Misinformation Classification Using BERT models

Kris Riedman
November 2023

Abstract

In this paper, I explore the ability of different BERT models to classify news articles as real or misinformation. The goal of this experiment is to find the best results using the smallest model and the least amount of training. By optimizing for efficiency, I can ensure that even those with limited compute resources will be able to fine tune a model and generate accurate results.

Introduction

In the highly interconnected modern world, spreading misinformation is easier than ever. Thankfully, we also live in a world where the tools to detect misinformation are cheap and accessible to anyone who wants to use them. With the prevalence of social media embedded in elections and public health, it is vital for citizens to have the tools necessary to separate fact from fiction.

Background

- PANACEA is an online tool created by a group out of Great Britain that “provides automated veracity assessment and supporting evidence for the input claim.” This may be useful for users who wish to understand why a claim is spurious and the facts surrounding it, but for others this may amount to information overload.
- Another common approach to propaganda detection is sentence level and fragment level propaganda detection as described in “Neural Architectures for Fine-Grained Propaganda Detection in News”. Again, this may provide too many details for the average user.
- A different team identified eighteen different propaganda techniques in their paper titled “Fine-Grained Analysis of Propaganda in News Articles”. Identifying spans of propaganda embedded in text can help with model explainability, but the nuance between different propaganda types may cloud the overall output.
- In their paper “Proppy: Organizing the News Based on Their Propagandistic Content”, Barron-Cedeno et.al. used a maximum entropy text classifier in a few experiments including a 4-way classifier: trusted vs. propaganda vs. hoax vs. satire, and a 2-way classifier: propaganda vs. non-propaganda.

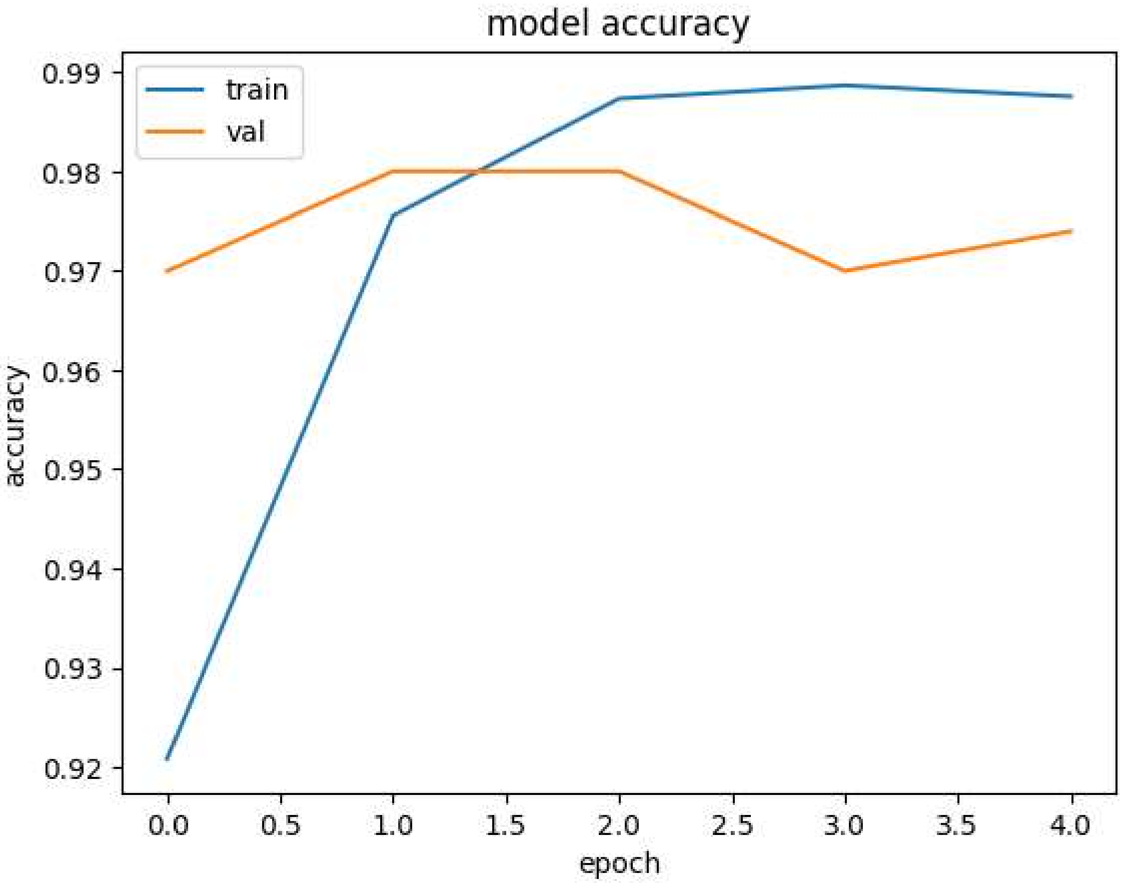
Methods

- Data is from “Misinformation and Fake News text” hosted on Kaggle
- 79k news articles collected from various outlets
    - 35k from reputable sources
    - 44k from questionable sources

- Also have an subset of data from Russian news outlets
    - 6k articles
- Classify articles as real or misinformation
- Baseline model is BERT Base cased
- Optimize to a smaller model with limited retraining of existing models using Lora.

Results and discussion

The baseline model shows 98% accuracy after training for 5 epochs on 5000 articles. Split 50/50 between legitimate articles and misinformation. Training took approximately 40 mins using Google Colab GPU. I am somewhat skeptical about the initial results from the baseline model but a cursory investigation has failed to find any glaring issues. One possible issue I can see is how the data was collected. The author scraped news articles from various sources and labeled them based on the source not necessarily on the content of the article. Another issue may be with the actual samples. The model may be learning certain phrases that are only present in one set but not the other instead of learning the general style of the articles.



Code: [riedmank/w266_final_project (github.com)](https://github.com/riedmank/w266_final_project)

Next Steps

- Add Lora to baseline model
- Evaluate baseline model on third dataset sourced from Russian news sites
- Try other Bert models to see if I can get similar results on a smaller model with and without Lora
    - BERT-Tiny
    - BERT-Mini
    - BERT-Small
    - BERT-Medium
- Compare performance of each model using a couple of metrics
    - Training time
    - Accuracy

Appendix

- Neural Architectures for Fine-Grained Propaganda Detection in News - D19-5012.pdf (aclanthology.org)
- Fine-Grained Analysis of Propaganda in News Articles - D19-1565.pdf (aclanthology.org)
- PANACEA: An Automated Misinformation Detection System on COVID-19 - 2023.eacl-demo.9.pdf (aclanthology.org)
- Proppy: Organizing the News Based on Their Propagandistic Content - barron-et-al_ipm2019_proppy.pdf (qcri.org)
- Dataset - Misinformation & Fake News text dataset 79k (kaggle.com)
- Lora - 2023-fall-main/materials/walkthrough_notebooks/peft/lora/Lora_(Roberta_Large)_pytorch.ipynb at master · datasci-w266/2023-fall-main (github.com)
- BERT models - google-research/bert: TensorFlow code and pre-trained models for BERT (github.com)