Quora Insincere Questions Classification

Sanya Zaveri
AU1920064
sanya.z@ahduni.edu.in
Ahmedabad University

Mihir Pathak
AU1920138
mihir.p@ahduni.edu.in
Ahmedabad University

Malav Doshi
AU1940017
malav.d@ahduni.edu.in
Ahmedabad University

Parth Shah
AU1940065
parth.s5@ahduni.edu.in
Ahmedabad University

***Abstract*—The internet today has become an unrivalled source of information where people converse on content based websites such as Quora asking doubts and sharing knowledge with the world. A major arising problem with such websites is the proliferation of toxic comments or instances of insincerity wherein the users instead of maintaining a sincere motive indulge in spreading toxic and divisive content. The obvious choice in confronting this situation is detecting such content beforehand and preventing it from publishing online.**

***Keywords*—Text Classification, Tokenization, Logistic Regression, Lemmatization, Confusion Matrix**

I. Introduction

Quora is a website where a community of users can ask and answer questions. They may also create Q&A blogs and analyze statistics about the users as well. The relevance of question is often taken as into consideration and is necessary to prevent an unnecessary or redundant question that might affect different set of people.
Quora Insincere questions arises when people ask questions that is intended to make statement rather than look for helpful answers. Questions consisting of an non-neutral term, disparaging, inflammatory or is not grounded in reality can be classified as Insincere questions.

II. Related Work

There have been many contributions in the domain of text classification. It is due to the business value it adds by classifying text on the basis of various parameters. Another such paper proposed a model for classifying tweets [cite]. The author had implemented a Logistic Regression model of machine learning for classifying tweets according to their the topic they belong. The system transformed the tweets into vector which is acceptable by the model. The confusion matrix showed an accuracy of around 92%. This system uses just one algorithm which does not give any evidence that this model has performed the best. The proposed system will implement 4 models which will enable to analyze which is performing better.

III. Understanding the Data

A. Data Analysis

The Data consists of *test.csv, train.csv* and *embeddings.zip*. The train.csv consists of the question id, questions text and classification into *0* or *1*, where 0 denotes a sincere question and 1 denotes an insincere question. This csv file consists of about 1.3 million entries.
We use the following method for data analysis:

1) Bar-Graph
On visualising the given training set on a bar graph (Figure. 1), we observe the data has 1.3 million entries, in which 93% of entries are sincere questions, while the rest questions are insincere questions.

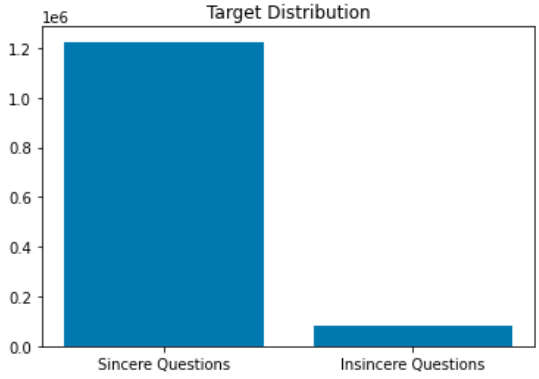

Fig. 1. Distribution of Questions into Sincere and Insincere

2) WordCloud
WordCloud helps us to observe nature of sincere and insincere questions. In Fig. 2. it is observed that ords such as *India and Indian*, which have essentially the same root, hence we need to apply lemmatization in our training set. Similarly words such as *will* shows the need for removing the stop words to only consider relevant words.

Fig. 2. WordCloud of Insincere Questions

B. Data Preprocessing

1) Removing Stop Words
Stop words act like connectors in any sentence. The words do not add any value of their own. Exampl include *a, the, is* and *are*. Removal of stop words hel us to remove low-level information and shifts our foc to value adding words. The NLTK library is used remove stop words, includes the list of stop words f different languages.
2) Lemmatization
Root words are the word from which other words a derived. For eg, *eat* is the root word for *eating, to eat* R ducing the inflected words while ensuring the root wo belongs to the selected language (unlike Stemming) known as Lemmatization. Wordnet Lemmatizer from t NLTK library is used for this purp

C. Data Balancing

1) Discarding the majority
Data has 1.3 million entries of which 93Since the model would be trained better on a balanced data set, we remove the excess of sincere questions so now it matches the 7Drawback: Too much data is being discarded.
2) Oversampling
The 7% data of insincere questions would be over sampled by duplication to match the 93
3) SMOTE (Synthetic Minority Oversampling Technique)
This is an oversampling method, but here the samples are not simply duplicated but they are generated using k-nearest neighbors.
4) Ensemble
Just like random forests, we split the data in multiple sets and train the model multiple times in permutations with different sets. This would result in us getting multiple models for each permutation. From here onwards we would simply choose the best model based on majority vote.

