TECHNICAL DOCUMENT

NLP Group B Assignment

IE University Business School
Natural Language Processing

Table of Contents

Business Understanding

The business objective of the project is to *leverage text classification to rid the online forum Quora of toxic content*. This is important because toxic content deters current and new users from frequenting the platform. As we should strive to maintain safety in the world, we should also strive to maintain safety in our platforms.

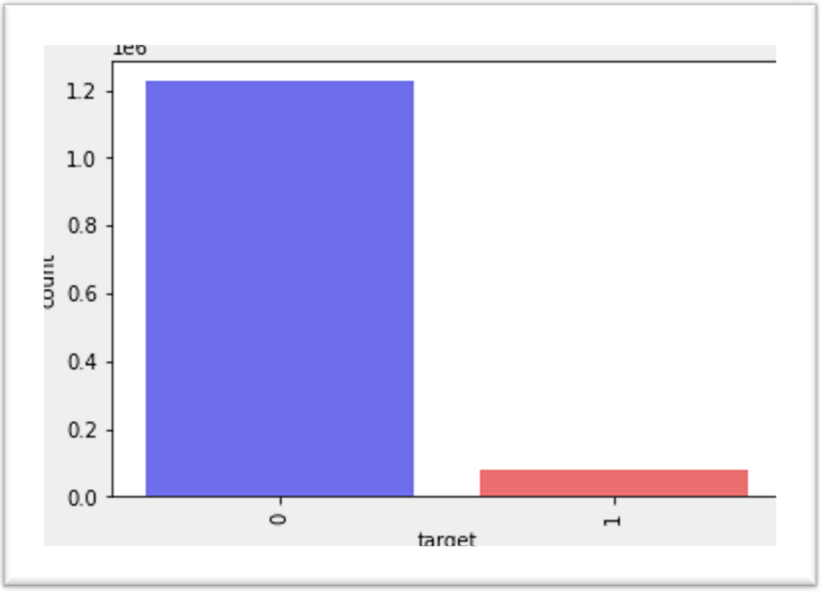
Data Understanding

The dataset is taken from Kaggle: [Quora Insincere Questions Classification](). The purpose of Quora posting this competition on Kaggle is to having users build the best algorithms to detect toxic content to improve online conversations. 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 question is insincere:

- Has a non-neutral tone
- Is disparaging or inflammatory
- Isn't grounded in reality

Basic Exploratory Data Analysis

Whereas the target class has around 60k records, the non-target has 1.2 million records. This is a clear example of a class imbalance classification.



Data Preparation

The training data includes the questions asked and an identified indicating sincerity (target = 1). Given that we need to create a Text Classifier, we want to preprocess the dataset with the ktrain library. Ultimately, we process the input questions based on the BERT encoder.

Create the Transformer Model

First, we will preprocess the dataset with ktrain library, create a Text Classifier based on a Transformer model.

Then we start with a Language Model pre-trained using a massive dataset. The model is BERT where it applies the idea of Self-Attention (instead of an RNN architecture) to learn the sequential information of textual contents.

It is needed to fine-tune this model, so we will re-train the model to our specific dataset.

```
trn, val, preproc = text.texts_from_df(df,'question_text',preprocess_mode='bert',label_columns='target',verbose=True, maxlen=32) # Process the input questions based

model = text.text_classifier('bert', trn, preproc=preproc) # Create a text classifier that uses the BERT-based representations created before

learner = ktrain.get_learner(model, train_data=trn, val_data=val,  batch_size=128) # Creates the learning process to fine-tune bert and train the classifier.
```

```
['not_target', 'target']
         not_target  target
912287          1.0     0.0
1121384         1.0     0.0
374662          1.0     0.0
512106          1.0     0.0
506620          1.0     0.0
['not_target', 'target']
         not_target  target
408923          1.0     0.0
851978          1.0     0.0
210992          1.0     0.0
201699          1.0     0.0
379167          1.0     0.0
```

Modelling

We created a text classifier that uses the **Bert**-based (Bidirectional Encoder Representations from Transformers) representations created before. Being this a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. **BERT** makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.

```
learner.validate(class_names=['Normal Question','Abnormal Question'])
```

```
                  precision    recall  f1-score   support

   Normal Question       0.98      0.98      0.98    122386
 Abnormal Question       0.72      0.68      0.70      8227

          accuracy                           0.96    130613
         macro avg       0.85      0.83      0.84    130613
      weighted avg       0.96      0.96      0.96    130613

array([[120240,   2146],
       [  2607,   5620]])
```

Evaluation

To evaluate the model, we decided to understand the performance of the model using the accuracy and F-measure. Fortunately, our model does a pretty good job identifying insincere messages from sincere messages.

Here is an example on how our model performs when provided with two different statements:

- Example – 1: *"Can women vote?"*
    - In this case our model did not detect this question as insincere, since for instance, still there are some countries where women can still not vote, and this might be under some specific context.
- Example – 2: *"Can we prevent women from voting?"*
    - In this case our model identified this question as insincere, as it has non-neutral tone and might be offensive

Deployment

Group B decided to leverage Streamlit.py to deploy our model. As a result, anyone with the code can install streamlit and launch in their local host! Please feel free to input questions to learn if your input is insincere or sincere.