A Study of the Differences between Real and Fake News

Team Members

Claudia Hidrogo
Madeline Odom
Jordan Sims
Xiomara Winkler

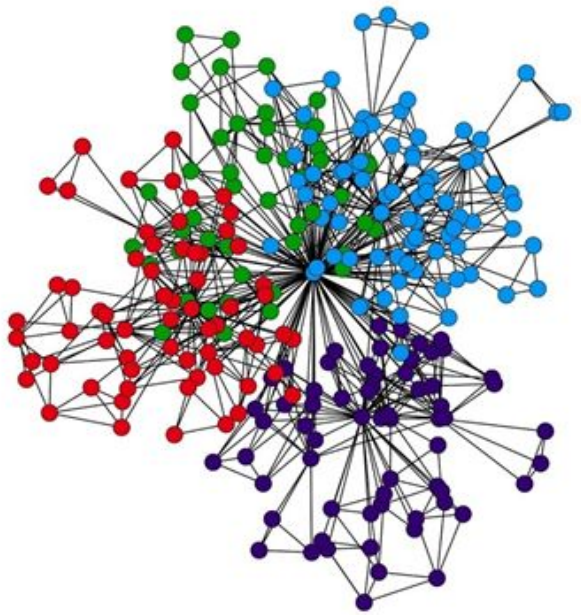
Applied Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Description

Prior Work

As a result of the spread of misinformation in recent years, there is now an abundance of research on this topic. Notably, the BotSlayer software developed by researchers at Indiana University Bloomington. The BotSlayer is a cloud tool that lets journalists, researchers and citizens alike track and detect potentially inaccurate information on Twitter in real time.

https://cnets.indiana.edu/blog/2020/09/18/new-botslayer-tool-to-expose-disinformation-networks/#more-9228



Datasets

Source: Harvard Dataverse	**URL:** https://dataverse.harvard.edu
A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles	
Source: Common Crawl	**URL:**https://github.com/huggingface/datasets/blob/master/datasets/cc_news/cc_news.py
News dataset containing news from news sites all around the world.	
Source: Kaggle: Clément Bisaillon	**URL:** https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset?select=True.csv
Dataset consisting of fake and real news articles.	
Source: Kaggle: Samrat Sinha	**URL:** https://www.kaggle.com/samrat96/fake-news-detection?select=test.csv
Fake news detection dataset consisting of train and test data for fake news model development.	
Source: Zenodo: Radu Prodan, Prateek Agrawal, Pawan Verma	**URL:** https://zenodo.org/record/4561253#.YWNCkNrMKUl
Generic word embedding dataset for fake news detection consisting of real and fake news from popular news sources.	

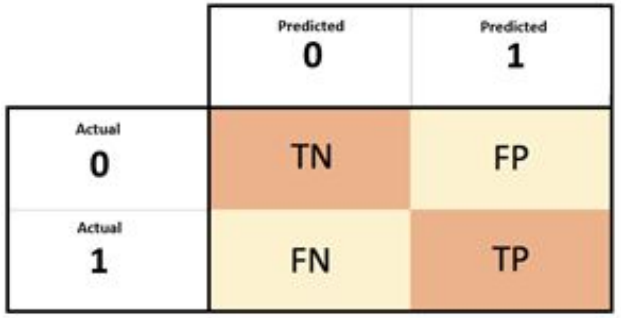
Data Preprocessing

Data Cleaning	Data Integration	Data Transformation	Data Reduction
Identify and remove/drop all the words in each news article that do not have any significant meaning in the text such “as”, “by”, etc.	Concatenate the datasets containing the fake news articles with the real news articles to form a large dataset.	Convert the list of words into tokens (numbers) for easy manipulation of the data.	Reduce the list by obtaining only the unique words contained in each news article.

List of Tools

pandas	For data frame manipulation
numpy	For numeric analysis
matplotlib	For data visualization
plotly	To create interactive plots – dynamic visualization
WordCloud	To represent frequency of words
keras	For deep learning
nltk	For language processing

Evaluation

Confusion Matrix	For binary classification of the news articles. True meaning that the article is not fake, and false meaning the opposite.	
Accuracy and precision	Both metrics are used to predict correctness of the labeled news articles.