BT5153 Topics in Business Analytics
Final Project Report

Quora Insincere Questions Classification

Group 4: Data Doctors

Han Bing	A0186003N
Li Liping	A0186040M
Monica Ravipudi	A0186119Y
Shubhanshu Gupta	A0185998X
Yagna Srikanth Akella	A0176603E

INTRODUCTION

As a platform that connects "people who ask questions" and those "who contribute unique insights and quality answers", Quora empowers people to share their knowledge with each other. As with any other massive opinion-sharing websites, Quora faces the need to handle toxic, divisive and misleading content, in order to provide its users a sense of security while sharing their own knowledge to the questions posted.

PROBLEM DEFINITION

One of the specific challenges regarding this influential problem is to weed out insincere questions. To tackle this problem, Quora has currently employed both manual review and machine learning techniques. However, they are seeking for more scalable methods to identify insincere questions with more efficiency. Their objective is "to combat online trolls at scale, in order to uphold their policy of 'Be Nice, Be Respectful' and continue to be a safe place for knowledge sharing and growing".

What is an Insincere Question?
A general definition of insincere questions would be those founded upon false premises, or that intend to make a statement instead of looking for helpful answers. Some characteristics that can signify that a question is insincere:

1) 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
2) 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
3) Isn't grounded in reality
    - Based on false information, or contains absurd assumptions
4) Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

Therefore, our group aims to develop suitable models to predict whether a question asked on Quora is sincere or not. In other words, our objective is to "**Identify and flag insincere questions in the Quora website**".

DATASET DESCRIPTION

The training dataset contains 1.31 million Quora questions asked, and their corresponding labels as sincere or insincere. We divided the dataset into training and validation sets in 4:1 ratio, to

develop scalable classification models that can apply to other questions asked. We aim to increase accuracy and scalability of our models.

Training Set: 1,048,000 questions

Validation Set: 252,000 questions

Data Fields:

Variable Name	Description	Data Type
qid	Unique question identifier	String
question_text	Quora question text	String
target	A question labeled "insincere" has a value of 1, otherwise 0	Boolean

PRE-PROCESSING

We performed the following steps for data preprocessing:

- Same case conversion; Null values check; Alphanumeric elements and stop words removal; Missing rows removal; Punctuation removal.
- Padding sentences to 200 words - Long sentences are trimmed to 200 words and shorter ones are imputed with zero indexes.
- Tokenization: Break down the sentence into unique words.
- Indexing: Put the words in a dictionary-like structure and give them an index each. Number of unique words in a dictionary was limited to 20000 words.

EXPLORATORY DATA ANALYSIS

We performed basic exploratory data analysis before performing machine learning techniques. We verified the distribution of sincere and insincere questions in the training dataset and noticed that close to 94% of questions in the dataset are sincere and only 6% are insincere questions. This points to a huge imbalance in the dataset which has to be dealt at the later stages of our analysis.

We then verified on the same features for both sincere and insincere questions. Below is the summary of the EDA.

Feature	Sincere Question	Insincere Question
Average Syllables	17	24
Average Lexicons	12	17
Average length	68	98
Average Syllables per word	1.42	1.44
Average letters per word	4.66	4.77

We have observed from the above Descriptive analysis that:

- Average number of Syllables, lexicons, length is higher for an insincere question than that of a sincere question.

- Average syllables per word, letters per word of an insincere question is close to that of a sincere question.

We have also analyzed the readability features of the text in sincere and insincere questions. Below are the readability indices we analysed:

1) **The Felsh Reading Ease Formula:** Flesch Reading Ease Formula is considered as one of the oldest and most accurate readability formulas. The formula for the Felsh reading ease score test is:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

The FRES score for insincere questions in our training data is 70. This implies that the text is plain English and easily understood by 13-15 year old students based on the index definition. The FRES score for sincere questions is 75. This implies that the text is fairly easy to read.

