Machine Learning Capstone Proposal

Title: Quora Insincere Questions Classification

Jeevan Anand Anne
December 9th, 2018

Proposal for the Capstone project

I decided to work on a current ongoing competition on Kaggle. It is related to natural language processing(NLP), which is very close to my current job and is the best candidate to explore more on NLP techniques.

Domain Background

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. This is very important for very good user experience.

My main motivation is to explore advanced nlp techniques while working on this project.

Problem Statement

Predicting whether a question asked on Quora is sincere or not. It is a supervised classification problem.

Datasets and Inputs

Dataset: https://www.kaggle.com/c/quora-insincere-questions-classification/data. It shared by "Quora, Inc., http://www.quora.com" in Kaggle via competition.

Also, they have shared training data with labels, test data and pre-trained embeddings.

Dataset Information: Training data has 3 attributes and about 1.3M obdersvations. Test data has 2 attributes and about 56.4K observations.

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

Solution Statement

This is supervised classification problem. And can be solved using simple to advanced machine learning algorithms: 1) Logistic Regression 2) Random Forests 3) Xgboost 4) Catboost 5) LightGBM etc.

Above mentioned algorithms will be choosed based on empirical evaluation. I will try to start modeling from basic techniques to advanced techniques.

Befor start applying machine learning algorithms, there are some preprocessing techniques to be applied to reduce the vocabulary. Like: a) With/without Stopword removal b) Contractions (don't - do not etc.) c) case folding (all chars to lowercase) d) Stemming and Lemmatization (to reduce the words to its roots)

Next, will be Creating tf-idf matrix to create features from text, then create a classification model, evaluate model on the validation set (splitting the training data to train and validation sets). Will be considering this as base model. Also will be trying to create word embeddings from the text using pre-trained embedding models and use them as features in the machine learning algorithms. Also, the class distribution is highly imbalanced, will try to take care of this via different techniques (oversampling and sample weighting etc.).

Benchmark Model

The given dataset is a typical supervised learning text classification problem for which simple model logistic regression or Naive Bayes will perform much better. So we will pick these machine learning models as benchmark and try to beat the benchmark with feature engineeing (features created from the text). We will also be trying ensemble methods if that doesnot improve the score.

Evaluation Metrics

Few common evaluation metrics are: a) Precision b) Recall c) Also, looking at False Positives and False Negatives might help d) Most importantly F1 score, as we have high class imbalance

Project Design

Will follow below workflow as part of my analysis and building models: **a) Exploratory Data Analysis:** I will try to visualize the distributions of the text (for example: length of the text etc), class distributions etc. **b) Preprocessing Techniques:** Stopword removal, Contractions, Casefolding, Stemming and Lemmatization etc. **c) Feature Engineering:** TF-IDF, Word Embeddings using pre-trained embedding models etc. Also, other features like sentiment score etc. Other dimensionality reduction techniques (PCA etc.). **d) Build Models:**

Above mentioned supervised classification algorithms **e) Model Evaluation:** Evaluating different models based on the evaluation criteria (F1 score for example) **e) Model Tuning:** To improve the accuracy **f) Testing:** To predict using trained model on unseen data and evaluate

References

1) https://www.quora.com/What-is-an-insincere-question
2) Data has been shared in Kaggle for a competition by Quora Inc., http://www.quora.com