# COVID-19 Misinformation Detection

Caitlin Moroney

12/2/2020


Figure 1: Twitter adds warning labels to tweets.

# Dataset

- 560 tweets, perfectly balanced classes
- sample of 282,201 users in Canada
- tweets posted between January 1 - March 13, 2020
- manually labeled as “reliable” or “unreliable”

Table 1: Misinformation rules from Boukouvalas et al. (2020)

| **Linguistic Feature** | **Example from Dataset** |
|---|---|
| Hyperbolic, intensified, superlative, or emphatic language | e.g., 'blame', 'accuse', 'refuse', 'catastrophe', 'chaos', 'evil' |
| Greater use of punctuation and/or special characters | e.g., e.g., 'YA THINK!!?!!?!', 'Can we PLEASE stop spreading the lie that Coronavirus is super super super contagious? It's not. It has a contagious rating of TWO' |
| Strongly emotional or subjective language | e.g., 'fight', 'danger', 'hysteria', 'panic', 'paranoia', 'laugh', 'stupidity' or other words indicating fear, surprise, alarm, anger, and so forth |
| Greater use of verbs of perception and/or opinion | e.g., 'hear', 'see', 'feel', 'suppose', 'perceive', 'look', 'appear', 'suggest', 'believe', 'pretend' |

# Methodology

Pipeline:

- raw text
- word embeddings
  - word-word co-occurrence matrix
  - latent variable methods
- tweet embeddings
- classification
- evaluation

# Word-Word Co-occurrence Matrix

# Latent Variable Methods

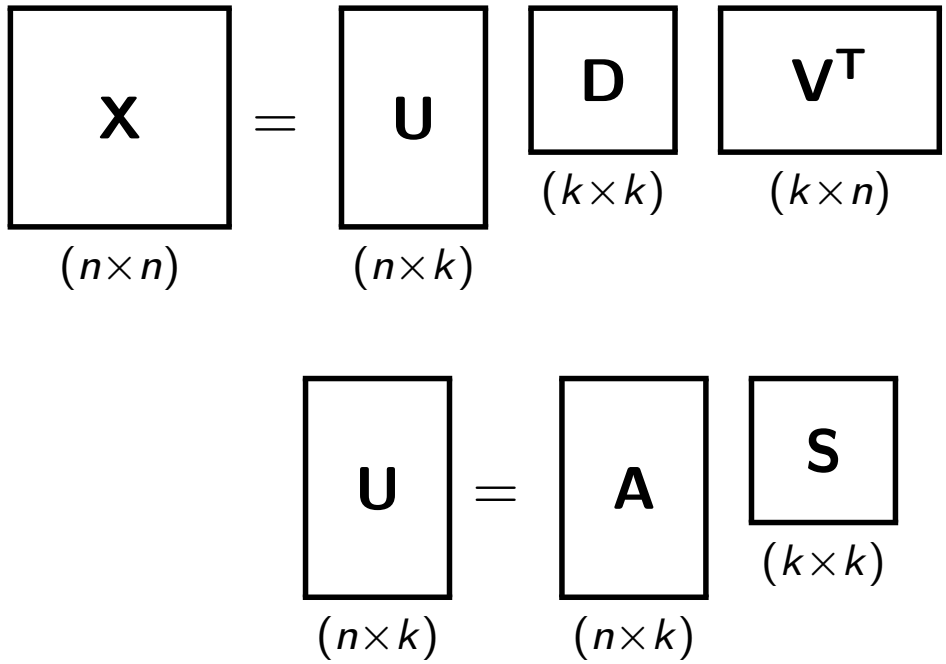


Figure 2: Truncated Singular Value Decomposition followed by Independent Component Analysis.

# Tweet Embeddings

$$\mathbf{v}_i = \frac{1}{T_i} \sum_{j=1}^{T_i} \mathbf{e}_j$$

# Classification

# Evaluation

# LIME: Local Explainability

# ICA: Global Explainability

# Example

Tweet 170: CNBC ADVICE NOW: Coronavirus is the flu. Wash your hands. Book a vacation. We'll look back on this and laugh.

