicate Questions on Social Media Forums".
[DOI:10.35940/ijitee.D1362.029420]

[10]MUHAMMAD UMER,MUHAMMAD AHMAD,SALEEM ULLAH,GYU SANG CHO,AND ARIF MEHMOOD "Duplicate Questions Pair Detection Using Siamese MaLSTM".[DOI:10.1109/ACCESS.2020.2969041]

[11] Seema Rania , AvadheshKumarb , Naresh Kumarc , Sanjay Kumard "Deep Neural Model for Duplicate Question Detection Using Support Vector Machines (Svm)"[Vol.12 No.6 (2021),
4024-4033]
[12]Random Forest Algorithm: https://www.javatpoint.com/machine-learning-random-forest-algorithm - 18 Aug 2022
[13] https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html 2 Sep 2022