CMPE 255 – Large Scale Analytics Project Report

Title: Quora Insincere Question Classification

Prof: Gheorghi Guzun

San Jose State University

Spring 2018

Team Members:
Ankit Joshi (013724412)
Mohit Gahlot (013753454)
Shashwat Jain (013707148)

INTRODUCTION

A problem today for major websites is, how to handle toxic and divisive content?
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 remove insincere questions, those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

In this Project we developed models that identify and flag insincere questions. With this model, Quora can develop more scalable methods to detect toxic and misleading content.

In this project we will be predicting whether a question asked on Quora is sincere or not.
An insincere question is defined as a question intended to make a statement rather than look for helpful answers. Some characteristics that can signify that a question is insincere are:

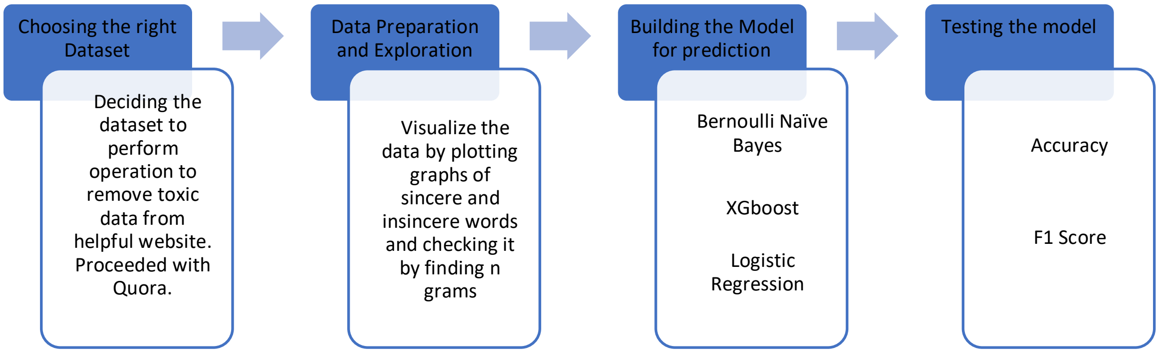
- Has a non-neutral tone
    - Has an exaggerated tone to underscore a point about a group of people?
    - Is rhetorical and meant to imply a statement about a group of people
- Is disparaging or inflammatory
    - Suggests a discriminatory idea against a protected class of people, or seeks confirmation 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.

SYSTEM DESIGN AND IMPLEMENTATION

System Design:

The objective of this project was to segregate sincere and insincere questions posted on Quora. This was achieved and observed based on various factors. In the dataset, there were multiple files with train dataset having 3 columns of Question ID, text and Target as sincere or Insincere. We decided to make use of these different approach to better predict the questions target. Three algorithms were followed to analyze and to predict the output.

The flow of the project is as follows:



Choosing the right Dataset

Deciding the dataset to perform operation to remove toxic data from helpful website. Proceeded with Quora.

Data Preparation and Exploration

Visualize the data by plotting graphs of sincere and insincere words and checking it by finding n grams

Building the Model for prediction

Bernoulli Naïve Bayes

XGboost

Logistic Regression

Testing the model

Accuracy

F1 Score

Fig 1. Workflow of the Project

To benchmark, various classification algorithms were used. They are discussed in the next section.

Algorithms Used:

After implementing sparse matrix following algorithm were performed on the data to find the target of the question to be sincere and insincere:

a) **Naive Bayes Classifier:**
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

b) **XGBoost:**
XGBoost is an ensemble classifier which is an implementation of gradient boosted classifiers. The algorithm used by XGBoost is a variant of gradient boosting. In Gradient Boosting, weak learners (like Shallow Decision Trees) iteratively run on the dataset. With each iteration, the next learner learns from its predecessors to predict the errors of the prior models. These models are then added together.
The model uses the gradient descent algorithm to minimize its error. XGBoost, of late, has almost become a silver bullet and is extensively used in competitions. By using weak learners as base estimators, it overcomes overfitting and aggregation reduces the bias of the weak learners. Thus it is able to overcome the bias-variance tradeoff.

Implementation: XGBoost algorithm is implemented by the XGBoost library. XGBoost is a fast algorithm and provides parallel tree boosting. We have used the XGBClassifier model provided by the XGBoost library. It provides 3 base learners: gbtree, gblinear, and dart.

c) **Logistic Regression:**

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Tools:
Anaconda - Jupyter Notebook
We used Jupyter Notebook because it has many inbuilt libraries and can be used readily.

