Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 8, Issue 01, January -2021

Quora Insincere Questions Classification

Daksh Suraj Rawat[1], Akshay Mungekkar[2], Prateek Nima[3]

[1]Electronics and Telecommunication, UCOE Mumbai
[2]Information Technology, UCOE Mumbai
[3]Information Technology, UCOE Mumbai

Abstract- *Internet has opened up a new horizon of opportunities. It has answers to all questions, and even if answer is not found there are sites where you can ask any question and people from all over the world answer it. Quora is one such question answer website. Here anyone can ask question from any domain, which are then answered by others. People misuse this boon and create a havoc by asking questions that are not to be asked in a public forum. In this project we propose to develop a system that recognizes these "insincere" questions so that appropriate actions can be taken against them. The questions are classified as insincere if they have a non-neutral tone, is disparaging, inflammatory or the questions are in negative sexual aspects and if it is not grounded to reality. In this project a model of text classification will be developed using Natural Language Processing. Based on the models, Decision Tree performed the best followed by Random forest. The metrics used for judging the result include F1 score, recall, precision and Area under Curve.*

Keywords *- Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, Tokenization, TF-IDF, Stemming, Lemmatization.*

I. INTRODUCTION

The internet has gained popularity over the years. It has now become one of the most important things that the human race thrives on. The main reason for this popularity is the way internet simplifies tasks that were very difficult before. One of the most popular use of internet is to search for solutions to questions. This feature of the internet has helped in many ways. But this was soon upgraded to asking questions on website where people from all over the world could help in an- swering. These type of websites are known as question forums. They have gained a lot of popularity due to their easy to use and understand methodology. Quora, Stack Overflow, Yahoo Answers are some examples on question forum websites.

Stack Overflow has mostly technical questions where people post questions or errors in the code and other people from all over the world read it, analyze it and try to provide a solution. Yahoo Answers and Quora work on the similar principles. Both websites can be used to ask simple, personal, professional questions. Some people use these websites for getting advice about career of a personal opinion. This al- lows in forming a community and helping each other through tough times.

But these forums have a huge drawback of people misusing it. People tend to ask questions that do not sound proper. They may be targeted at a group of people or make no sense. These questions tendto create a havoc, thus disturbing the main cause of creating these kind of website which is to help each other. It is therefore of an utter importance to classify these questions and remove them before they harm the reputation of the website. In this paper, we propose classification models to classify insincere questions that will help in maintaining the integrity of the website.

The website from which the data is collected is Quora. It is one of the most popular Question Forum. It believes in motivating people to learn from each other. It is important for the website to make sure it is safe for people from all over the world to share knowledge and ask questions. Quora had previously made use of Manual Reviews to distinguish between sincere and insincere questions. But, with the use of Machine Learning Models this process can be made more efficient.

Supervised Learning Models are implemented for classification of the questions. The models developed include Naive Bayes, Logistic Regression, Support Vector Machine (SVM), an ensemble of Naive Bayes and Logistic Regression and Random Forest. All the models provide good accuracy rates. But, since the data is imbalanced F1 score and ROC will be used as the parameters for judging the performances of the models. Term Frequency–Inverse Document Fre- quency (TF-IDF) and Document Term Matrix (DTM) is used for counting the occurrences of words in the questions. Further, Lemmatization and Stemming is used for reducing the forms of inflection and remove the endings to result in a base word.

The first part of this paper introduces the topic with an overview of the whole proposed project. Rest of the paper is structured as follows. Section II discusses the work undertaken previously for similar kind of data or using similar methodologies. It further points out the pros and cons of the same. Section III explains the approach used and methodology pro- posed in detail. This project follows the CRISP-DM methodology which is explained in this section. Also, the details of implementation have been discussed in detail. Section IV highlights the methods of Evaluation and states the results of the models. It gives a direct comparison thus shedding light on which model has the best performance. Section V is the last section which puts forth the overview of the project along with the results. It further

discusses the future work that can be implemented to improve the existing models.

