

Quora Insincere Questions Classification Project

Introduction

Research 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.

Data



The datasets for this project are available on Kaggle. They are:

Training & Testing data:

- For each question the unique question identifier and question text are provided. For training data we also have the label that indicates whether the question is insincere or not.

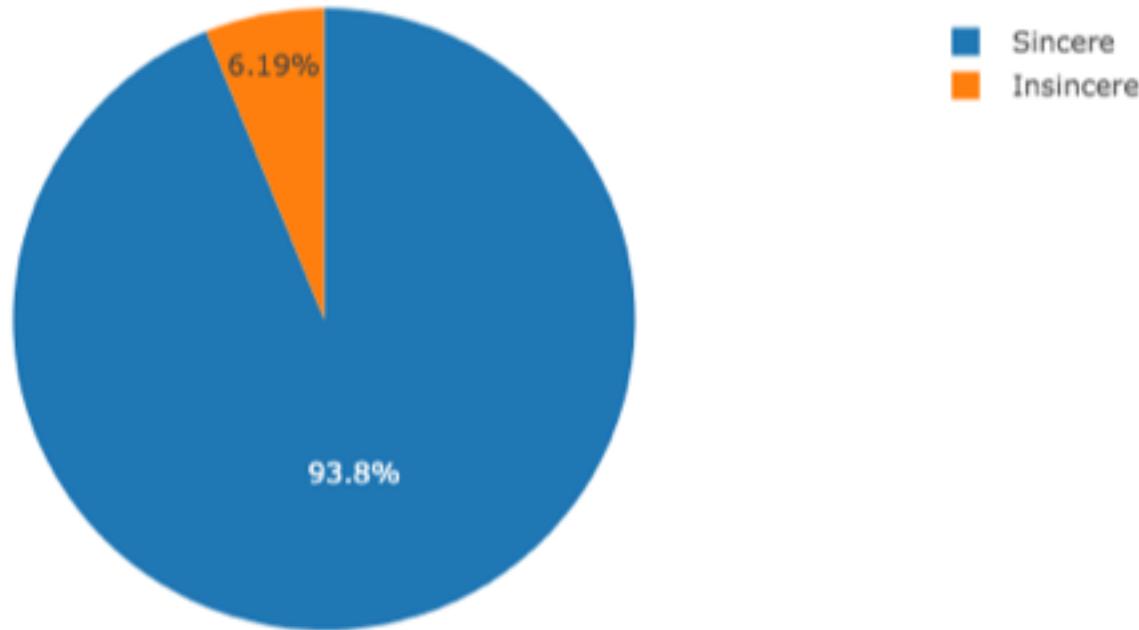
Embeddings:

- The following embeddings are provided for text processing:
 - GoogleNews-vectors-negative300
 - Glove.840B.300d
 - Paragram_300_sl999
 - Wiki-news-300d-1M

