ITCS 6156

Project Proposal

Quora Insincere Questions Classification

Detect toxic content to improve online conversations

Team Members

Member Name	Student ID
Akhil Morampudi	(801138186)
Bharadwaj Aryasomayajula	(801151165)
Mahanth Mukesh Dadisetty	(801034945)

Introduction

Summary

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.

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. A key challenge is to weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

Abstract

Machine learning models have been very successful when dealing with text classification problems. In this project we are tasked with classifying questions from a dataset provided by the popular website Quora, as ‘sincere’ or ‘insincere’. The large-scale dataset, provided by the website Kaggle, contains over 1,300,000 questions with labels to train our models on. A separate test set that contains over 300,000 unlabeled questions is used by Kaggle to test our model. We implement three different models, logistic regression, Naive Bayes, and recurrent neural networks, and use different pre-trained word embedding’s to achieve the best results

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.

Deliverables & Project Flow:-

Data fields

- qid - unique question identifier
- question_text - Quora question text
- target - a question labelled "insincere" has a value of `1`, otherwise `0`

This is a Kernels-only competition. The files in this Data section are downloadable for reference in Stage 1. Stage 2 files will only be available in Kernels and not available for download.

Demonstration of the project:

Methods and Approach

Machine Learning based approach

Data Pre-processing and cleaning: We will be following the standard data pre-processing techniques used for solving NLP based tasks like spelling correction, removing stop words, punctuations and other tags followed by lemmatization.

Pre-processing Data Before building our models, we perform text pre-processing methods to clean up our data set. We remove empty space and delete invisible characters by replacing them with blank strings. We delete stop words, and correct the most common misspelled words,

Lemmatization *is a process of replacing the word with its root word using a known word dictionary.* It converts a word to its root form but with one difference i.e., the root word in this case belongs to a valid word in the language. For example the word caring would map to ‘care’ and not ‘car’ as the in case of stemming.

TF-IDF

It stands for term frequency — inverse document frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

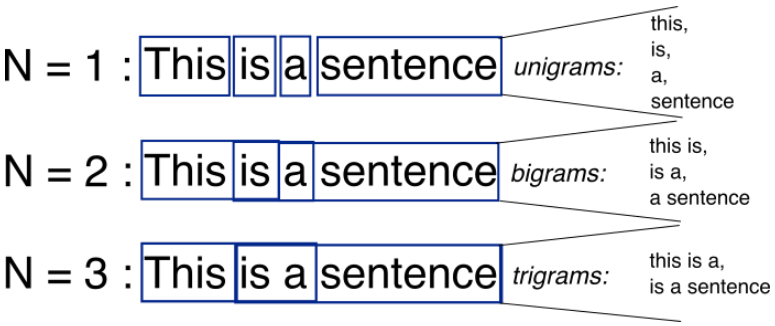
tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Term frequency: probability of finding a word in the document.

Inverse document frequency: defines how unique is the word in the total corpus.

TF-IDF is the multiplication of TF and IDF values. It gives more weightage to words which occurs more in the document and less in the corpus.

N-grams:



N-grams are the combination of multiple words used together, Ngrams with N=1 are called unigrams. Similarly, bigrams (N=2), trigrams (N=3) and so on can also be used.

N-grams can be used when we want to preserve sequence information in the document, like what word is likely to follow the given one. Unigrams don't contain any sequence information because each word is taken individually.

Text Data Vectorization:

The process of converting text into numbers is called text data vectorization. Now after text preprocessing, we need to numerically represent text data i.e., encoding the data in numbers which can be further used by algorithms**.**

Bag of words (BOW):

It is one of the simplest text vectorization techniques. The intuition behind BOW is that two sentences are said to be similar if they contain similar set of words.

Consider these two sentences:

S1: Without music life would be a mistake

S2: Radiohead are a great music band

In NLP tasks, each text sentence is called a document and collection of such documents is referred to as text corpus.

Interpreting the Results — Classification Report

- **Precision** is the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- **Recall** is the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- **F1-Score** is the harmonic mean of precision and recall.
- **Support** is the number of true results for each class.

Project Planning

Tasking:

Member Name	Responsibilities
Mahanth Mukesh (801034945)	1. Project Report 2. Infra Setup 3. Data Cleaning & Pre-processing 4. Modelling 5. Training & Prediction 6. Performance Analysis
Bharadwaj (801151165)	1. Project Report 2. Data Cleaning & Pre-processing 3. Feature Engineering 4. Modelling 5. Performance Analysis 6. Poster
Akhil Morampudi (801138186)	1. Project Report 2. Infra Setup 3. Data Cleaning & Preprocessing 4. Feature Engineering 5. Modelling 6. Training & Prediction

Differences (How our idea is different from others):

- Text pre-processing is necessary since it effectively clean the noise.
- Word Embeddings used in this project outperforms TfidVectorizer when dealing with language in such a sentiment classification problem.
- Models of Neural Network might get a better performance with emotional problem related to Humour

The Novel Expectation:

We will develop models that identify and flag insincere questions. To date, Quora has employed both machine learning and manual review to address this problem. With your help, they can develop more scalable methods to detect toxic and misleading content.

Final Goal.

What questions this project can answer?

- Building a model that can classify various sincere questions based on the dataset and definitions given and achieve a high accuracy by computing the Recall score and report it as metric.
- Advanced Level Approach: Develop attention based mechanism that allows determining the relative weights of words in a text that contribute to the classification decision. We would be visualizing the obtained results and demonstrate.

CONCLUSIONS:

This project focuses on detecting toxic content of Quora questions using machine learning methods. After trying various combinations of models and parameter adjustments, we found that using word embedding's in the period of data processing and using neural networks to fit the data yielded the best performance. Neural network works well on the natural language processing problem

References:

[1] C. Liu, Y. Sheng, Z. Wei, and Y. Yang, "Research of text classification based on improved tf-idf algorithm," in 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Aug 2018, pp. 218–222.

[2] O. Aborisade and M. Anwar, "Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers," in 2018 IEEE International Conference on Information Reuse and Integration (IRI), July 2018, pp. 269–276.

[3] Mungekar, Akshay, et al. "Quora Insincere Questions Classification."

[4] https://towardsdatascience.com/a-gentle-introduction-to-natural-language-processing-e716ed3c0863

[5] https://towardsdatascience.com/quora-insincere-questions-classification-d5a655370c47

[6] Quora Insincere Questions Classification.‖ [Online]. Available: https://kaggle.com/c/quora-insincere-questions-classification. [Accessed: 29-Jul-2019]