542 Machine Learning Summer 1 Final Project
Fake News/Article Detection

Team Members: Bhagyasri Kota, Ezelia Franson, Shweta Baindur

Abstract:
Fake news has become a huge problem in these new age technology times. Mainly with the advent of several media platforms, news is manipulated to spread one's own agenda or mislead the population. So we aim at using the techniques of machine learning in identifying the false/fake data since not all platforms have an algorithm that does this work. The main goal of this project is to build a model which can classify if a news article is fake or not. For this, we aim at using well established as well as modern classification techniques like Naive Bayes or Bert models. This is not just a problem with English news media but all over the world. But for the scope of this project, we aim at training a model based on English dataset and test the efficiency.

Approach:
The first and foremost task is collection of data for which we found a pre-existing dataset at the below mentioned url:
https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
We will be performing an initial analysis of the dataset to check for genre, bias as well as to identify salient features that can be used for training. Based on the findings, we will either try to train a general model which can be used for fake news detection or genre specific if enough data is available. Since it is a text based dataset, we would be performing preprocessing of the data which might include cleaning up punctuation, removing of stop words, tokenizing etc., using packages such as NLTK, spacy. Our primary aim is to generate a model which can recognize fake data from real data for which we intend to try models such as Logistic Regression, naive Bayes as well as Bert based models. Recent pre-trained multilingual models such as mDeberta, XLM-Roberta which are transformer based models and said to have outperformed existing models can also be used for this task. Based on the model trained, tfidf and countvectorizer or auto tokenizer from transformers will be used to convert the text to vectors and input to the model.

Future Scope:
Based on the progress we make, further development can be done by training the models on different language datasets or can be used to test and see if models trained on a particular language data set can give good results on an unknown language dataset.