Max Henkel-Wallace

Professor Dube

Advanced Topics in Data Science

12 April 2022

Project Proposal: Fake News Detection with RNNs

1. Title: “Hyperparameter tuning Fake News Detection using RNNs (LSTMs)”
2. Group Members: Max Henkel-Wallace (Alone)
3. Goal/Objective: The goal of this project is to train a RNN to detect fake news based on a dataset compiled by the university of Victoria and classified as credible or fake news by Politifact. We are using RNNs as they represent the best NLP tools at our disposal, which we need in order to create a credible meaning vector from the whole sentence rather than simply a static word. From here, I would like to do some hyper parameter experiments in order to discern the optimal dimensionality hyper parameter.
4. Challenges: The biggest challenge will be processing the data into a format that the RNN can analyze efficiently. NaNs in the data is an issue, as is edge case data. Finally finding hardware that is suitable for the runtime requirements of the algorithm and processing is another challenge.
5. Approach/Techniques: My plan is to fully one hot encode each word of the data, and remove a few things such as periods and commas, and then run it though the RNN, specifically an LSTM in order to handle the time series structure of the data (sentences) and to avoid the vanishing gradient problem that often occurs in RNNs.

6. Implementation Details: I plan to use Tensorflow Kerns mainly, and specifically the LSTM function. Additionally I would like to use general python libraries in order to clean the data. I will do some development locally on my machine as tensorflow works with my GPU, (torch does not) and then once I have it working move it over to Google Colab in order to do hyper parameter tuning with my free HPC credits.
7. Demo plan: For my demo plan I will run the model at the best hyperparemeter sets and then have it evaluate certain headlines and texts. I may even write headlines and have it evaluate it.
8. References

https://towardsdatascience.com/fake-news-detector-with-deep-learning-approach-part-ii-modeling-42b9f901b12b (I found this article quite helpful for many aspects of implementation, though I will also tune hyperparameters)

Dataset (University of Victoria compiled, fake news as per Politifact analysis): https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php

"Detecting Fake News using Machine Learning: A Systematic Literature Review" – https://arxiv.org/pdf/2102.04458.pdf

"Universal Sentence Encoder" – https://arxiv.org/abs/1803.11175

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" – https://arxiv.org/abs/1810.04805