Model Info Sheet

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

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

*Zeng, Li et al (2016)*

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

*#Unconfirmed: Classifying Rumor Stance in Crisis-Related Social Media Messages*

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

**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 terribly general one – just that their classifier achieves 88% accuracy in detecting rumor stance (i.e., promoting or debunking rumor statements) over the specified dataset (a manually-labeled dataset of some 4K tweets).*

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 clearly describe their sampling process. In greater detail: they filtered tweets retrieved by the Twitter Streaming API for rumors about the Sydney hostage crisis. They organized those tweets into five different rumor narratives, then manually assigned each tweet to one of three mutually-exclusive categories (for, against, neutral). Further details about their manual curation process in 12). Would personally like to know more about the process by which ‘a comprehensive, low-noise sample of tweets related to a particular rumor story’ was extracted from the larger tweet set, and whether or not this process introduced any bias into final tweet set.*

*The population is the corpus of all tweets available through the Twitter Streaming API over the course of the study. (A more stringent defn of the study population might be the set of all tweets pertaining to the Sydney hostage crisis, pre- filtering by rumor narrative.)*

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.”

*Random forest classifier seems to do better than other classifiers on both pooled and rumor-specific datasets; the pooled model seems to outperform baselines(?) on stance prediction.*

**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).

*Unspecified? But possibly captured by canonical procedure for X-fold cross-validation*

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

*So as to avoid bias towards more popular tweets, exact duplicates (assuming this means retweets?) were removed from the dataset.*

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).

*Duplicates not allowed, and were deleted from the dataset. Unsure if quote tweets were included in this count.*

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.

*Authors equalized size of affirm/deny sets before model training (without this step, ‘affirm’ test sets would’ve been disproportionately large) s.t. baseline accuracy would be 50/50. Authors stemmed tweet text.*

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.

*Unspecified, though possibly captured by canonical procedure for X-fold cross-validation*

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

*Feature selection (more details in 21), pre-processing on data (equalizing sizes of affirm/deny test sets, pruning dataset for tweets aligning with one of five rumor narratives), model selection (logistic regression, Gaussian naive Bayes, random forest), 10x cross-validation.*

*Some model-tuning to determine optimal number of trees (authors settled on 30).*

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

*Unspecified, though possibly captured by canonical procedure for X-fold cross-validation*

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

*Cross-validation over 10 folds mentioned.*

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

*Cross-validation over 10 folds mentioned.*

**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?

*Yes – test set is appropriate for the rumor in question. Hard to say if the model employed by authors is generalizable to other rumors, or other social media platforms.*

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

*Selection process possibly introduces biases, since a specific set of keywords / queries were used to surface pertinent tweets, and authors seems to have specifically chosen rumor narratives for which a large corpus of tweets with discernible rumor stance was available (per the paper: ‘queries are designed to produce a comprehensive, low-noise sample of tweets related to a particular rumor story’), unclear if more popular rumors are also more likely to be more outlandish, or more likely to be affirmed/denied on the platform).*

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.

*Data are tweets pertaining to a long past event (Sydney hostage crisis), so this is true for any analysis done on current events / anything post-December 2014.*

**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.

part-of-speech features – use a Tweet NLP system for POS tagging *[authors identified POS tagging – particularly negation patterns – as most indicative of rumoring stance across all five rumor narratives]*

punctuation – authors extract number of exclam points, q marks, and interrobangs (should this be normalized w tweet length?)

twitter-element features – hashtags and external links (should maybe also include multimedia, such as vids and photos?)

LIWC features – style analysis: includes negation, swear words, negative emotion words, personal pronouns, etc

tweet sentiment features – authors used external sentiment classifier from MetaMind

N-grams – authors stemmed and lowercase words; also apply minimum freq threshold (tho don’t specify what this is)

*Again, unclear if these features will be generalizable to other rumors – even within the dataset of all tweets related to the Sydney hostage crisis, different narratives in the set of five prioritized different features.*

add’l notes:

early in life cycle -- rumor tends to get more traction

<https://dl.acm.org/doi/pdf/10.1145/2818048.2819964>

^ would like to know HOW dataset was constructed ??

email emma spiro + kate starbird about this dataset

nice to have: cases of failure; more detail about actual model training and testing