2) **Coleman – Liau Index:** This index relies more on the letters in the text. Below is the formula to calculate the Coleman – Liau index:

$$CLI = 0.0588L - 0.296S - 15.8$$

Where L is the average number of letters per 100 words;
S is the average number of sentences per 100 words.
The CLI for insincere questions is 9.2 and that for sincere questions is 8.13.

3) **Automated Readability index:** Automated index also depends on the letters in the text rather than syllables. Below is the formula to calculate this index.

$$4.71\left(\frac{\text{characters}}{\text{words}}\right) + 0.5\left(\frac{\text{words}}{\text{sentences}}\right) - 21.43$$

ARI for insincere questions is higher than that of sincere questions, similar to CLI. ARI for insincere questions is 8.2 and that for sincere questions is 6.2.

FEATURE ENGINEERING

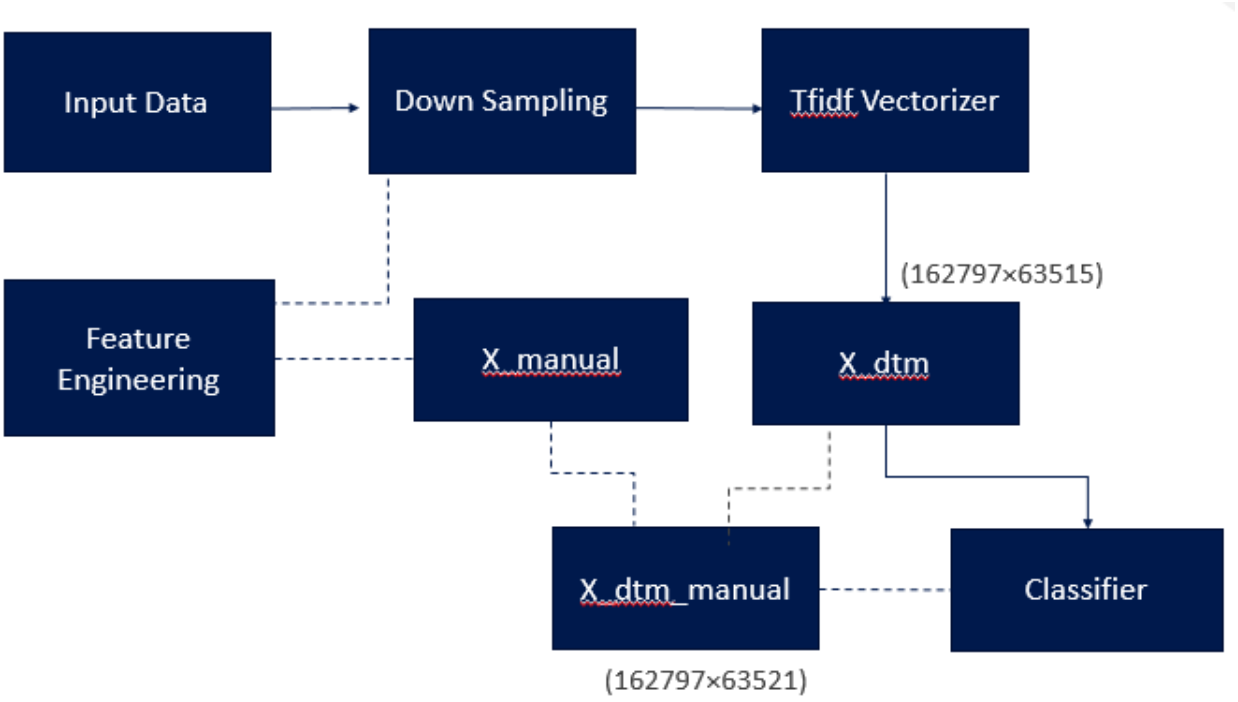
As part of feature engineering, we tried to understand whether physical attributes can have an influence on the target. We extracted features such as number of words, number of unique words, number of words in full capitals, number of uppercase words, number of characters, average of word length. Thus we have added 6 more features to our raw data:

question_text	target	num_words	num_unique_words	num_chars	num_words_upper	num_words_title	mean_word_len
How did Quebec nationalists see their province...	0	13	13	72	0	2	4.615385
Do you have an adopted dog, how would you enco...	0	16	15	81	0	1	4.125000
Why does velocity affect time? Does velocity a...	0	10	8	67	0	2	5.800000

MACHINE LEARNING MODELS

TF – IDF Vectorizer: To make TF-IDF vectorizer more scientific, we fit all our data to get all the features in the question text. Then we use this vectorizer to transform the data we get after down sampling, as well as the validation data. The number of features we get from TF-IDF vectorizer is about 60K.

