Model Info Sheet

**Section 1: Information about paper or report**

1) Author(s): Names of the authors of the paper or report

Lia Bozarth, Aparajita Saraf, Ceren Budak

2) Title of the paper or report which introduces the model

*Higher Ground? How Ground-truth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees*

3) DOI or permanent link to the paper or report (for example, link to arxiv.org webpage)

https://ojs.aaai.org/index.php/ICWSM/article/view/7278/7132

4) License: Under which license(s) are the data and/or model shared?

5) Email address of the corresponding author

lbozarth,apsaraf,cbudak@umich.edu

**Section 2: Scientific claim(s) of interest**

6) Does your paper make a generalizable claim based on the ML model? If yes, what is the scientific claim? For example, “Our ML model can be used to diagnose Covid-19 using chest radiographs of adult patients”.

*No – on the contrary, specificity is kind of the point of this paper. Authors are chiefly interested in how choice of ‘ground truth’ (fact-y) reference impacts perceived prevalence of misinformation on known fake news sites.*

7) Is the scientific claim made about a distribution or population from which you can sample? If yes: (a) what is the population or distribution about which the scientific claim is being made? (b) What is the sample used for the study? For example, “(a) Population: adult patients with symptoms of Covid-19. (b) Sample: We use a random sample of adult patients who present at a U.S. based hospital between April 2020 and June 2020”.

*Yes – news sites of interest clearly described.*

8) Does the scientific claim only apply to certain subsets of the distribution mentioned in Q6? For example, “Our model works on chest radiographs of U.S.-based adult patients and might not generalize to radiographs taken in other places or using different machines.”

*Yes – correlation scores specific to which pairs of real news / fake news sites are used and time period of interest.*

**Section 3: Train-test split is maintained across all steps in creating the model**

9) Train-test split type: How was the dataset split into train and test sets? (For example, cross-validation; separate train and test sets).

*Topic modeling step: unclear how news articles were chosen for both selection of seed words (base LDA) and guided LDA (using those seed words), though we do have some sense of how large this corpus is.*

10) Are there duplicates in the dataset? If yes, explain how duplicates are handled to ensure the train-test split.

*Unclear if there is e.g. syndication across news sites (esp. for legitimate news sites).*

11) In case the dataset has dependencies (e.g., multiple rows of data from the same patient), describe how the dependencies were addressed (for example, using block-cross validation).

*Authorship, geography, etc aren’t addressed (though, in some senses, these dependencies are what make the specificity of each pairwise comparison interesting).*

12) List all the pre-processing steps used in creating your model. For example, imputing missing data, normalizing feature values, selecting a subset of rows from the dataset for building the model.

*For the LDA step, news articles were lemmatized and stemmed, and stop words were removed. All news articles with < 100 or > 800 word tokens were removed from the data set.*

13) How was the train-test split observed during each pre-processing step? If applicable, use a separate line for each step mentioned in Q12.

*Unclear (though this is likely captured in the validation step the authors mention).*

14) List all the modeling steps used in creating your model. For example, feature selection, parameter tuning, model selection.

*List curation and selection, domain identification and comparison across sites, keyword selection, topic modeling, manual labeling (where needed)*

15) How was the train-test split observed during each modeling step? If applicable, use a separate line for each step mentioned in Q14.

*Again, a bit murky – I think a lot of this is assumed in standard LDA procedure.*

16) List all the evaluation steps used in evaluating model performance. For example, cross-validation, out-of-sample testing.

17) How was the train-test split observed during each evaluation step? If applicable, use a separate line for each step mentioned in Q16.

**Section 4: Test set is drawn from the distribution of scientific interest.**

18) Why is your test set representative of the population or distribution about which you are making your scientific claims?

*b/c claims are, in this case, tied to these sites by design! i.e. this is a meta-analysis of how choice of ground-truth sites affects perceived misinfo prevalence.*

19) Explain the process for selecting the test set and why this does not introduce selection bias in the learning process.

*Test sets are those lists of fake and real news websites already identified by some well-known publications – seems fairly likely that there might be a bias towards hot-button issues (lots of fake news articles on political topics).*

20) In case your model is used to predict a future outcome of interest using past data, detail how data in the training set is always from a date earlier than the data in the test set.

*The authors aren’t necessarily making predictions using the model trained on these datasets – however, it is a bit unclear to me how articles might be grouped within ‘domains,’ and how these groupings might affect chronology.*

**Section 5:** **Each feature used in the model is legitimate for the task**

21) List the features used in the model, alongside an argument for their legitimacy. A legitimate feature is one that would be available when the model is used in the real world and is not a proxy of the outcome being predicted. You can also include this list in an appendix and reference the relevant section of your Appendix here.

**Section 6: Findings from quantitative analysis**

Note: this paper was less similar to the podcast misinfo study than I’d previously thought, though we can glean some useful insights from its methodology. Some findings to highlight:

* Agenda-setting priorities (did fake vs real news sites have significantly different levels of coverage of certain topics?) – *Student’s t-test.* More coverage of sex scandals, conspiracy theories on fake news sites; less coverage of climate and economy
* Groundtruth difference (how do topics contribute to variance in agenda-setting across groundtruth pairs?) – PCA. Unshockingly, fake news sites assign higher priority to sex scandals and controversial issues + conspiracy theories
* List expansion (does increasing keyword set impact downstream analysis?) – authors collected tweets referencing ‘hillary’ in addition to those referencing ‘clinton’...lol. no significant difference observed