II. RELATED WORK

The following section will discuss the various works done in this area using similar methodologies. 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.

The most important part of text classification is pre- processing. It is important to use appropriate methods for accurate results. Word2vec was used along with TF-IDF for automating text classification [1]. Word2vec was used for vectorizing and TF-IDF for calculating the weight of the feature words. The approach used multiplied the weight of the word to the word vector. Further, word vectors are grouped to give accurate results. This methods will be used for pre- processing in the proposed paper.

Vectorization was done using TF-IDF and Bag of Words[2]. This was used by an author to classify tweets in the users on the basis of a given corpora. The author trained models of Logistic Regression and Naive Bayes after preprocessing and extracting features using vectorization. The total number of tweets were 46,895 which gave an accuracy of 91% for Logistic Regression and 89.9% for Naive Bayes. The implemented model has a very small dataset and hence a proper training is not provided for the models. The proposed project will implement models on a huge dataset thus enabling more training.

Another such paper proposed a model for classifying tweets [3]. 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.

Classifying text requires tokenizing words and using it to train the models. However, when a different language like Chinese is used for training the model, word tokenizing becomes a tough task. The author[4] has developed a model which is based on N- gram which relates the words rather than using word tokenizer. The model used for training is based on Logistic Regression. The proposed model is better than the previous approaches by at least 11%. Word tokenizer is used in the proposed model as the text is in English and will perform better than it performs with other languages.

Using ensemble approach for models may improve the accuracy of the model. This was tested by an author who used ensemble approach for classification based on keywords[5]. The research focuses on the automatic keyword based operations carried out in terms of keyword indexing, classification, clustering along with five different keyword extraction methods.

In addition, 2-way ANOVA has been used to validate the performed analysis. The study states that the use of ensemble approach consisting of Bagging based Random Forest method provided the accuracy around 93%. The proposed project will also implement an ensemble approach to compare with the normal machine learning models.

A comparative analysis of several ensemble approaches to explain the importance of ensemble approaches and their contribution to improve the accuracy of the model.[6]. Various SVM and Naive Bayes approaches are used and compared and a later stage in the paper. The author concludes by stating that this approach is better than the conventional approaches. The proposed project will also implement both basic and ensemble models to see which perfoms better.

Support Vector Machine Model was used to classify BBC documents into five categories[7]. The model was enhanced by using Chi-Squared along with SVM. Both Stemming(Lancaster Stemmer) and Lemmatization(WordNet Lemmatizer) were used and fed separately to the model. The results stated that Stemming provided better results and chi-squared was an added advantage as it improved the model. The proposed model will also use both Stemming and Lemmatizer for different models for a comparison of the results.

SVM was used and compared with k-Nearest Neigh- bour (kNN) for performing sentiment analysis of comments and opinion mining by classification[8]. This approach aimed to compare Supervised learning models with Unsupervised Learning Models to see which performs better. The end results demonstrate that SVM outperformed kNN. Emergency events were detected using one-class SVM method which uses irrelevant texts as training data for detecting relevant texts[9]. The author extracted length of the variables using n-gram to take the relations of words into account. After conducting two experiments for evaluation, it was seen that this method can detect emergency text successfully with good accuracy. The proposed model will implement SVM model which provided a good result here.

Text classification was used for identifying tweets related to suicides[10]. This was done with the motive of reducing the negative impact of tweets. The models used for this project included SVM, Naive Bayes and Random Forest. Decision tree performed the best in among the three models implemented. The F-measure ranged from 0.346 to 0.778. The proposed project will implement all these models to see which model detects the most insincere questions.

Another paper implemented SVM, kNN,Decision Trees, Multinomal Naive Bayes, Bernoulli Naive Bayes for classifying Turkish News Text[11]. The text was classified into five predefined classes and trained with different systems. The model which outperformed others was Multinomal Naive Bayes with an accuracy of 90%. The same model will also be used in the proposed project.

