Model Info Sheet

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

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

*Baly et al. (2018)*

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

*Predicting Factuality of Reporting and Bias of News Media Sources*

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

https://drive.google.com/file/d/1Cx4H4l4\_c7sdc7srqwZeBG0Ns6MCU3wT/view?usp=sharing

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

5) Email address of the corresponding author

**pnakov@qf.org.qa**

**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”.

*The experimental results show sizable performance gains over the baselines, and confirm the importance of each feature type.*

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 – text corpus is made available, and appears to be mostly news sites of varying political persuasions. Features are also made available in Github (you could conceivably perform their analysis on any set of URLs). Two datasets – containing 860 and 1067 news sites, respectively – are also available on Github.*

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.”

*Evaluation works for the dataset of 1067 fact-checked websites, no out-of-sample testing mentioned in paper.*

**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).

*Shockingly little detail about test-train split or hyperparameter tuning – the most the authors say is ‘we used the above features in a Support Vector Machine classifier, training a separate model for factuality and for bias.’ Also: ‘we tuned the SVM hyper-paremeters … using an internal cross-validation on the training set and optimizing macro-averaged F\_1.’*

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

*No duplicate URLs within the set of 1067 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).

*These are all sites for which bias and accuracy data are available on mediabiasfactcheck.com. I’m of two minds about their decision to choose sites that have already been evaluated:*

***good*** *because there are ground-truth evals to which they can compare their classifier results;*

***bad*** *because this bias / accuracy eval literally already exists for these sites! Unclear, then, if the evals they find (or even the dataset) are novel. Also, these news sites are well-known enough for these evals to exist on a third-party site. This might skew the contribution of things like the ‘has page’ feature in the ablation study*

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.

*Very little detail about this*

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.

*Again – very little detail about this!*

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

*Features were selected from multiple sources. There are 141 in all, and it is unclear how these were normalized (some of them appear to be booleans – does the source have a wikipedia page or not? – while others are not, i.e. the scaled eval of how far left or right a source is). The authors claim to average over all features for all articles, which seems bad for 1) oversmoothing, and 2) ignoring important differences in scale and weighting.*

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

*Very little detail about this*

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

*Ablation study was a rare methodological bright spot in the paper, though nature of the dataset used (relatively well-known sites) might mean that its results lend limited insight into actual effectiveness of classifier*

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

*Very little detail about this*

**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?

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

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.

**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.

For example, “Patient age: We include this feature in our ML model for hypertension diagnosis since patient age is easily available in a clinical setting”.

An example of a feature that should not be included (for illustration only; you do not need to include these in your model info sheet): “Anti-hypertensive drugs: We do not include the use of anti-hypertensive drugs as a feature in our ML model for hypertension diagnosis since that information is only available after diagnosis and would not be available when a new patient presents with symptoms of hypertension.”

***Note:*** *You do not need to list each feature used in your model here. However, you must provide an argument for the legitimacy of each feature included in your model to ensure that your model does not suffer from leakage due to illegitimate features. For example, “our model only uses data from the previous year as features. For instance, to predict civil war in 2017, we only use lagged features from the year 2016. Since these features are always available in advance of when we want to make predictions using our model, none of these features can lead to leakage.”*