Capstone Project 2: Milestone Report 1

Problem: Quora Insincere Questions Classification

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

Potential 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. 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
    - 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

The training data includes the question that was asked, and whether it was identified as insincere (target = 1). The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

Data Wrangling

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

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1306122 entries, 0 to 1306121
Data columns (total 3 columns):
qid              1306122 non-null object
question_text    1306122 non-null object
target           1306122 non-null int64
dtypes: int64(1), object(2)
memory usage: 29.9+ MB
```

Initial Findings

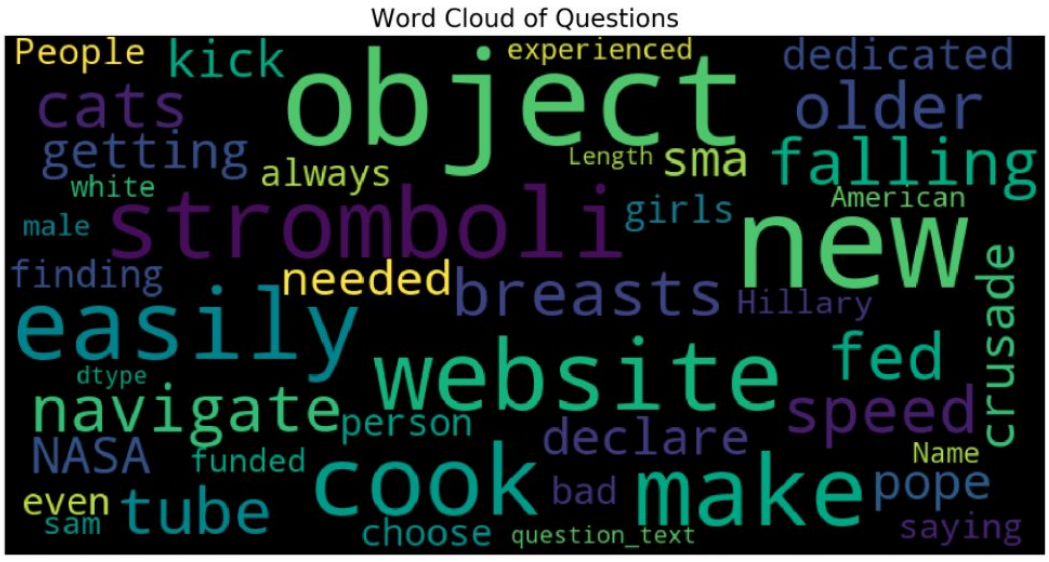
During exploratory data analysis, we ask the following questions to understand more about our data:

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?

and ask follow-up questions if needed.

