AUTOMATED DETECTION OF MISINFORMATION IN TWEETS ABOUT COVID-19

A PREPRINT

Caitlin Moroney *
Department of Mathematics and Statistics
American University
Washington, DC 20016
cm0246b@american.edu

November 28, 2020

ABSTRACT

Enter the text of your abstract here.

***K*eywords** COVID-19 · coronavirus · NLP · machine learning · misinformation

1 Introduction

Insert introduction text here.

The rest of the paper is organized as follows: Section 2 discusses our methodology; Section 3 presents the results of our experiments; Section 4 discusses our results and plans for future work.

2 Methodology

This paper explores a series of methods which seek to exploit linguistic features in raw text data in order to perform the automated detection of unreliable tweets. The experiments follow the same general framework with certain implementation details tweaked for each experiment. The first step in this framework is the application of NLP featurization methods to the raw tweet text. While different featurization methods are compared, all methods involve the use of NLP tools to represent the text with numeric features. Subsequently, latent variable methods are employed to reduce the dimensionality of the resulting X feature matrix (as well as to uncover latent variables). The latent variables are then used in the classification task. Finally, we evaluate the classification algorithms paired with the featurization methods with respect to performance and explainability. We use the typical performance metrics, including accuracy, F-score, precision, recall, and ROC-AUC. We use LIME to evaluate local explainability for non-interpretable methods. Furthermore, we present a new explainability framework for latent variable methods as well as a new explainability metric. Below, we explore in greater detail the methods used for featurization, latent variables, classification, and evaluation of explainability.

2.1 NLP featurization

In order to obtain features from the raw tweet text, we first employed standard preprocessing, to include removing stop words and punctuation as well as lemmatizing words. The two approaches we pursued involved (1) Bag-of-Words and (2) word embeddings, each followed by latent variable methods.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Optional.

2.1.1 Bag-of-Words

- Bag-of-Words + topic modeling
  - Vectorization: raw counts, tf-idf, binary
  - PCA + ICA

2.1.2 Word embeddings

- Word embeddings + topic modeling of word-context co-occurrence matrix
  - My embeddings
    - Laplace smoothing
    - raw counts, PMI, PPMI
    - PCA + ICA
  - BERT

2.2 Latent variable methods

- Topic modeling
- Word embeddings

2.3 Classification

- One-class SVM
- Binary SVM

2.4 Evaluation

- Performance
- Explainability
  - LIME (local)
  - ICA
    - global
    - local
  - Metric I came up with

3 Results

Report evaluation metrics for...

- One-class SVM
- Binary SVM

4 Discussion

- Interpret results
- Future work
  - Better ICA explainability tool?
  - Better explainability metric?
  - Other classifiers (e.g., neural nets)
  - Incorporate other features:
    - Part-of-speech tag counts
    - Punctuation counts

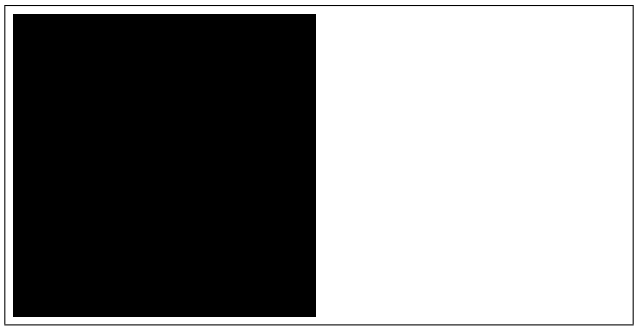


Figure 1: Sample figure caption.

* Use of all-caps
* Sentiment analysis

Misc. equation.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}$$

Paragraph Misc. text.

5 Examples of citations, figures, tables, references

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

some text (Kour and Saabne 2014b, 2014a) and see Hadash et al. (2018).

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

`https://www.ctan.org/pkg/booktabs`

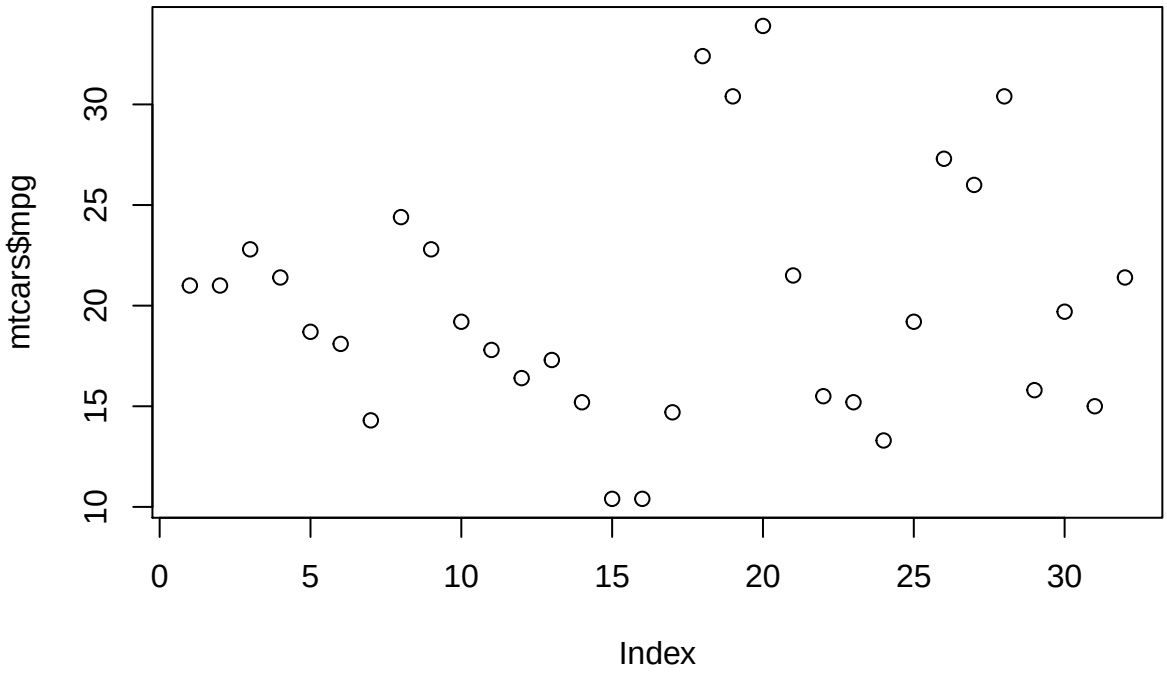
5.1 Figures

Misc. text. See Figure 1. Here is how you add footnotes. [^Sample of the first footnote.]

```
plot(mtcars$mpg)
```

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6



5.2 Tables

Misc. text.

See awesome Table~1.

References

Hadash, Guy, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. 2018. "Estimate and Replace: A Novel Approach to Integrating Deep Neural Networks with Existing Applications." *arXiv Preprint arXiv:1804.09028*.

Kour, George, and Raid Saabne. 2014a. "Fast Classification of Handwritten on-Line Arabic Characters." In *Soft Computing and Pattern Recognition (Socpar), 2014 6th International Conference of*, 312–18. IEEE.

———. 2014b. "Real-Time Segmentation of on-Line Handwritten Arabic Script." In *Frontiers in Handwriting Recognition (Icfhr), 2014 14th International Conference on*, 417–22. IEEE.