Paper Topic: ReCOVery: A Multimodal Repository for COVID-19 Credibility Research

Due on 9/15/2020 at 11:59PM

CS 396ISH, UMass Amherst, Fall 2020

This week, I read the paper, "ReCOVery: A Multimodal Repository for COVID-19 Credibility Research", from Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. The work was done primarily to facilitate research on combating fake news and low credibility sources. They first looked at 2,000 news publishers and then look at the credibility of the media being published on 2,029 news articles on covid-19 as well as 140,820 tweets on how news articles spread on Twitter. They

When looking at datasets related to covid-19, they utilized a covid-19 twitter dataset from Emily Chen's research on coronavirus as well earlier in 2020. When looking at "fake" news and rumor datasets, they founds datasets with various focuses where they only contain news content that can be full articles or short claims that contain social media information only. These datasets collected by other researchers previously would typically bring in human subjects to rate how true or false a statement was. Instead, the ReCOVery project created an approach to filtering news sites through their criteria.

1. Does not repeatedly publish false content, (22 points)
2. Gathers and presents information responsibly, (18 points)
3. Regularly corrects or clarifies errors, (12.5 points)
4. Handles the difference between news and opinion responsibly, (12.5 points)
5. Avoids deceptive headlines, (10 points)
6. Website discloses ownership and financing, (7.5 points)
7. Clearly labels advertising, (7.5 points)

This approach above was created by NewsGuard. They had also utilized Media Bias/Fact Check website and label each news media with one of six factual-accuracy levels based on the fact-checking results the news articles had published. They range from very low to very high factual accuracy.

Their criteria was this:

- Reliable A news site is reliable if its NewsGuard score is greater than 90, and its factual reporting on MBFC is very high or high.
- Unreliable A news site is unreliable if its NewsGuard score is less than 30, and its factual reporting on MBFC is below mixed.

To collect their COVID-19 News Content, they first determine whether the new article was about COVID-19 through the three phrases:

- SARS-CoV-2,
- COVID-19, and
- Coronavirus

The data was crawled on Newspaper Python library and each news article was assigned with a unique id and extracts the URL, publisher, publication date, author, news title, new image, country, political bias, and the scoring by NewsGuard and MBFC. To track news spreading on Social Media, they used Twitter Premium Search API.

On top of that, they then did some visualizations with the data to get some information of their dataset.

