Machine Learning Capstone Proposal

Quora Insincere Questions Classification Project

Fabian Reyes

December 2018

Domain Background

The digital transformation has completely changed our way of communicating between ourselves. Today most of our social relationships, work, and reading is done online. Although, online communication has been a great place for personal expression and insightful discussions, it has also been a place for toxic and divisive content. The threat of abuse and online harassment, which derives from this toxic content, has been a growing concern this past few years [1]. People who suffer from this divisive content usually stop expressing themselves and give up on seeking different opinions. On the long run, this can become an existential problem in freedom of speech.

Many well-known social and news media platforms have tackled this issue with online moderators, however, the immense amount of content that every day is generated makes this task extremely hard and most of the time futile. Hence, many platforms have turn to machine learning to identify, evaluate and support online moderators on this critical task. This has become a reality for big media platforms such as The Economist [2], Wikipedia [3], New York Times [4], and The Guardian [5]. All of them have partnered with Google and Jigsaw to experiment with their new Perspective API that helps them host better online conversations [6].

I personally feel engaged with this issue and related ones like fake news. I also find very exciting and challenging the possibility to teach a machine to understand language and its intricacies to help, solve and support human work. This task is part of a subfield called Natural Language Processing (NLP), which is concerned with the interactions between computers and human languages, in particular how to program computers to process and analyze large amounts of natural language data [7]. Recently I have started a project which involves NLP and I taking the chance to learn more about it through this project.

Problem Statement

Quora is a platform that empowers people to learn from each other, thus it's been a frequent target for abusive content. Quora wants to keep their platform as a place where users can feel safe sharing their knowledge with the world, therefore it wants to face this problem by hosting a Kaggle competition aimed at identifying insincere questions [8]. Quora wants kagglers to develop scalable methods and models that identify and flag insincere questions. By definition an insincere question is a question which is founded upon false premises, or that intend to make a statement rather than look for helpful answers.

Datasets and Inputs

The dataset considered for this project can be found and downloaded from the competition site [9]. The training data has 3 columns:

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

The criteria's to define an insincere question was based on the following characteristics:

- Has a non-neutral tone: exaggerated tone or rhetorical tone meant to make a statement about a group of people.

- Is disparaging or inflammatory: making discriminatory and/or harsh comments, or based on outlandish premise about a group of people.
- Isn't grounded in reality: based of false information or absurd assumptions.
- Uses sexual content for shock value, and not to seek genuine answers

As expected from this type of labeling, the ground-truth are not guaranteed to be perfect but suficient for the problem.

There competition also provides an unlabel test set, which is used to make predictions and check your performance in the leaderboard.

Solution Statement

I will test different supervised learning models (e.g Logistic Regression, Random Forest, Gradient Boosting) supported by a rich feature extraction process (e.g PCA, kmeans, LSA, LDA) and efficient text cleaning process. My aim is to develop a scalable data pipeline and a robust yet interpretable model that helps Quora not only identify insincere question, but to give insights on feature importance and context relevance.

Benchmark Model

I will construct two benchmark model:

- Floor model: This benchmark is my floor reference, it has no intelligence behind it. I will predict for all test cases the predominant label. All my expected results must be always above this score.
- Basic model: This second benchmark is my simple yet intelligent model, which I will try to beat with more complex models or hypertuned models.

There is also a leaderboard that orders each of the competition participants by their public test set performance [10]. This will also be consulted for comparison.

Evaluation Metrics

The main performance measure, as evaluated in the competition, is the F1-score, which is especially suited for this type of task. This score considers both precision and recall by performing a harmonic average from these two metrics. The F1-score reaches its best value at 1 and worst at cero. As a comparison, the best F1-socre up to today in the leaderboard is 0.711.

Project Design

I anticipate that this project can be divided into 3 parts:

1. Text cleaning process. The main objective of this part is to produce a clean and tidy document-term matrix (DTM) object (bag of words model), which is cleaned from stop words and irrelevant punctuation marks. I will also explore TF-IDF (term frequency - inverse document frequency), lemmatization or stemming procedures to improve downstream results. I will not only consider unigrams for the DTM construction, but also bigrams and trigrams. At the end of this task I will split training data into two sets, 70% training and 30% validation set. Only the training set will be used for the downstream tasks.
2. Feature Extraction. To support further modelling I will extract basic descriptive features from questions (number of character, number of words, presence of certain punctuation mark or particular words), and overall structure using PCA. Two interesting non-supervised learning algorithms are Latent Semantic Analysis (LSA) and Latent Dirichlett Allocation (LDA), both of them yield topics constructed from the DTM. This could aid the classification task.

3. Modelling. Beforehand i do not know which learning algorithms is best, thus I will test some of the following:
    - Logistic Regression (base learner)
    - Gaussian Naive Bayes
    - Classification trees such as Random Forest
    - Gradient Boosting (particularly the XGboost or LightGBM)

I will support my modelling with learning curves and cross-validation (5-fold) methodology.

References

[1] Nobata, Chikashi & Tetreault, Joel & Thomas, Achint & Mehdad, Yashar & Chang, Yi. (2016). Abusive Language Detection in Online User Content. 145-153. 10.1145/2872427.2883062.

[2] https://medium.economist.com/help-us-shape-the-future-of-comments-on-economist-com-fa86eeafb0ce. Help us shape the future of comments on economist.com

[3] https://meta.wikimedia.org/wiki/Research:Detox. Research:Detox.

[4] https://www.nytco.com/press/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/. The Times is Partnering with Jigsaw to Expand Comment Capabilities

[5] https://www.theguardian.com/technology/series/the-web-we-want. The web we want.

[6] https://www.perspectiveapi.com/#/. Perspective API.

[7] https://en.wikipedia.org/wiki/Natural_language_processing. Natural Language Processing.

[8] https://www.kaggle.com/c/quora-insincere-questions-classification. Quora Insincere Questions Classification.

[9] https://www.kaggle.com/c/quora-insincere-questions-classification/data. General Description of Data.

[10] https://www.kaggle.com/c/quora-insincere-questions-classification/leaderboard. Leaderboard.