Construction of Pipeline: Then we build a pipeline for all our models. For each algorithm, we run two results, one is without feature engineering, and another one is with feature engineering. Below is a flow chart of the pipeline:



Dealing with Imbalanced Dataset: Since our dataset is an imbalanced one, we need to do sampling in our data to make sure that our models are reliable and not biased. We choose down sampling, since the number of rows of raw data is more than 1 million in training dataset. By using

down sampling, not only will our training data become a balanced dataset, but also the size of the training dataset will decrease significantly, leading to significant reduction of computational power. In the pipeline with feature engineering, we union the features from TF-IDF and the features from feature engineering. Then we throw our sparse data frame into different classifier.

Accuracy Measure: We have picked F1 score to check the accuracy of the models, which is one of the famous accuracy measures for a classification problem. Since F1 Score is the weighted average of Precision and Recall, it takes both false positives and false negatives into account. Below is the formula to calculate F1 score from Precision and Recall.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially when handling data with uneven class distribution. Accuracy works best if false positives and false negatives have similar cost, but when the cost of false positives and false negatives are significantly different, it's better to look at both Precision and Recall. Obviously, in our case, the cost of false negatives is greater than that of false positives.

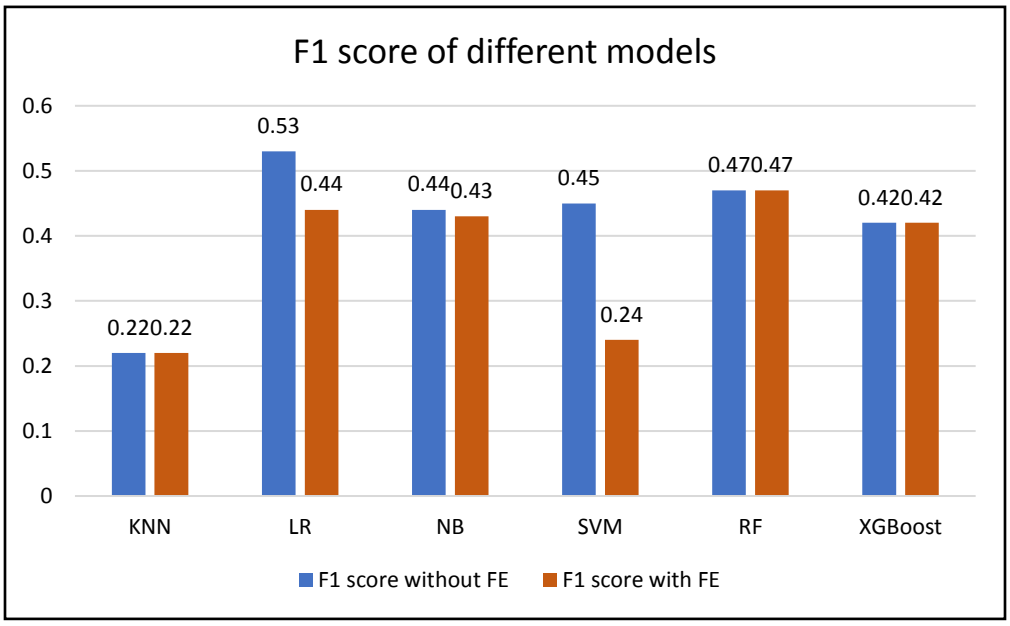
We have considered the following baseline models on which we trained our data after down-sampling:

- KNN
- Logistic Regression
- Naive Bayes
- Support Vector Machine

We also trained the dataset using the below Ensemble Models :

- Random Forest
- XGBoost

The graph below shows the result of each algorithm.