IV. Feature Engineering

We extract features from the raw data set using domain knowledge we have. We do this as having more defined features will help the model improve its' quality of results.
features added = ['num words','num unique words','len question text', 'len char question text', 'len word question text','num stopwords','common words','len mean words']
Now upon plotting correlation matrix which is Fig. 3, we see there is high correlation matrix between many features. We infer that there are many redundant features in our feature engineering. Hence we drop the features that are redundant for our model training.
We also include features using methods known as Word2Vectorization and TF-IDF, in order to add some mode features to the data. Fig. 4 shows the final correlation matrix after dropping all the insignificant and redundant features.

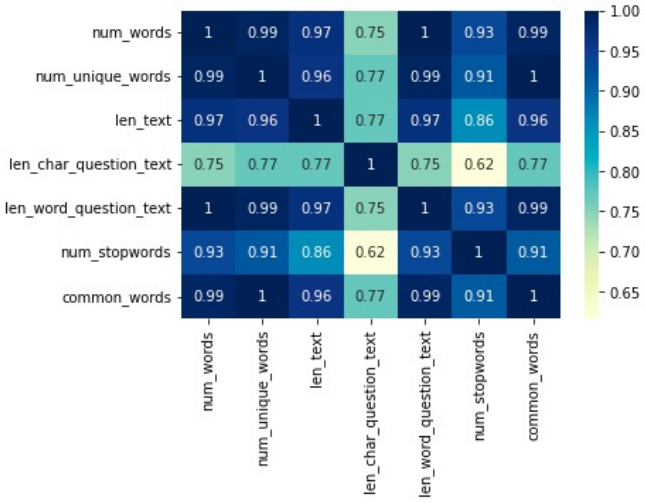

Fig. 3. Correlation between existing features

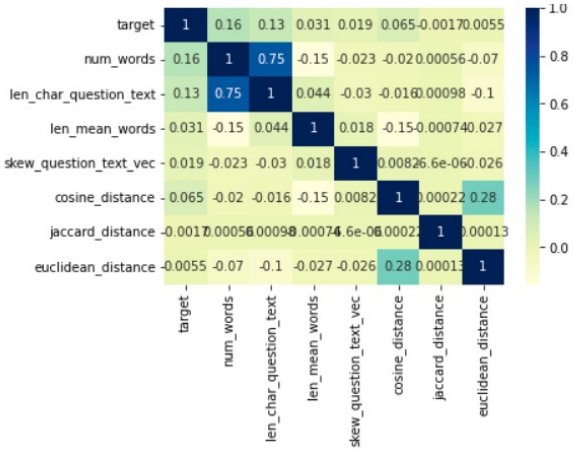

Fig. 4. Correlation between the final features

V. Modeling

We used various model for binary classification of our data. Models such as SVM, logistic regression, LDA, and Random Forrest are used to find the best predictions. After the model is fitted to the given data set, it is used to predict the labels for the unseen future data. X test and y test are compared to get accuracy of the model. We use various scores such as accuracy and f1 - scores and various graphs such as PR curve for comparison of our models. We present our various findings from these models and based on the comparison we would like to select the best suitable curve for our data.

1) **Logistic Regression**
Logistic Regression is a model where the coefficients are learned during the training of the model. Here the logistic regression model is imported from sklearn linear model and the hyper parameters used are inverse regularization and solver sag.