Two different text vector representations which included a simple bag of word and an embedded global vector which is

more refined[3]. Multiple models were implemented to see which vector representation works well with what kind of model. It was stated that while embedded global vector worked well with Logistic Regression, Random Forest and SVMs, bag of words worked well with Multinomal Naive Bayes. Bag of words will be used for this project as it works well with majority of models.

III. METHODOLOGY

The most critical task of a successful project is implementing it using an appropriate methodology. There are various methodologies like KDD and CRISP-DM which allow efficient development of project. This project will use CRISP-DM which is short for Cross- Industry Process For Data Mining. It is a methodology which is an extended version of the KDD approach. In this approach we can update the model at any stage and make changes to the previous stages at any stage of the project. This allows scope for improvement in the project. The following figure 1 shows the CRISP- DM Methodology along with the steps involved.

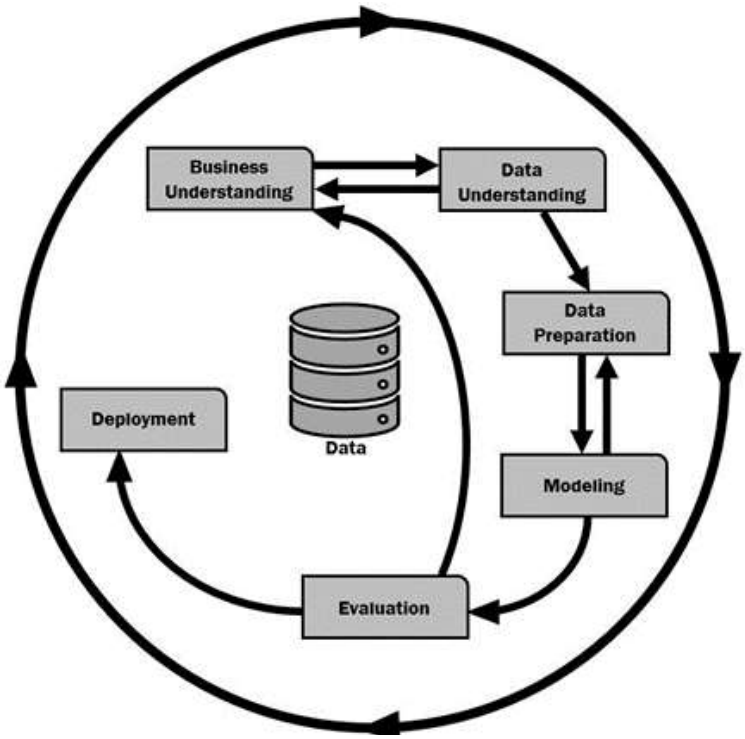

Fig. 1. CRISP-DM Methodology1

A) Business Understanding

Before initiating any project it is important to under- stand what value it adds to the Business. Quora has been one the top websites for asking and answering questions. It also has many categories in which you can ask question. This gives user a better understanding of what category the question belongs to. But,some people take wrong advantage of this website. *https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome.*
They post questions that are not appropriate to post on inter- national website. These questions degrade the quality of business. Hence, this project attempts to predict this insincere questions thus, upgrading the value of business. This also helps in gaining users trust about the website.

B) Data Understanding

The data is collected from a website known as Kaggle. It is website which has many datasets that are available for public use. The dataset is available at the link:*https://www.kaggle.com/c/quora-insincere-questions-classification/data*.
The training dataset is used for this project. It has three parameters as follows:

•Q ID : Unique question ID assigned to each question.

• Question Text : The actual question.

•Target : Either 1 or 0. 1 denotes that the question is insincere while 0 denotes that it is sincere. The sincerity of questions is decided on the basis of following parameters:

- Has a non-neutral tone
- Is inflammatory or disparaging

- Is not grounded in reality.
- Uses sexual content like incest, bestiality or pedophilia for shock value and not to seek answers of sort.

It is very essential to prepare the data before actually using it. Some basic analysis along with Feature Engineering to understand the distribution of data were conducted in order to get an idea of the data set. The following are the various analysis conducted.

