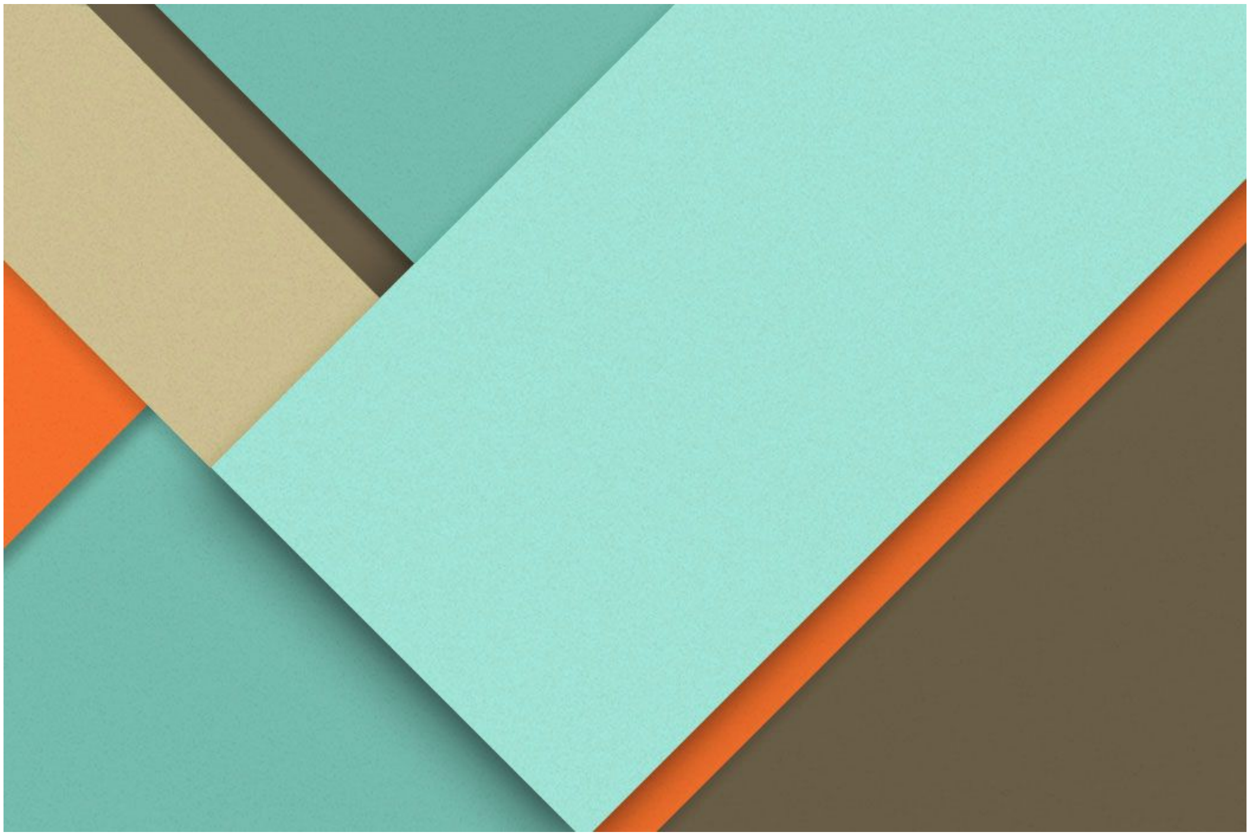

Quora Insincere Questions Classification Kaggle Challenge

11.15.2018

Tara Boyle
Thinkful Final Capstone Proposal

Overview

Quora is a service that helps people learn from each other. On Quora, people ask and answer questions - and a key challenge in providing this type of service is filtering out insincere questions. Quora is attempting to filter out toxic and divisive content to uphold their policy of "Be Nice, Be Respectful".

An insincere questions is defined as a question intended to make a statement rather than look for helpful answers. According to the Kaggle competition characteristics of an insincere questions include:

- Having a non-neutral tone:
    - Having an exaggerated tone to underscore a point about a group of people.
    - Are rhetorical and meant to imply a statement about a group of people.
- Are disparaging or inflammatory:
    - Suggest a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype.
    - Make disparaging attacks/insults against a specific person or group of people.
    - Are based on an outlandish premise about a group of people.
    - Are Disparaging against a characteristic that is not fixable and not measurable.
- Aren't grounded in reality:
    - Based on false information, or contains absurd assumptions.
- Using sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers.

Goals

1. Identify and flag insincere questions using machine learning.
2. Maximize F1 Score by accurately predicting whether a question is sincere or not.

Specialization

As this is a natural language processing problem, the specialization is advanced NLP. However, to develop the best possible model, I also anticipate using deep learning methods with TensorFlow and Keras.

Data Source

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

I will access it at the above link.

Techniques and Models and Packages

Techniques

I expect to try several techniques including:

- Text Pre-processing
    - Removal of punctuations
    - Removal of stop words
    - Stemming
    - Lemmatization
- Feature Engineering
    - Topic Modeling
        - LDA
        - LSA
    - N-Grams
    - Word Embeddings
        - Glove
        - Word2Vec
        - Self-Trained Embeddings
- Dimensionality reduction
    - PCA

Models

I expect to try using several models including:

- Naive Bayes (Baseline)
- Logistic Regression (Baseline)
- Decision Trees
- Random Forest
- Keras Sequential
- Keras Model

Packages

I intend to use the following Python packages:

- Spacy
- NLTK
- Gensim
- Keras
- TensorFlow

Methods

I expect to follow six basic steps in developing my predictive model:

1. Define Problem

   I will investigate and characterize the problem to develop a clear goal.

   In this case the goal is the accurate classification of sincere and insincere questions.

2. Analyze Data

   I will conduct exploratory data analysis to better understand the available data.

   I will use Seaborn to visualize the training dataset, including the distribution of classes.

3. Prepare Data

   I will clean and transform the data to prepare for modeling.

   Specifically, I will clean the text data by utilizing common text cleaning methods including: removing special characters, lowercasing, stemming, lemmatization.

4. Evaluate Algorithms

   I will use the competition's chosen evaluation metric of F1 score to evaluate a number of standard algorithms. I will then select a subset of those algorithms to investigate further.

5. Improve Results

   I will use algorithm tuning, including hyperparameter tuning, and ensemble methods to improve algorithm performance.

6. Present Results

   Lastly, I will finalize my model, make predictions, and present my results.

Expected Difficulties

Large Dataset

The training data has over one million rows. I expect there will be challenges in dealing with the large dataset. Challenges may include running into memory errors and excessive processing times.

To combat the large size of the dataset, there are several techniques to try, including using smaller samples of the data for training and dimensionality reduction. I anticipate feature selection and engineering as well as model optimization will be important.

Imbalanced Dataset

The dataset is highly imbalanced, with only 6% of samples belonging to the target (insincere) class. I anticipate this will cause challenges with recall. Maximizing recall, or true positive rate, could be a difficulty here due to the small number of insincere samples. I anticipate resampling techniques and data augmentation could improve model performance.

Value of Solution

As stated in the competition description, an accurate solution can help Quora develop more scalable methods to detect toxic and misleading content and combat online trolls at scale. This solution will help Quora to uphold their policy of ‘Be Nice, Be Respectful”.