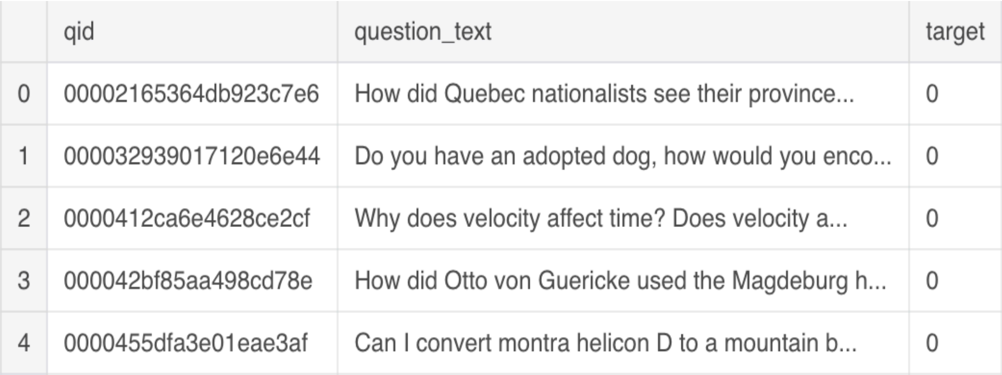
EXPERIMENTS/ PROOF OF CONCEPT EVALUATION

A) **Dataset:**

The Quora Insincere Question Classification dataset was obtained from Kaggle (https://www.kaggle.com/c/quora-insincere-questions-classification). Mainly it contains 2 CSV files (train.csv and test.csv). The total size of the dataset is 6 GB. The table below contain details pertaining to number of instances for each attribute in respective files.

CSV FILE NAMES	#Rows	#Columns
Train.csv	1306122	3
test.csv	56370	2

Total training records - approximately 1306122, Insincere questions – approximately 80810.



	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0

Fig 2. Dataset containing the data

The train data has 3 columns question id (qid), question_text and target. Target = 0 means sincere question and target = 1 means insincere question.
The test data set has 2 columns- question id and question_text.

Data Set Analysis:
Perform n-gram operation to find the maximum occurring words in the data. Used bigrams and trigrams for the evaluation of the same and used ngrams 1, 2 and 3 in tf idf vectorization.

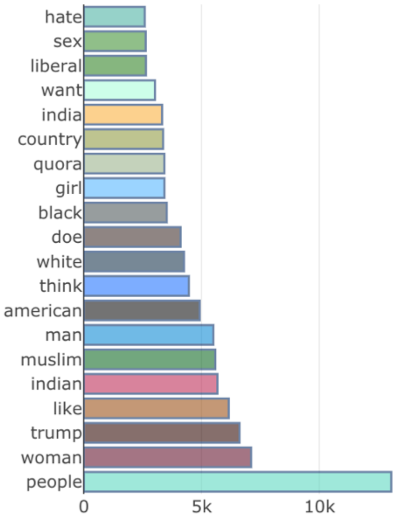


Fig. 3 Top Sincere Words

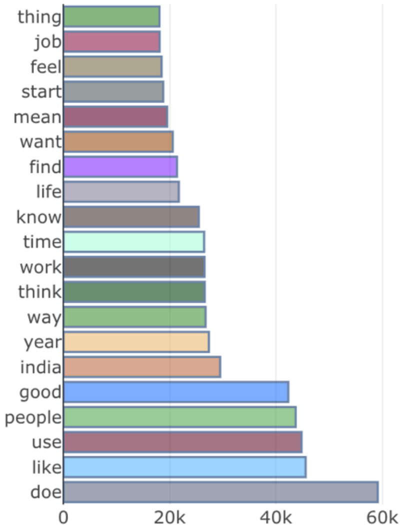
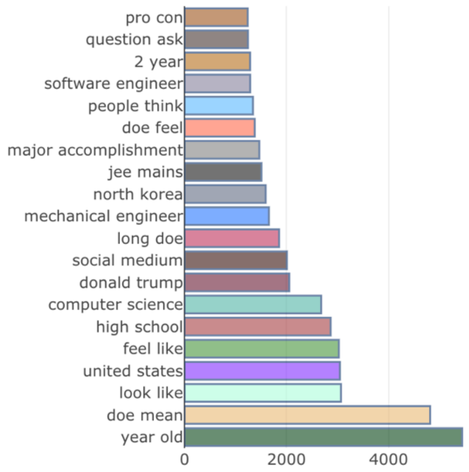