Blue bars represent the results without feature engineering

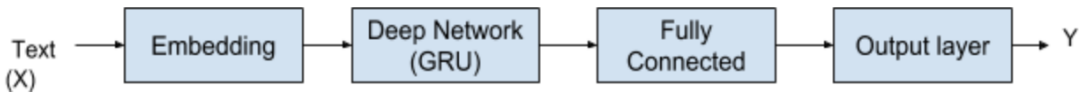
Red bars represent the results with feature engineering

We noticed that feature engineering didn't have the desired effect on the model performance. This might be because of the large number of independent variables used in the model prediction. We also notice that Logistic Regression and SVM perform worse with added feature engineering variables. We notice that the Logistic Regression model performs the best in the models tried so far. This confirms the famous known norm that sometimes simple models perform the best.

EMBEDDINGS

The primary question here is: Why do we need to use word embeddings when we already have used TF-IDF and count vectorizer? To understand the reason, we need to understand why word embeddings are useful. A word embedding converts a word to an n-dimensional vector. Words which are related such as 'house' and 'home' map to similar n-dimensional vectors, while dissimilar words such as 'house' and 'airplane' have dissimilar vectors. In this way the 'meaning' of a word can be reflected in its embedding, a model is then able to use this information to learn the relationship between words. The benefit of this method is that a model trained on the word 'house' will be able to react to the word 'home' even if it had never seen that word in training.
This proves that word embeddings could be very useful, since the entire premise of our project is to understand the underlying meaning and context of words to help us distinguish the Quora question text from insincere to sincere. Whereas, TF-IDF matrix captures no meaning and instead, just maps a word to value.

We used 3 different types of word embeddings: **GloVE, Paragram** and **FastText**. **GloVE** is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase linear substructures of the word vector space. Here we used Common Crawl pre-trained word vectors, which consisted of 840B tokens, 2.2M vocabulary, and 300d vectors. Next, we used **Paragram** and **FastText** embeddings. The **FastText** embeddings is a collection of 1 million-word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset.
After a brief description of word embeddings, we will understand how word embeddings work.



