A
Project Report
On

Testing Portal For Quora Insincere Questions Classification
B. Tech. - Semester VII

Prepared By

Aniket Rajani(IT-94)
Hitenkumar Rathod(IT-97)

DEPARTMENT OF INFORMATION TECHNOLOGY
FACULTY OF TECHNOLOGY,
DHARMSINH DESAI UNIVERSITY,
COLLEGE ROAD, NADIAD- 387001

December, 2019

A
Project Report
On

Testing Portal For Quora Insincere Questions Classification

In partial fulfillment of requirements for

System Design Practice

B. Tech. - Semester VII

Submitted By:

1. **Aniket Rajani**
2. **Hitenkumar Rathod**

Under the Guidance of

Prof. N. P .Desai

DEPARTMENT OF INFORMATION TECHNOLOGY

FACULTY OF TECHNOLOGY, DHARMSINH DESAI UNIVERSITY
COLLEGE ROAD, NADIAD- 387001

CANDIDATE’S DECLARATION

We declare that pre-final semester report entitled **“Testing Portal for Quora Insincere Questions Classification**” is our own work conducted under the supervision of the guide Prof. N. P .Desai.

We further declare that to the best of our knowledge the report for B. Tech. VII semester does not contain part of the work which has been submitted either in this or any other university without proper citation.

Aniket Rajani
Student ID: 16ITUOS112

HitenKumar Rathod
Student ID:16ITUBS054

Submitted To:

Department of Information Technology,
Faculty of Technology,
Dharmsinh Desai University, Nadiad
Gujarat.

DHARMSINH DESAI UNIVERSITY
NADIAD-387001, GUJARAT

CERTIFICATE

This is to certify that the project carried out in the subject of Software Design Project , entitled "Testing Portal for Quora Insincere Questions Classification" and recorded in this report is a bonafide report of work of

1) Aniket Rajani Roll No. IT-94 ID No: 16ITUOS112

2) Hitenkumar Rathod Roll No. IT-97 ID No: 16ITUBS054

Of Department of Information Technology, semester VII, They were involved in Project work during academic year 2019 -2020.

Prof. N. P .Desai.,
Department of Information Technology,
Faculty of Technology,
Dharmsinh Desai University, Nadiad
Date:

Prof. (Dr.) V . K. Dabhi,
Head , Department of Information Technology,
Faculty of Technology,
Dharmsinh Desai University, Nadiad.
Date:

ACKNOWLEDGMENT

We would like to express our profound gratitude to our project guide, Prof. Nikita Patel Desai, as well as our head of department, Prof. (Dr.) Vipul Dabhi, who gave us this resplendent opportunity to work on this project. Working on this project led us to research extensively in this field, which subsequently helped us grow our knowledge immensely.

With Regards
Aniket Rajani
Hiten Rathod

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1 PROJECT DETAIL

An existential problem for any major website today is how to handle toxic and divisive content. Quora wants to tackle this problem head-on to keep their platform a place where users can feel safe sharing their knowledge with the world.

Quora is a platform that empowers people to learn from each other. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers. A key challenge is to weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

In this project, we will develop models that identify and flag insincere questions

1.2 PURPOSE

The main purpose of this project is weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers.
In this Project our task is to predict whether a question asked on Quora is sincere or not.Some characteristics that can signify that a question is insincere:

- Has a non-neutral tone
  - Has an exaggerated tone to underscore a point about a group of people
  - Is rhetorical and meant to imply a statement about a group of people
- Is disparaging or inflammatory
  - Suggests a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
  - Makes disparaging attacks/insults against a specific person or group of people
  - Based on an outlandish premise about a group of people
  - Disparages against a characteristic that is not fixable and not measurable
- Isn't grounded in reality
  - Based on false information, or contains absurd assumptions

- Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

1.3 SCOPE

To date, Quora has employed both machine learning and manual review to address this problem. With the help of this project, they can develop more scalable methods to detect toxic and misleading content.The project can help Quora uphold their policy of “Be Nice, Be Respectful” and continue to be a place for sharing and growing the world’s knowledge

1.4 OBJECTIVE

The objective is to predict whether a question asked on Quora is sincere or not. Our main goal is to develop more scalable methods to detect toxic and misleading content.The project can help Quora uphold their policy of “Be Nice, Be Respectful” and continue to be a place for sharing and growing the world’s knowledge.

1.5 TECHNOLOGY AND LITERATURE REVIEW

Front End:

HTML, MaterializeCSS Framework

Backed End:

Flask

Language:

Since, Python is a General-Purpose Programming Language ,the language used for the project is Python. Since Python provides many inbuilt Packages for Data Manipulation, Data Wringling, Exploratory Data analysis and Machine Learning.It is one of the popular languages used in Data Science

2. PROJECT MANAGEMENT

2.1 FEASIBILITY STUDY

2.1.1 Technical Feasibility

This project is Machine Learning and Software based Project.

Each of the technologies used in the project are freely available to learn and technical skills required are manageable between the team members. Time limitation of project development and ease of implementing using this technologies are synchronized.

2.1.2 Time Schedule Feasibility

The estimated time to complete this project would be around 3-4 months. One of the team members is working on the software/Web-App that will interact with the users and other team Members are working on Machine Learning Algorithms. Since parallelism of tasks is achieved the project can be completed in a reasonable amount of time

2.1.3 Implementation Feasibility

The main tools and technologies which are used in our project are :

- Flask
- Jupyter(IPython Notebook)
- Scikit-Learn
- Pandas
- Numpy
- Matplotlib
- Seaborn
- nltk
- TextBlob

Team members are skilled in above technologies and there is no such risk involved that can act as an obstacle while implementing this project .So it is expected that the project will be implemented and completed in a reasonable amount of time.

2.2 PROJECT PLANNING

2.2.1 Project Development Approach and Justification

Our project "Testing Portal for Quora Insincere Questions Classification" the important part is machine learning. So, How a typical machine learning project work :

Machine learning project life cycle is divided into six step

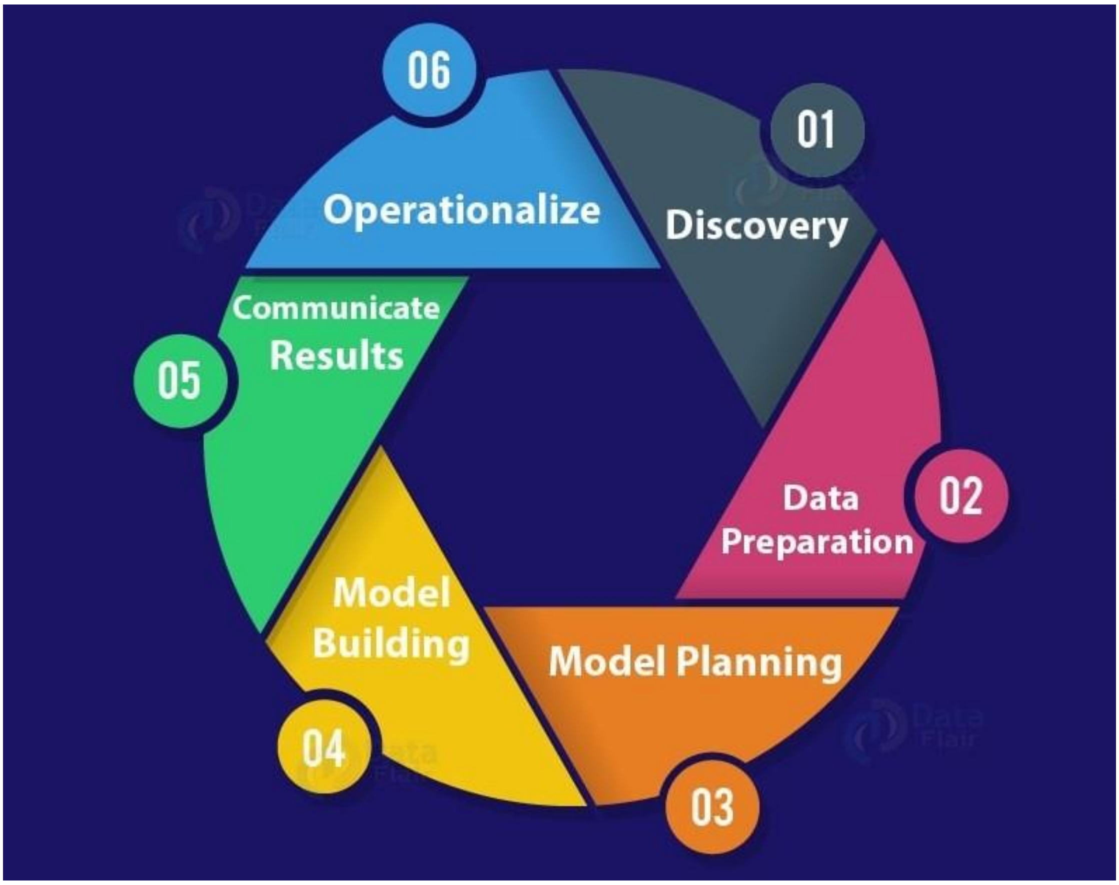


Figure 1 MACHINE LEANING LIFE CYCLE

1. Discovery

2. Data Preparation

3. Model Planning

4. Model Building

5. Communicate results

6. Operationalize

1. Discovery:

In this stage we understand the problem statement, so thorough study of the business model is required to fully understand the concept. In this project our task is to find whether the Question asked on Quora by the user is Sincere or Insincere

2. Data Processing:

This stage is also known as Data munging. It is the most important aspect of Data Science life cycle for any valuable insights to pop up. It includes following steps:

· Data Cleaning: As a part of Data Cleaning stopwords (i.e frequent words) were removed from the Questions. No other Inconsistency in data was obersved by us.

