Research report - Hyperpartisian NLP

Florian Biebel Daan Middendorp

May 2019

1 Introduction

This report is part of a project at the Technische Universität Berlin, called Advanced Projects at the Quality and Usability Lab. During this project, the problem of hyperpartisian news detection will be solved using natural language processing. Hyperpartisian news is news that is extremely biased or sharply polarized in favor of a political party.

This problem comes from a task in the Semantic Evaluation[1] competition which is held simultaneously during this project. This makes it possible to use the dataset provided to the participants of the competition. The goal of this project is to achieve a deeper understanding of the inner working of such a system and learn how to build such a system ourselves.

2 Existing approaches

2.1 Preprocessing

The dataset from SemEval contains a thousands of articles. These articles cannot be fed into a form of artificial intelligence immediately. Several approaches are using the following preprocessing:

2.2 LSTM

Long short-term memory a form of an artificial neural network. The difference with a recurrent neural network is that LSTM networks are able to use information from the past to understand the input. E.g. a video frame can be better understood if the previous video frames are also taken into account. This is also valueable for text processing, because hyperpartisian news detection requires a deeper understanding of the link between words.

[1] `https://pan.webis.de/semeval19/semeval19-web/`

2.3 CNN

Convolutional neural networks have proven to be useful for multiple NLP tasks, such as sentiment analysis or summarization. CNN have the ability to extract n-gram features from input sentences to create a semantic sentence representation to be used in downstream tasks. Such ability comes from the convolutions creating an evermore abstract representation of the input, but still conserving a micro-context due to how convolutions work. This particular usage of CNN is called Sentence Modeling and could prove useful for hyperpartisan news detection.

2.4 Naive Bayes

The Naive Bayes method utilizes Bayes' Theorem to classify text and is used in text and topic classification. Additionally, it relies on the 'bag of words' (BOW) representation to function. For the binomial classification a positive and negative class are used—in our case hyperpartisan and non-hyperpartisan—to each of which is a BOW assigned. Depending on the count of positive and negative words, the text is classified as such. This approach might be useful depending on the chosen BOWs.

2.5 Support Vector Machine

If the wordt are modelled as words, is is possible to draw them in a multidimensional space. Support Vector Machines are able to draw a line in the space with creates the best separation between two types of categories. Therefore it is really suitable as a binary classifier. In this case both classes would be *hyperpartisian* or *non-hyperpartisian*.

2.6 Logistic Regression

3 Used tools

4 Proposed solution

5 Architecture