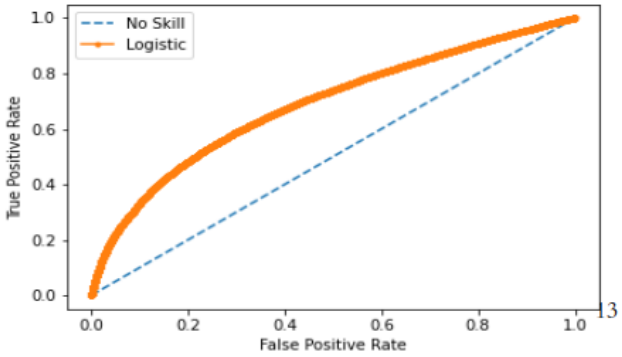

Fig. 5. ROC curve for logistic regression

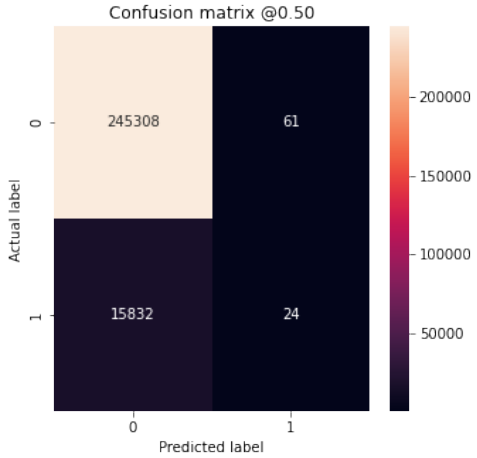

Fig. 6. Confusion matrix for logistic regression

2) **Linear Discriminant Analysis**

We will be using a LDA model that is provided from the sklearn.discriminant_analysis library. respectively. Following are the value of different scores for the purpose of model selection.

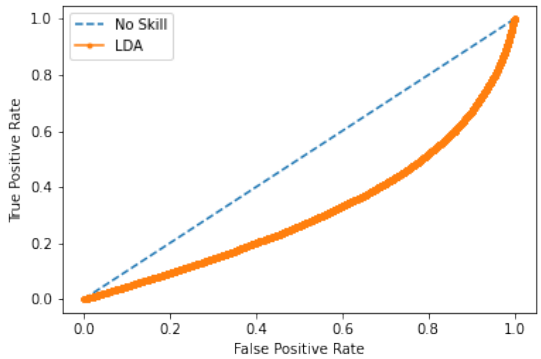

Fig. 7. ROC curve for LDA

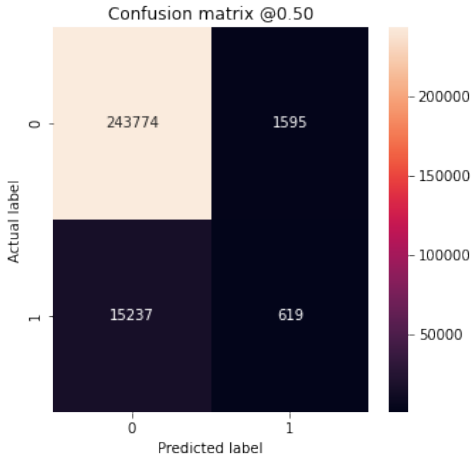

Fig. 8. Confusion matrix for LDA

3) **Support Vector Machine**

Support Vector Machine (SVM) is a model that learns from the training data and assigns categories to new data. SVC is imported from the sklearn.svm model.

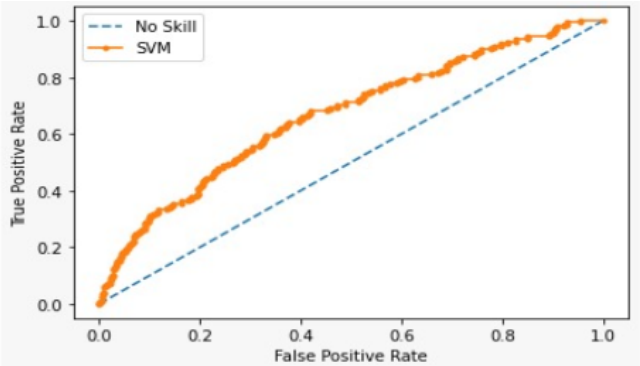

Fig. 9. ROC curve for SVM

4) **Random Forest**

Random forests is a group of decision trees working together to give output. In this multiple decision trees are developed during training. Imported RandomForestClassifier from sklearn.ensemble model.These tress are then used to produce output using the mode of the decision trees. In this project various multiple trees are developed using various combinations of words. The mode of these trees is taken to produce the output of random forest.

