Course Project: News Stance Detection

Qiang Zhang, Bill Lampos

February 23, 2018

1 Task Definition

In context of news, a claim is made in a news headline, as well as in the piece of text in an article body. Quite often, the headline of a news article is created so that it is attractive to the readers, even though the body of the article may be about a different subject/may have another claim than the headline.

Stance Detection involves estimating the relative perspective (or stance), of two pieces of text relative, i.e. do the two pieces agree, disagree, discuss or are unrelated to one another. Your task in this project is to estimate the stance of a body text from a news article relative to a headline.

The goal in stance detection is to detect whether the headline and the body of an article have the same claim. The stance can be categorized as one of the four labels: "agree", "disagree", "discuss" and "unrelated". Formal definitions of the four stances are as:

- "**agree**" – the body text agrees with the headline;

- "**disagree**" – the body text disagrees with the headline;

- "**discuss**" – the body text discusses the same claim as the headline, but does not take a position;

- "**unrelated**" – the body text discusses a different claim but not that in the headline.

2 Dataset

We will be using the publicly available FNC-1 dataset[1]. This dataset is divided into a training set and a testing set. The ratio of training data over testing data is about 2:1. Every data sample is a pair of a headline and a body. There are 49972 pairs in the training set, with 49972 unique headlines and 1683 unique bodies. This means that an article body can be seen in more than one pair.

"unrelated" data takes the majority (over 70%) in both sets while the percentage of "disagree" is less than 3%. The percentage of "agree" and "discuss" are less than 20% and 10%, respectively. Severe class imbalance exits in the FNC-1 dataset.

FNC-1 implements an official baseline[2] that may be helpful to read files, and to split the train dataset into a training subset and a validation subset.

3 Involved Subtasks

The course project involves several subtasks that are required to be solved. This is a research oriented project so you are expected to be creative and coming up with your own solutions is strongly encouraged for any part of the project.

- Split the training set into a training subset and a validation subset with the data number proportion about 9:1.The training subset and the validation subset should have similar ratios of the four classes. Statistics of the ratios should be presented.

---

[1] https://github.com/FakeNewsChallenge/fnc-1/
[2] https://github.com/FakeNewsChallenge/fnc-1-baseline

- Extract vector representation of headlines and bodies in the all the datasets, and compute the cosine similarity between these two vectors. You can use representations based on bag-of-words or other methods like Word2Vec for vector based representations. You are encouraged to explore alternative representations as well.

- Establish language model based representations of the headlines and the article bodies in all the datasets and calculate the KL-divergence for each pair of headlines and article bodies. Feel free to explore different smoothing techniques for language model based representations.

- Propose and implement alternative features/distances that might be helpful for the stance detection task. Describe feature meaning and extraction process.

- Choose two kinds of representative distances/features that you think may be most important for stance detection and plot the distance distribution for the four stances. Comment on why you think these are the important features and try to validate their importance using the data.

- Using the features that you have created, implement a linear regression and a logistic regression model using gradient descent for stance classification. The implementations of these learning algorithms should be your own.

- Analyse the performance of your models using the test set. Describe the evaluation metric you use and explain why you think would be suited for this task. Feel free to use alternative metrics that you think may fit. Compare and contrast the performance of the two models you have implemented. Analyse the effect of learning rate on both models.

- Explore which features are the most important for the stance detection task by analysing their importance for the machine learning models you have built.

- Do a literature review regarding the stance detection task, briefly summarize and compare the features and models that have been proposed for this task.

- Propose ways to improve the machine learning models you have implemented. You can either propose new machine learning models, new ways of sampling/using the training data, or propose new features. You are allowed to use existing libraries/packages for this part.

4 What to submit

You are expected to submit all the code you have written, together with a written report up to 5 pages. Your report should describe the work you have done for each of the aforementioned steps. Unless otherwise stated above, all the code should be your own and you are not allowed to reuse any code that is available online. You are allowed to use both Python and Java as the programming language.

5 Deadline

The deadline for submitting your project is midnight on April 6th.