ECE 580 Project Proposal

---

1) Team Members (NetID):

- Rebecca Du (rrd17)
- Anish Parmar (avp30)

2) Problem Statement:

- "Distinguishing fake from real news based on existing political articles"

  The creation of the internet has enabled individuals to an unprecedented stream of real-time information. However, that does not mean all of the information is genuine. The recent decade has seen its fair share of scandals and disasters resulting from the propagation of fake news, and the need to distinguish between fact and fiction has never been more urgent. Thus, creating an ML model designed for this very task not only addresses the issue of fake news, but aids in solving it by identifying false articles.

3) ML Task(s)

- Binary classification of news articles. The model will effectively predict whether an article is:
  - Real (authentic & trustworthy)
  - Fake (misleading & deceptive)

4) Data Involved

- ISOT Fake News Kaggle Dataset (Public) (2016-2017)
  - 2 CSV files: "True.csv" and "Fake.csv".

Characteristic	Description	
	True.csv	**Fake.csv**
Size	21,417 articles	23,502 articles
Features	- Title - Text - Subject - Date	- Title - Text - Subject - Date

Origin (Both Real World Sources)	- Reuters.com (news site)	- Unreliable websites flagged by fact-checking US organization Politifact - Wikipedia
Topic	- Mostly political + world news	- Mostly political + world news

5) Models to Consider

Logistic regression is a simple model that could potentially work well with text features with methods like TF-IDF matrices or word embeddings. Additionally, random forests can capture non-linear patterns in textual data. Support Vector Machines (SVMs) are another approach to consider as it is primarily used for classification. Naive Bayes is a well-known method that is proven to be efficient with text-based data and is another potential avenue worth pursuing.

Convolutional neural networks (CNNs) may not be the best choice for this topic since they are primarily used in image processing tasks. Though Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) have proven to be useful with sequential text analysis, it is worth noting that RNNs and LSTMs generally require significant computational resources.

6) Expected Results

Common success metrics for binary classification problems include accuracy, precision, recall, and AUC-ROC. Accuracy would be calculated by dividing correct predictions (true positives + true negatives) and dividing them by all predictions (true positives + true negatives + false positives + false negatives). Precision would serve as an indicator of how many articles that are classified as fake by the model are actually fake (true positives / (true positives + false positives)). Meanwhile, recall measures how many fake news articles are correctly labeled (true positives / (true positives + false negatives).

The goal of our model is to achieve an accuracy of at least 90% on the test set, a threshold which should be attainable by fitting models such as Naive Bayes and SVMs.

7) Relevant Literature

- Detecting Fake News Using Machine Learning: A Systematic Literature Review (2021)
    - This paper provides a thorough review of various existing methods of fake-news classification, and we intend to use it (and the other literature cited within it) to guide our approach to the project.