Review Analysis Interview:

Subject09

**7/17/2019 11:00am**

# Summary

**Interviewee demographics**

*Male*

*Ph.D. in computer science1 year working on reviews*

*1 year of data science experience*

*Postdoctorate at Company 2*

# *Reports to other researchers*

**What are the goals of your analysis?**

*The goal is to create high-quality sentence classifiers for several tasks: whether it is funny, a fun fact, a tip, or a spoiler (in the case of media).*

*Downstream, these may be used in an application.*

**Describe a recent review analysis project.**

*Spends most of his time iteratively cleaning data and collecting crowdsourced (figure8) labels.*

*The first step is to take a random sample (uniform distribution) of tens or hundreds (out of millions) and label them himself.*

*Goals of labeling:*

* *try to find patterns in keywords that occur for positive and negative labels.* 
  + *Subject computes top-k most frequent words in two groups and compares them*
* *Make sure the task is well-defined*
* *Once he has crowd-sourced labels, he labels some himself to audit the results*

*Then he filters out sentences that are too long or too short, and further filters based on the keywords he discovered.*

*Sparsity is an issue and one of the goals of the data preparation is to make the labeling more focused*

*Suspects about half of the reviews don’t contain substance relevant to his tasks.*

**Tools**

*Python.*

*Uses jupyter notebook to compute features of sentences, and excel to groupby labels or keywords. GGplot or excel for viz.*

**Important aspects of the data (CSV)**

Tokens

stemmed tokens

word distributions

labels (his own and crowdsourced)

Votes from yelp users (funny, useful, etc)

Sentence level (not review)

**What about modeling the data?**

*Performance of the models used is not a real concern—*

Interviewer note: this could be due to the fact that the focus is application of existing pre-trained models, instead of developing new ones.

*If the classifier is failing, the problem is in the data and/or the model, and the interviewee examines the data first: it should be separable. If it’s not, either the labels or the task definition need to be fixed.*

*Interviewee prefers command-line tools to train the model so that it can run on the office server.*

**Collaboration**

*Weekly meeting with other researchers where they discuss problems. Verbally presents experiments connecting dataset characteristics to the predictions.*

**Features Wish-list**

*Easily comparing two sets of sentences*

* *By label*
* *By dataset (hotel reviews, restaurant reviews, book reviews, etc)*

*Features of interest in these datasets:*

* Word distributions (KL divergence) → how to compare them?
* *Why does the model perform differently on different datasets?*

*Tools that would make data curation (prep + labeling) for modeling easier*

*Provide me a measure (or tools) that would tell me the data prep or labeling quality is good enough to stop*

*Error analysis tools. (model misclassification).*