· Data Transformation: Standardization or Z-Score Normalization was applied on Numerical Features to transform the data.

3. Model Planning:

After proper understanding and cleaning of the data suitable model is selected. Selecting a model is totally depended on data type which is been extracted. This step involves Exploratory Data Analysis to understand the relation between variable and to see what data can tell us. Also key variables are selected.

Various techniques such as Histogram, Box Plot, WordCloud,Bar Graph were used for Exploratory Data Analysis.Some of the key insights from the data were:

1. From the WordCloud the words like Muslim,racist,american,Donald Trump,Indian,Pakistan,Gay,Christian and sex seems to be the most important words for declaring a question as an insincere question, Because if the question consists of any of these words than there is high probability of a question being insincere.

2. The length of the insincere questions was more than sincere questions.

3. Number of words for insincere questions are more as compared to sincere questions.

4. Number of sincere questions are much more in number as compared to insincere questions

4. Model Building:

Based on the Observations of the data after the Exploratory Data Analysis Phase Model building is to be carried out.

Various Machine Learning Models like Naïve Bayes, Logistic Regression and SVM were used. Since this is a text Classification Problem, Naïve Bayes is one of the algorithm which is most popular for this problem.

5. Communicate results:

Keys Observations, Conclusions and findings are to be communicated to the technical team.

6. Operationalize:

Final reports code, and technical documents are delivered by the team.

2.2.2 Project Plan

- Understand business requirement : define the problem
- Data Preparation : Cleaning , pre-processing
- Exploratory data analysis
- Modeling , Evaluation , Interpretation
- Deployment
- Testing
- Optimization : improve models , more data , optimize code

2.2.3 Milestones and Deliverables

We deliver web based machine learning system. This contain interesting functionality for end users. In this system end users can perform analysis on the question that is to be asked on Quora with different machine learning algorithms.

2.2.4 Roles and Responsibilities

1. Roles:

Understanding the requirements, purpose, goals and the scale of project

- Finalizing the project problem definition
- To study various machine leaning algorithm
- Studying and understanding of various HyperParameters used in algorithm
- Train machine learning model on Quora Insincere questions dataset.
- Designing the Graphical User interface part of the project
- Deploy machine leaning model using flask.
- Demonstrate and suggested modification.
- Prepare a final report and presentation

2. Responsibility:

As only two member were involved in the whole team each of them had to perform all the tasks as the project proceeded through its different phases. This helped each one to develop all kinds of skill in all the phases.

2.2.5 Group Dependencies

Once resources are allocated, the next step is to identify dependencies between tasks. Dependencies come in many forms: compare model functionality implement after model train.
We have also identified dependencies among the modules and sub modules in our project. Then we have divided our work as per dependency.

3. SYSTEM REQUIREMENTS STUDY

3.1 STUDY OF CURRENT SYSTEM

Currently, Quora uses various Machine Learning and Deep Learning based Algorithms like LSTM,RNN,GRU etc to detect toxicity of the questions asked.

3.2 PROBLEMS AND WEAKNESS OF CURRENT SYSTEM

In the Current System, there is no such testing Portal where users can check the toxicity level of the Question i.e before actually posting the question on quora based on the policies and terms and condition of Quora.

3.3 USER CHARACTERISTIC

Users can be anyone who are active on Quora, Who use Quora on regular basis, who frequently answer’s the Questions on Quora, Who Frequently asks/post Questions on Quora.

3.4 HARDWARE AND SOFTWARE REQUIREMENT

3.4.1 Hardware Requirement

- RAM :- Atleast 8GB
- Processor :- AMD or Intel
- Operating system :- Any OS like Window, Linux, Mac

3.4.2 Software Requirement

- NumPy
- Pandas
- Scikit-learn
- Matplotlib
- Jupyter(IDE)
- Flask

- Web Browser

3.5 CONSTRAINTS

3.5.1 Higher order language requirements:

The higher level language used here is Python and its packages are:

- Flask
- Numpy
- Scikit-Learn
- Matplotlib
- Pickle

3.5.2 Reliability requirement:

The Reliability of overall program depends on the reliability of the separate Predict function machine learning model.

4. SYSTEM ANALYSIS

4 .1 REQUIREMENT OF NEW SYSTEM

R1 : Login

Input : Enter Login Credentials Email and Password

Output : Verified by Unique Email and Password

R1.2 : Registration

Input : Enter user details like Username , Email , Password

Output : Verified by giving Unique Id and password and Register the User

R2: Testing

Input : enter question , machine leaning algorithm and type of text vectorization

Output : Question is sincere or insincere

R3 : Compare Two Model

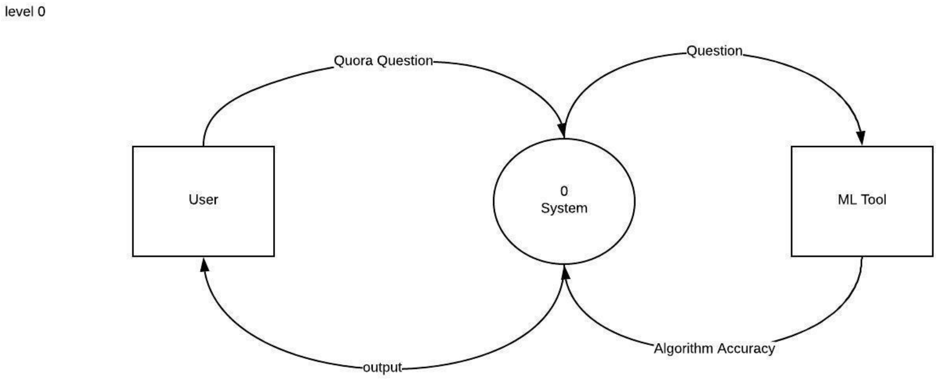
Input : enter question and select model

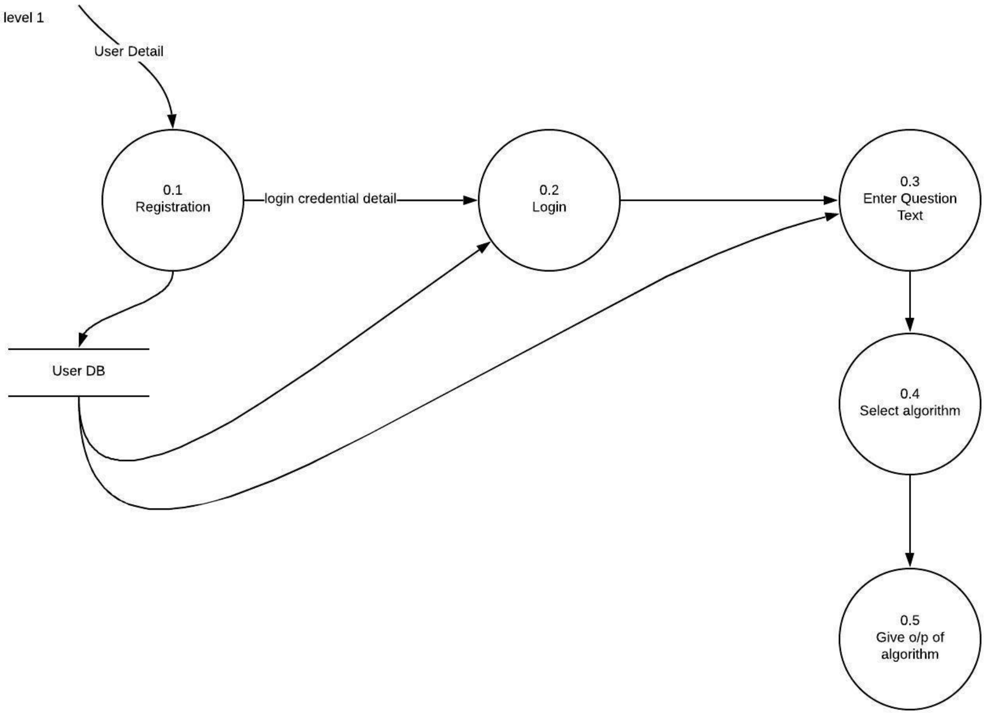
Output : comparison of two model

4.2 FEATURE OF NEW SYSTEM

The New features of the system will allow user to test different machine leaning Models. Also users will be able to compare two machine leaning model and analyze different parameters of different Algorithms by Observing the results of the same question on 2 different models and Algorithms.