Fig. 4 Top Insincere Words



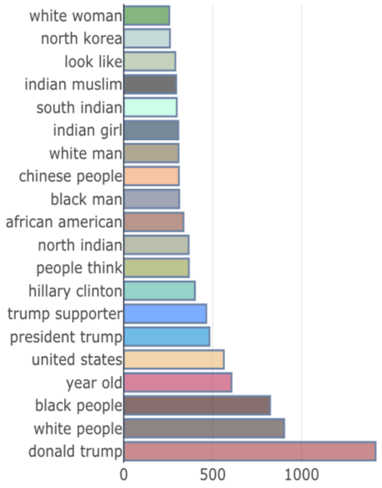
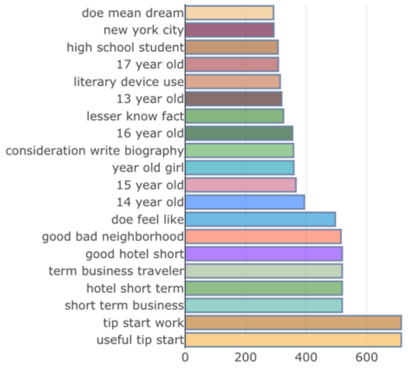


Fig. 5 Top 20 Sincere Bigrams

Fig. 6 Top 20 Insincere Bigrams



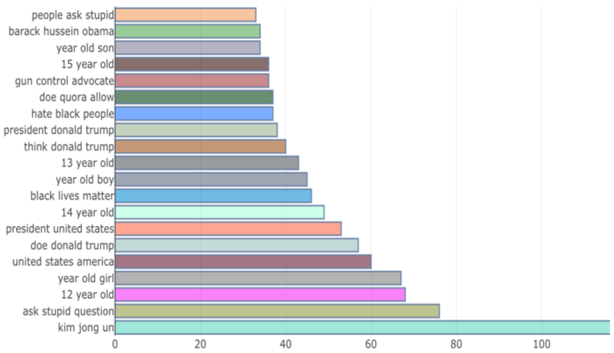


Fig. 7 Top 20 Sincere Trigrams

Fig. 8 Top 20 Insincere Trigrams

B) Data Preprocessing

- **Balancing the Data** The data provided is unbalanced with number of sincere questions making up 93.8% of the dataset. To overcome this issue, data is created by under sampling the sincere questions and making them equal to the number of insincere questions.
- If there are negation words, like "wouldn't", transform it as "would not".
- Convert the misspelled words to proper spelling (also British English to American English)
- Remove the stop words as we don't want these to be part of our model.

- Remove non-alphabetic characters
- Convert the texts to lowercase.
- Filter out all punctuation, plus tabs and line breaks, except the ' character in a text question.
- For vectorization of each question text, **tf-idf** vectorization was used. tf-idf stands for Term Frequency - Inverse Document Frequency. Each word in a document is given a weight which is proportional to its frequency in the question text but inversely proportional to the frequency in the corpus.

C) Methodology followed

- The data, once preprocessed was split into test-train ratio of 20:80.
- Used tf idf vectorizer library and made sparse matrix of the text data.
- The terms are defined as n-grams where n can be 1, 2 or 3. (Unigram, Bigram and Trigram)
- Used 2 approaches
  a. Without using word embedding and using tf idf vectorizer.
  b. With using word embedding and not using tf idf word vectorizer.
     Using word tokenization and lemmatization extra as the preprocessing part.

  and 3 different algorithms on that, Bernoulli Naïve Bayes, XGBoost and Logistic Regression.
- Analyzed the result with both the approaches.

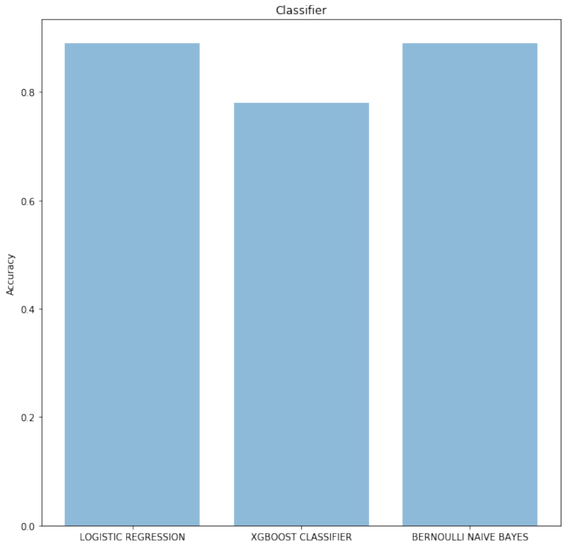
D) Results & Evaluation

The results are tabulated below for each approach:

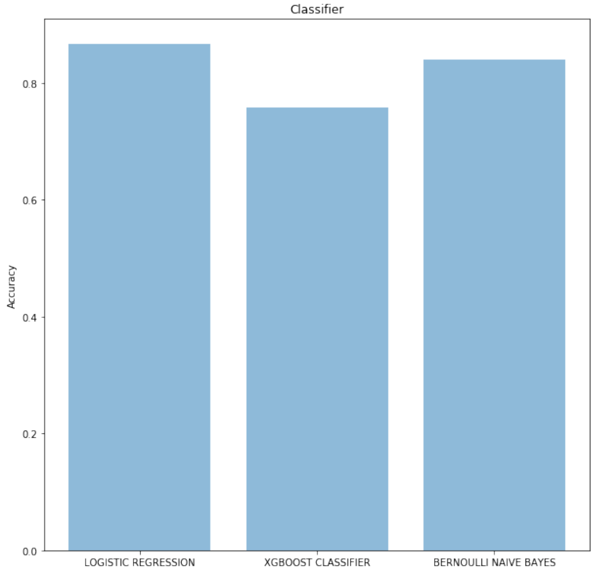
Algorithm	Approach 1		Approach 2	
	F1 Score	**Accuracy**	**F1 score**	**Accuracy**
Bernoulli Naïve Bayes	0.83	0.84	0.86	0.89
XGBoost	0.71	0.75	0.75	0.78
Logistic Regression	0.87	0.86	0.88	0.89

Accuracy Plots:

Approach 1:

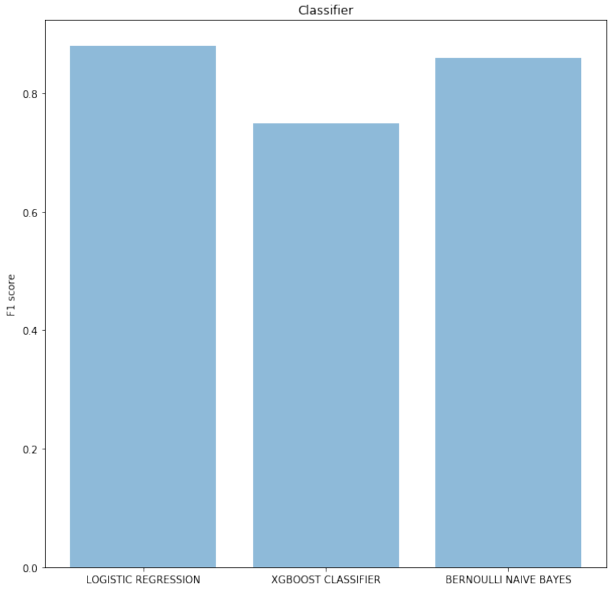


Approach 2:

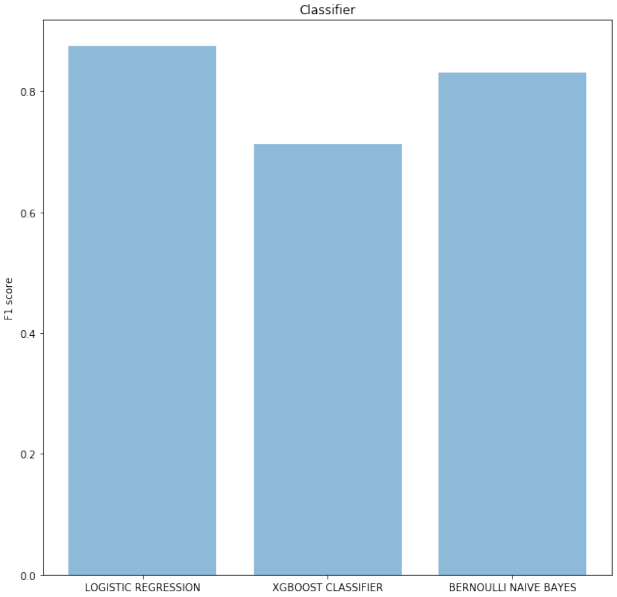


F1 score plot:

Approach 1:



Approach 2:



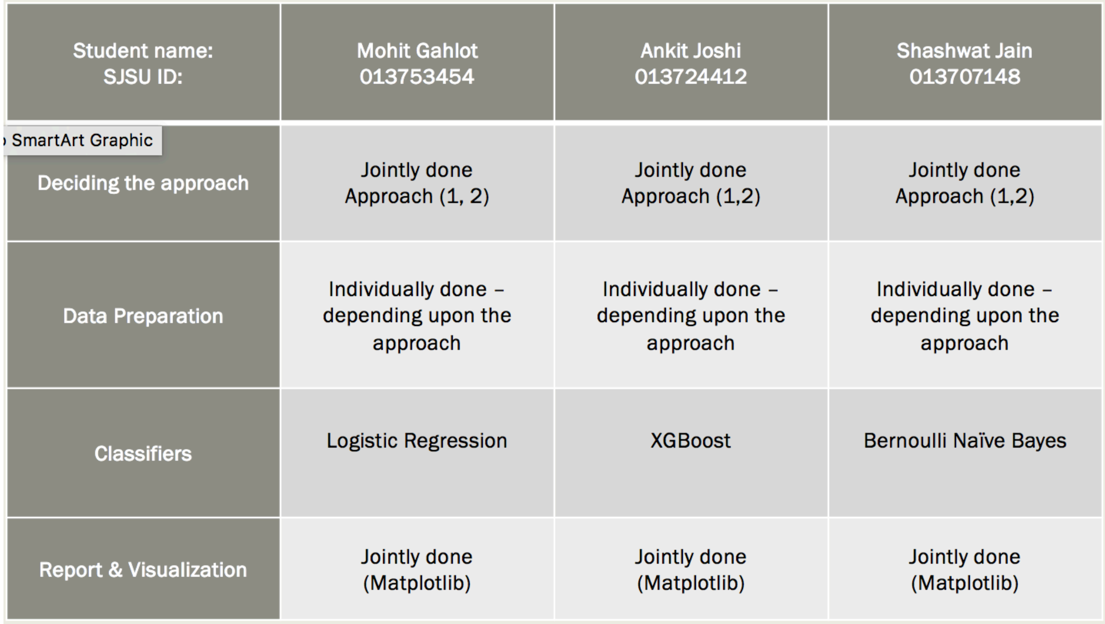
Analysis:
As observed from the experimental results, the approach of XGBoost classifier yield to the accuracy of 0.75. Upon using Bernoulli Naïve Bayes classifier, the accuracy boost up to around 0.84 and finally by using Logistic Regression we got accuracy of 0.86.

DISCUSSION & CONCLUSION

- Looking at the volume of data, it was decided to include as many attributes as possible from the available data. This led to dividing the methodology into three different approaches, which later reconfirmed our belief that more relevant data can lead to better predictions.
- The basic initial idea was to first make a sparse matrix and use different classifiers or approaches to predict the question target as sincere and insincere.
- Decided to try and test by using different classifiers and individually came up to solution of using Bernoulli Naïve Bayes, XGBoost and Logistic Regression and observed best accuracy with Logistic Regression.
- Dataset was imbalanced hence creating problem in finding the desired accuracy.
- Overcame this problem by under sampling the data to make sincere and insincere questions count equal.
- There were difficulties in compilation of the program as it takes a lot of time to train the data and hence took time to decide the best approach.
- We tried working on hpc and on Kaggle to train the data on multiple machines and hence decreased the computation time.
- The aim of the project was to predict accurately the target of the question. The model built can be used to further predict the target of the question to be sincere or insincere and can help Quora to filter out the correct post to be posted, which was implemented successfully.

Project Plan/ Distribution

The project had three different approaches towards the same goal of predicting the target of the question. The tasks were equally distributed amongst the team and the distribution is as follows:


SmartArt Graphic

Student name: SJSU ID:	Mohit Gahlot 013753454	Ankit Joshi 013724412	Shashwat Jain 013707148
Deciding the approach	Jointly done Approach (1, 2)	Jointly done Approach (1,2)	Jointly done Approach (1,2)
Data Preparation	Individually done – depending upon the approach	Individually done – depending upon the approach	Individually done – depending upon the approach
Classifiers	Logistic Regression	XGBoost	Bernoulli Naïve Bayes
Report & Visualization	Jointly done (Matplotlib)	Jointly done (Matplotlib)	Jointly done (Matplotlib)

Github Link for the project:
https://github.com/Shashwatjain31/Quora-Insincere-Question-Classification

References:

https://xgboost.readthedocs.io/en/latest/ https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math- behind-xgboost/

http://www.tfidf.com/ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

https://www.kdnuggets.com/2017/10/understanding-machine-learning-algorithms.html