DATS-6312 NLP for Data Science

Fake and Real News Dataset

Prof. Amir Jafari

Group Proposal

Aasish Kumar, Greeshmanjali Bandlamudi, Pranay Bhakthula

Project Proposal

The goal of our project is to classify real and fake news on the internet. As the scourge of “fake news” continues to plague our information environment, attention has turned toward devising automated solutions for detecting problematic online content. But, to build reliable algorithms for flagging “fake news,” we will need to go beyond broad definitions of the concept and identify distinguishing features that are specific enough for machine learning. We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

Two CSV files make up the dataset. The first file, "True.csv," contains more than 12,600 reuter.com articles. The second file, "False.csv," has almost 12,600 items culled from various fake news sources. The following information is included in each article: the title, the text, the type, and the date the article was published. We focused on gathering articles from 2016 to 2017 to match the false news data acquired for kaggle.com. The information gathered was cleaned and processed, but the punctuation and errors found in the bogus news were left in the text. The real news contains 21417 articles, and we have two types in that one is world-news with article size of 10145 and the other is political news with 11272. The fake news contains 23481 articles, and we have government-news with article size 1570, middle east with size 778, US news with 783 sizes, left-news has 4459 size, politics has 6841, news has 9050 article size.

Dataset:
https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Schedule for the project

Project Steps	***Duration***
Data Preprocessing	03/28/2022
Modelling	04/04/2022
Evaluation	04/18/2022
Presentation and Report	04/25/2022