**Interviewee -** (Subject05, Subject03), 07/29/2019 @ 1pm

**Name**: Subject05

**Title:** Senior data scientist

**Company:** Company 3

**Education:** PhD in IEOR Operation research (industrial/mechanical engineer)

**Experience in data analysis:** since Master’s 14 yrs ago

**Experience in review analysis**: 1 year (since joining company 3)

Prior experience with image classification for deep learning 4yrs. Before that, mechanical engineering sensor data processing

**Gender** M

**Intern name**: Subject03

**Title** intern

**Company** company3

**Education** 3rd year PhD student in Electrical Engineering

**Experience in data analysis:** 3 years

**Experience in review analysis:** 1 month

No prior NLP experience

Prior experience with image classification for deep learning 4yrs. Before that, mechanical engineering sensor data processing

# Interview

## Walk us through a recent review analysis task

Most of the work on this task carried out by the intern

### Pipeline

1. Data cleaning (includes filtering), rule based foul language detection and removal of matching reviews
2. Initial manual labeling for real vs. fraudulent reviews
3. Classifier model training to detect whether the review is real or fraudulent
4. Use trained model to classify a test set of reviews (accepted reviews are sent to downstream tasks)
5. Human validation of rejected reviews. Training data is augmented with these examples

### Preparing/debugging the training data

* Filter based on length and remove inappropriate language, punctuation and stop-words (automated, rules based)
  + works well, no need for improvement
* Automoderation ML model detects and partitions fraudulent or unhelpful interviews
  + built on prior project, recall could be improved but largely it works
  + 70 features, character level fine grained information
  + supervised model using human labeling in-house

### Debugging the model

Problem: how to incorporate fraudulent reviews as negative training data

* tune parameters, try to come up with new features using intuition
* classification of features into types, some features are purely intuitive, some are based on summary statistics
* try other models
* manual review
* statistical analysis

currently looking to improve recall

### What is the goal of the analysis?

* + Improve a naive bayes + xgboost model to classify reviews as real or fraudulent
  + Explore reviews to see which variables are important in making a review helpful for employees seeking jobs
  + Figure out company culture from reviews

### What kinds of data sources and formats do you use?

* .csv file
* Review data is acquired from company’s users

### What kind of data preparation tasks?

* cleaning, remove unhelpful examples
* anonymize information about specific individuals (other than top management)

### Do you often need to look at the raw data points?

Yes, when coming up with new feature ideas

### Is scalability a problem?

* half a year’s data 200,000 reviews
* some feature-extraction scripts do not scale well
* For the most part not an issue

### What tools and languages do you use?

* + Python (for modeling)
    - Pandas
    - NLTK
    - Gensim
    - Tensorflow
    - Fast Test
    - scikit-learn
  + Java (for production)

### What are the downstream applications?

Using the review as part of a summary for a company, providing this as a service

Future modeling tasks

### How do you present your results? (To whom?)

Jupyter notebooks

#### How do you collaborate with others on your work?

Stash repository

data files stored on a server

Small team, really just one person working on the project + a supervisor working on a related project

### What are the bottlenecks you run into or things you spend the most time on? (Are they different?)

* data cleaning takes the most time
  + feature engineering
* time constraints are not significant
* Challenge: improving system recall

### What features/tools do you wish you had?

* better feature extraction
* The intern is pretty happy with what’s available
* Tracking labeling history