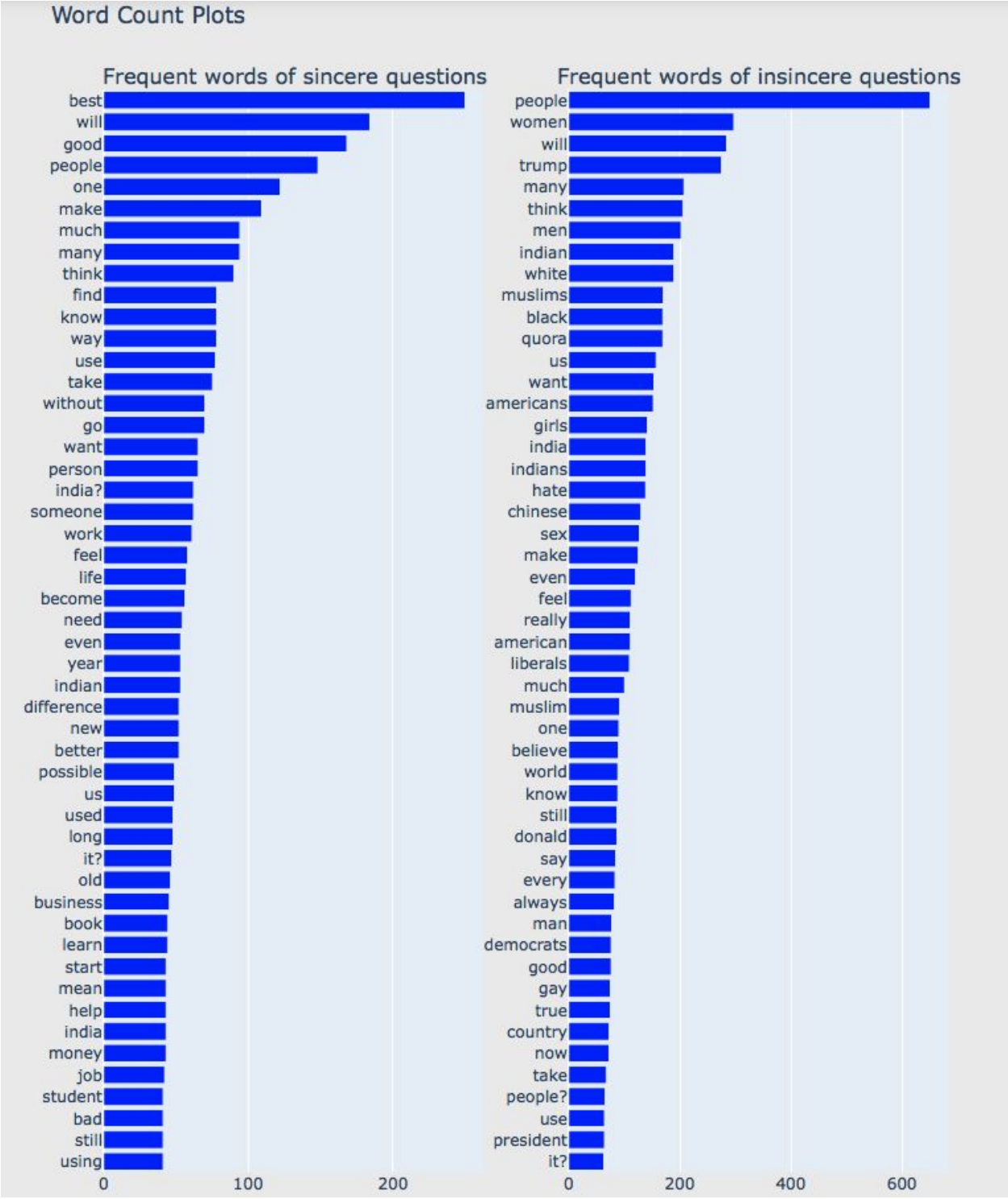
Initial findings are:

1. Then we check which words occur most in questions. The most common two are “object” and “new”.



There seem to be a variety of words in there. Maybe it is a good idea to look at the most frequent words in each of the classes separately.

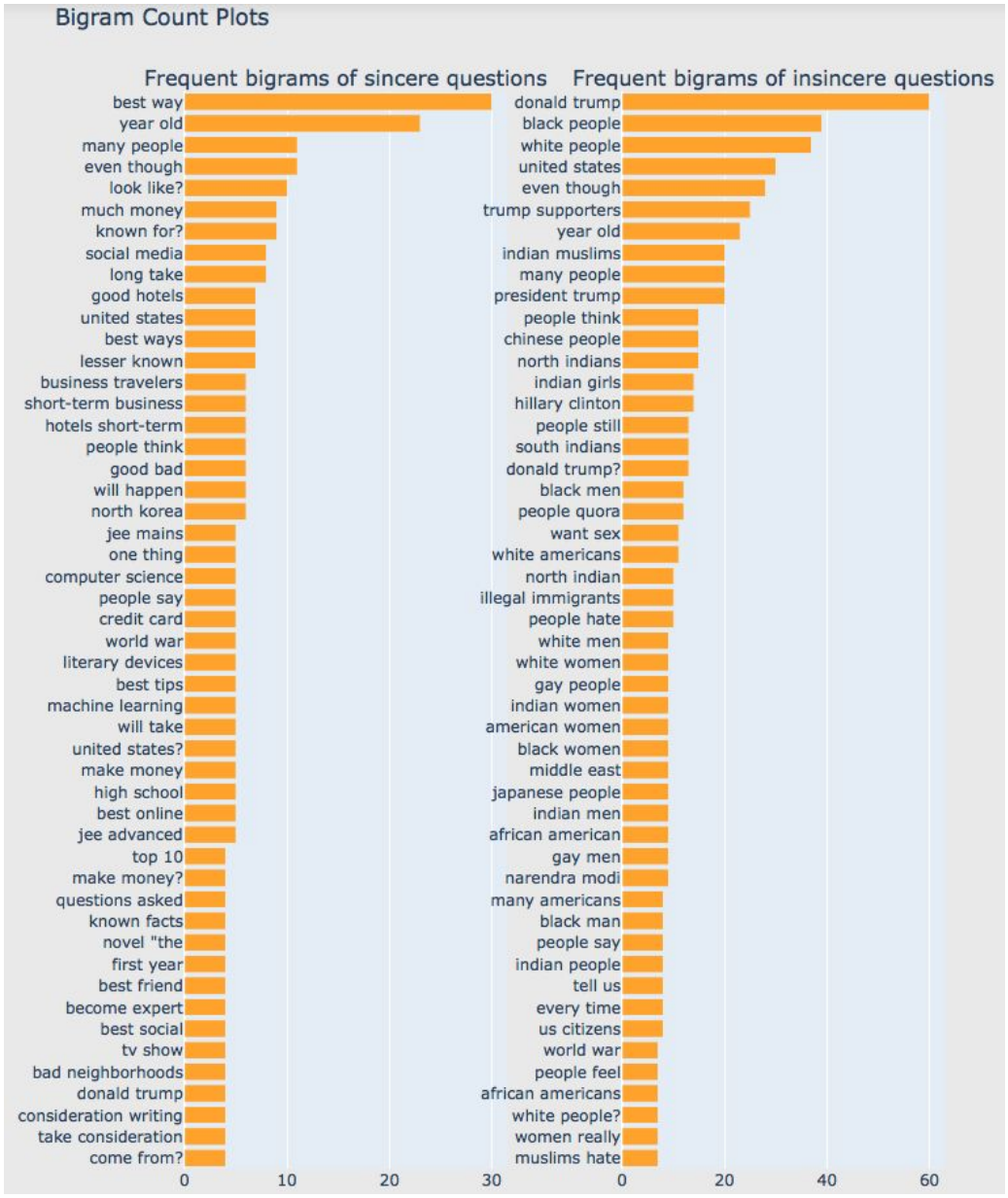
2. Word Frequency



Observations:

- Some of the top words are common across both the classes like 'people', 'will', 'think' etc.
- The other top words in sincere questions after excluding the common ones at the very top are 'best', 'good' etc.