Exploratory Analysis

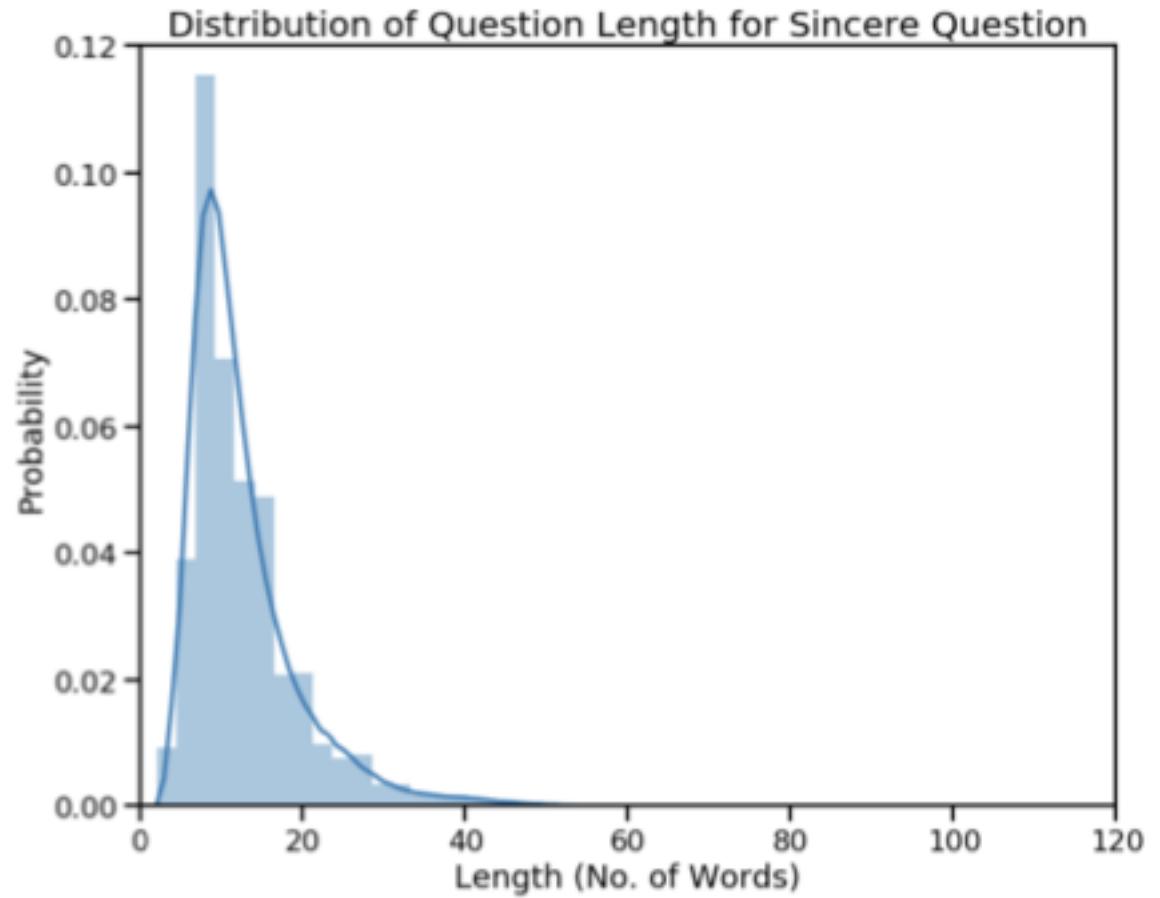
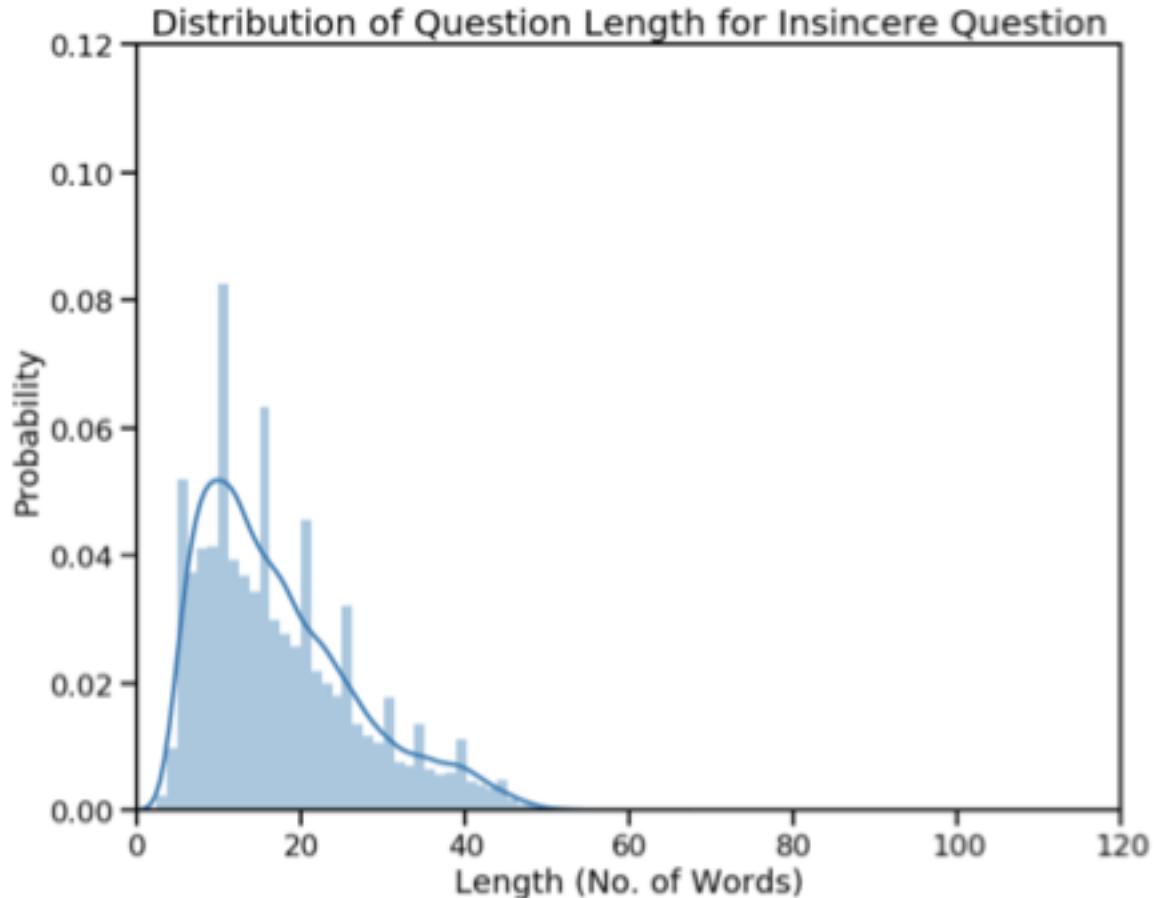
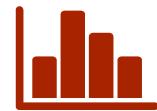


