Model Info Sheet for Detecting and Preventing Leakage in ML-based Science (beta)

**About model info sheets**

Completing this model info sheet requires the researcher to provide precise arguments to justify that predictive models used for making scientific claims do not suffer from leakage. It is inspired by the model cards introduced by Mitchell et al.[[1]](#footnote-1)

Model info sheets are intended to accompany the paper or report that introduces the model: for instance, as an appendix or supplemental material.

This is a beta version of our model info sheet template. We are soliciting feedback and will continue to update the template. For more information about model info sheets and to obtain the latest version of the template, see [reproducible.cs.princeton.edu](https://reproducible.cs.princeton.edu/). For feedback or questions, contact: [sayashk@princeton.edu](mailto:sayashk@princeton.edu)

The model info sheet starts on the next page. After filling it out, save it starting from that page. To cite the paper that introduces the model info sheets, use the bibliography file available at [reproducible.cs.princeton.edu/citation.bib](https://reproducible.cs.princeton.edu/citation.bib)

Model Info Sheet

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

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

*M.J. Wubbels*

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

*Comparison of Neural Networks and Word Embeddings for Sentiment Prediction*

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

*N.A.*

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

5) Email address of the corresponding author

*m.j.wubbels@tilburguniversity.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”.

If there are multiple claims, list each claim in a new line, along with a claim number.

*These models can be used to predict sentiment on customer reviews from Amazon.*

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”.

If there are multiple scientific claims, list your answer for each claim in a new line, corresponding to their claim number in Q6.

***Note:*** *A difference between the population and the set from which the sample is drawn could highlight potential generalizability failures, which are related to but distinct from leakage.*

*Customer reviews on Amazon gathered in 2018.*

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.”

If there are multiple claims, list your answer for each claim in a new line, corresponding to their claim number in Q6.

*The models work on customer reviews from Amazon on which they were trained and tested, their generalizability is therefore unknown.*

**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).

*Separate train and test sets.*

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

*There were no duplicates in 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).

*The dataset did not contain dependencies.*

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.

*Rows with missing text data were removed.*

*Ratings were transformed: 1 or 2 stars was considered “negative”, 4 or 5 stars was considered “positive.*

*Text data was cleaned: special characters and punctuation were removed. The text was made lowercase and stop words were removed.*

*Data was split in train-test.*

*The words were tokenized and padded to a length of 100 tokens.*

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.

*The data was split after removing missing data and cleaning the text, since this would be the exact same procedure and would have the exact same outcome before or after splitting.*

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

*Model training, parameter tuning, and model evaluation.*

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

*The data was further split before any of the modelling steps, into training and validation data.*

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

*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.

*The data was split during each of the evaluation methods.*

**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?

*The scientific claims are based on the test set.*

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

*The test set consists of amazon reviews, which the scientific claims are based on. There are no claims being made on different populations.*

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.

*N.A.*

**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.

*Text reviews: We include this feature in the models for predicting sentiment since this the main feature the prediction is based on.*

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”.

***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.”*

1. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In *Proceedings of the conference on fairness, accountability, and transparency*, 2019. [↑](#footnote-ref-1)