1) Bar Graph

A simple bar graph was built to understand the distribution of data. As it can be seen that insincere question comprise of only 6.19% of the total dataset. This analysis show that the data is skewed towards sincere questions.

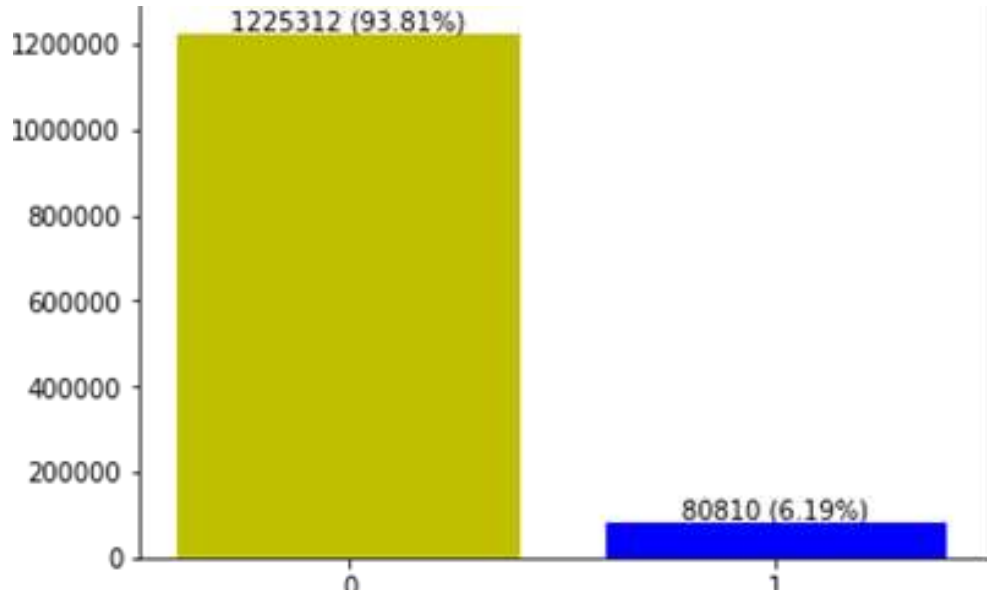

Fig. 2. Data Distribution

1) Common Words

Four graphs were created for this. Bigram was used to make a graph with two words. The following graphs show the common words in the whole data.