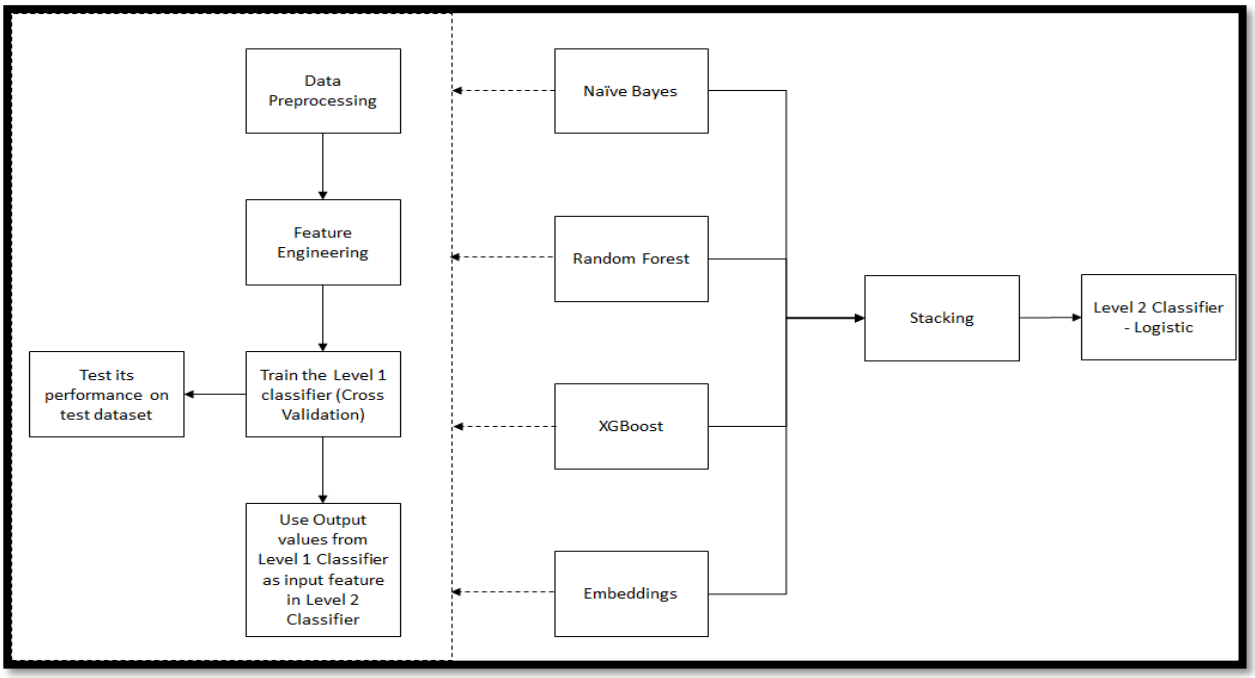
As per the workflow above, we have the embedding layer, deep network layer, fully connected layer, and output layer. As we know, word embedding is a layer that represents text where words that have the same meaning, have a similar representation. In other words, it represents words in a coordinate system where related words, based on a corpus of relationship, are placed closer to each other. The **embedding layer** stores a lookup table to map the words represented by numeric indexes to their dense vector representations. Then comes the **Deep Network layer**, which takes the sequence of embedding vectors as input and converts them into a compressed representation that captures all the information in the sequence of words in the text. We have used GRU as our deep network. Subsequently, we come to the **Fully Connected Layer**. This layer takes the deep representation from GRU and transforms it into the final output classes or output scores. Then finally, we have the **Output Layer**. In this output layer, we use Sigmoid activation function for binary classification (sincere or insincere). **Glove embeddings gave us 0.60 F1 score, Paragram**

gave 0.61, FastText delivered 0.62. Hence, we considered 0.62 as the best score from the embeddings.

STACKING

Stacking is an ensemble model, where a new model is trained from the combined predictions of two (or more) previous models. The predictions from the models are used as inputs for the new model and combined to form a new set of predictions. We picked the Naive Bayes, Random Forest, XGBoost and Embeddings to input into level 2 classifier for Stacking. We used Logistic Regression Model as the level 2 classifier. This is to increase the predictive force of the entire model. This leads us to a **final F1 score of 0.64.**

A brief overview of Stacking is shown in the flow chart below:



RESULTS ANALYSIS

Based on the predicted labels, word clouds for sincere and insincere questions can be drawn to showcase the difference between these two types of questions. Words appearing in sincere questions include “engineering”, “electrical”, “java”, “android”, “internship” etc., which are relatively normal and related to topics such as job and study. However, words for insincere questions include “republicans”, “democrats”, “muslims”, “christians”, “blacks”, “feminists”, which are more related to race, politics, religion, and gender.