4.3 DATA FLOW DIAGRAM (DFD)





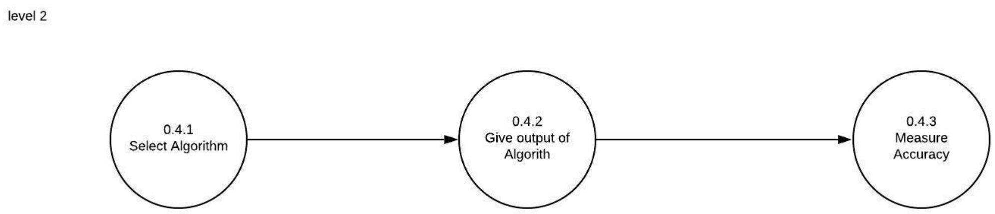


Figure 2 DATA FLOW DIAGRAM

4.2 SEQUENCE DIAGRAM

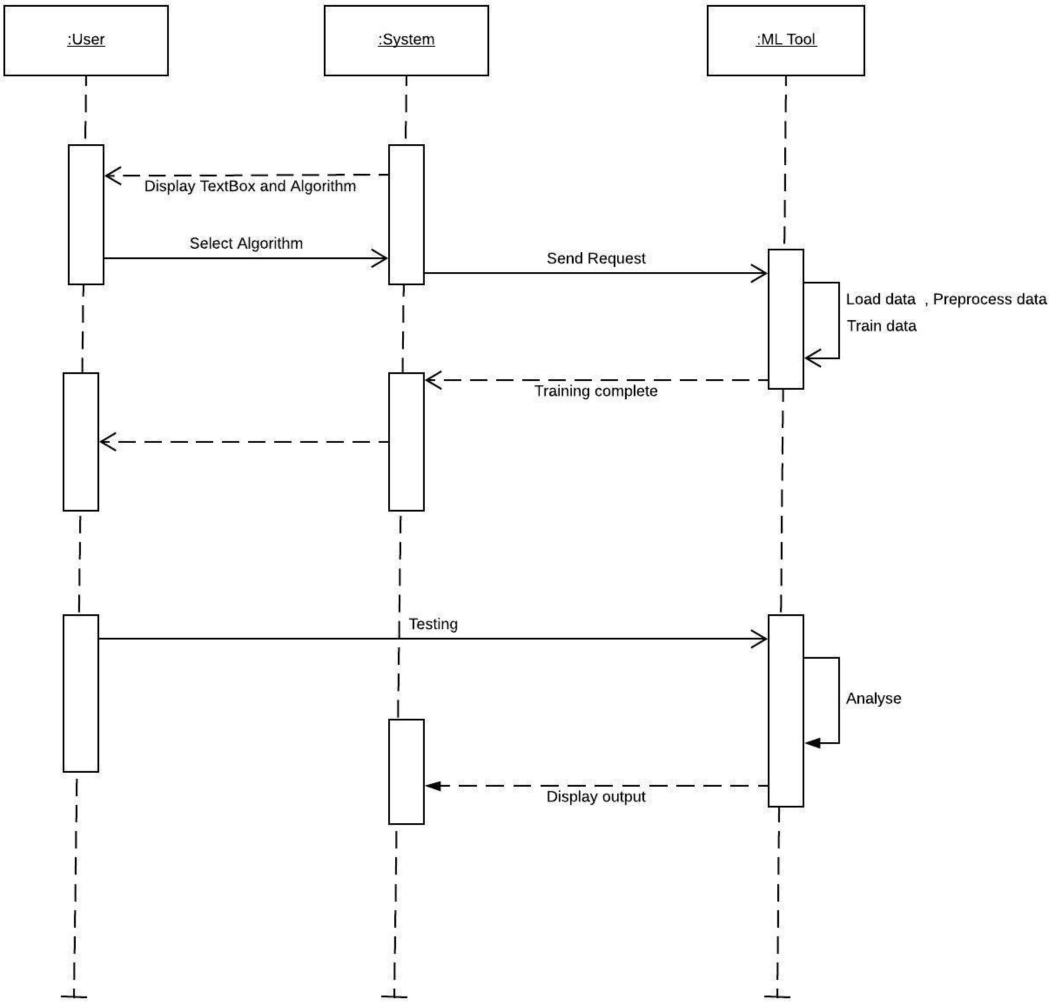


Fig 3 SEQUENCE DIAGRAM

4.3 ACTIVITY DIAGRAM

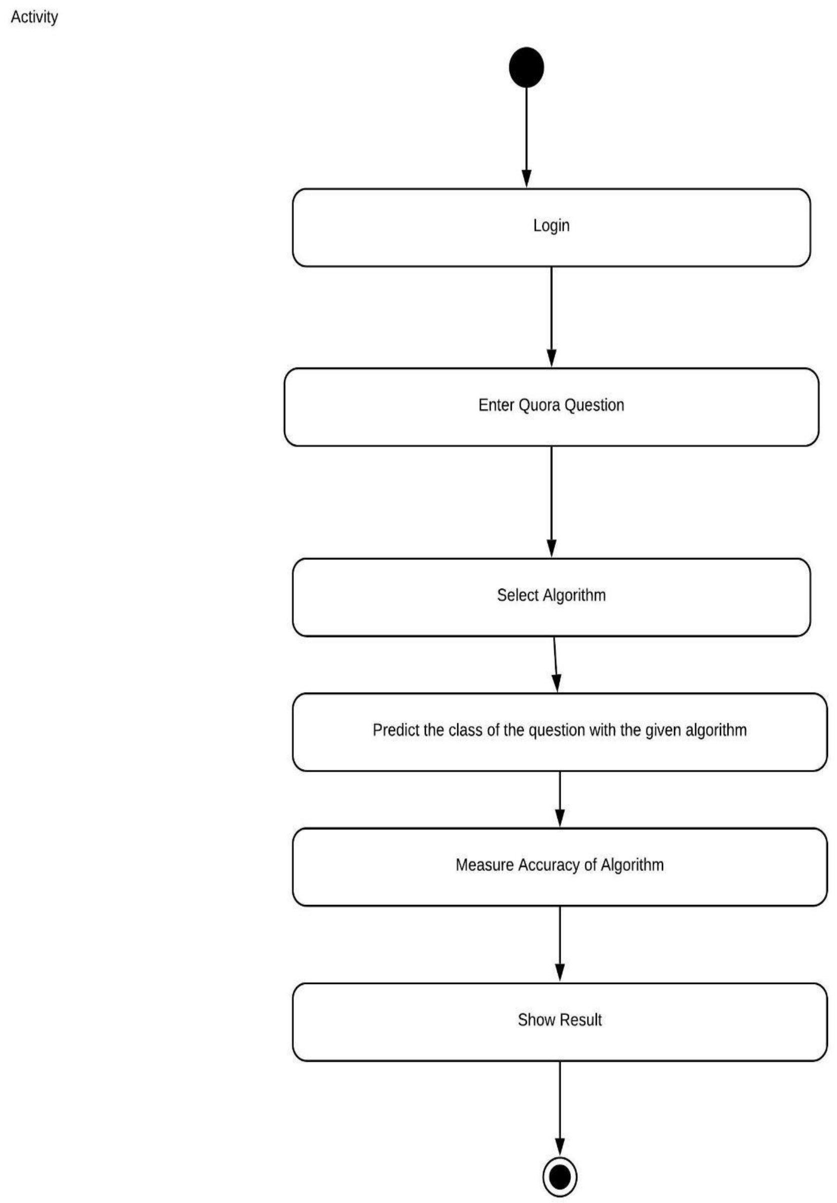

Activity
Login
Enter Quora Question
Select Algorithm
Predict the class of the question with the given algorithm
Measure Accuracy of Algorithm
Show Result

Fig 4 ACTIVITY DIAGRAM

4.4 USECASE

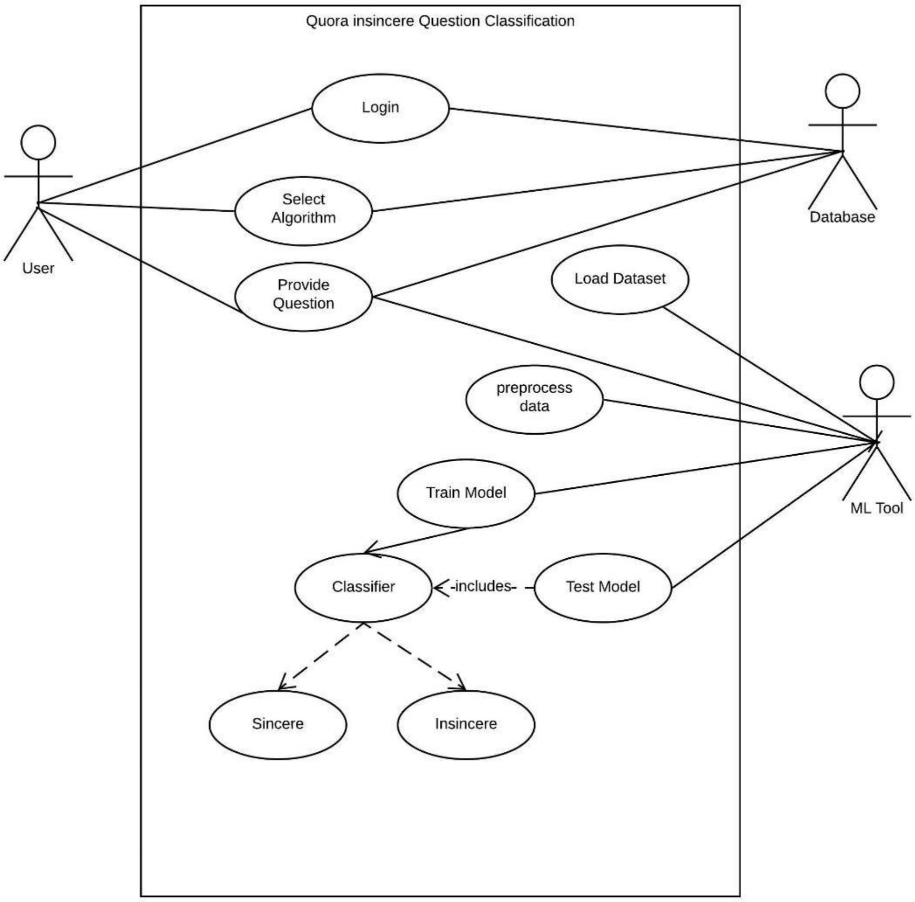


Fig 5 USECASE

5. Machine Learning Algorithms Used:

5.1 Naive Bayes

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Using Bayes theorem, we can find the probability of **y** happening, given that **x** has occurred. Here, **x** is the evidence and **y** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive. The variable **y** is the class variable

X is given as,

$$X = (x_1, x_2, x_3, \ldots.., x_n)$$

Here x_1,x_2….x_n represent the features, i.e they can be mapped to outlook, temperature, humidity and windy. By substituting for **X** and expanding using the chain rule we get,

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remain static. Therefore, the denominator can be removed and a proportionality can be introduced.

$$P(y|x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

In our case, the class variable(**y**) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we need to find the class **y** with maximum probability.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Using the above function, we can obtain the class, given the predictors

5.2 Logistic Regression

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

TASK IN LOGISTIC REGRESSION IS : To find W and b to discover a plane such that it separates positive and negative point i.e we have to find (π). Here, in above image_1 we can find out the distance from any point to plane Pi(π). Also we assumed W is a unit vector and normal to plane. So now comes to interesting part via seeing diagram in below picture i.e if we calculate: 1⇒ Distance from positive point to plane (π) it will be positive , i.e di=W(transpose) Xi>0 because (W and Xi are on same side).

