Quora Insincere Questions Classification Project Milestone Report

Objective:

Quora is a knowledge sharing platform for asking questions and connecting with others who contribute unique insights and quality answers. Unfortunately, insincere questions are posted from time to time and these are the questions that are based upon false premises, intend to make a statement rather than look for helpful answers or even try to insult against a specific group of people. Such questions are against Quora's core values and might pose a threat to the Quora community. The goal of this project is to identify insincere questions based on the question wordings so that Quora can develop scalable methods to detect toxic and misleading content.

Research Question: Identify insincere questions based on how the questions are worded.

Data:

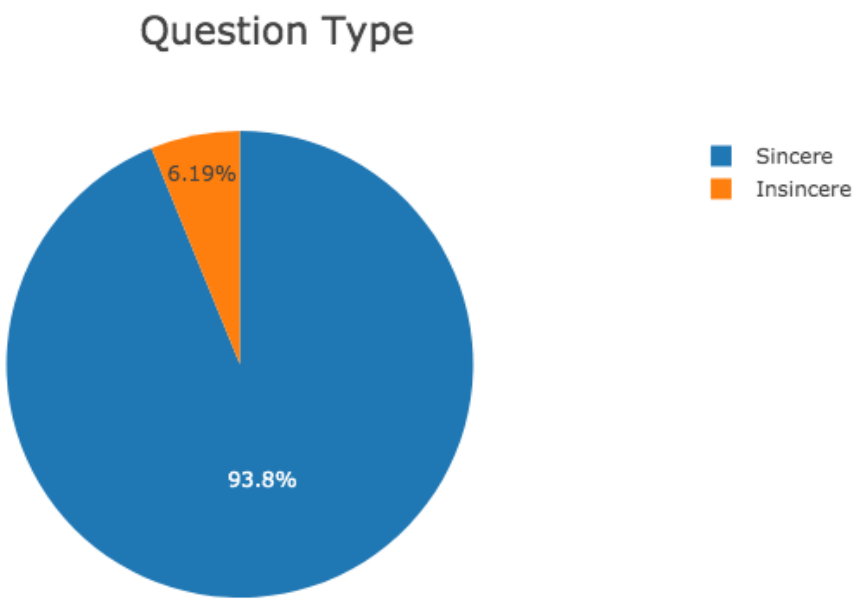
The datasets for this project are available on Kaggle (https://www.kaggle.com/c/quora-insincere-questions-classification/data)
These datasets are:

- train.csv: training data. For each question the unique question identifier, question text and label (whether the question is insincere or not) are provided.
- test.csv: testing data. It contains the same fields as train.csv's except that test_users.csv does not include the label.

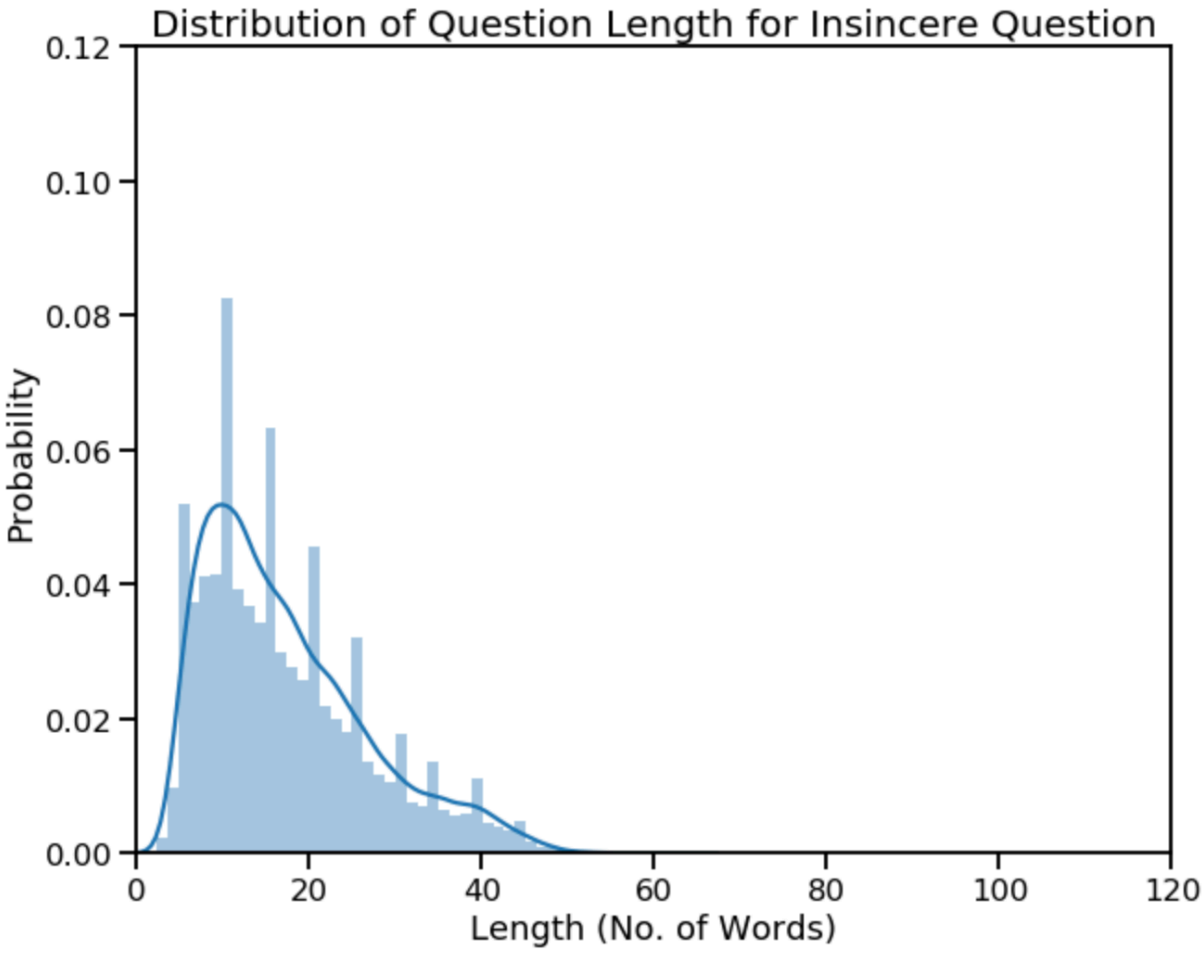
Data Wrangling:

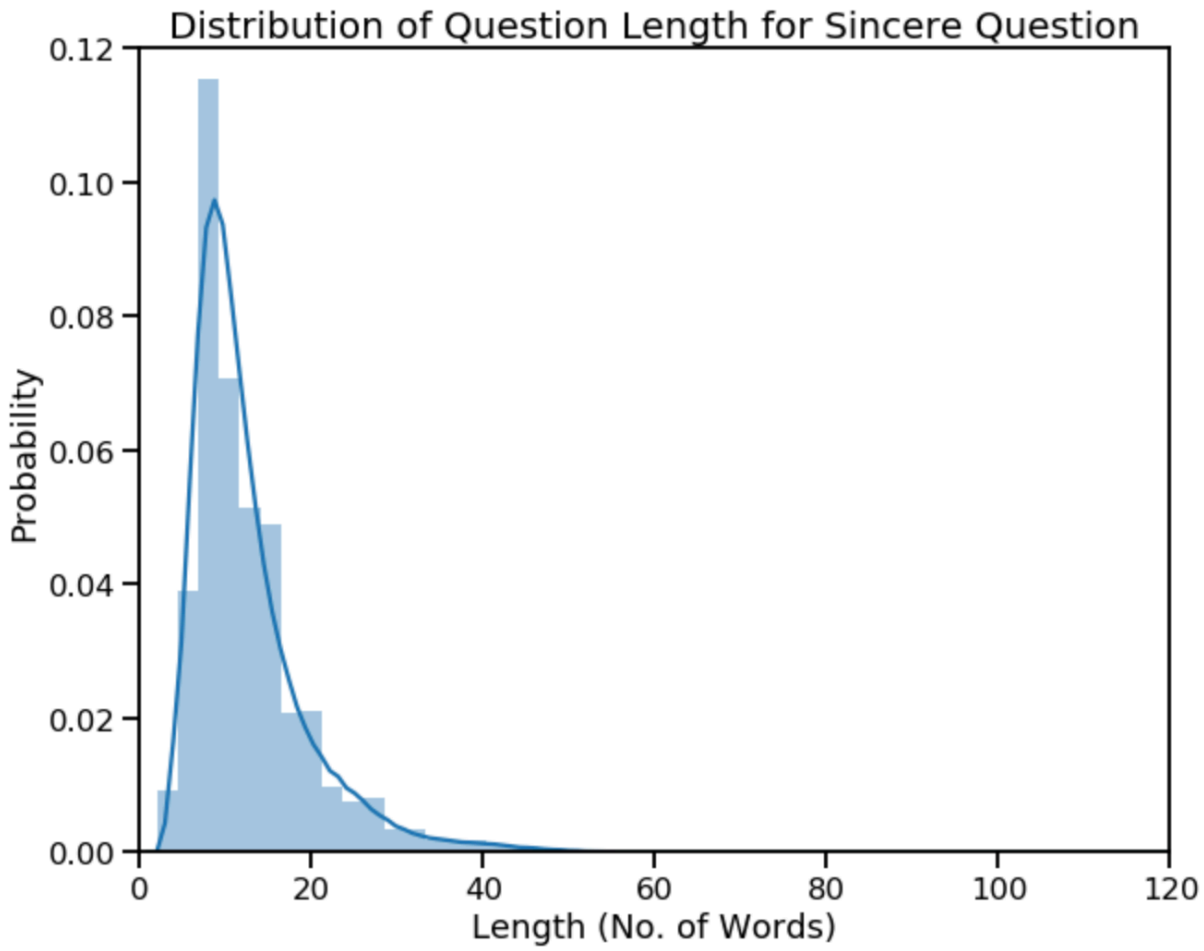
- Change contracted forms (e.g. I'm) to long forms (e.g. I am).
- Remove special characters, punctuations and stop words.
- Tokenize and Lemmatize the words.
- After removing special characters, punctuations and stop words, some questions become NaNs. Replace these NaNs with string 'na' for embedding purpose.