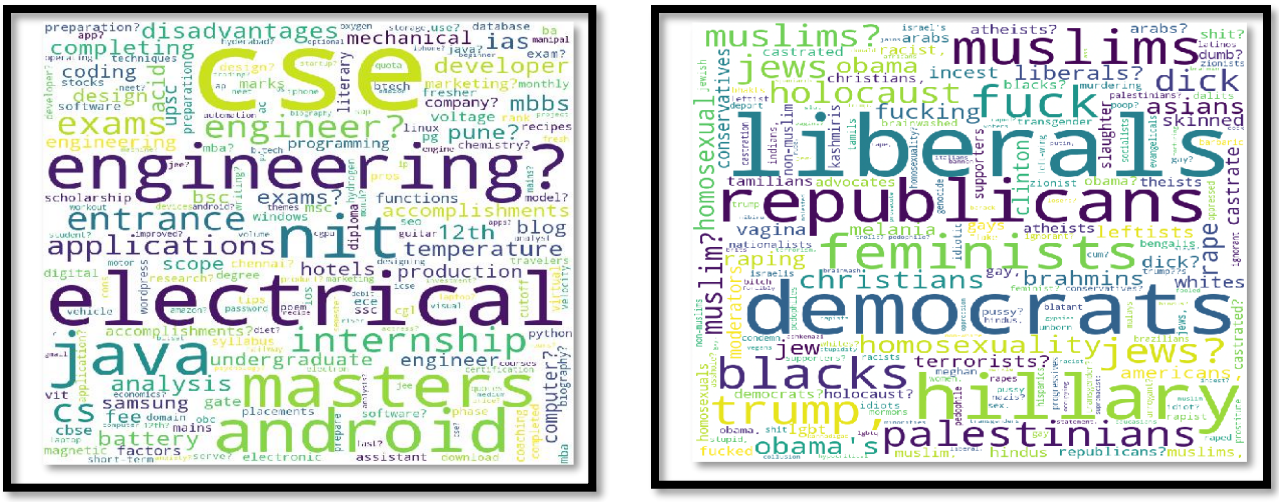


Figure: Word Count of Predicted Sincere Questions (Left) and Insincere Questions (Right)

A ranking of frequent words in sincere and insincere questions would manifest their difference more clearly. Words that are common across both include “people”, “would”, “like”, which are frequently used in daily conversation. Words specifically for sincere questions include “best” and “good”, demonstrating a positive attitude; And words specifically for insincere questions include “trump”, “women”, “white” and “americans”.

Further exploring word count for 2 grams, we found frequent 2-grams for sincere questions include “computer science” and “high school”, and those for insincere ones include “donald trump”, “hillary clinton”, and “white people”. The results for 3-grams are also plotted to provide a clearer sense. Frequent 3-grams for sincere questions include “advice would give”, “what’s best way”, “useful tips someone”, “best way learn”, demonstrating that sincere questions are mostly truly asking for suggestions and tips; And those for insincere ones include “united states america”, “president donald trump”, and “kim jong un”.

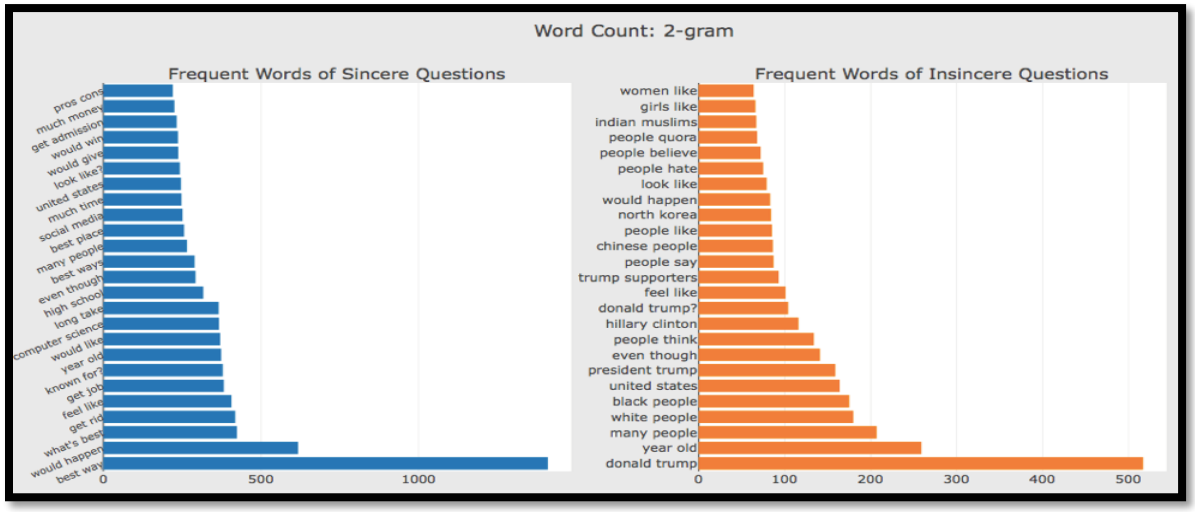


Figure: 2-gram Frequent Words of Predicted Sincere Questions (Left) and Insincere Questions (Right)

It can thus be concluded that there is a significant difference between the predicted sincere and insincere questions, with insincere ones more related to racism, politics, nationalism, religious discrimination, and sexism.

