Model Info Sheet

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

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

*Helmstetter, Stefan et al. (2018).*

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

*Weakly Supervised Learning for Fake News Detection on Twitter*

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

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

5) Email address of the corresponding author

**stefanhelmstetter@web.de**

**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”.

*Not a strong one – but suggests that a coarse-grained per-source labeling for large datasets might be a sufficient proxy for a finer-grained per-tweet labeling of a small dataset.*

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 – the authors describe the news sites they use and their sampling methodology, though it’s unclear if their precise dataset is still accessible (the Twitter API only returns a set number of most recent tweets for every user). They do claim that ‘the process of acquiring the dataset can be repeated, gathering a large-scale, up-to-date dataset at any time.’*

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.”

*Since the API returns a set number of tweets per account (thus making it difficult to guarantee a similar distribution of tweets per account within the same timeframe), it seems likely that the sampled tweets for different accounts had very different proportions of tweets pertaining to news stories (e.g., a sample from a very prolific account might not even include tweets from February 2017, on the earlier end of sampling time frame).*

*I wonder if the classifier was preferentially trained for news stories occurring in the later portion of the testing window, since there was likely more coverage for those months.*

**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).

*Appreciated that the 116 ‘gold-standard’ tweets – selected from a pre-labeled sample provided by politifact – were not from any of the misinfo/real news sources used in the training set.*

*To generate negative examples of fake news: the authors selected 116 tweets that were closest to fake news tweets in the positive (for fake news) class, then removed those from the training dataset.*

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

*Unclear if retweets were double-counted.*

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).

*Again, unclear if there were double-counts of retweets, or (maybe?) the edge case in which one org on the misinfo / real news list retweeted a tweet from another org on either list.*

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.

*Semi-manual curation of gold standard positive and negative misinfo tweets;*

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.

*In the case of the positive tweets, at least, authors make selections from out-of-sample sources.*

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

*Curation of fake / real news sites from existing publications;*

*development of positive and negative examples of fake news;*

*feature selection (user- and non-user specific);*

*training (five different models);*

*cross-validation on training set or validation against gold standard*

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

*Some of this info is elided (probably assumed in the process of working with well–known models).*

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

*Cross-validation in two settings: firstly, on noisy training set. Secondly, training on training set and validating against gold standard dataset.*

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

*Training set came from out-of-sample sources.*

**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?

*Assuming that the population is the set of websites and associated twitter accounts tested here, yes.*

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

*Bias very possible, given that these news sets were curated years ago and are probably mostly well-known ones (if they’re still around).*

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.

*Unclear what the temporal relationship is between gold standard tweets and training set tweets.*

**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.

*Feature categories:*

*user-level features – pulled from user info generated by API*

*tweet-level features – pulled from API; very time-sensitive feats omitted (likes, retweets)*

*text features – BOW and TF-IDF*

*topic features – LDA and HDP*

*sentiment features – TextBlob and SentiWordNet*