CAPSTONE PROJECT 2

QUARA INSINCERE QUESTIONS CLASSIFICATION

PROBLEM

An existential problem for any major website today is how to handle toxic and divisive content.

CLIENTS

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.

By using sentiment analysis, platforms like Quora, they can develop more scalable methods ***to detect toxic and misleading content*** to improve online conversations.

DATA

The training set readily available on Kaggle.

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.

The training data has 1,306,122 rows and 3 columns which are qid, question_text, and target.

DATA COUNTINUED

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

The training data includes the question that was asked, and whether it was identified as insincere (target = 1).

DATA WRANGLING

- Datas don't have NaN values and '', datas don't need to be cleaned.

EDA

- During exploratory data analysis, we ask the following questions to understand more about our data:

1. How to handle unequal portiation of sincere and insincere questions?
2. How to drop the number of text from 1,306,122 to 1,000?
3. Word frequency?
4. Usually how many tokens are there in a question?
5. and ask follow-up questions if needed.

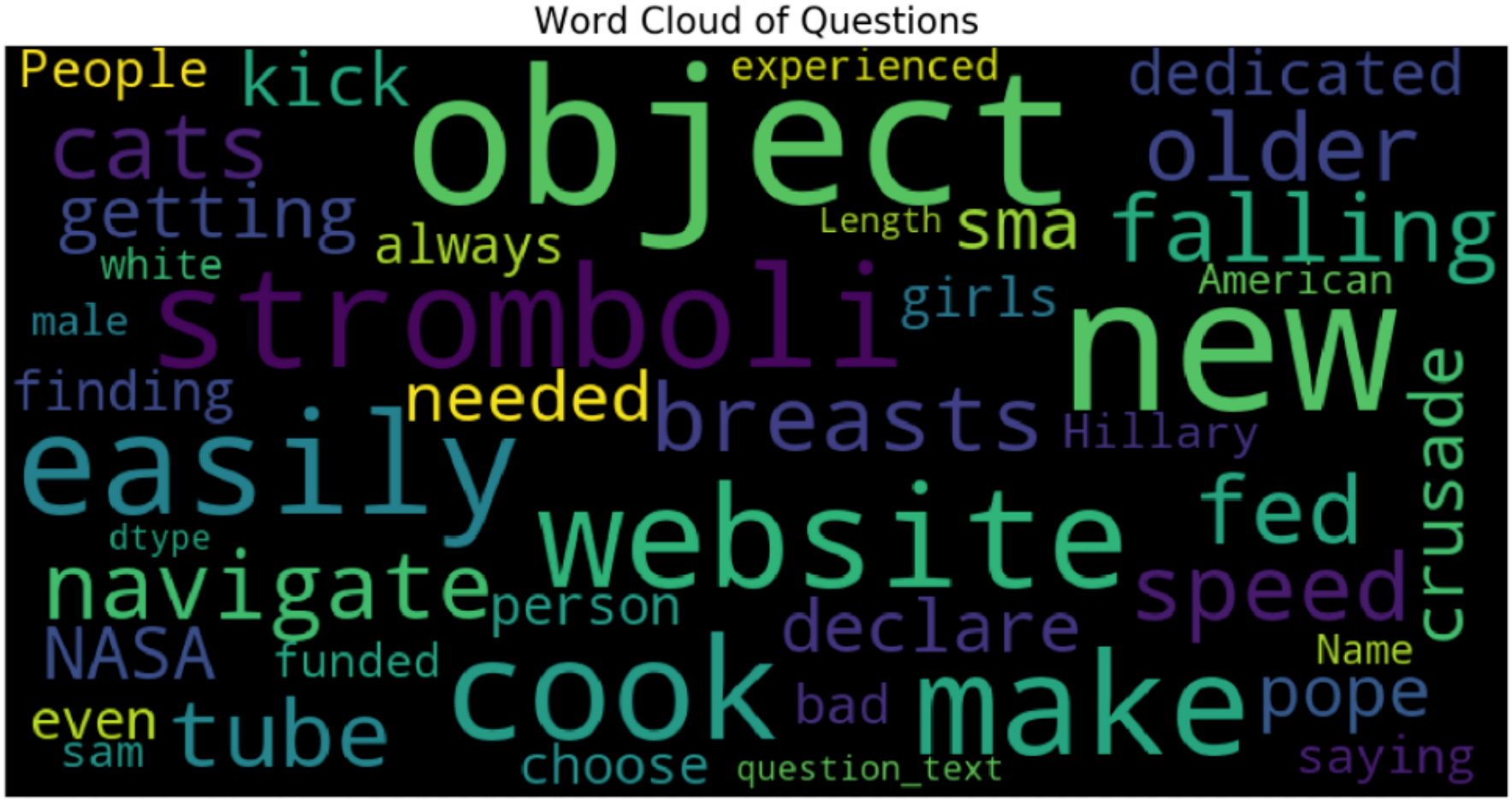
EDA CONTINUED

Numbers of sincere questions: 1225312

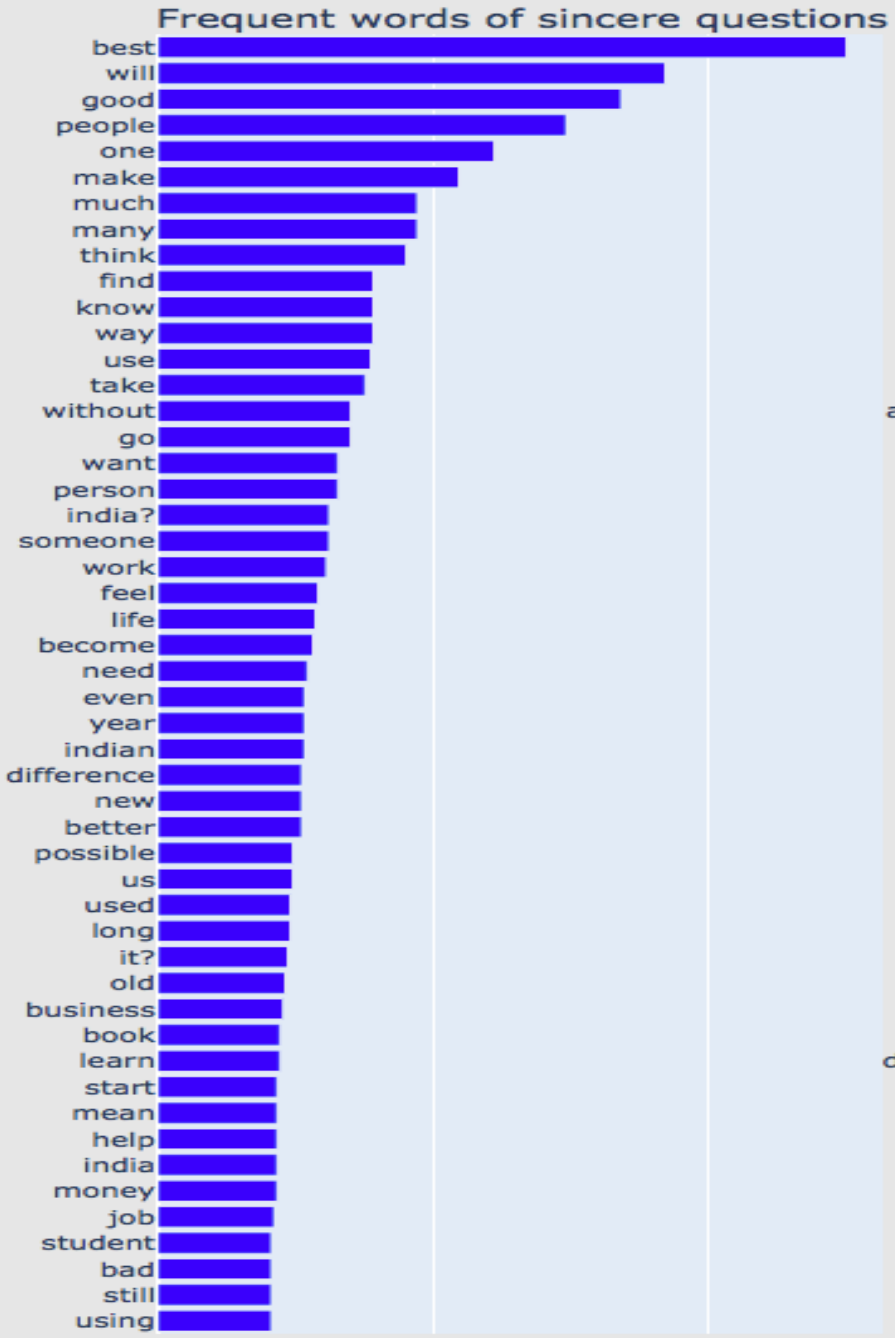
Numbers of insincere questions: 80810

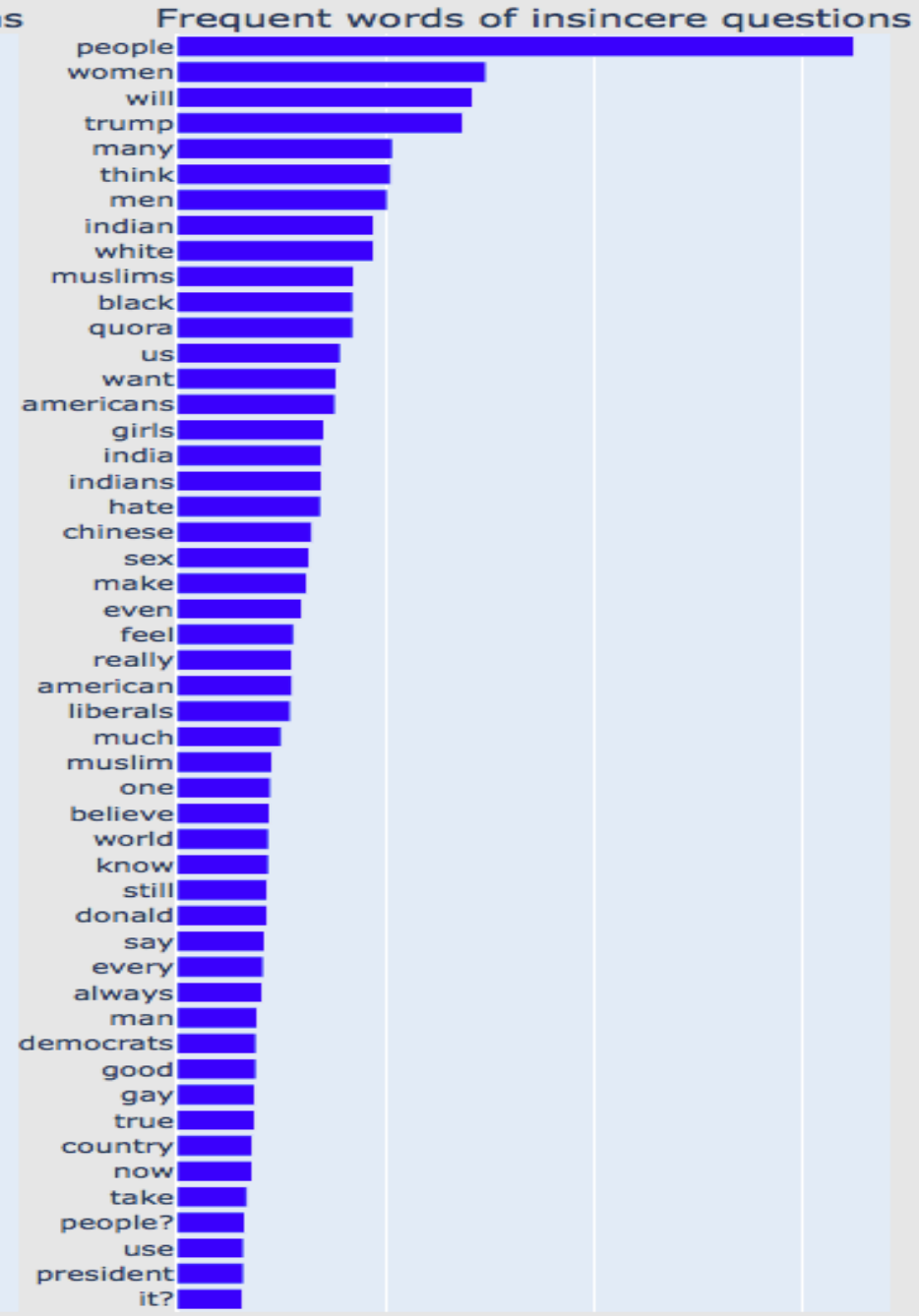
- Proportion very unbalanced, we have to try to scale datas. We will pick 5000 random sincere and 5000 insincere questions from the dataset.

