Quora Insincere Questions Classification

Introduction

Quora is a question answer website, anyone can ask question from any domain, which

are then answered by others, in this project we propose to classify these "insincere" questions, and to remove stop words, numbers, punctuations, common words and converted to lower case. so that appropriate actions can be taken against them.

This preprocessed train data from Quora Insincere Questions competition 2018.

Dataset description

The dataset is collected from a website known as Kaggle, it is a website which has many datasets that are available for public use. The dataset is available at the link: https://www.kaggle.com/rizdelhi/clean-quora-train-data, it contains approximately 1.3 million rows and 4 columns:

- First column "**Unnamed: 0**" it is an index with "**int64**" data type.

- Second column "**qid**" it is unique question ID assigned to each question, with "**object**" data type.

- Third column "**question_text**" it is the actual question with "**object**" data type.

- Fourth column "**target**" target =0 for sincere questions and target=1 for insincere questions, with "**int64**" data type.

Predicate questions

Predict whether a question asked on Quora is sincere or not?

We consider that of the following:

- Has a non-neutral tone

- Is inflammatory or disparaging

- Uses sexual content

Tools

We will use **jupyter notebook**, and some libraries such as **pandas** and **numpy** for visualization and calculation, and **Seaborn** for making statistical graphics in **Python**.

Right now, these tools we've thought about, but as the project progresses, we can use more tools.

Steps to achieve the goal

Exploratory data analysis for analyzing datasets to summarize their main characteristics, using statistical graphics and other data visualization methods, then apply **machine learning model** to predict our target.

Prepared by

Haya Almadhi

Amwaj Hantool