Covid-19 Fake News Classifier

By
Jeffery Braga,
Matthew Goodman,
Doris Chia-ching Lin

Introduction

Our team hopes to train a classifier to identify whether COVID-19 articles hold misinformation. Identifying cases of misinformation is important during a health pandemic, as the population's response could impact the health of those at risk.

If we are successful, our classifier could help curtail the spread of misinformation, and it could even indirectly save lives.

Background

- Information presented directly impact the choices people make regarding their own health. We must acknowledge that both our societal personal choices affect the well-being and possible infection of others.
- As such, the ability to distinguish between the true and false information about health, safety, and other COVID-19 related news would be an invaluable tool. Within the short time-frame, a new wave of false information can be said to impact both public and personal health in a direct sense.
- Natural Language Processing (NLP) to build a classification system for accurately separating true articles from those which are false. Of the variety of algorithms available for classification, our focus is on the application of logistic regression and random forests.
- The system is trained on news articles occurring before the timeframe of the pandemic, primarily before the 2016 United States Presidential election. Now, it is evaluated in its ability to correctly classify articles relating to COVID-19 pandemic.

Approach

- Use the Natural Language Toolkit (NLK; Loper & Bird, 2002) and Scikit-Learn (Pedregosa et al., 2011) packages.

- Two datasets
    - *Train Dataset* - 21,498 true articles and 23,490 fake articles. (politics, world news, and local news)

    - *Test Dataset* - Team-gathered an additional 100 COVID-19 articles from various sources. The dataset consisted of 91 true articles and 9 fake articles.

Approach

- Data Preparation
  - Pre-process: merge the all fields into a single data field. Then converted to lowercase letters.
  - Remove stop words and stem each text sample.

Approach

- Training and Testing the Model
  - To train the classifier, our team used a 80% train and 20% test split on the training dataset. We decided to assess the performance of the classifier on the training dataset to ensure the model would be capable of successfully classifying a new and foreign dataset.
  - Two algorithms:
    - logistic regression algorithm. $P(Y_n = 1 \mid x_n) = \sigma(w * x_n)$
    - random forests.

Demo Time

Result & Evaluation

We assessed the classifiers with three performance metrics:

1) recall;
2) precision;
3) f-measure.

For the 80/20 split on the training dataset, both models performed similarly.

Algorithm	Recall	Precision	F-Measure
Logistic Regression	0.9928	0.9907	0.9918
Random Forests	0.9949	0.9967	0.9958

Performance metrics on the training dataset

Algorithm	Recall	Precision	F-Measure
Logistic Regression	0.8901	0.9529	0.9205
Random Forests	0.6593	0.9524	0.7792

Performance metrics on the COVID test dataset

Result Tables

Future Development

- Enhancement for social media posts
- Recompile more data and evaluate the model more.

Thank you!!