Some important parameters for Logistic Regression Algorithm are:

1. **Alpha** :It is a Float Constant that multiplies the regularization term. Defaults to 0.0001 Also used to compute learning_rate when set to 'optimal'.
2. **learning_rate** :It is the amount that weightds are updated during training.
3. **Penalty**:The penalty (aka regularization term) to be used. Defaults to 'l2' which is the standard regularizer for linear SVM models. 'l1' and 'elasticnet' might bring sparsity to the model (feature selection) not achievable with 'l2'.
4. **Fit_prior** : It is a boolean value Whether to learn class prior probabilities or not. If false, a uniform prior will be used.

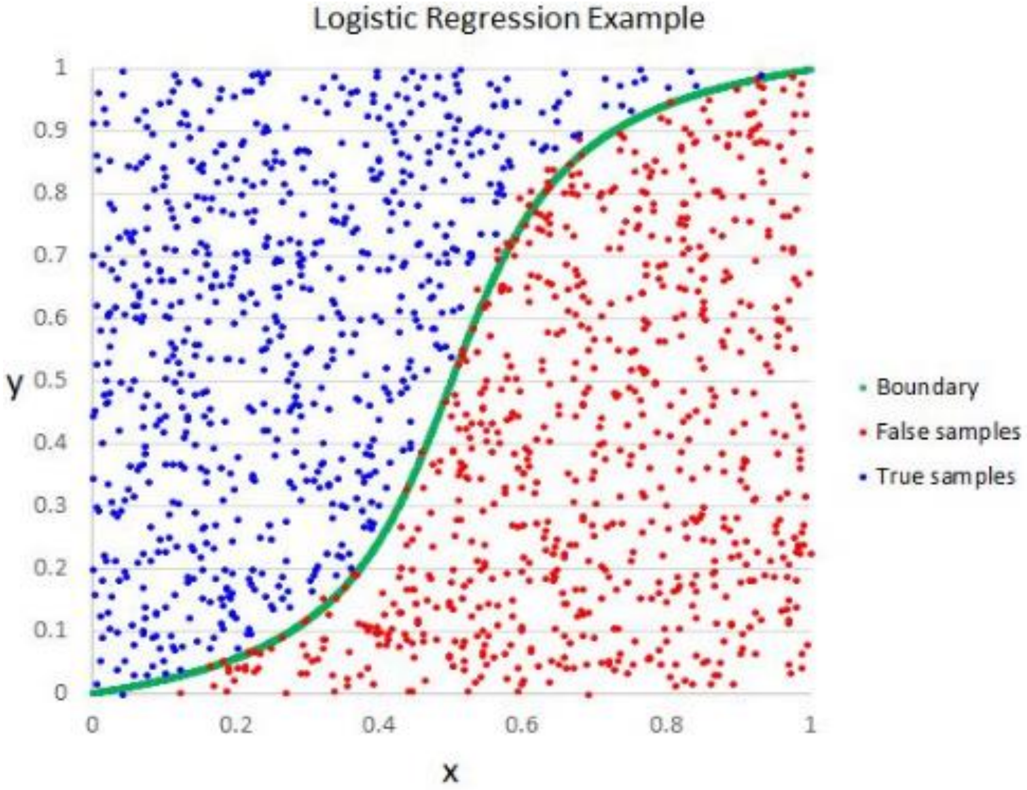
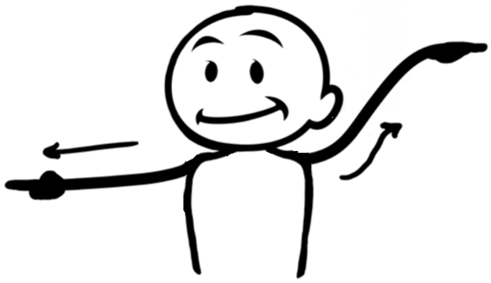


Fig 6 LOGISTIC REGRESSION EXAMPLE



$$y = \frac{1}{1+e^{-x}}$$

5.3 Support Vector Machines

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate

Some important parameters for SVM Algorithm are:

1. **Alpha** :It is a Float Constant that multiplies the regularization term. Defaults to 0.0001 Also used to compute learning_rate when set to 'optimal'.
2. **learning_rate** :It is the amount that weightds are updated during training.
3. **Penalty**:The penalty (aka regularization term) to be used. Defaults to 'l2' which is the standard regularizer for linear SVM models. 'l1' and 'elasticnet' might bring sparsity to the model (feature selection) not achievable with 'l2'.
4. **Fit_prior** : It is a boolean value Whether to learn class prior probabilities or not. If false, a uniform prior will be used.
5. **kernel** : Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable. If none is given, 'rbf' will be used. If a callable is given it is used to pre-compute the kernel matrix from data matrices; that matrix should be an array of shape

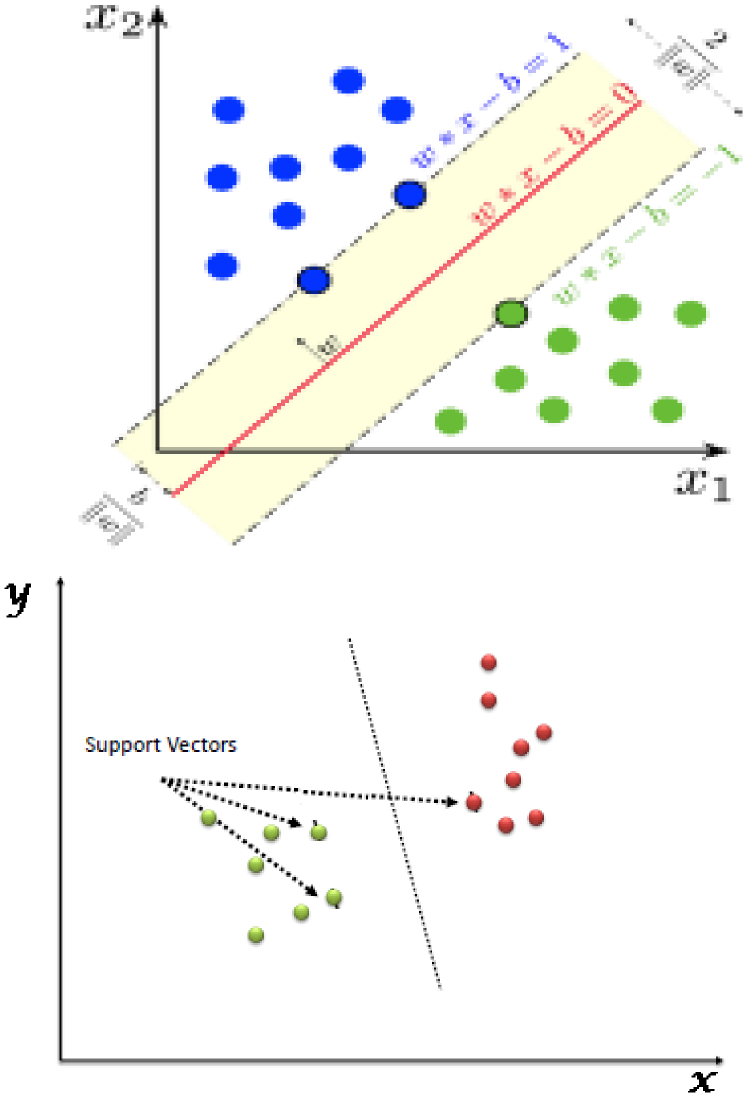


Fig 7 SVM EXAMPLE

6. IMPLEMENTATION PLANNING

6.1 IMPLEMENTATION ENVIRONMENT

Currently, System is multi user. Users can interact with the System using GUI developed using various Frameworks
Python high level packages are used for develop testing functionalities.

6.2 MODULE SPECIFICATION

1. Testing phase

It will provide testing functionality for different machine leaning algorithms.
Users need to enter question, select algorithm and select Text vectorizer then the predicted output will be displayed. Type Of Question(i.e Sincere or Insincere) and various Other Features are displayed in the output.

2. Comparison of Model

It will provide comparison of two machine learning model.Users need to enter question, select algorithm and select text vectorizer and output will displayed. HyperParameter values, Feature Values, Type of Question Predicted by both the Models is displayed in the output that will help the user to compare 2 Models.

3. Data analysis

It will provide information about different plots used for Data Analysis. Various Plots like boxplot ,barplot , WordCloud etc were used for Exploratory Data Analysis.

4. Models

It gives the information of machine leaning algorithms i.e how actually the algorithm works and other information of the HyperParameters used while Implementing the Algorithms.