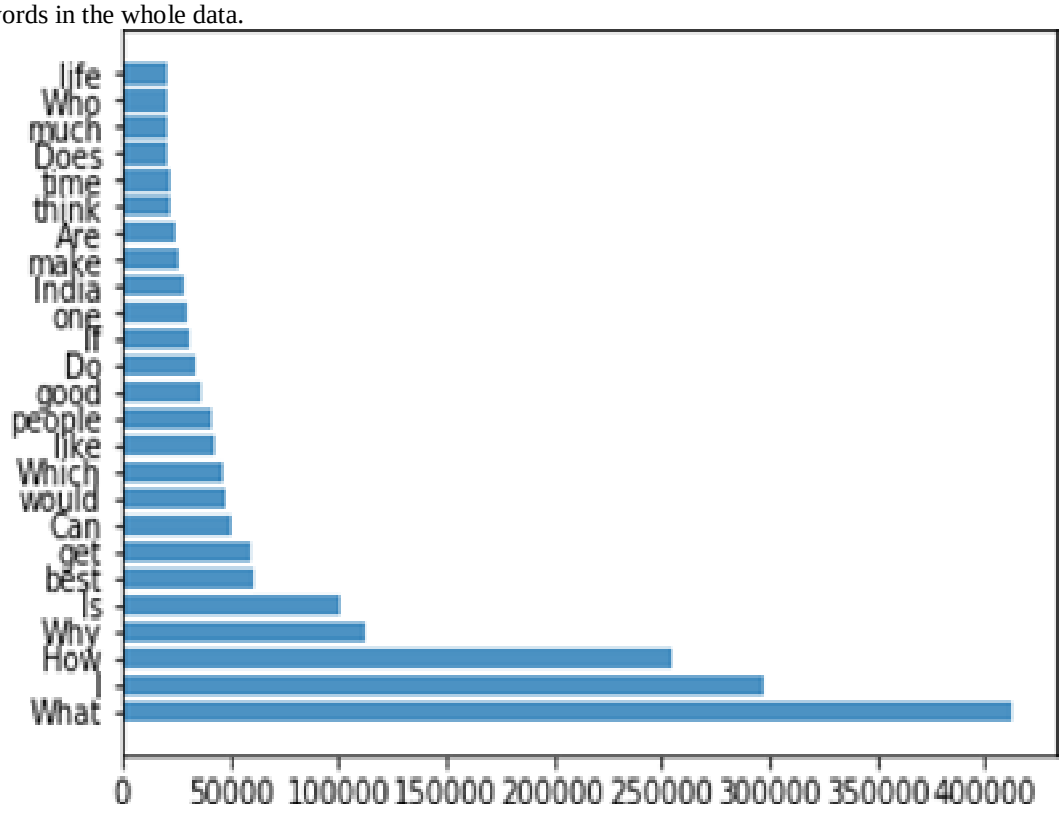

Fig. 3. Insincere Questions

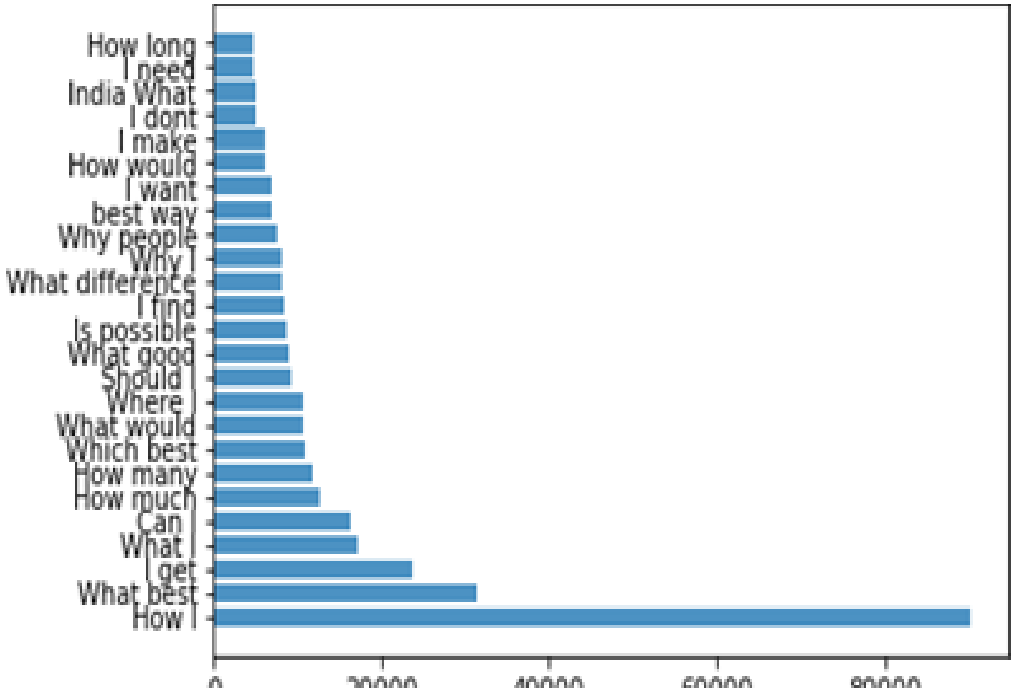

Fig. 4. Insincere Questions using Bi-gram

1) WordCloud

A Word Cloud was used to highlight the most common words in each category. The following figure shows the Word Cloud of insincere questions. These graphs were also created using bigram for both sincere and insincere questions.

