Learning to predict the stance of news articles DRAFT

Justin Ty z5232245

Introduction

The prevelance of fake news on the media has been an increasing issue in recent times. A Fake News Challenge (FNC) was created in 2017 with the attempt to address this issue using automated systems. The goal of the challenge was to build an AI system that can automatically detect the stance of news articles. This project for COMP9417 was an attempt to build an end-to-end process which can take raw news headlines and articles and predict their stance.

Implementation

Dataset and Labels

The provided dataset from the challenge contained news article bodies and headlines. The goal of the challenge is to build a classifier that can detect the stance between the body and headline. The stance of the dataset can be one of the following:

- Agree - If the headline and article body agrees
- Disagree - If the headline and article body disagree
- Discusses - If there was a discussion of the topic
- Unrelated - If the headline and body were not related.

Feature Engineering

The main approach for the features was to use a bag of words representation of words in the article. A TFIDF Vectorizer was used to generate this bag-of-words. This tranformation made more frequent words and stop words less important.

A large vocabulary with a one-hot-encoding resulted in a very sparse dataset. In order to improve performance and reduce noise, dimensionality reduction was applied to the bag of words using Latent Semantic Indexing.

Additionally, some feature engineering scripts were borrowed fromt he original challenge. These include. Cooccurence of words between were the main features created by the original utility functions from the fake news challenge. The borrowed features from the original repository include the count of refutes and overlaps between the article body and headline.

Scoring

Some scoring utility functions were borrowed from the challenge repository in order to compare this project's models against the baselines. This is referred to as the FNC scores.

Hanselowski (2018) noted that the original scoring functions were not appropriate for validating the document-level stance detection task. Instead, it was their recommendation to use F1 score with macro averaging. This is so that the scores are not affected by majority classes.

For this project, the FNC, F1, and accuracy scores were recorded. The detailed logs of the application were also saved to in the archive that shows their confusion matrices.

Experimentation

The baseline from the competition used a Gradient Boosting Classifier with 200 estimators.

Another simple benchmark was a Simple Decision Tree. This model was prone to overfitting to as its training scores are very high yet its validation scores are low. The learning curves in the appendix confirm this.

To avoid the overfitting of a single decision tree learner, a randomised cross validation search was used to tune its hyper parameters. This method allowed a single tree to be optimised and tuned by k fold cross validation. The hyper paramters include the tree's max depth and minimum items for a split.

Boosted ensemble methods seem to work very well especially for an unbalanced dataset such as this. Additional ensemble methods were also used such as Adaboost, Random Forest were used. These methods show comparable results against the boosted baseline.

In addition, XGBoost was also trained and tested. It has the benefit of being fast and scalable with the focus on speed and performance. It also contains regularisation parameters which help against overfitting. If the data is structured properly, it can be very efficient for sparse data which is suitable for a bag-of-words approach.

Neural networks proved to be popular and effective in the past competition. Talos, the winner of the FNC combined both Gradient Boosting and a Convolutional Neural Network.

A simple neural network was also implemented in this project using a dense network in Keras. It was found that the scores for the neural network in this project does not score as well as the other ensemble models. This may be due to the lack of quality features. This could be futher improved with more background in NLP feature engineering.

A summary of FNC scores, F1 Scores, and Accuracy is illustrated below.

Conclusion and further recommendation

This project has shown an end-to-end process of feature engineering and predicting the stance of news articles.

Futher improvements could be done with this project. Areas of improvement can be done with better featuring engineering in the NLP space.

Handling sparse data such as Light GBM.

References

Andreas Hanselowski , et al (2018) *A Retrospective Analysis of the Fake News Challenge Stance Detection Task*, : Research Training Group AIPHES Computer Science Department, Technische Universität Darmstadt. https://arxiv.org/pdf/1806.05180.pdf

Melanie Tosik, et al (2018) *Debunking Fake News One Feature at a Time*, : Department of Computer Science New York University. https://arxiv.org/pdf/1808.02831.pdf

Andreas Hanselowski , Avinesh PVS , Benjamin Schiller , Felix Caspelherr , Debanjan Chaudhuri, Christian M. Meyer , Iryna Gurevych (2018) *A Retrospective Analysis of the Fake News Challenge Stance Detection Task*, : Research Training Group AIPHES Computer Science Department, Technische Universität Darmstadt.
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760230.pdf

Sean , Baird, Doug Sibley, Yuxi Pan (2018) *Talos Targets Disinformation with Fake News Challenge Victory,* Talos Intelligence
https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html

Fake New Challenge http://www.fakenewschallenge.org/

Tianqi Chen and Carlos Guestrin. (2016). *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 785-794. DOI:
https://doi.org/10.1145/2939672.2939785