CMPE257

2019/5/5

SAY

Samuel Yang

Paper - Draft 1

0. Dataset(s), Scraping, Enrichment

Dataset: Lair Lair Fake News Datasets

Enrichment: Political Social Media Datasets

https://www.kaggle.com/crowdflower/political-social-media-posts

1. what did you try

I tried to first analyze the correlation between the bias (label) and other attributes like (source, status, topics, …). I used several modeling algorithms like Linear Regression, Logistic regressions and Random Forest. After that, I trained a LDA model on my enriched dataset and use it as my Political-Bias detector.

2. what worked,

Logistic regression model and Random Forest Model works pretty well and ended up with the accuracy of nearly 90%. LDA model works as well, it some how can distinguish how bias a post or news is.

3. what did not work,

Linear regression doesn't work properly here since the label only has two values. LDA model's accuracy isn't quite as good since it only has around like 10% accuracy. Before I cleaned up null-values, it does have a 50% accuracy.

4. what alternatives did you try,

I tried to figure how I can clean the data more accurately to increase my accuracy but also to prevent from over-fit. I might try to not clean all the null-value data but replace them into some specific values instead. Also, changing the chosen features

can effect how sparse my data will be which will further more make adjustments to LDA model's accuracy.

5. **what did you research, what references (eg code) did you study or leverage (code)**

   For Dataset:

   https://www.kaggle.com/crowdflower/political-social-media-posts

   For Machine Learning Algorithms:

   https://scikit-learn.org/stable/index.html

   For LDA:

   https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

6. **what steps did you take as a team to find data enrichment sources, label data, feature engineering (e.g., NLP life cycle: stemming, lemmatization, spelling check, part of speech, classifications, Naive Bayes, decision tree, SVM, TF-IDF, LDA etc)**

   I did spell checking, NLP life cyclye, lemmatization, stemming, Naïve Bayes, and Random forest, Logistic Regerssion, Linear Regression on my dataset.