CONTEXT DRIVEN APPROACH TO DETECTING CROSS-PLATFORM COORDINATED INFLUENCE CAMPAIGNS

Project Team: Anthony Chen, Ken Zeng, Jennifer Rodriguez-Trujillo, Frances Yuan
Project Mentor: Zhouhan Chen

Center of Data Science, New York University

Introduction

In an information-based economy influenced by a constant in and out flow of news, there is importance in preventing the detrimental effects of low credibility URLs. Our approach towards identifying harmful content is through building an ML model, using an open source labeler: Media Bias Fact Checker (MBFC), and Twitter data.

Research Question

Can We Determine Whether a Source Contains Fake News By Solely Analyzing User Interactions?

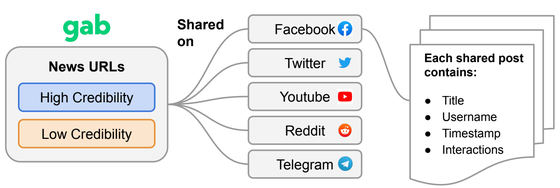
- Existing approaches focus on website content or features from a single social media platform. Our methods utilize **cross-platform** information from five major social media platforms.
- Our approach predicts source credibility using **context based** features such as aggregated user interactions, user activity, and domain information.
- We want to assess the viability of ML classifiers to uncover new sources of fake news.

Data Collection

The initial data used for the model was collected from Zhouhan's PhD work **Information Tracer API** python library. Information Tracer provides us fine-grained intelligence about how URLs, keywords, and hashtags spread online to journalists, and consumers. Through this data collection, we quantify patterns in the spread of content, and highlight suspicious behaviors that are then assessed through **MBFC**.

Due to the volatile nature of our data in terms of major geopolitical events during the time span of **January 2022 until August 2022**, it was necessary to iteratively adjust or add features during our modeling process.

We Collect Interactions on 5 Major Social Media Sites



In addition, we also use **WhoIs**, **IpAddress**, and **Twitter API** in order to allocate other open sources of information in relation to the domains at hand as well as to attain further user details, such as suspension status.

User Statistics From Twitter

Account suspended
Twitter suspends accounts that violate the Twitter Rules. Learn more

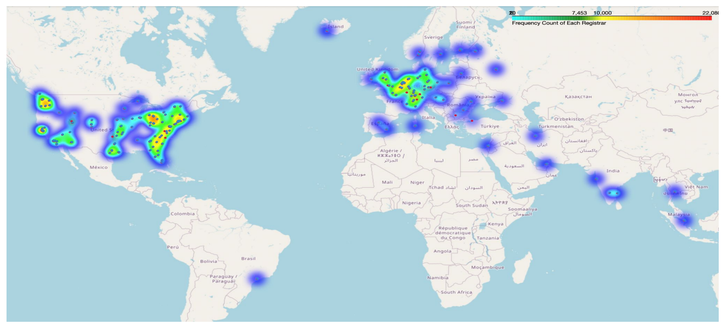
Additional Sources

Media Bias / Fact Check

Whois
Identity for everyone

Data Overview

Where Are These Fake News Sources Hosted?



- IP Address and WhoIs sites permitted us to explore latent features in relation to domains
- **34,705** IP Addresses correspond to *Low Credibility*
  - **571** reside in **California** followed by **80** in **Virginia**
  - Preferred Hosting Provider **: myshopify.com**
- **29,666** IP Addresses correspond to *High Credibility*
  - ***147*** reside in **California** followed by **17** in **Virginia**

- Our map demonstrates the location of Hosting Providers labeled as *Low Credibility* by Frequency
  - ColorBar: Closer to Red, higher the frequency
- Max Frequency Count: **22080**
  - Hosting Provider with Highest Frequency: GoDaddy.com, LLC
- Min Frequency Count : **1**
- Standard Deviation : **9706.07**

Methodology

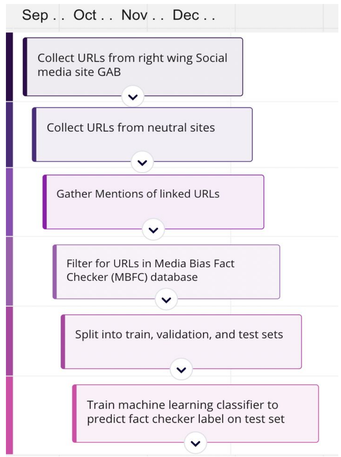
How Do We Train ML Models to Detect Suspicious URLs?

Our ML approach is to utilize all scraped urls from **right wing social media platform Gab,** and with the assistance of a fact checker source, **MBFC**, to split our data into three groups: URLs with high, low credibilities or are unlabeled.

