ML1010

Project Proposal

The Fake News Challenge

7th April 2019

Mark Lewis

Salman Amin

PROJECT PROPOSAL

Introduction

Fake news has received a lot of media attention lately. It has been blamed as the key weapon used by Russia to influence the 2016 U.S. elections and turned the attention of regulators towards social media platforms such as Facebook and Twitter. The New York Times has formally defined it as a "made-up story with an intention to deceive."

Problem Definition

We take our inspiration from the Fake News Challenge to apply machine learning techniques to classify news articles with the objective of developing a semi-automated process to determine the authenticity of news. Specifically, we will use Natural Language Processing to classify a news article into one of four categories. The Fake News Challenge website poses this as a Stance Detection problem, which "involves estimating the relative perspective (or stance) of two pieces of text relative to a topic, claim or issue." In this project, we will explore if the content of an article agrees, disagrees, discusses or is unrelated to the headline.

We chose this problem instead of classifying a news article as either real or fake due to several limitations as explained by fakenews.org. First, "truth labeling" news is extremely difficult due to its inherent complexity. Second, labelled data is scarce, copyrighted, diverse, unstructured, and may be contested as being unbiased.

The Input and Output of our model would be:

Input: a headline and a body text – either from the same news article or from two different articles

Output: Classify the stance of the body text relative to the claim made in the headline into one of four categories:

1. **Agrees:** the body text agrees with the headline
2. **Disagrees:** the body text agrees with the headline
3. **Discusses:** the body text discusses the same topic as the headline, but does not take a position
4. **Unrelated:** the body text discusses a different topic than the headline

Dataset

The labelled dataset we will use for this project is available at the following link: https://github.com/FakeNewsChallenge/fnc-1

The dataset consists of two csv files: Stances.csv and Bodies.csv.
Stances has two columns called articleBody and articleID. Each row contains the full text of the article corresponding with its ID that matches the article ID column in Bodies.csv.

Bodies.csv has three columns called Headline, articleID, Stance, where Stance could have one of four values: unrelated, discuss, agree, disagree.

Data Preparation and Exploration

In the dataset there are 49972 rows of data. The distribution of each stance are as follows.
Unrelated: 0.7313
Discuss: 0.1783
Agree: 0.0736
Disagree: 0.0168094

The repository includes data split into training and test sets. It also includes code for pre-processing text, splitting data carefully to avoid bleeding of articles between training and test, k-fold cross validation, scorer, and most of the tools we will need to write to experiment with this data.

Feature Engineering

For feature engineering, we can turn stance values into numerical values. Other features created would be Basic Count Features, TF-IDF features, SVD features, Word2Vec Features and Sentiment Features. The hand-crafted features include word/ngram overlap features and indicator features for polarity and refutation. The scripts to create these features are provided in a repository.

A baseline using hand-coded features and a GradientBoosting classifier is available on Github.

With these features and a gradient boosting classifier, the baseline achieves a weighted accuracy score of 79.53% (as per the evaluation scheme described above) with a 10-fold cross validation. The baseline is used only as a reference to evaluate other models.