6.3 CODING STANDARDS

Coding conventions are set of guidelines for a specific programming language that recommend programing language style, practices and methods for each aspect of a program written in that language. These conventions usually cover file organization, indentation, comments, declarations, statements, white space, naming conventions,
Programming practices, programming rules of thumb, architectural best practices, etc.
Software programmers are highly recommended to follow these guidelines to help improve the readability of their source code. Coding conventions are only applicable to human maintainers and peer reviewers of software project. Coding conventions are not enforced by compilers.

7. TESTING

7.1 TESTING STRATEGY

Once source code has been generated, software must be tested to uncover as many errors as possible before delivery to customer. Your goal is to design a series of test cases that have a high likelihood of finding errors. Software testing techniques provide systematic guidance for designing tests that (1) exercise the internal logic of software components, and (2) exercise the inputs and outputs domains of the program to uncover errors in program function, behavior and performance.

During early stages of testing, a software engineer performs all tests. However, as the testing process progresses, testing specialists may become involved. Reviews and other activities can and do uncover errors, but they are not sufficient. Every time the program is executed, the customer tests it! Therefore, you have to execute the program before it gets to the customer with the specific intent of finding and removing all errors. In order to find the highest possible number of errors, tests must be conducted systematically and test cases must be designed using disciplined techniques.

Testing Objective:-

- Testing is a process of executing a program with the intention of finding an error.
- A good test case is one that has a high probability of finding an as-yet undiscovered error.
- A successful test is one that uncover an as-yet undiscovered error.

7.1.1 Unit Testing

Unit testing is a software development process in which the smallest testable part of an application, called units, are individually scrutinized for proper operation. Unit testing is often automated but it can also be done manually. This testing mode is a component of Extreme Programming (XP), a pragmatic method of software development that takes a meticulous approach to building a product by means of continual testing and revision.

Unit testing involves only those characteristics that are vital to the performance of the unit under test. This encourages developer to modify the source code without immediate concerns about how such changes might affect the functioning of the units or the program as a whole. Once of whole of the units in a program have been found to be working in the most efficient and error free manner possible, larger components of the program can be evaluated by means of integration testing. I tested each single part of the entire application. I tested each and every module individually.

7.1.2 Sub System Testing

After testing each unit, we move on to larger units called sub system. In subsystem testing I tested the whole user side as one system. On the user side all the modules like Prediction API, API, etc. were tested together to see if there was any error or bug found.

7.1.3 System Testing

After testing all the sub-system, it is time to test the whole system. System testing of software is testing conducted on a complete, integrated system to evaluate the system's compliance with its specified requirements

7.1.4 Acceptance Testing

Acceptance testing can be connected by the end user, customer, or client to validate whether or not to accept the product. Acceptance testing may be performed as part of the hand-off process between any two phases of development. The acceptance test suite is run again the supplied input data or using an acceptance test script to direct the tester. Then the results obtained are compared with the expected results. If there is a correct match for every case, the test suite is said to pass.

7.2 TESTING METHODS

The verification activities fall into the category of static testing. During static testing, you have a checklist to check whether the work you are doing is going as per the set standards of the organization. These standards can be for coding, integrating and deployment. Reviews, Inspections and Walkthroughs are static testing methodology. Dynamic testing involves working with the software giving input values and checking if the output is as expected. These are the validation activities. Unit test, integration test, System and acceptance tests are few of the dynamic testing methodologies. Alpha & beta testing: the alpha test is conducted at the developer's site by a customer. The software is used in a natural setting with the developer "looking over shoulder" of the user and recording errors and usage problems. Alpha test are conducted in a controlled environment. The beta testing is conducted at one or more customer site by the end-user of the software. Unlike alpha testing, the developer is generally not present. Therefore, the beta test is a "live" application of the software in an environment that cannot be controlled by the developer. TESTING

7.2.1 Black Box Testing

Also known as functional testing. A software testing techniques where by the internal working of the item being tested are not known by the tester. For example, in a black box test on software design the tester only knows the inputs and what the expected outcomes should be and not how the program arrives at those outputs. The tester does not ever examine the programming code and does not need any further knowledge of the program other than its specification.

- The advantages of this type of testing include:
- The test is unbiased as the designer and the tester are independent of each other
- The tester does not need knowledge of any specific programming languages
- The test is done from the point of view of the user, not the designer
- Test cases can be designed as soon as the specifications are complete.
- The disadvantages of this type of testing include:
- The test can be redundant if the software designer has already run a test case
- The test cases are difficult to design
- Testing every possible input stream is unrealistic because it would take an inordinate amount of time: hence many program paths will go untested.

7.2.2 White Box Testing

Also known as glass box, structural, clear box and open box testing. A software testing technique whereby explicit knowledge of the internal workings of the item being tested are used to select the test data. Unlike black box testing, white box testing uses specific knowledge of programming code to examine outputs. The test is accurate only if the tester knows what the program is supposed to do. He or she can than see if the program diverges from its intended goal.

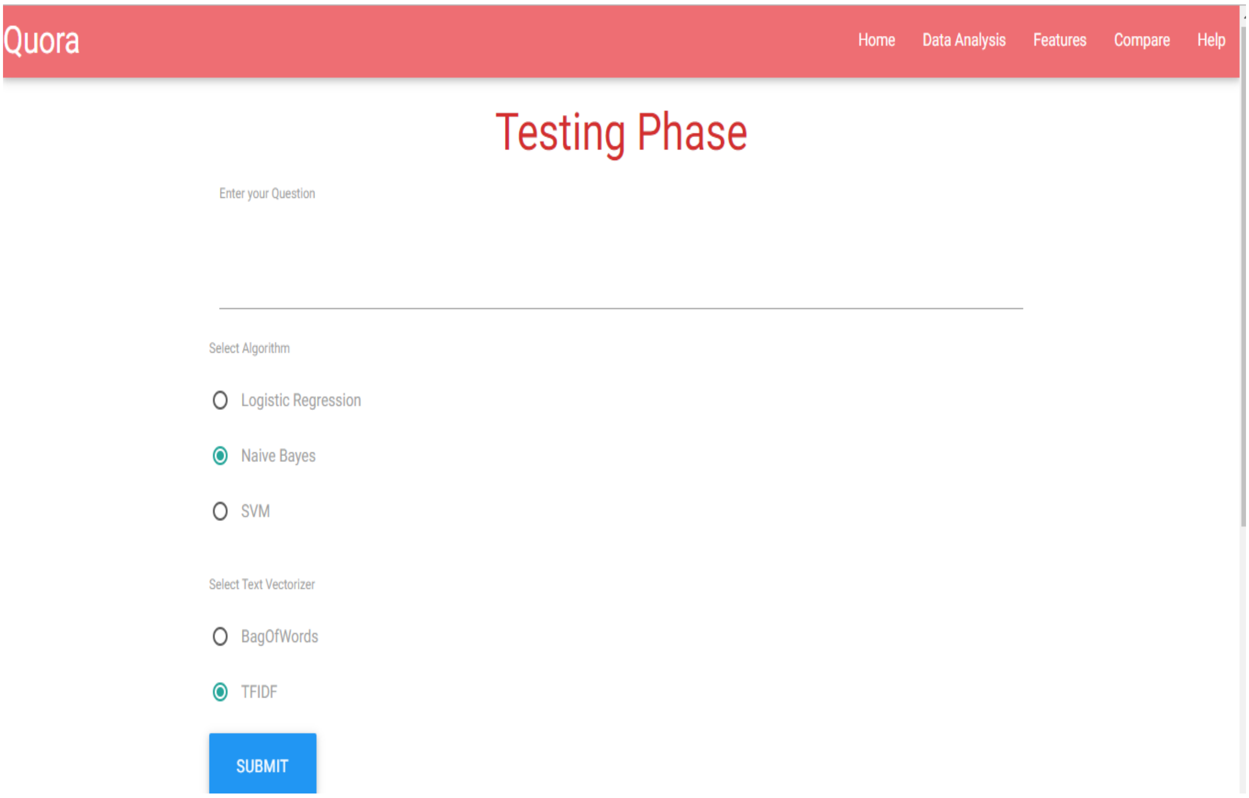
7.2.3 Design of Test Cases

To minimize the number of errors in software, a rich variety of test design methods have evolved for software. These methods provide the developer with a systematic approach to testing. More important, methods provide a mechanism that can help to ensure the completeness of test and provide the highest likelihood for uncovering errors in software. An engineering product can be tested in one of the two ways: (1) knowing the specified function that product has been designed to perform, tests can be conducted that demonstrate each function is fully operational while at the same time searching for errors in each function: (2) knowing the internal workings of a product, tests can be conducted to ensure that “all gear mesh”, that is, internal oppression are performed according to specifications and all internal components have been adequately exercised. Here are the test cases that we had made for our application.

