EECS731 – Introduction to Data Science

Fall 2020

Semester Project Proposal

Team Name: News Muse

Team Members:

1. Ishrak Hayet
2. Sai Damaraju
3. Sushmitha Boddi Reddy
4. Madhu Peduri

Overview of the Project:

Our project will be focused on applying various data science approaches on fake news. Fake news is a falsified article or story intended to mislead the audience. People or online bots use manipulated news content as a tool to spread propaganda, influence a network of people, and to gain socio-political or economic benefits. Additionally, owing to its inherent dynamics, social media has become a fertile ground for spreading fake news.

However, we believe that a piece of news is at its core, a piece of data. And, using a data science lifecycle effectively, we can identify fake news and formulate methods to determine its impact from the spatiotemporal and social metadata. At first, we can perform some exploratory data analysis to get an idea about the context or distribution of the content. Then, we can preprocess the data and work on identifying or generating suitable features from the news text and the metadata. Eventually, we can harness powerful classical and modern classification, clustering, and regression techniques to reveal information from the data that would otherwise be hidden initially. Finally, we can evaluate the performance of our data science pipeline both quantitatively and qualitatively.

Datasets:

1. Kaggle Fake News Dataset: https://www.kaggle.com/c/fake-news/data
2. FakeNewsNet (Twitter content): https://github.com/KaiDMML/FakeNewsNet
3. Fakeddit (Reddit content with text and images): https://github.com/entitize/Fakeddit

Problem Areas and Approaches:

1. Classification of fake news along with anomaly detection:

   We can use the different labeled datasets to build one or more classification models to predict whether a given news content is fake or not. Then, we can perform anomaly detection to analyze the classification results.

2. Clustering of news based on content and metadata:

   We can use clustering algorithms to create neighborhoods of similar news and try to find whether specific fake news items are similar based on metadata

3. Regression of news impact based on content metadata:

   We can quantify the impact of a news (tweet or reddit post) using the retweet counts (twitter), follower counts (twitter), comment counts (reddit), upvote counts (reddit). Then, we can couple these quantitative data with the textual or image content to predict the impact of an unseen news item using regression analysis.