EDA CONTINUED

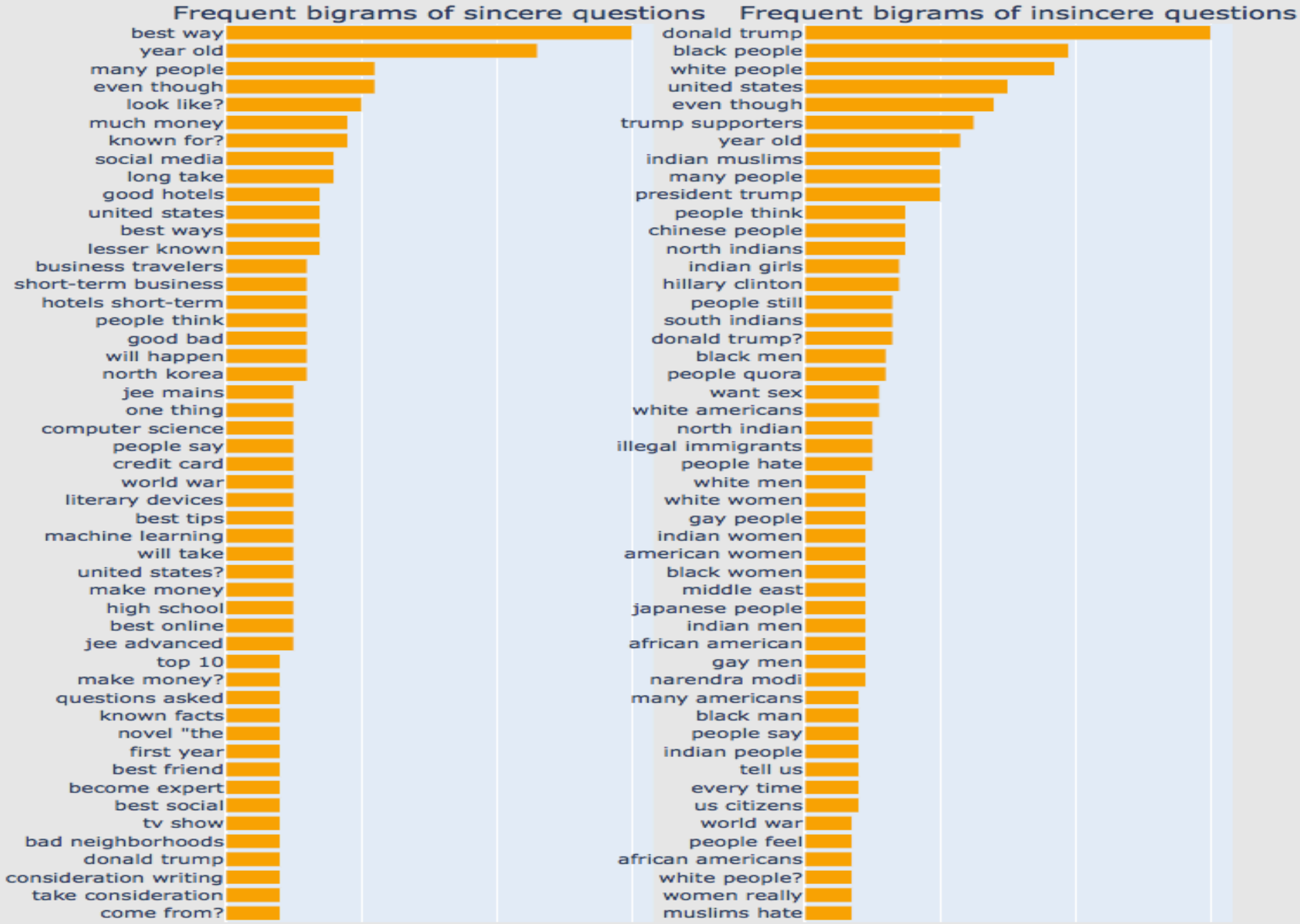


Word Count Plots

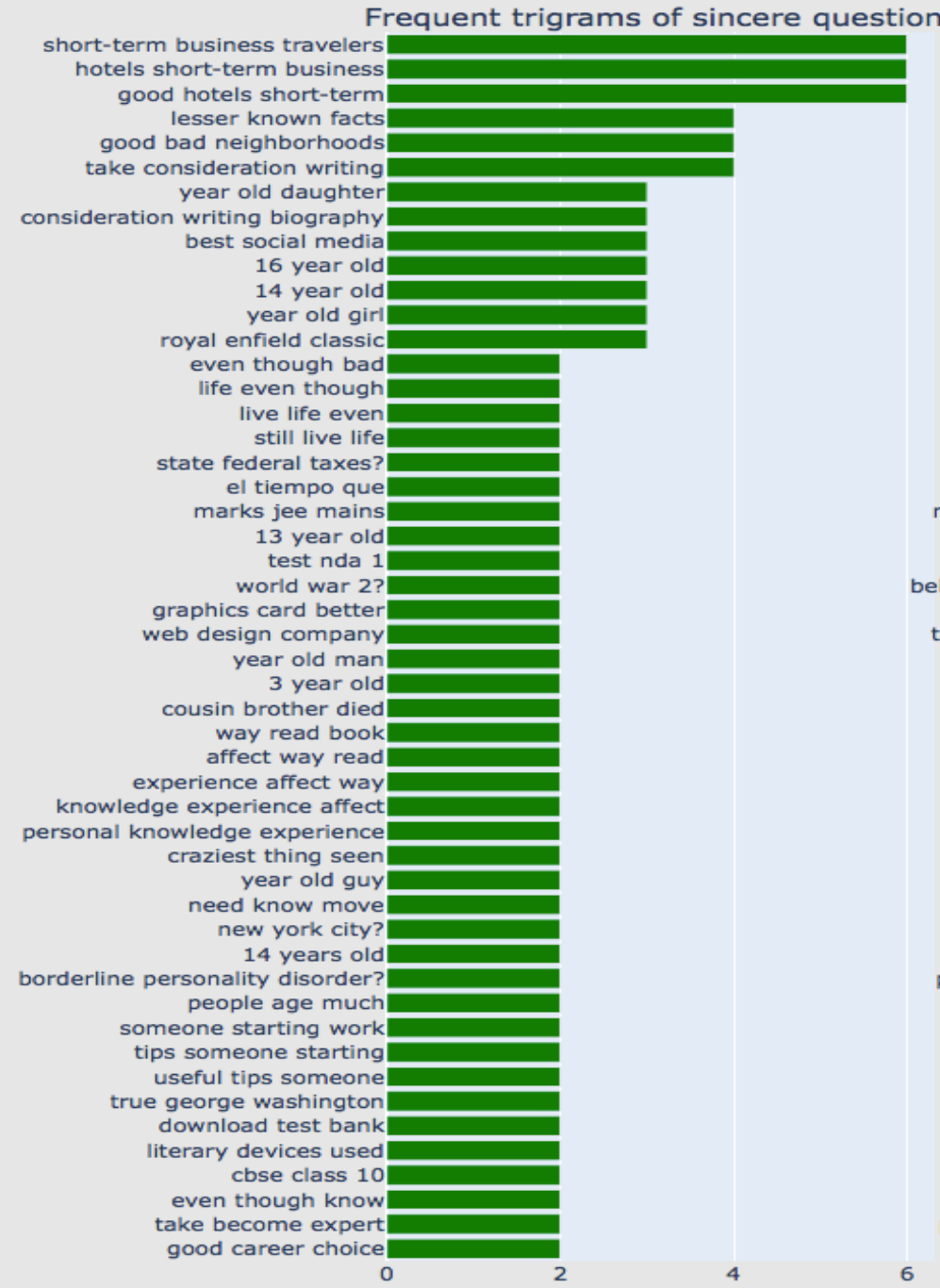


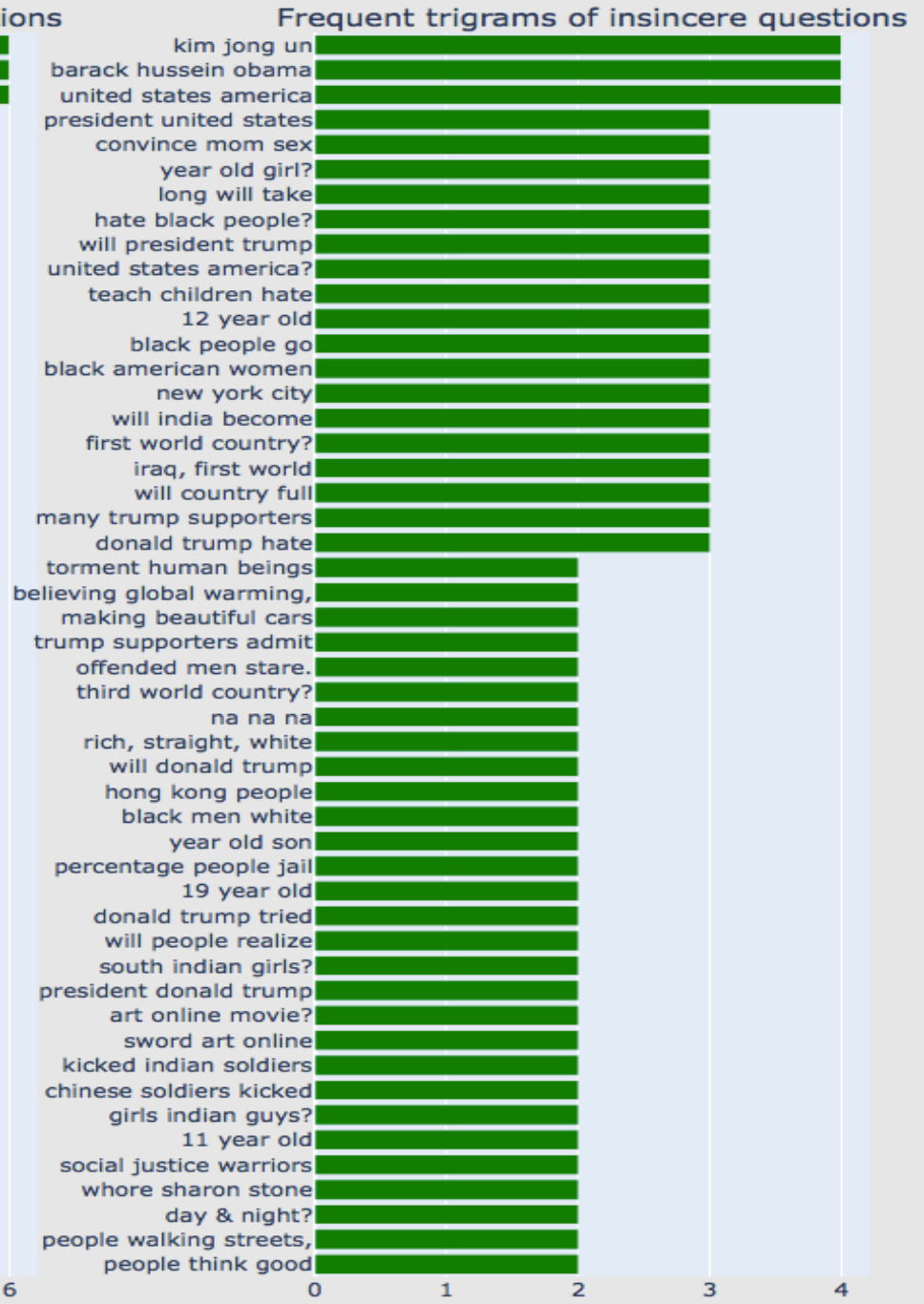


Bigram Count Plots

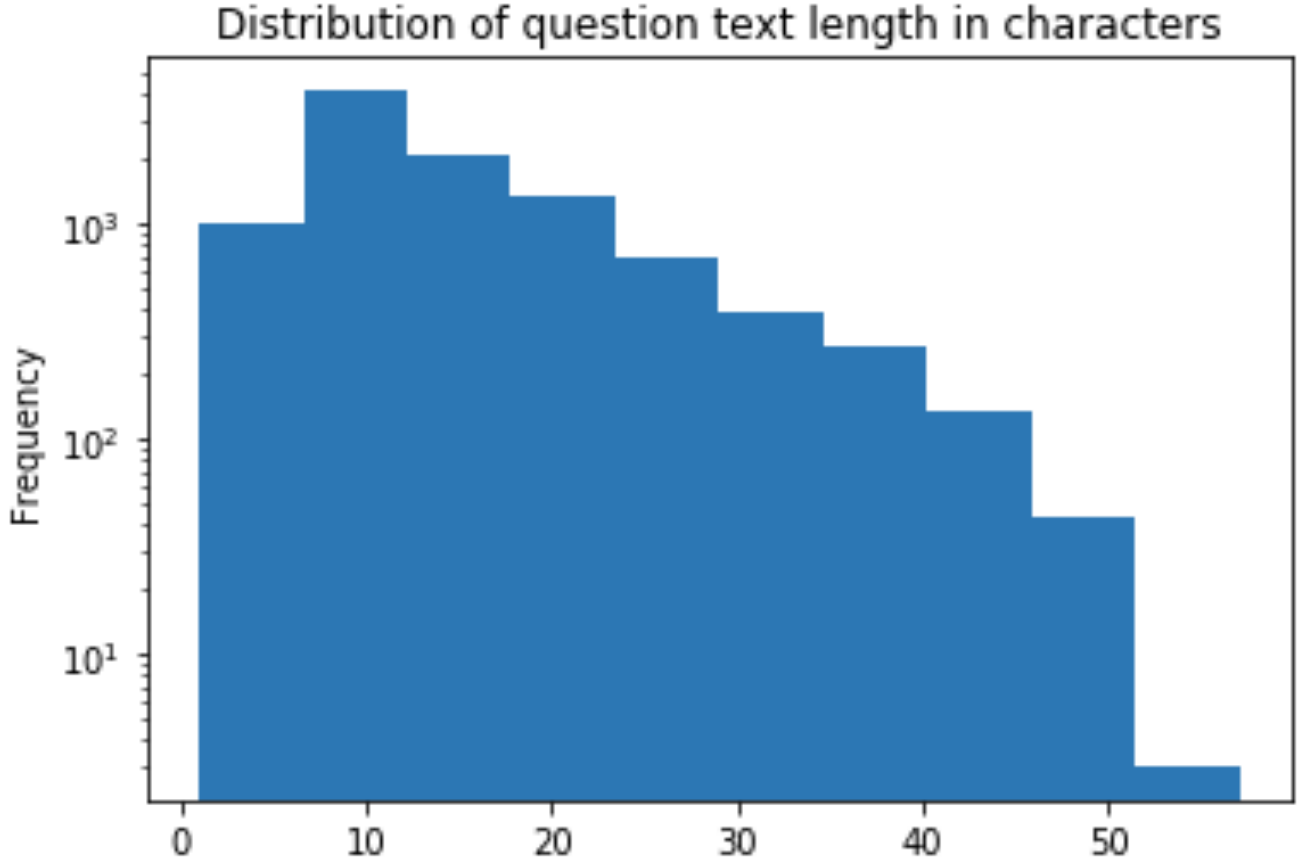


Trigram Count Plots