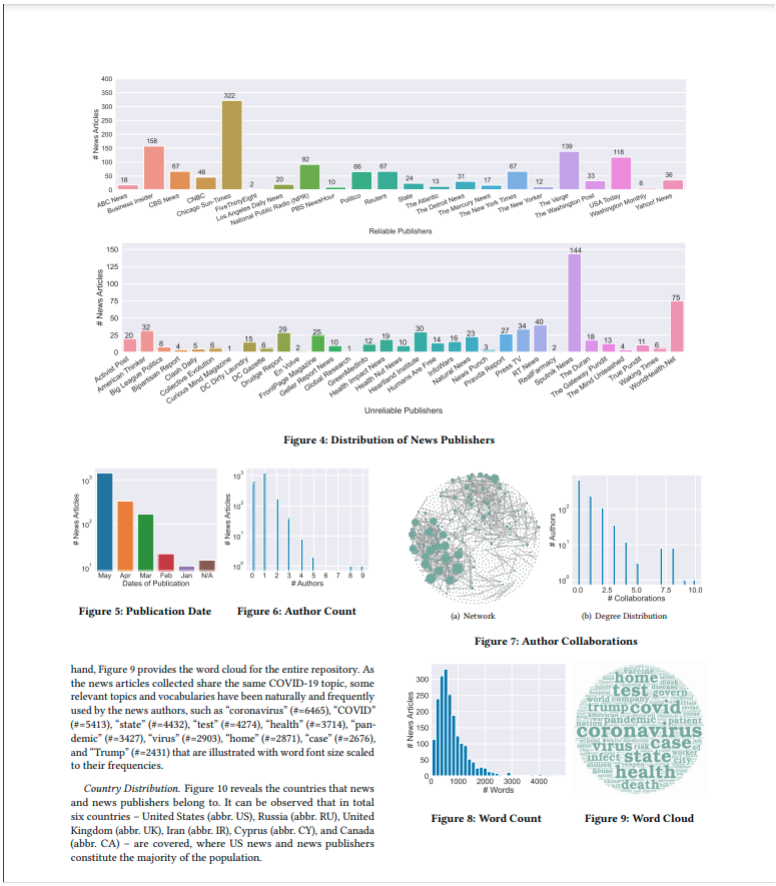


Figure 4: Distribution of News Publishers

Figure 5: Publication Date

Figure 6: Author Count

(a) Network (b) Degree Distribution

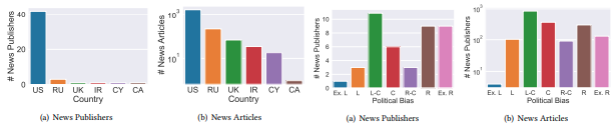
Figure 7: Author Collaborations

hand, Figure 9 provides the word cloud for the entire repository. As the news articles collected share the same COVID-19 topic, some relevant topics and vocabularies have been naturally and frequently used by the news authors, such as "coronavirus" (#=6465), "COVID" (#=5413), "state" (#=4432), "test" (#=4274), "health" (#=3714), "pandemic" (#=3427), "virus" (#=2903), "home" (#=2871), "case" (#=2676), and "Trump" (#=2431) that are illustrated with word font size scaled to their frequencies.

Country Distribution. Figure 10 reveals the countries that news and news publishers belong to. It can be observed that in total six countries – United States (abbr. US), Russia (abbr. RU), United Kingdom (abbr. UK), Iran (abbr. IR), Cyprus (abbr. CY), and Canada (abbr. CA) – are covered, where US news and news publishers constitute the majority of the population.

Figure 8: Word Count

Figure 9: Word Cloud



(a) News Publishers (b) News Articles

Figure 10: Country

(a) News Publishers (b) News Articles

Figure 11: Political Bias

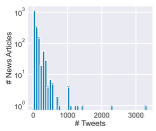


Figure 12: Spreading Frequency

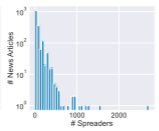


Figure 13: News Spreaders

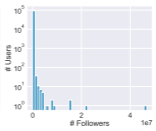


Figure 14: Follower Distribution

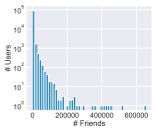


Figure 15: Friend Distribution

They then formed some methods to predict COVID-19 New Credibility. They utilized some single-modal methods such as LIWC (Linguistic Inquiry and Word Count) , RST (Rhetorical Structure Theory), Text-CNN, and SAFE models. LIWC is a psycholinguistic lexicon where given a news story, LIWC would count the words in the texting falling into one or more of the 93 linguistic, psychological and topical caategoical. RST organizes a piece of content as a tree and gets the rhetorical relation among its phrases and sentences. Text-CNN is Convolutional Neural Network for Text classification. SAFE is a neural network implementation that utilizes news multimodal information for fake news detection. The results are below:

Table 2: Baselines Performance in Predicting COVID-19 News Credibility Using ReCOVery Data

Method	Reliable news			Unreliable news		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
LIWC+DT	0.779	0.771	0.775	0.540	0.552	0.545
RST+DT	0.721	0.705	0.712	0.421	0.441	0.430
Text-CNN	0.746	0.782	0.764	0.522	0.472	0.496
SAFE	0.836	0.829	0.833	0.667	0.677	0.672

When implementing their code, they randomly divided into training and testing datasets of 80/20. They evaluate the prediction results based on precision, recall, and the F_1 score. On methods where they relied on traditional statical learners, they used classifiers such as Logistic Regression, Naive Bayes, KNN, Random Forest, Decision Trees, and Support Vector Machines.

I ended up utilizing this article for a couple of reasons. I think some of the ideas utilized on news articles related to scoring the news articles may come in handy. Additionally, I think that it might be interesting to explore along with identifying political ideology in COVID-19 news, if there is some relation to reliable and unreliable news based on that. As I further go into building my ML/Deep Learning models, some of the preprocessing methods mentioned here as well as how they utilized their ML models would be good to know.

This is my link: Paper Link.