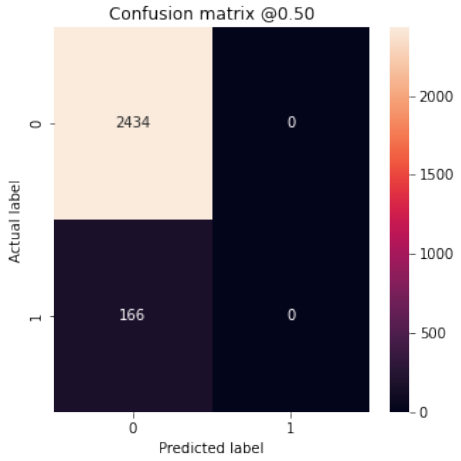

Fig. 10. Confusion matrix for SVM

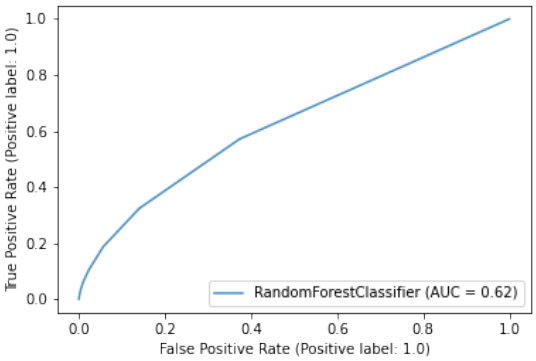

Fig. 11. ROC curve for Random Forest

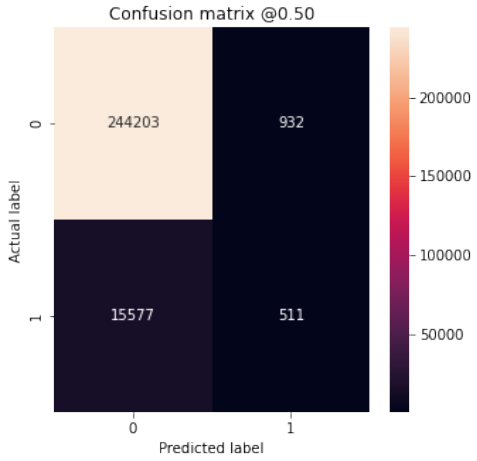

Fig. 12. Confusion matrix for Forest

VI. Inferences

A. Evaluation

Due to data imbalance the evaluation is not focused on Accuracy, rather it is focused on other metrics like F1 score, Area Under Curve, Precision and Recall. These metrics are explained below.

1) Accuracy: Accuracy can be said to the measure of the closeness of the output to a certain value. It does not work well with imbalanced data. Hence, in this project other metrics are used for evaluation.
2) F1 - score: F1 score is the one that is calculated by combin- ing the precision and recall measures. It is the harmonic mean of the two. It results nearly the same as the average of the two measures when they are closely related.
3) Precision: Precision is ratio of correctly predicted outcomes to the total predicted outcomes. It is not dependent on the accuracy of the model. It is therefore a measure that be used in case of class imbalance.
4) Recall: It is the ratio of correctly predicted outcomes to the total outcomes. It is also known as the sensitivity of the model.

Model	Accuracy	F1- Score	Precision	Recall
LogRes	0.939	0.901	0.899	0.939
LDA	1.000	0.912	0.901	0.936
Random Forrest	0.938	0.912	0.905	0.938
SVM	0.936	0.905	0.940	0.936

Table 1: Comparison between various models

References

[1] S. T. Indra, L. Wikarsa and R. Turang, "Using logistic regression method to classify tweets into the selected topics," 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016, pp. 385-390, doi: 10.1109/ICACSIS.2016.7872727.

[2] O. Aborisade and M. Anwar, "Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 269-276, doi: 10.1109/IRI.2018.00049.

[3] "Sklearn.linear_model.logisticregression," scikit. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed: 20-Mar-2022].

[4] "Precision-recall," scikit. [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html. [Accessed: 20-Mar-2022].

[5] Y. Liu, J. Niu, Q. Zhao, J. Lv and S. Ma, "A Novel Text Classification Method for Emergency Event Detection on Social Media," 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2018, pp. 1106-1111, doi: 10.1109/SmartWorld.2018.00192.

[6] "Quora insincere questions classification," Kaggle. [Online]. Available: https://www.kaggle.com/c/quora-insincere-questions-classification. [Accessed: 20-Mar-2022].