Capstone Project 2: Project Proposal
Ahmet Katmer

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.

Approach:

1. Apply machine learning model to detect toxic content to improve online conversations.

Deliverables:

1. Code
- data cleaning
- data exploration analysis & data storytelling
- machine learning model
2. Presentation Slide Deck