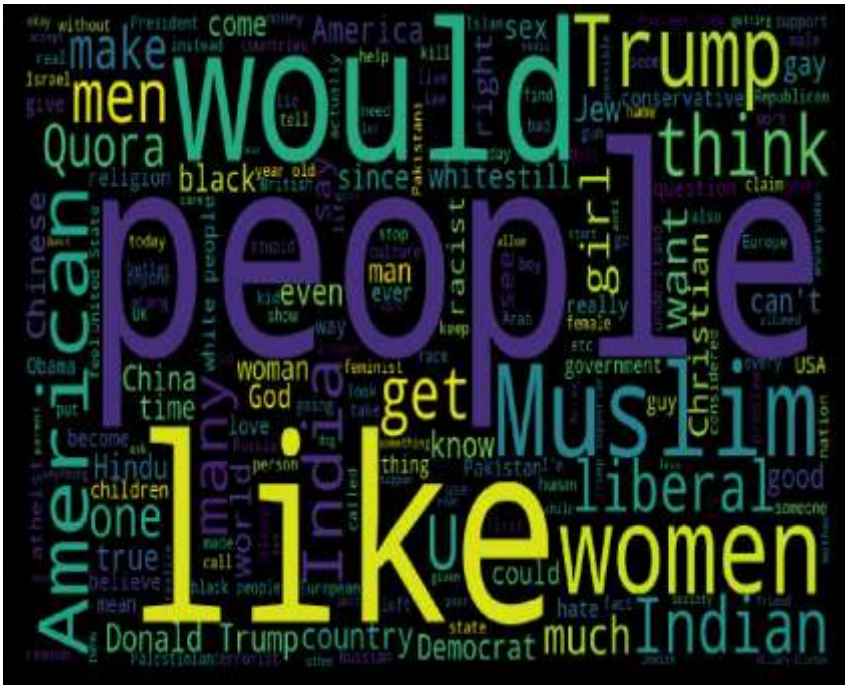

Fig. 5. Word Cloud of Insincere Questions

A) Data Preparation

1) Word Tokenization

The first step for preparing data was splitting it into words. It is the most essential step fortext processing. The sentences are divided into words. The Natural Language Toolkit (NLTK) package is used for this purpose. It divides the words along with special characters treated as a word. This helps in efficient splitting of sentences into a list of words separated bycommas. This array is then passed for removing Stop Words.

2) Removal of Stop Words

Stop words are the words that are used to add meaning to the sentence like ”a”, ”an” ”in”,”of”. They do not have any value of their own. It is therefore important to get rid of them as theycan influence the results of the models. The NLTK package has stopwords for various languages. The stopwords in English language were used for this project. After splitting the sentences and removing the stopwords, the next task is to identify the root of all words. This is explained in the following point.

3) Finding the Root Word

Root word are the words from which other words are made. Like dance is the root word for dancing, dances, danced etc. Using root words instead of the words itself result in a better accuracy rate. There are two approaches to achieve this. One is Lemmatization and the other is Stemming.

- Lemmatization

It is a method which groups words as a single term based on their inflection. Wordnet Lemmatizer is used for this purpose. The NLTK package has a Word Net Lemmatizer which lemmatizes words.

- Stemming

This is other method used for linking words to the root words. It is a process of identifying the variants of the base words. There are two methods of stemming used in this project as follows[2].

- Porter Stemmer

This stemming approach was published in 1980 used for English language initially

. Later many other languages stemmers were also released for use in Snowball, which is a framework for stemming algorithms. It is also one of the most commonly used stemmer.

- Lancaster Stemmer

This approach was developed in the late 1980s, which is a repetitive stemmer and also has very strict rules. The rules pro- vide a strong stemming process which replace or remove the ends of a string.

Both these approaches are correct in their own methods. But, normally lemmatization is preferred over stemming due to its analysis on the basis of morphology of the text. One major difference is that lemmatizer takes a speech parameter as “pos” if not supplied, the default is noun. In this project both methods are used to analyse which one performs better.

[2] https://en.wikipedia.org/wiki/Stemming#Hybrid approaches

4) Text Vectorization

Text Vectorization helps in converting textual data into a series of numeric factors for process- ing it further. The two most popular approaches for this are TF-IDF and DTM. This project uses both these approaches to compare the result.