CONCLUSION

Our modeling techniques provided best F1-score of 0.64 through stacking model. Quora should be able to utilize our work (with more data) and stand to their policy of "Be Nice, Be Respectful" in providing safe platform for users to share knowledge with the world. We can enhance accuracy rather than interpretability by varying parameters of work.

Business Value:

Our work can add value to many organizations across industries. Social Media websites such as Twitter, Facebook, Instagram can use our work to ensure and avoid insincere/toxic comments or discussions in their social platforms.

Our work can also help the e-commerce websites where users write reviews on various products. Be it Amazon, Ali Baba or even review websites such as IMDB, Grab Food, Food Panda can also find value in our work to create a friendly environment for the customers and vendors all the time.

Various organizations that gather information and feedback of various potential applicants for job opportunities or providing feedback about their experience working with the organization can benefit with a moderator mode of platform that filters toxic words about the firm which might create a negative remark about the organization in public.

Advantages and Limitations:

Major advantages from our project are:

- ✓ Ensure safe platform to all users and enhance knowledge sharing with the world
- ✓ Application across different industries and organizations

However, we found that our work has its own limitations such as:

- The data is not balanced and so the possibility of false positives need to be handled with care.
- Restriction of computation - Since our data has about 60K features, it is time consuming to perform comprehensive parameter tuning.

Future Scope:

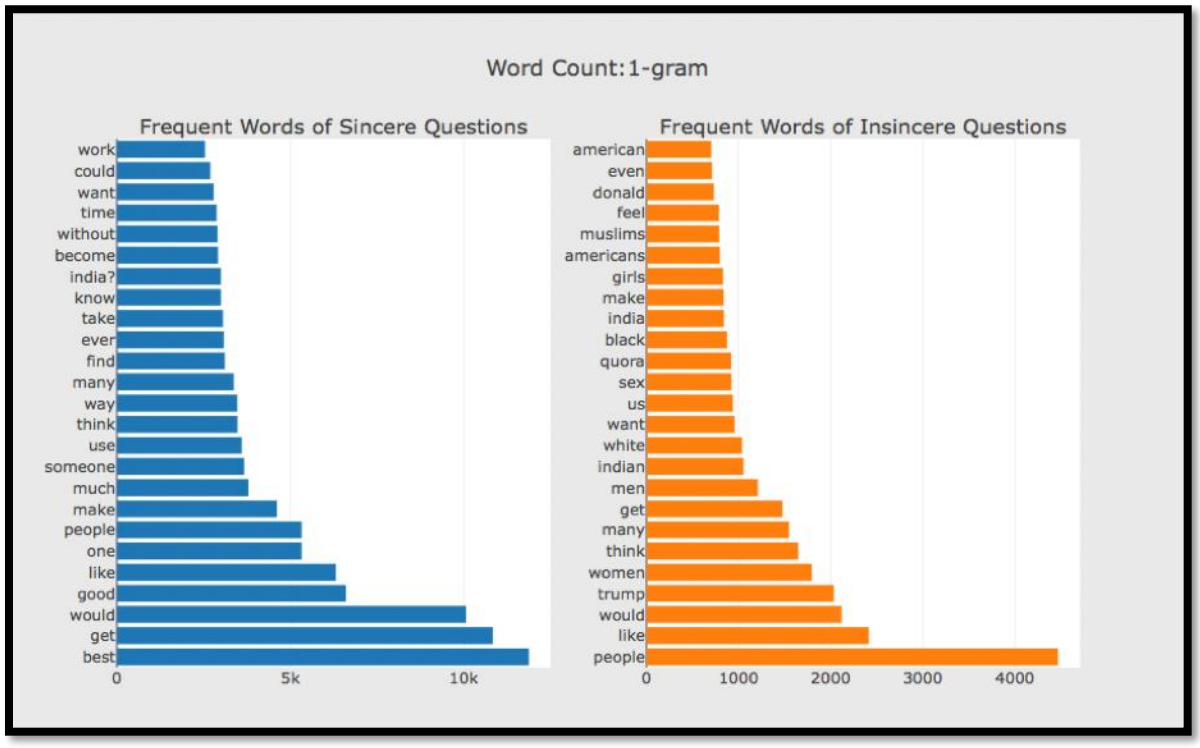
- As part of future work, we can consider to ensemble more deep learning models to further enhance F1 score and reliability of modeling work we have performed.
- We would also like to gather more balanced data and perform the same methodology to observe if the performance of our techniques can improve with more balanced data.