7.3 TEST CASES

Sr. no	Module	Input	State	Excepted Output	Actual Output	Test result
1	Testing Phase	Valid Question, Algorithm Selected TextVectorizer Selected	Class Prediction	Success	Success	Pass
2.	Testing Phase	Blank Question field Algorithm Selected TextVectorizer Selected	Testing Phase	Failure	Failure	Pass

8. USER MANUAL



Testing Phase

Enter your Question

Select Algorithm

○ Logistic Regression

● Naive Bayes

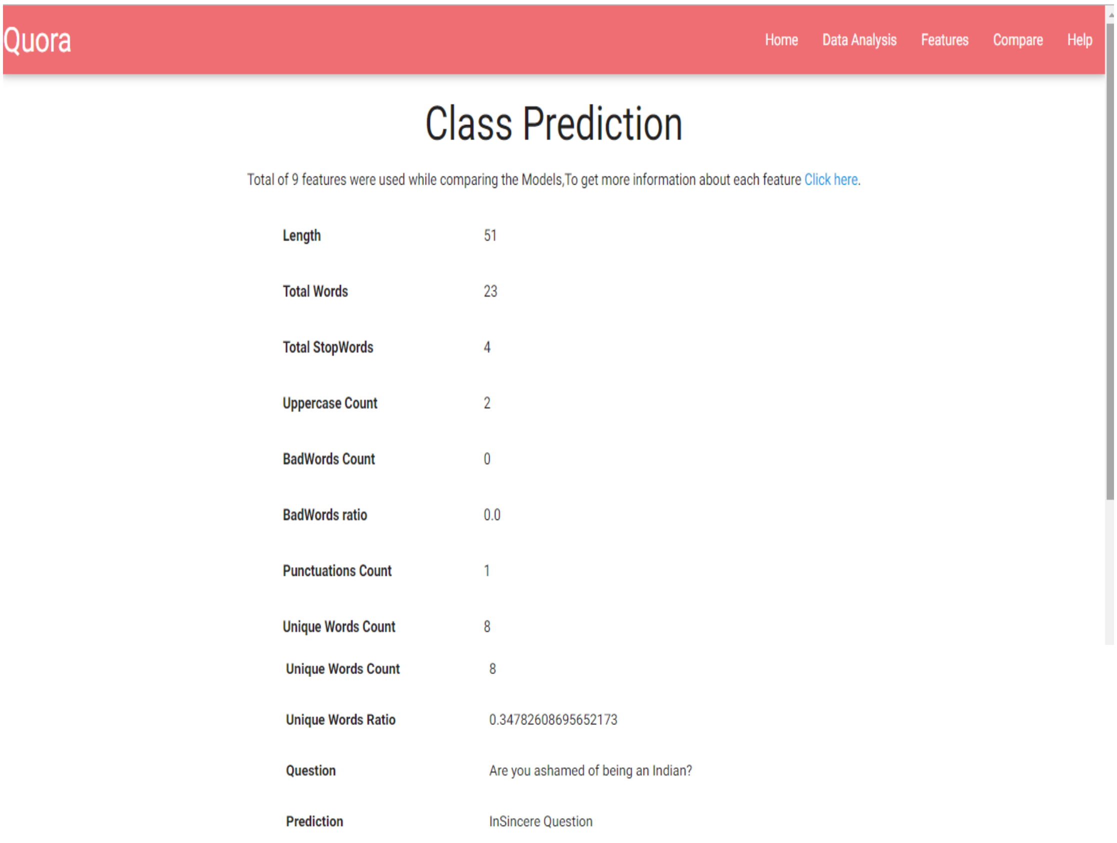
○ SVM

Select Text Vectorizer

○ BagOfWords

● TFIDF

Fig 8 TESTING PHASE

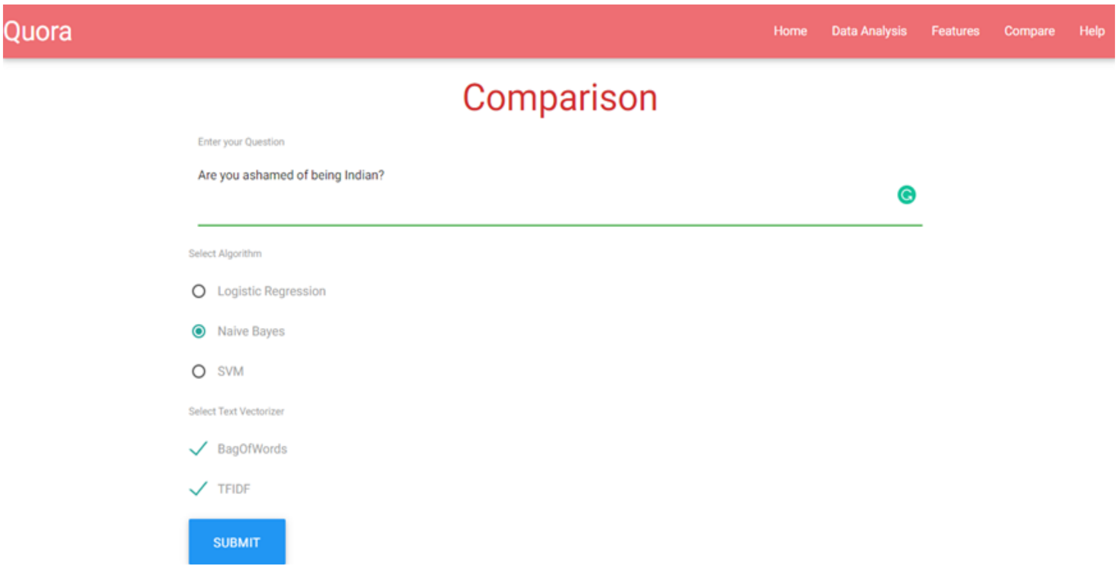


Class Prediction

Total of 9 features were used while comparing the Models,To get more information about each feature Click here.

Length	51
Total Words	23
Total StopWords	4
Uppercase Count	2
BadWords Count	0
BadWords ratio	0.0
Punctuations Count	1
Unique Words Count	8
Unique Words Count	8
Unique Words Ratio	0.34782608695652173
Question	Are you ashamed of being an Indian?
Prediction	InSincere Question

Fig 9 OUTPUT OF TESTING PHASE



Comparison

Enter your Question

Are you ashamed of being Indian?

