Project Proposal

MSc program: Data and Web Science, 2020 - 2021
Course: Text Mining and Natural Language Processing
Team Members: George Georgiou, Panagiotis Papaemmanouil, Theodoros Konstantinidis

The problem addressed and Motivation

For this Course Project, we choose to tackle a famous Kaggle Competition called **"Quora Insincere Questions Classification"**, also known as "The Quora challenge II". This competition was held 2 years ago and more than 4.000 teams participated in it.

Quora (www.quora.com) is an American question-and-answer website where questions are asked, answered, followed, and edited by Internet users, either factually or in the form of opinions. Its owner, Quora Inc., is based in Mountain View, California, United States. The company was founded in June 2009, and the website was made available to the public on June 21, 2010. In 2020, the website was visited by 590 million unique people a month *(Source: Wikipedia)*

A key challenge for Quora is to ensure that they create an inclusive and respectful community for their users. This means that any form of toxic or troll content has no place on the platform. To make sure that this will be the case, Quora has to take certain actions to block any kind of "bad" content such as racism, sexual, troll, or toxic information from being posted online. These actions include careful 24/7 surveillance from their employees, of all the traffic of their platform, like questions, answers, and comments posted.

In this project, we will develop machine learning models that will automatically identify and flag insincere questions and toxic content posted on the quora website.

Data Involved

The training dataset was provided by Quora and contains 1.306.122 questions in the English Language, along with labels denoting if the question was "insincere" or not. The labeling process has been done by human experts.

Dataset structure

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

There also exists a test dataset with no target, which will be used for the final evaluation of our solution.

Modeling steps and Methods

The first step to work on this problem will be *Feature engineering*. We will employ various NLP techniques to create Features from textual data, like embeddings and classic text mining features (e.g. TF-IDF).

Some available pre-trained embeddings that we will explore, are the following.

- **GoogleNews-vectors-negative300** - https://code.google.com/archive/p/word2vec/
- **glove.840B.300d** - https://nlp.stanford.edu/projects/glove/
- **paragram_300_sl999** - https://cogcomp.org/page/resource_view/106
- **wiki-news-300d-1M** - https://fasttext.cc/docs/en/english-vectors.html

We will also try to utilize the ***Hugging Face NLP framework*** (https://huggingface.co/), which contains many pre-trained NLP models for various tasks. After that, we will experiment with various binary classification Machine Learning algorithms, using the ***scikit-learn ML Framework***.

To evaluate our implementation, we will make use of the **F1-score** between the predicted and the observed targets in the test set.

Competition Link
https://www.kaggle.com/c/quora-insincere-questions-classification/overview