REFERENCES

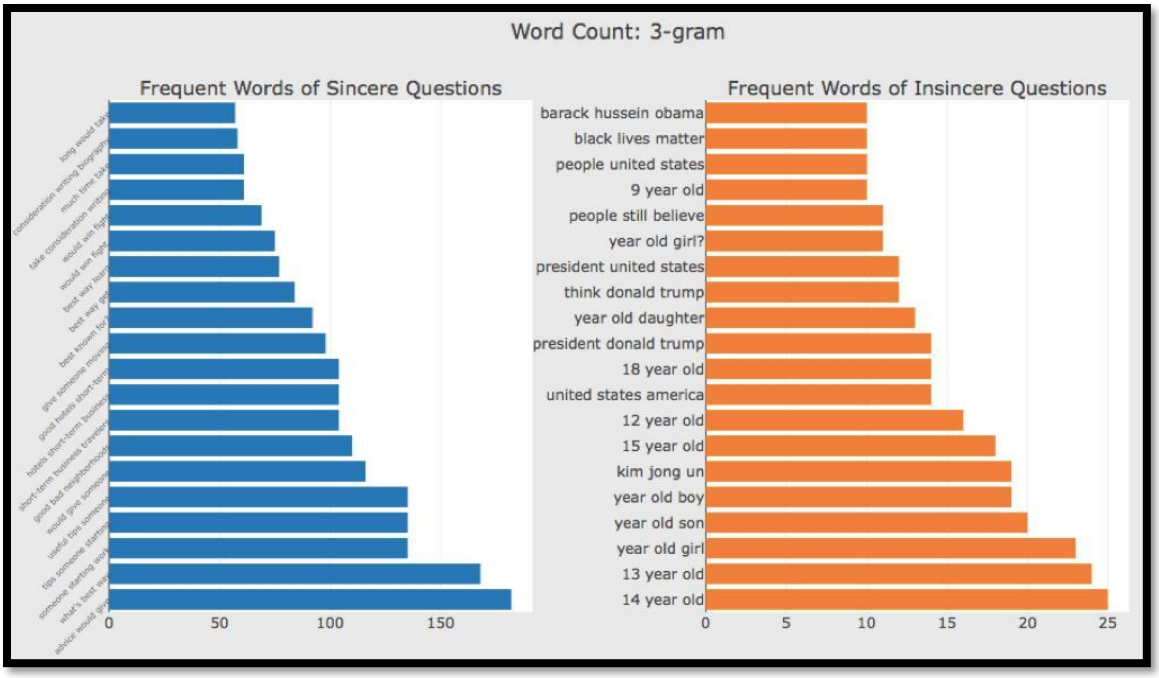
1. https://www.kaggle.com/c/quora-insincere-questions-classification/data
2. https://towardsdatascience.com/beyond-word-embeddings-part-2-word-vectors-nlp-modeling-from-bow-to-bert-4ebd4711d0ec

Appendix

Some of the additional work we have performed that we would like to discuss as part of our project is about the word count of 1-gram and 3-grams performed on the Quora questions in the dataset available.



Appendix 01: 1-gram Frequent Words of Predicted Sincere Questions (Left) and Insincere Questions (Right)



Appendix 02: 3-gram Frequent Words of Predicted Sincere Questions (Left) and Insincere Questions (Right)