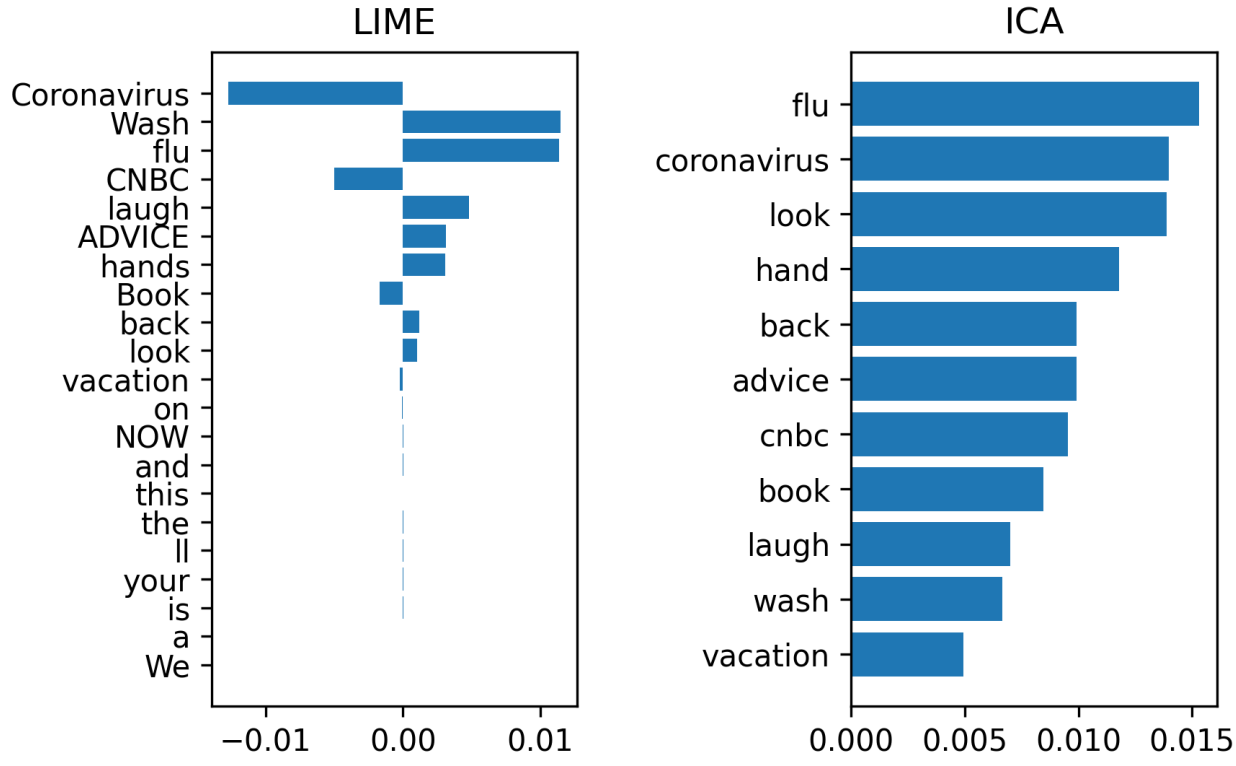


Figure 3: Comparing LIME and ICA explainability.

# Results

# One-class Classification

# Binary Classification

Table 2: Binary SVM Results

| Dimensions | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|
| 50 | 0.903 | 0.804 | 0.801 | 0.818 | 0.804 |
| 100 | 0.911 | 0.796 | 0.793 | 0.817 | 0.796 |
| 150 | 0.906 | 0.795 | 0.792 | 0.810 | 0.795 |
| 200 | 0.901 | 0.800 | 0.798 | 0.815 | 0.800 |
| 250 | 0.904 | 0.807 | 0.804 | 0.827 | 0.807 |
| 500 | 0.908 | 0.789 | 0.785 | 0.814 | 0.789 |

# Conclusion

Future work:

- local ICA explainability
- different word embeddings (e.g., BERT)
- different classifiers (e.g., neural net)

# Slide with R Output

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

# Slide with Plot