•Count Vectorizer

The Document Term Matrix (DTM) is a matrix with the occurence of each word in a given sentence. The rows belong to the sentences while the columns belong to the terms. It is a package in python called sklearn with a function of Count Vectorizer which provides DTM. It can also be called as the bag of words representation of ques- tions. This process helps in converting text to numeric values for training the models. Another such method is TF-IDF which is explained below.

•TFIDF

Term Frequency- Inverse Document Fre- quency are the two parts of the TF-IDF algorithm. Term Frequency is the frequency of word in the text, it can be expressed in formula as follows [12]:

$$TF_{i,j} = \mathbf{n_{i,j}} / \sum \mathbf{n_{k,j}}$$

where, j is the document and i is the term. $n_{i,j}$ is the frequency of repetition of a word in the document. The denominator is the sum of all words frequency.

Inverse Document Frequency (TDF) is the importance of a word denoted in degrees. It can be expressed as a formula as follows:

$$IDF_i = \log((N/n_i+1) + con)$$

where, N is the number of documents. ni isthe number of documents with the feature word. To prevent the denominator from becoming 0, 1 is added. And con acts as the constant to avoid IDF from becoming 0.

Hence, the formula for TF-IDF is :

$$TFIDF = TF_{i,j} x IDF_i$$

Once the texual data is converted to vectors, thenext step is to split them as training and testing data. This is explained below.

5) Splitting

The data is split in the ratio of 80:20, where the former comprises the training set while the latter belongs to the testing set.

A. Modeling

1) Logistic Regression

Logistic Regression is a model where the co- efficients are learned during the training of the model. It does not assume anything. It gives the probability estimates of all classes. For Logistic Regression we will be using both TF-IDF and DTM.

2) Naïve Bayes

This model predicts the output as 0 which is sincere or 1 which is insincere. The approach used for this says that is the probability of ex- pression id greater than 1, we conclude detecting it as insincere. For this, the probability of a single word is calculated first. The using matrix multiplication, we calculate the probability of the question as a whole. DTM will be used for the calculation of probability as it has already built a matrix with word occurences for a certain question.

3) Logistic + Naïve Bayes

This model combines the theoretical coefficients of Bayes and learned coefficients of Logistic Regression to make a new Ensemble Model. It gets the best features of both the models with the aim of improving the results. Forth is DTM will be use as it is vectorizer that can be used as both Logistic Regression and Naive Bayes.

4) Support Vector Machines

Support Vector Machine (SVM) is a model that learns from the training data and assigns categories to new data. It learns from the data if a certain type of data belongs to one category or other. Later, it checks weather the data belongs to one category or other based on the proper- ties learned from training data. In this project, the data is trained based the words that occur repeatedly for a certain category. It understands the occurrence of words then classifies data if it is sincere or insincere.

5) Decision Tree

This model decision based on various outcomes. At a time only two outcomes are tested, it is passed to the one it fits in. In this project, decision tree will be used to identify if a question is sincere or not based on its textual content.

6) Random Forest

Random forests is a group of decision trees working together to give output. In this multiple decision trees are developed during training. 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.

The next section will discuss the various evaluation metrics taken into consideration.

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) Precision

Precision is ratio of correctly predicted outcomes to the total predicted outcomes. It is not de- pendent on the accuracy of the model. It is therefore a measure that be used in case of class imbalance.

3) Recall

It is the ratio of correctly predicted outcomes to the total outcomes. It is also known as the sensitivity of themodel.

4) F1Score

F1 score is the one that is calculated by combining 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.

A. Deployment

After the data is trained, the testing data is fed into the models to see how well the models perform. This testing data is used to calculate the evaluation values. Based on the evaluation metrics, Decision Tree performs the best followed by Random Forest.