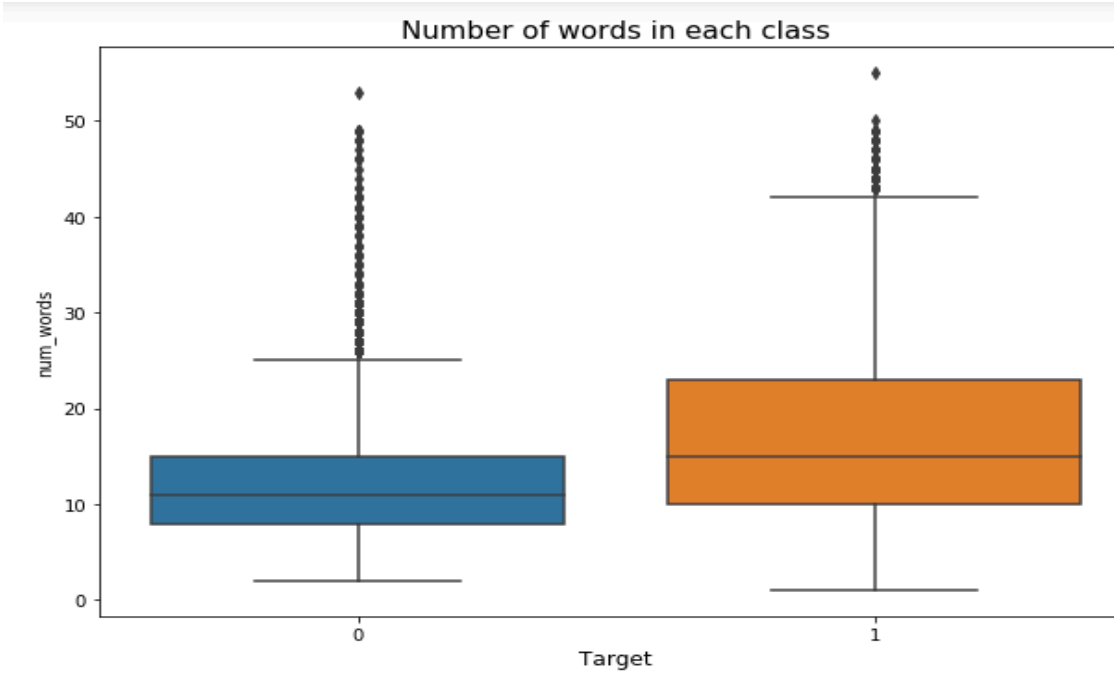


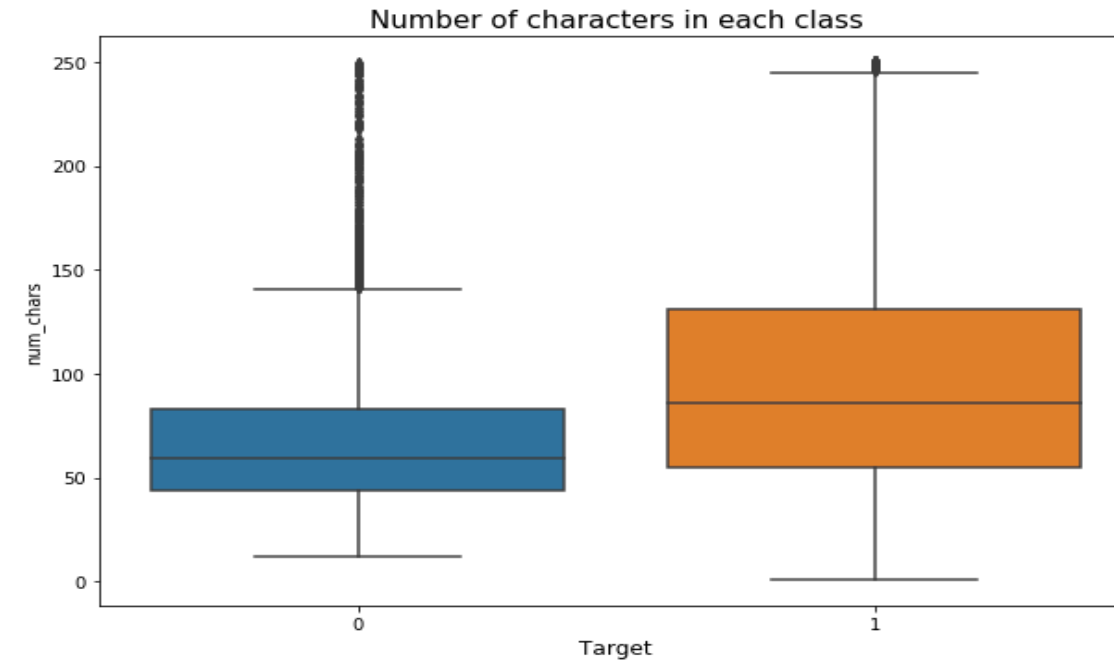


- We also analyze the length of a question. After removing unwanted tokens, stop words & applying lemmatization, most of the questions have about 7 ~ 17 words long.



- We can see that the insincere questions have more number of words as well as characters compared to sincere questions.
- So this might be a useful feature in our model.





MODELING

- With insights based on exploratory data analysis (EDA), we start to train predictive models.
- As for modeling, we identified four Machine Learning models which are

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosting

CONCLUSION

1. In general Logistic regression has the best accuracy for prediction.
2. Training accuracy is always higher than test accuracy, which can be improved with more data.

NEXT STEPS

1. Apply dimmensionality reduction on texts to see if accuracies can be further improved.
2. Apply pre-trained model for embedding.
3. Using more data to improve test accuracies.