Final Project Proposal

DS4440 - Practical Neural Networks - Pablo Kvitca - November 2020

Event Aware Adversarial Neural Networks for Text-based Fake News Detection

With the spread of access to social media and the ability to stories and opinions, the power of information has shifted from an expensive and limited resource to a cheap and abundant one. On the surface, this is a positive change but several issues have come up around the reliability and trustworthiness of information (and sources) being shared. There is an enormous interest in the ability to control public opinion, leading to the spread of intentionally deceptive and misleading content being created and shared. To fight this, we would like to use Machine Learning models, but they must be accurate, unbiased, and applicable to the systems they would be applied to. I propose to investigate a system for the detection of fake news detection based on the architecture presented in Wang et al.'s paper: *EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection*.[1][2]

In their paper, they take the intuition that neural network models for fake news detection may overfit and learn features specific to the events seen during training. Because of the ever-renewing nature of news' stories and social media posts, this makes it hard for the models to be extensible to future/unseen content. They create an adversarial network that adjusts against learning event-specific features. I aim to use this same idea and architecture with some changes to achieve a similar goal:

- First, their model uses only Chinese language data from the Twitter and Weibo platform. Since I am not familiar with that language, I will be using a different dataset of **English articles (rather than short Tweets).**
- Second, they only use a Convolutional Neural-Network applied to word-embeddings on the text data. Instead, I will explore alternatively using **LSTM** and Transformer layers. **I will explore combining both LSTM and CNN layers and trying them by themselves**.

[1] *EANN Paper: https://dl.acm.org/doi/10.1145/3219819.3219903*

[2] *EANN Code: https://github.com/yaqingwang/EANN-KDD18*

- Third, their model uses a multi-modal framework using both text and images as inputs. However, I believe this limits the system to be only properly applicable to Tweets with images. **I will be using only the text content of articles, and possibly some metadata**.
- Fourth, I will **explore the bias in the resulting model** through error analysis. The most interesting bias to check is favor/disfavor for specific ideological or political groups (for example, bias against conservative speech). Since the dataset does not have annotated data about this, I will randomly sample a small part of the dataset and manually tag those samples for different properties where it might be interesting to analyze the bias. This will include a bias for/against an ideological group, usage of slang language, the gender of the author, and possibly others.
  To analyze the bias I will compare the accuracy and scores, false positive and false negative rates for each group.

Dataset

I will use *FakeNewsCorpurs* as my dataset. This is presented on *FakeNewsCorpus* this is sourced from data from http://www.opensources.co/ and contains just over 9.000.000 samples. It contains 10 different labels (*fake, satire, bias, conspiracy, state, junksci, clickbait, unreliable, political,* and *reliable*). I will try both using the provided labels and bundling them together into *true*/*fake* cases.

Metrics

I will use the basic *f1-score* and *accuracy* score to check for the general performance of my models. For the models that I choose to do error analysis for bias, I will also the classical *false positive, false negative, true positive,* and *true negative* rates for each class.

Hypothesis

My hypothesis is that I will be able to gain some accuracy improvement over a baseline model that does not use the adversarial network architecture. Additionally, I expect that the bias will be lower for the model with the adversarial network (as compared to the baseline).

Anticipated Difficulties/Outcomes

I expect I might have difficulty using all of the data for training since computation resources are limited. I might decide to use just part of it. I don't expect any issues with the implementation, but I have not used PyTorch for any adversarial networks yet.

Extra Idea (if reasonable and time allows?)

The proposed *EANN* model treats each individual Tweet as a separate event. This is useful since we do not need to annotate each Tweet for which event it refers to. However, I believe some there could be value in trying to also adjust against the model learning events in the sense of Tweets referencing the same real-world event. This comes from my intuition that the model could overfit on tweets about certain events. For example, lots of Tweets talking about elections (through the world) could be annotated as "false", so the model might learn to reality heavily on related words (such as "votes", "elections", "polls") and have a bias on the topic.
Unfortunately, to my knowledge, there are no annotated Tweet datasets for both event labels and misinformation labels. Instead, I will explore treating each *day* as an event. I expect this to be a reasonable way to "cluster" Tweets together without using annotated event data. The adversarial model would then aim to determine what day a Tweet was created.