# Quora Insincere Questions Classification

## Problem Statement Formation

What data wrangling methods and classification techniques can be applied to the Quora dataset in order to filter out insincere content with an aim to provide a fake news free knowledge based share and learn forum?

## Context

An existential problem for any major website today is how to handle toxic and divisive content. Quora wants to tackle this problem head-on to keep their platform a place where users can feel safe sharing their knowledge with the world.

An effective data oriented solution needs to be devised by developing models that identify and flag insincere questions that keeps a check on misinformation that can spread at a viral peace if not kept in check.

This is a chance to combat online trolls at scale in order to help Quora uphold their policy of “Be Nice, Be Respectful” and continue to be a place for sharing and growing the world’s knowledge

## Criteria for Success

Filtering out insincere questions posted by general users that can be identified or classified as

* Having non neutral tone
* Having disparaging and inflammatory content
* False information or contains absurd assumptions
* Suggesting a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
* Using sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

## Scope of solution space

Quora is a free form forum which invites contributions from everyone on topics ranging from current affairs to latest technology trends to personal view points in the realm of sports, politics etc.

The content is not moderated before it can be published and hence, is available for general view to the public.

Scope of the solution would be to perform data classification over the dataset available for the scope of this problem, which mainly consists of word embeddings from various digital channels and sources like Wikipedia, Google etc as well as questions pulled from Quora postings in the range of ~432K Quora postings (split into training and test datasets)

## Constraints

* The dataset is a representative sample of the questions posted in Quora and is not indicative of the entire population i.e. all Questions posted on Quora **(Sample bias)**
* 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.
* Need to be vary of **Confirmation bias** when building the model since there could be predetermined assumption involved when building the model, so it is important to welcome inputs from multiple Machine learning and Data Science experts when trying to solve this problem.

## Stakeholders

* Quora management team
* Quora Data science and Machine learning experts

## Data Sources

<https://www.kaggle.com/c/quora-insincere-questions-classification/data>

## What is the problem you want to solve?

## As indicated in the Problem statement

## Who is your client and why do they care about this problem?

**Client** – Quora

Considering the problem statement which is to identify and flag insincere questions, till date - Quora has employed both machine learning and manual review to address this problem. With the help of Data Science experts and enthusiasts who can submit their solutions, Quora intends to develop more scalable methods to detect toxic and misleading content.

## What data are you using? How will you acquire the data?

**File descriptions**

* train.csv - the training set
* test.csv - the test set
* sample\_submission.csv - A sample submission in the correct format
* Embeddings

**Data availability** – The data is available in the Kaggle community as indicated in the Data Sources section above

* The data will be acquired from the data source which has the files to be imported.
* Various Data wrangling techniques and text pre-processing will be applied to the csv files that contain textual data for initial pre-processing and cleaning
* Natural language processing techniques will be applied to do word pre-processing on the embeddings

## Method and Solution

* Classification models will be applied to identify the insincere questions to weed them out
* Various ML models and NLP techniques will be applied to improve the accuracy and score

## Deliverables

* A GitHub repo containing the work you complete for each step of the project, including:
  + A slide deck
  + A project report