Select Algorithm

○ Logistic Regression

● Naive Bayes

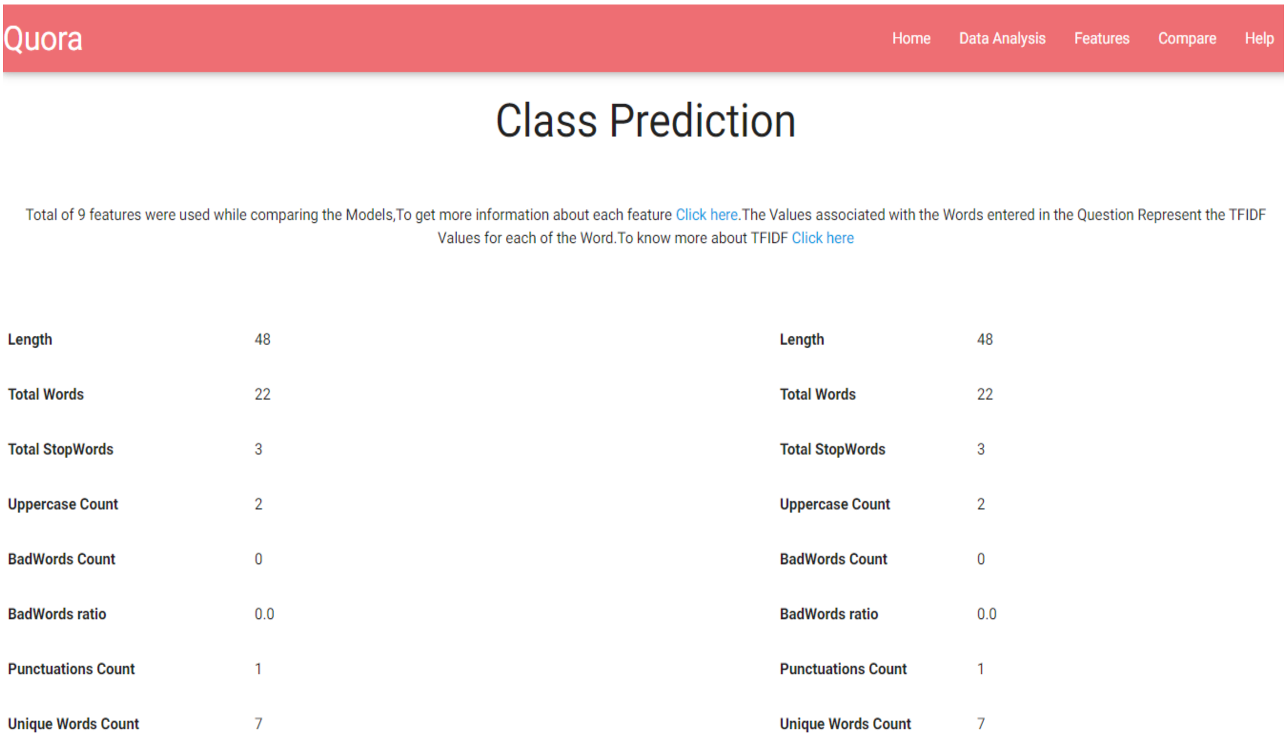
○ SVM

Select Text Vectorizer

✓ BagOfWords

✓ TFIDF

Fig 10 COMPARISON

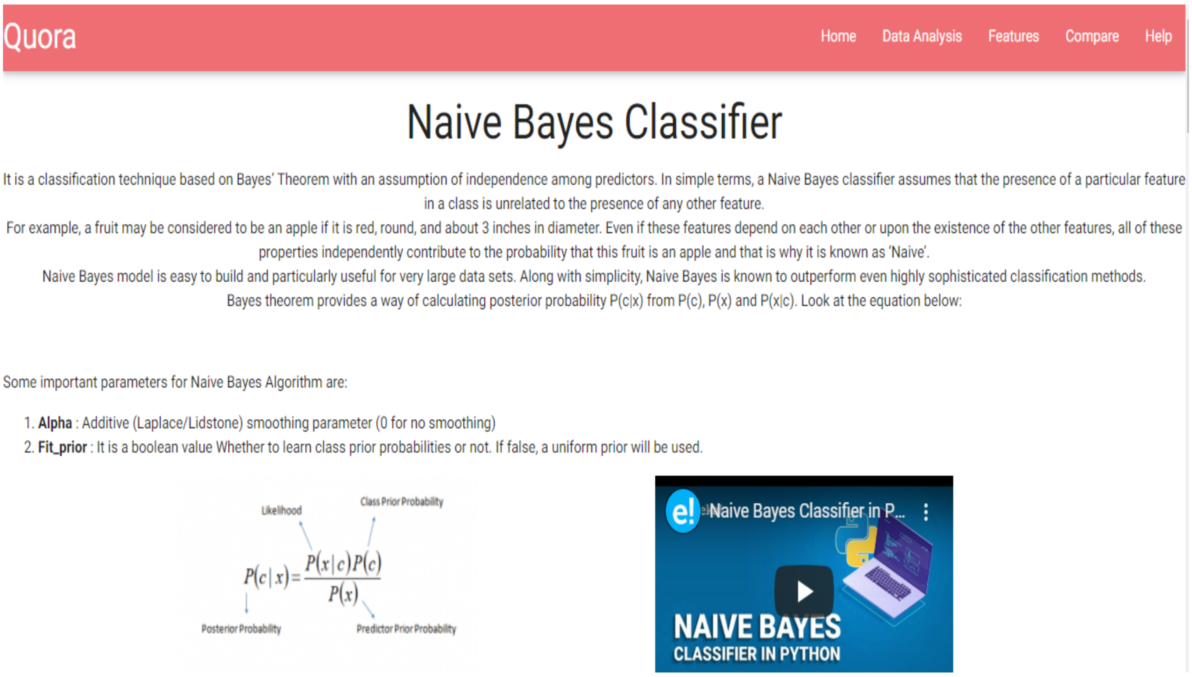


Class Prediction

Total of 9 features were used while comparing the Models,To get more information about each feature Click here.The Values associated with the Words entered in the Question Represent the TFIDF Values for each of the Word.To know more about TFIDF Click here

Length	48	**Length**	48
Total Words	22	**Total Words**	22
Total StopWords	3	**Total StopWords**	3
Uppercase Count	2	**Uppercase Count**	2
BadWords Count	0	**BadWords Count**	0
BadWords ratio	0.0	**BadWords ratio**	0.0
Punctuations Count	1	**Punctuations Count**	1
Unique Words Count	7	**Unique Words Count**	7

Fig 11 COMPARISON OUTPUT



Naive Bayes Classifier

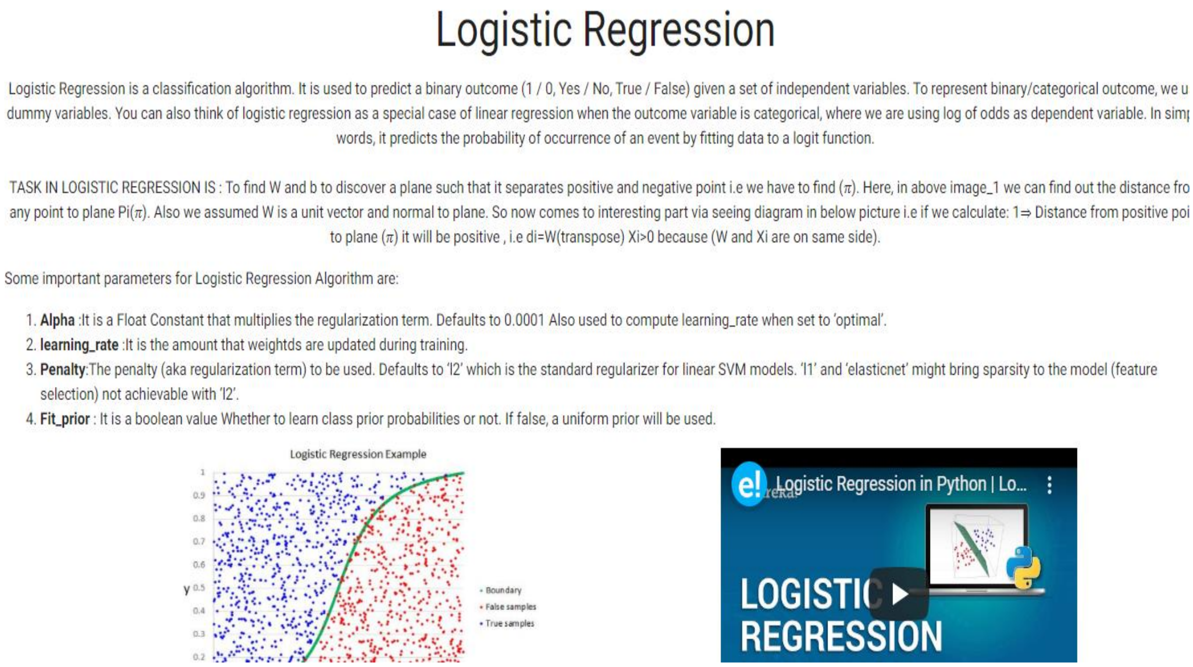
It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.
Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

Some important parameters for Naive Bayes Algorithm are:

1. **Alpha** : Additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing)
2. **Fit_prior** : It is a boolean value Whether to learn class prior probabilities or not. If false, a uniform prior will be used.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Fig 12 Naïve Bayes Detail



Logistic Regression

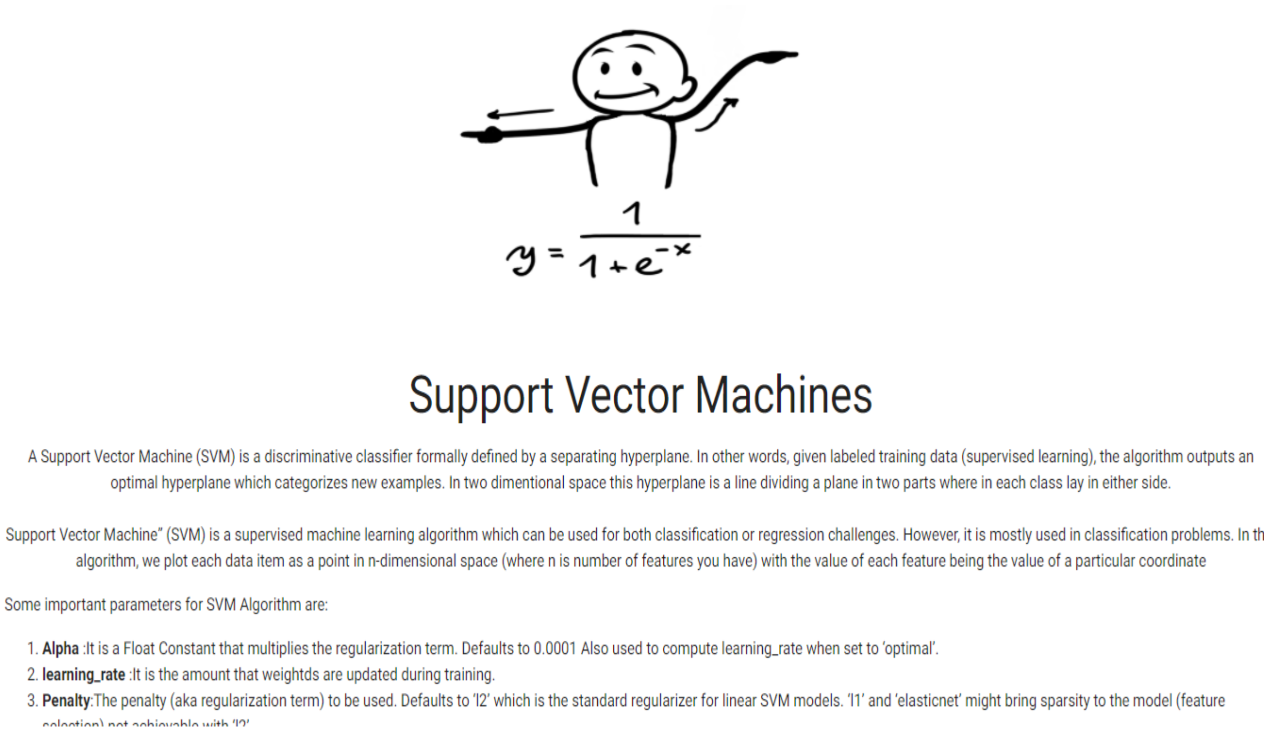
Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcome, we u dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simp words, it predicts the probability of occurrence of an event by fitting data to a logit function.

TASK IN LOGISTIC REGRESSION IS : To find W and b to discover a plane such that it separates positive and negative point i.e we have to find (π). Here, in above image_1 we can find out the distance fro any point to plane Pi(π). Also we assumed W is a unit vector and normal to plane. So now comes to interesting part via seeing diagram in below picture i.e if we calculate: 1⇒ Distance from positive poi to plane (π) it will be positive , i.e di=W(transpose) Xi>0 because (W and Xi are on same side).

Some important parameters for Logistic Regression Algorithm are:

1. **Alpha** :It is a Float Constant that multiplies the regularization term. Defaults to 0.0001 Also used to compute learning_rate when set to 'optimal'.
2. **learning_rate** :It is the amount that weightds are updated during training.
3. **Penalty**:The penalty (aka regularization term) to be used. Defaults to 'l2' which is the standard regularizer for linear SVM models. 'l1' and 'elasticnet' might bring sparsity to the model (feature selection) not achievable with 'l2'.
4. **Fit_prior** : It is a boolean value Whether to learn class prior probabilities or not. If false, a uniform prior will be used.