Question Type



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

Exploratory Analysis

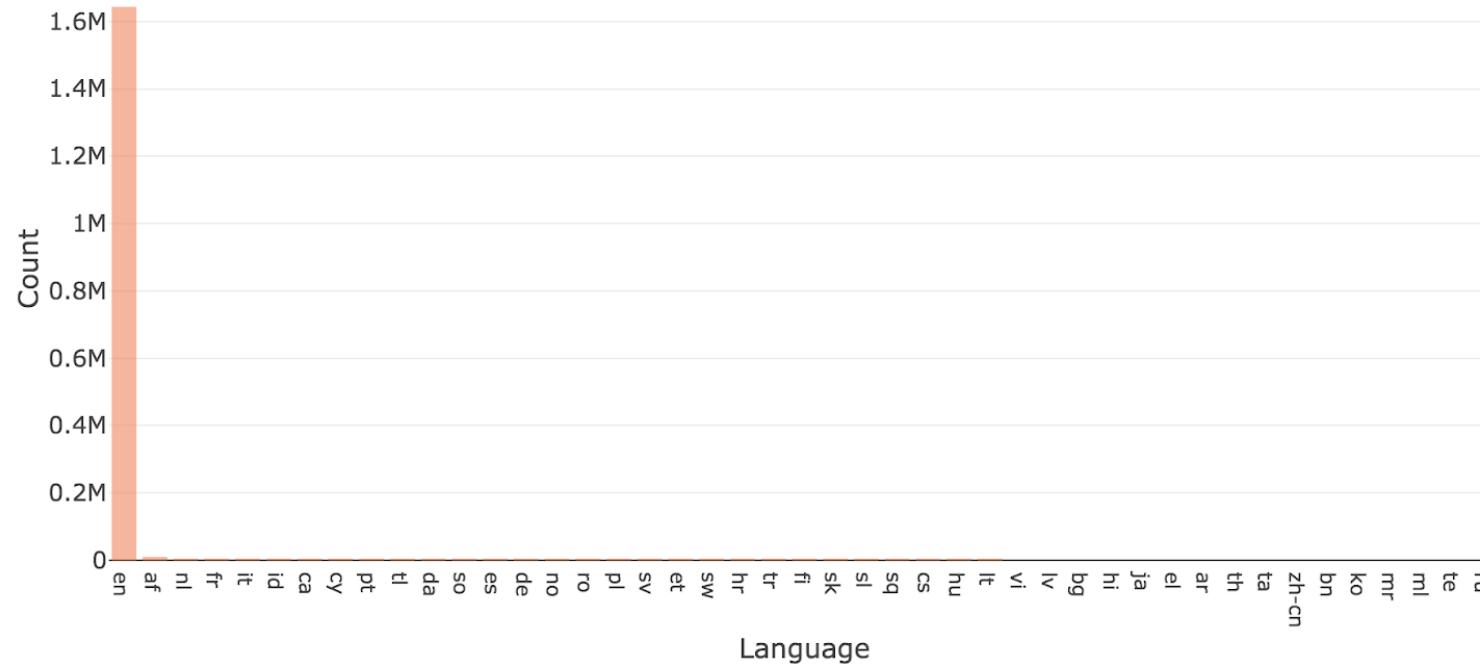


- The distributions of question length between insincere and sincere questions are quite different; insincere questions have a lower and broader peak.

Exploratory Analysis

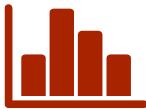


Distribution of Language



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

Exploratory Analysis



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.