IV. EVALUATION AND RESULT

Various evaluation metrics are considered as the data is highly imbalanced. The accuracy level cannot be considered for judging the best model as even if all questions are to be considered sincere the accuracy will be above 90%. Hence, F1 score acts as the main metric for evaluating the performance of the models. As seen in the figure below, Naive Bayes performed using TF-IDF, Logistic Regression using Count Vectorizer and their ensemble method gives the highest F1 score. Thus, they are considered as the best performing models. The most poor performance is given by Random Forest Model which uses Count Vectorizer and Lemmatization who has an F1 score of just
0.41. In most cases Count Vectorizer has performed better as all models have been implemented using all methods of pre-processing and the ones with high performance have been mentioned in the report. As seen, the accuracy of all models is very high but this is due to imbalance in the dataset. To overcome that other metrics have been considered to give accurate results.

V. CONCLUSION

Overall it can be said that the project is able to classify between sincere and insincere questions. It used multiple models which follow supervised learning algorithm. An ensemble model is also implemented in the project which combines the advantages of both Naive Bayes and Logistic Regression Model. According to various metrics, it can be stated that

Decision Tree perform the best followed by Random Forest. Other models also perform considerably well and there is no marginal difference between the metrics of different models.

A. Future Work

For future implementation, this project can be implemented using various deep learning models, LSTM etc. The problem of data imbalance can be solved by first training the data with equal number of sincere and insincere questions.

REFERENCE

[1] C. Liu, Y. Sheng, Z. Wei, and Y. Yang, "Research of text classification based on improved tf-idf algorithm," in 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Aug 2018, pp. 218–222.

[2]O. Aborisade and M. Anwar, "Classification for authorship of tweets by comparing logistic regression and naive bayes classi- fiers," in 2018 IEEE International Conference on Information Reuse and Integration (IRI), July 2018, pp. 269–276.

[3]S. T. Indra, L. Wikarsa, and R. Turang, "Using logistic regres- sion method to classify tweets into the selected topics," in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Oct 2016, pp. 385–390.

[4]S.-J. Yen, Y.-S. Lee, J.-C. Ying, and Y.-C. Wu, "A logistic regression-based smoothing method for chinese text catego- rization." Expert Systems With Applications, vol. 38, pp. 11 581– 11 590, 2011.

[5]A. Onan, S. Korukog˘lu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification." Expert Systems With Applications, vol. 57, pp. 232 – 247, 2016.

[6]"A comparison of several ensemble methods for text categorization," in IEEE International Conference onServices Computing, 2004. (SCC 2004). Proceedings. 2004, Sep. 2004, pp. 419–422.

[7]A. Wibowo Haryanto and E. K. and, "Influence of word normalization and chi-squared feature selection on support vector machine (svm) text classification," in 2018 International Seminar on Application for Technology of Information and Communication, Sep. 2018, pp. 229–233.

[8]C. Alfaro, J. Cano-Montero, J. Go´mez, J. Moguerza, and F. Ortega, "A multi-stage method for content classification and opinion mining on weblog comments." Annals of Operations Research, vol. 236, no. 1, pp. 197 – 213, 2016.

[9]Y. Liu, J. Niu, Q. Zhao, J. Lv, and S. Ma, "A novel text classification method for emergency event detection on social media," in 2018 IEEE SmartWorld, Ubiquitous In- telligence Computing, Advanced Trusted Computing, Scal- able Computing Communications, Cloud Big Data Comput- ing, Internet of People and Smart City Innovation (Smart- World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Oct 2018, pp. 1106–1111.

[10]F. CHIROMA, H. LIU, and M. COCEA, "Text classification for suicide related tweets," in 2018 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2, July 2018, pp. 587–592.

[11]F. Gu¨rcan, "Multi-class classification of turkish texts with ma- chine learning algorithms," in 2018 2nd International Sympo- sium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Oct 2018, pp. 1–5.

[12]P. Liu, H. Yu, T. Xu, and C. Lan, "Research on archives text classification based on naive bayes," in 2017 IEEE 2nd In- formation Technology, Networking, Electronic and Automation Control Conference (ITNEC), Dec 2017, pp. 187–190.