Fig 13 LOGISTIC REGRESSION DETAIL



Support Vector Machines

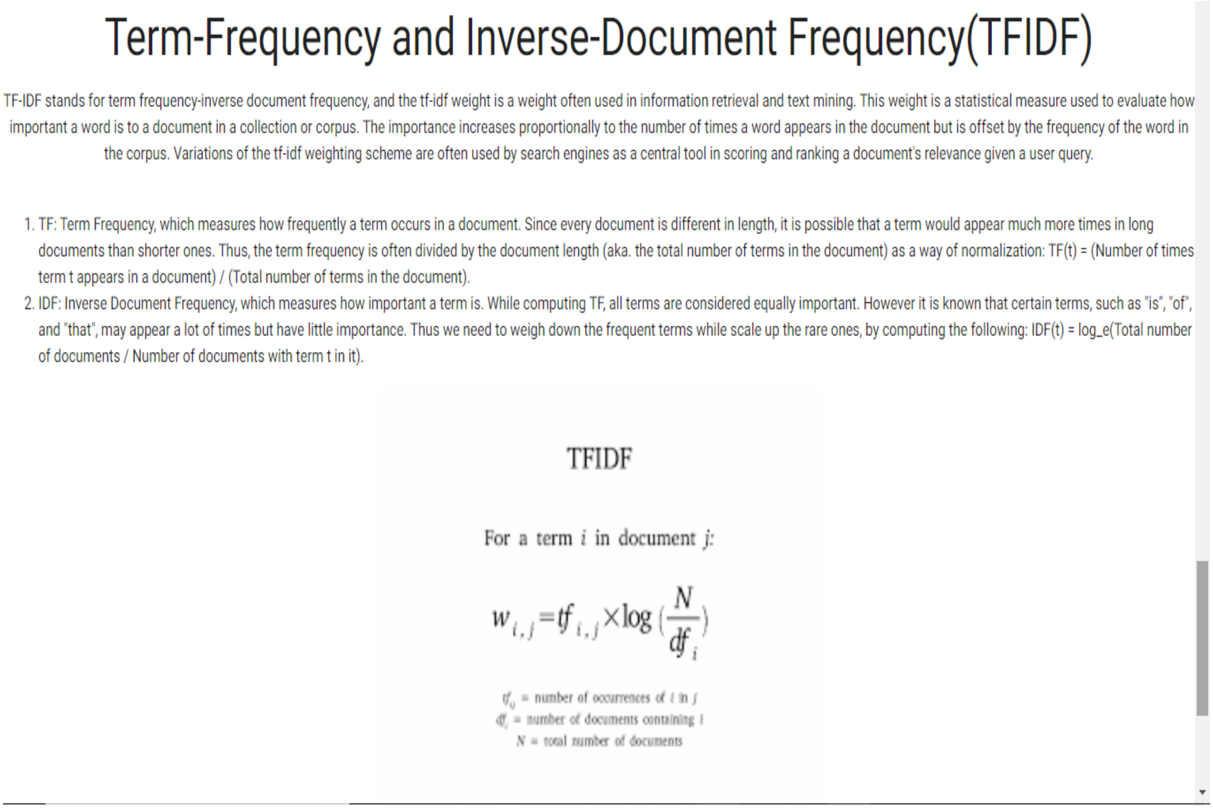
A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In th algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate

Some important parameters for SVM Algorithm are:

1. **Alpha** :It is a Float Constant that multiplies the regularization term. Defaults to 0.0001 Also used to compute learning_rate when set to 'optimal'.
2. **learning_rate** :It is the amount that weightds are updated during training.
3. **Penalty**:The penalty (aka regularization term) to be used. Defaults to 'l2' which is the standard regularizer for linear SVM models. 'l1' and 'elasticnet' might bring sparsity to the model (feature selection) not achievable with 'l2'

Fig 14 SVM DETAIL



Term-Frequency and Inverse-Document Frequency(TFIDF)

TF-IDF stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

1. TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization: TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).
2. IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following: IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

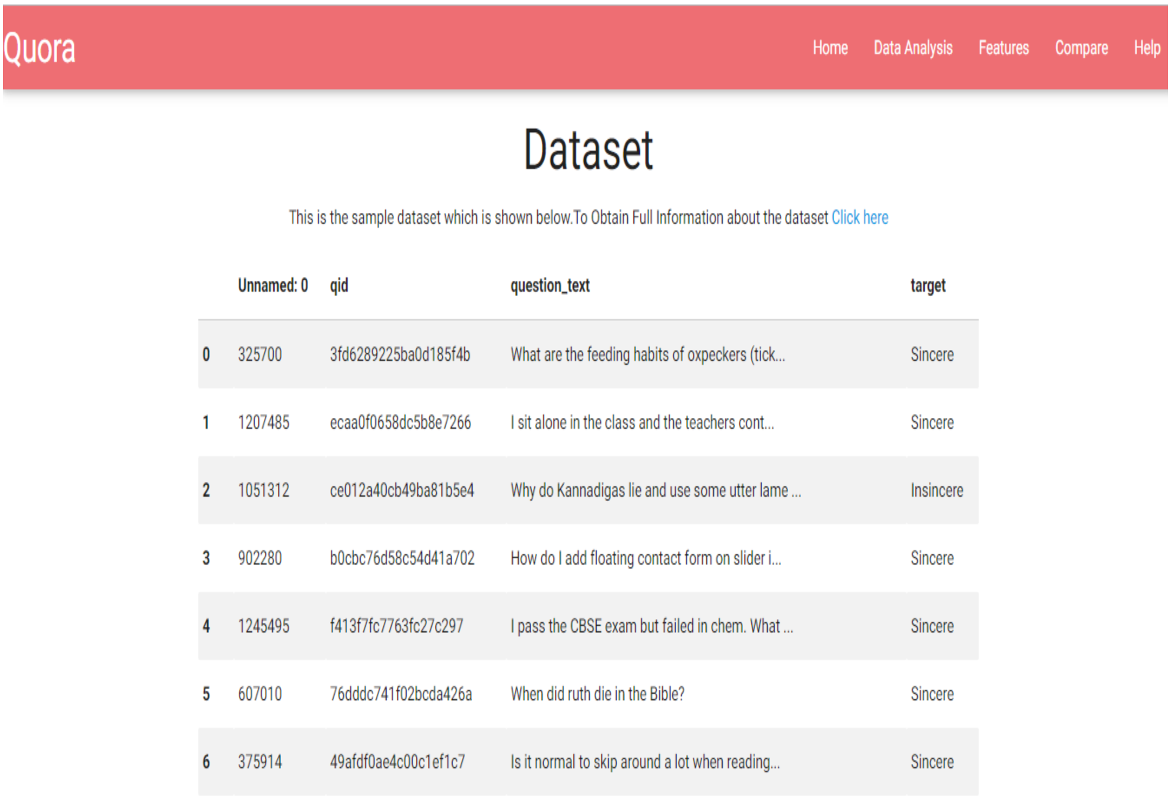
TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Fig 15 TFIDF DETAIL



Dataset

This is the sample dataset which is shown below.To Obtain Full Information about the dataset Click here

	Unnamed: 0	qid	question_text	target
0	325700	3fd6289225ba0d185f4b	What are the feeding habits of oxpeckers (tick...	Sincere
1	1207485	ecaa0f0658dc5b8e7266	I sit alone in the class and the teachers cont...	Sincere
2	1051312	ce012a40cb49ba81b5e4	Why do Kannadigas lie and use some utter lame ...	Insincere
3	902280	b0cbc76d58c54d41a702	How do I add floating contact form on slider i...	Sincere
4	1245495	f413f7fc7763fc27c297	I pass the CBSE exam but failed in chem. What ...	Sincere
5	607010	76dddc741f02bcda426a	When did ruth die in the Bible?	Sincere
6	375914	49afdf0ae4c00c1ef1c7	Is it normal to skip around a lot when reading...	Sincere

Fig 16 DATASET

9. LIMITATION AND FUTURE ENHANCEMENT

- The F1-Score obtained by Naïve Bayes Model was 0.44 whereas the F1 Score obtained by Logistic Regression Model was 0.52 and for SVM it was 0.48.When the Model was tested on certain unseen Questions it was observed that some of the Sincere Questions were also predicted/Classified as Insincere Question.

- More Algorithms and Text Vectorization Techniques can be added so that users can test the Insincerity of the Question by tuning the algorithms and various hyper parameters of the algorithm.

- Comparison of more than 2 algorithms can also be implemented that will help the users to decide which algorithm is best.

10. CONCLUSION AND DISCUSSION

10.1 CONCLUSION

The main aim of this project was to create a User-Friendly Testing Interface so that Users can detect the toxicity level of the Question. So this project can help users to detect the Sincerity of the Question before actually posting the Question on Quora. Because if the Question is Insincere and is detected by Quora the Question is Generally removed from the Website.

10.2 DISCUSSION

10.2.1 Self-Analysis of Project Viabilities

In our opinion, this project has served the goal that we set when we started. It provides a service that is not provided to users currently. Besides, there are numerous other improvements and enhancements possible to this software.

10.2.2 Problem Encountered and Possible Solution

Some of the issues we faced and solved during the project development are as follows:-

1. Deployment of Machine Learning Model and Creating an API for Deployment using Flask.

2. Integrating Web App with the Model

10.2.3 Summary of Project Work

We have completed our project work using software engineering and system analysis and design approach following the for software development. We have done our work with planned scheduling pertaining the time constraints and result oriented progress in project development.

REFERENCES

- https://www.tutorialspoint.com/flask/index.htm
- https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- https://en.wikipedia.org/wiki/Logistic_regression
- https://en.wikipedia.org/wiki/Support-vector_machine
- https://materializecss.com/getting-started.html

/