Exploratory Analysis:

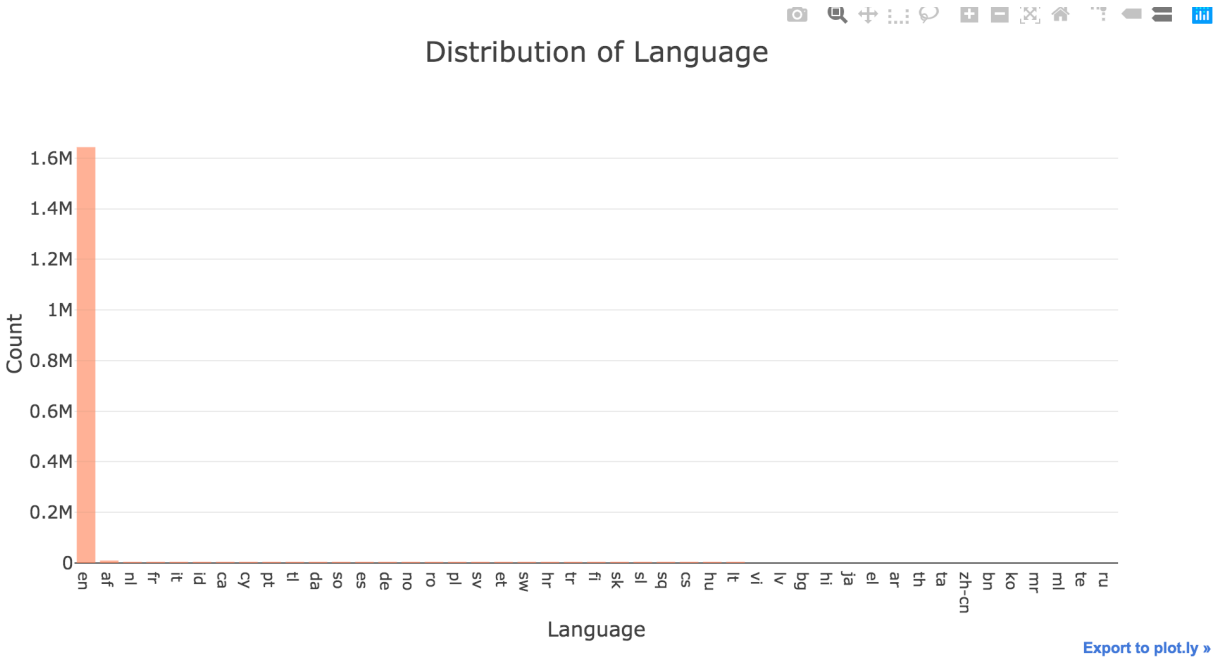


- We have an imbalance dataset with only 6% of questions labeled as ‘insincere’ and 94% as ‘sincere’.


Distribution of Question Length for Insincere Question
Probability
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
120
Length (No. of Words)



- The distributions of question length between insincere and sincere questions are quite different; insincere questions have a lower and broader peak.



- Most questions are in English. Since we only have very few questions in other languages, it is safe to ignore them.

Word-cloud for Insincere Questions

Word-cloud for Sincere Questions

- The word-cloud for insincere questions is mainly made up of words related to nationality/ race and gender. On the other hand, the word-cloud for insincere questions is more diverse and consists of different topics like knowledge, career, relationships, etc.