After labeling a certain portion of URLs:

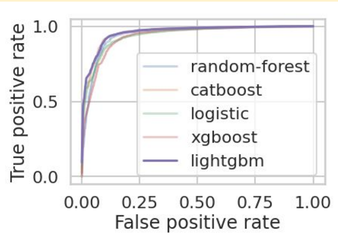
- We aggregated all text interactions from social media as **TF-IDF vectors**
- Collected **user statistics** about Twitter users who engaged with our URLs
  Examples of additional features we extracted from Twitter users:
  - Percentage of users that were suspended by Twitter
  - Percentage of verified users
  - Median number of favourites (likes), followers, friends, etc.
- Added more URLs with Neutral or High Credibility due to a data imbalance problem: Most of our URLs had Low Credibility

Our data was further split into a training, validation and testing sets for our machine learning pipeline. With a trained ML algorithm, we can provide suspicion labels for originally unlabeled URLs either individually or in a cluster form. After identifying these URLs, they can be further flagged for **further human evaluation**.



Results

Model Performances (ROC-AUC Curves)

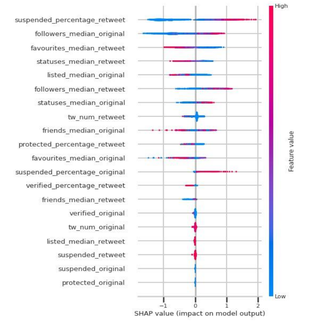


In total, we evaluated 5 widely used machine learning algorithms, consisting of a logistic regression alongside 4 state-of-the-art decision-tree based ensemble classifiers. With only Twitter user statistics, we achieved an AUC of **0.92.** Using text-only features, we achieved an AUC of **0.94.** With the combination of the previous mentioned features, we achieved the **optimal test AUC of 0.96**

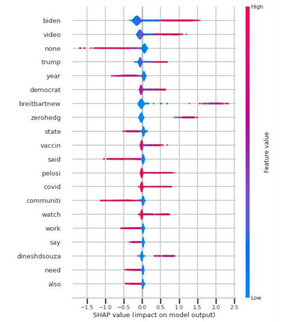
	Accuracy	F1-score	AUC
Random-Forest	0.925	0.953	0.945
Catboost	0.931	**0.967**	0.957
Logistic Regression	0.906	0.94	0.941
XGBoost	0.910	0.944	0.934
Lightgbm	0.930	0.956	**0.960**

Discussion

Top Twitter User Statistics Features

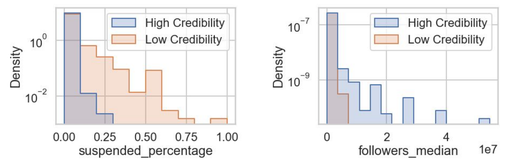


Top Text Features



We used Shapley additive explanation (SHAP) plots to analyze the most significant inputs in our LightGBM model's results. A large number of **suspended users retweeting** is a strong indicator that a URL has low credibility. This is similar to a high word frequency of **'biden'** being mentioned.

Key Indicators of URLs' Credibility



Future Considerations

Through the implementation of our model, we have come across resolutions and recommendations on how we can better adapt to the fast-paced realm of the web. After analyzing the data collection, data wrangling, and feature selection steps, we have observed ways to further improve upon and extend the scope of the project.

- **Unforeseen Power-Transfer & Updates to Twitter's platform:**
  - Rise in concerns of the stability of a data provider (concept drift)
  - Brings our attention to the importance of building durable, future-proof Machine Learning (ML) models
- **Language (non-English):**
  - Domains are not solely from the U.S, since they may also stem from other countries
  - Improve our ML models by training on non-English content
- **Media Format:**
  - ML models should further be applied to investigate the spread of content through multimodality, non-text features (Videos, Images, Apps etc.)

Acknowledgements & References

We would like to thank Yinjie Huang, Annie Franco, Joshua Tucker, and Haohan Chen for feedback throughout our project. Special thanks to Zhouhan Chen for guidance and support.

1. DeepLearning.AI. "Iteration in AI Development: AI News & Insights." *Iteration in AI Development | AI News & Insights*, Iteration in AI Development | AI News & Insights, 25 Oct. 2022, https://www.deeplearning.ai/the-batch/iteration-in-ai-development/
2. "Media Bias/Fact Check News." *Media Bias/Fact Check*, 22 July 2021, https://mediabiasfactcheck.com/.
3. "Whois Search, Domain Name, Website, and IP Tools - Who.is." *WHOIS Search, Domain Name, Website, and IP Tools Who.is*, https://who.is/.
4. Zhouhan, C. "Information Tracer: Python Client to Interact with Information Tracer." *GitHub*, https://github.com/zhouhanc/informationtracer.