- The other top words in insincere questions after excluding the common ones are 'women', 'trump', 'men' etc.

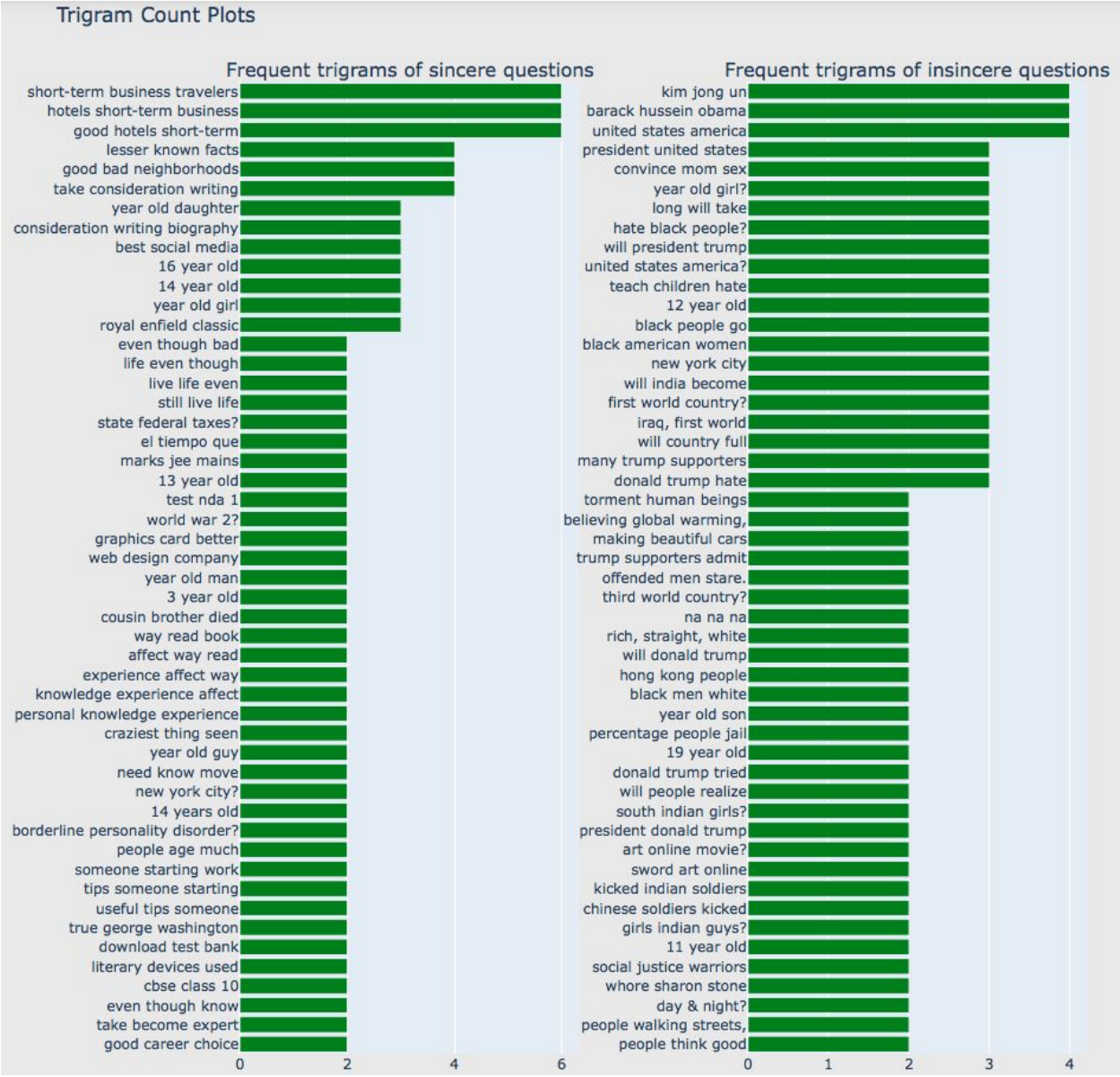
Now let us also create bigram frequency plots for both the classes separately to get more idea.



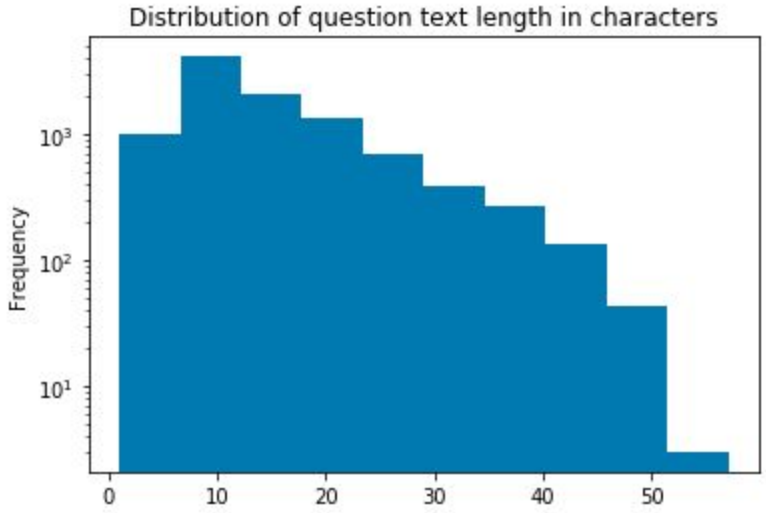
Observations:

- The plot says it all. Please look at the plots and do the inference by yourselves.

Now let us look at the trigram plots as well.



3. 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.



4. 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.