BT5153 TOPICS IN BUSINESS ANALYTICS

PROJECT PROPOSAL

QUORA INSINCERE QUESTIONS CLASSIFICATION

By: DATA DOCTORS(Group 4)

HAN, BING (A0186003N)

LI LIPING (A0186040M)

MONICA RAVIPUDI (A0186119Y)

SHUBHANSHU GUPTA (A0185998X)

YAGNA SRIKANTH A (A0176603E)

INTRODUCTION

As a platform that connects "people who ask questions" and those "who contribute unique insights and quality answers", Quora empowers people to share their knowledge with each other. As with any other massive opinion-sharing websites, Quora faces the need to handle toxic, divisive and misleading content, in order to provide its users a sense of security while sharing their own knowledge to the questions posted.

One of the specific challenges regarding this influential problem is to weed out insincere questions. To tackle this problem, Quora has currently employed both manual review and machine learning techniques. However, they are seeking for more scalable methods to identify insincere questions with more efficiency. Their objective is "to combat online trolls at scale, in order to uphold their policy of 'Be Nice, Be Respectful' and continue to be a safe place for knowledge sharing and growing".

Therefore, our group aims to develop suitable models to predict whether a question asked on Quora is sincere or not.

Problem Statement:

A general definition of insincere questions would be those founded upon false premises, or that intend to make a statement instead of looking for helpful answers.

Interesting Aspects:

Some specific characteristics that can signify an insincere question include:

1) 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

2) 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

3) Isn't grounded in reality
    - Based on false information, or contains absurd assumptions

4) Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

DATA DESCRIPTION

We are using datasets provided by Quora as a part of Kaggle Competition. The dataset can be downloaded from Kaggle. Since Quora does not have any official API, the questions asked on Quora can be gathered by a python scraper. Beautiful soup library can be used to scrap all the question titles. The questions can be tagged as insincere and sincere, based on Quora policies as listed here.

Dataset Description:

The training dataset contains 1.31 million Quora questions asked, and their corresponding labels as sincere or insincere. We divided the dataset into training and validation sets in 4:1 ratio, to develop scalable classification models that can apply to other questions asked. We aim to increase accuracy and scalability of our models.

Training Set: 1,048,000 questions

Validation Set: 252,000 questions

Dataset Fields:

Variable Name	Description	Data Type
Qid	Unique question identifier	String
Question_text	Quora question text	String
Target	A question labeled “insincere” has a value of 1, otherwise 0.	Boolean

PROPOSED METHODOLOGY

What we hope to mine from the dataset:

Through Natural Language Processing, we want to identify the word clouds that appear more frequently for insincere questions, such that we can use << a,b,c,d >> models in generating a pattern to predict insincere questions posted by any user in Quora.

The only data field that we use from the dataset is "Question_text". However, this field is important for us to classify a question as Sincere/Insincere, as it contains texts that comprise of words.

Possible textual machine learning techniques:

1. Naive Bayes
2. SVM
3. Random Forest
4. Deep Neural Network

To achieve this, we need a combination of NLP and Neural networks. As text cannot be processed by Tensorflow, we need to first process the text and convert it into a bag of words model. The pre-processing steps involve tokenizing, stemming and forming the list of all stemmed words in the corpus. Using the stemmed set of words, we have to convert each document into a bag-of-words model and then feed it into the Tensorflow Deep Neural Network.

5. Random Multi-Model Deep Learning (RMDL)

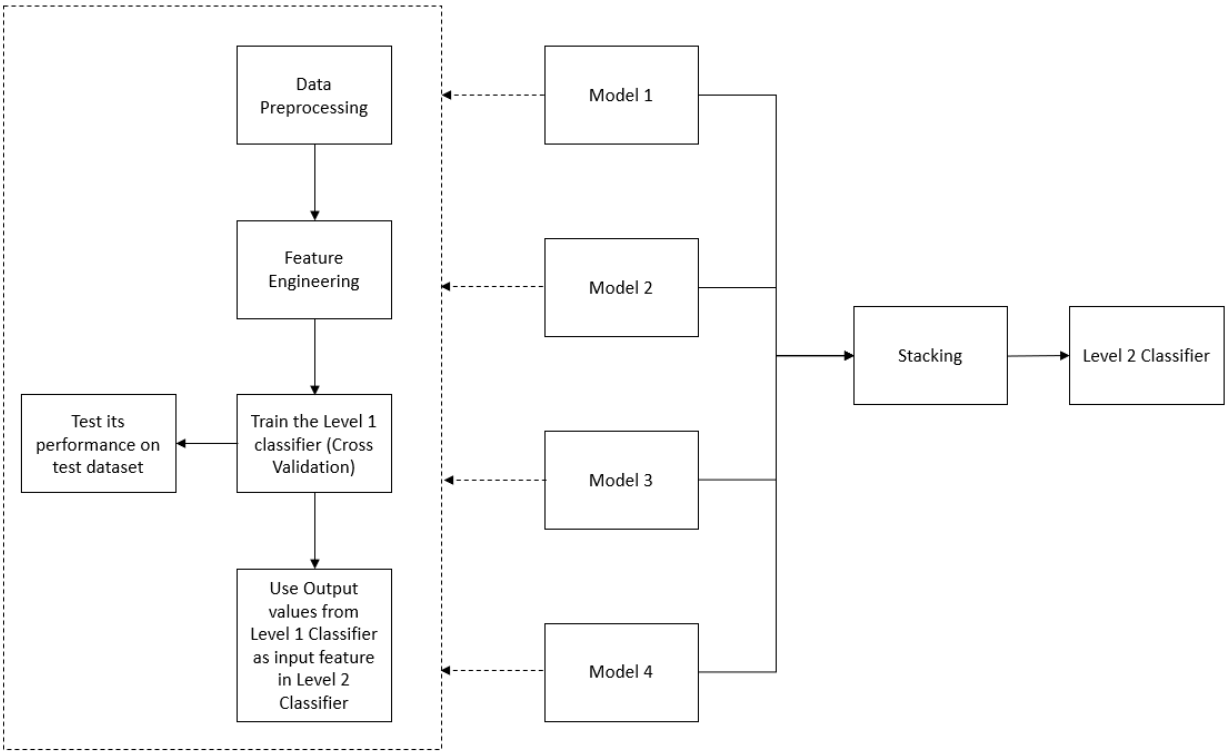
This helps to solve the problem of finding the best deep learning structure and architecture, while simultaneously improving robustness and accuracy through ensembles of deep learning architectures. RDML can accept as input a variety of data, including texts, videos, images, and symbols.

Approach Overview:

1. **Feature Engineering:** Since the dataset consists of question texts and labels, we plan to engineer several features to help us understand the characteristics of a text that construe a question as insincere or sincere. Therefore, we plan to investigate:
    a. Meta features such as number of words, stop words, average length of the words in the text
    b. Topic modeling
    c. Frequent term analysis in sincere and insincere questions
    d. Semantic quality of question texts

e. Apply word embeddings to group words/text that share the common context

2. **Modeling:** Create various Classifier models **(**Naïve Bayes, SVM, Random Forest, CNN, Logistic Regression, Bidirectional Grated Recurring networks (GRU) etc.**)** and determine which models better define the problem.
3. **Stacking:** Stack the selected models and use the stacked model as the final classifier.



REFERENCES

1. Kaggle Competition: https://www.kaggle.com/c/quora-insincere-questions-classification
2. Maeve Duggan. 2014. Online harassment. Pew Research Center.
3. Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE Transactions on Signal Processing45.11 (1997): 2673-2681.