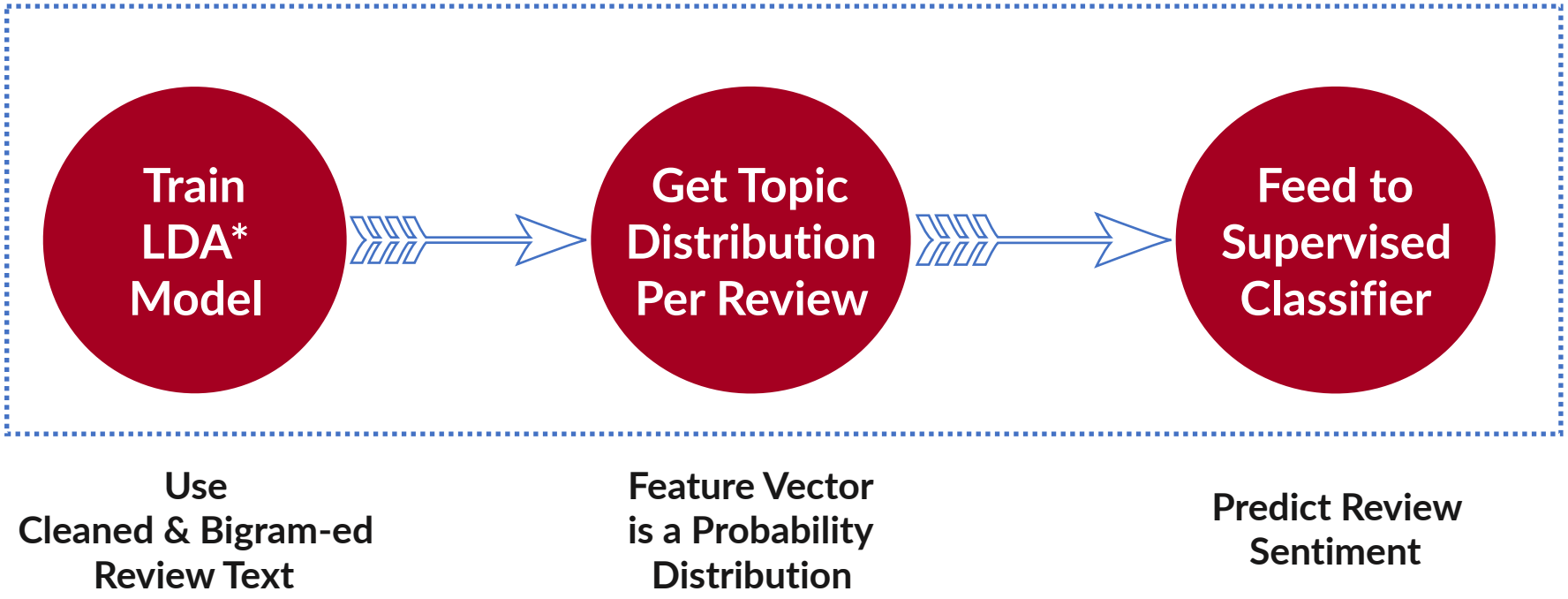
Topic Model Classification

via Yelp Reviews

Marc Kelechava

Can Topic Model distributions predict review sentiment on unseen text?

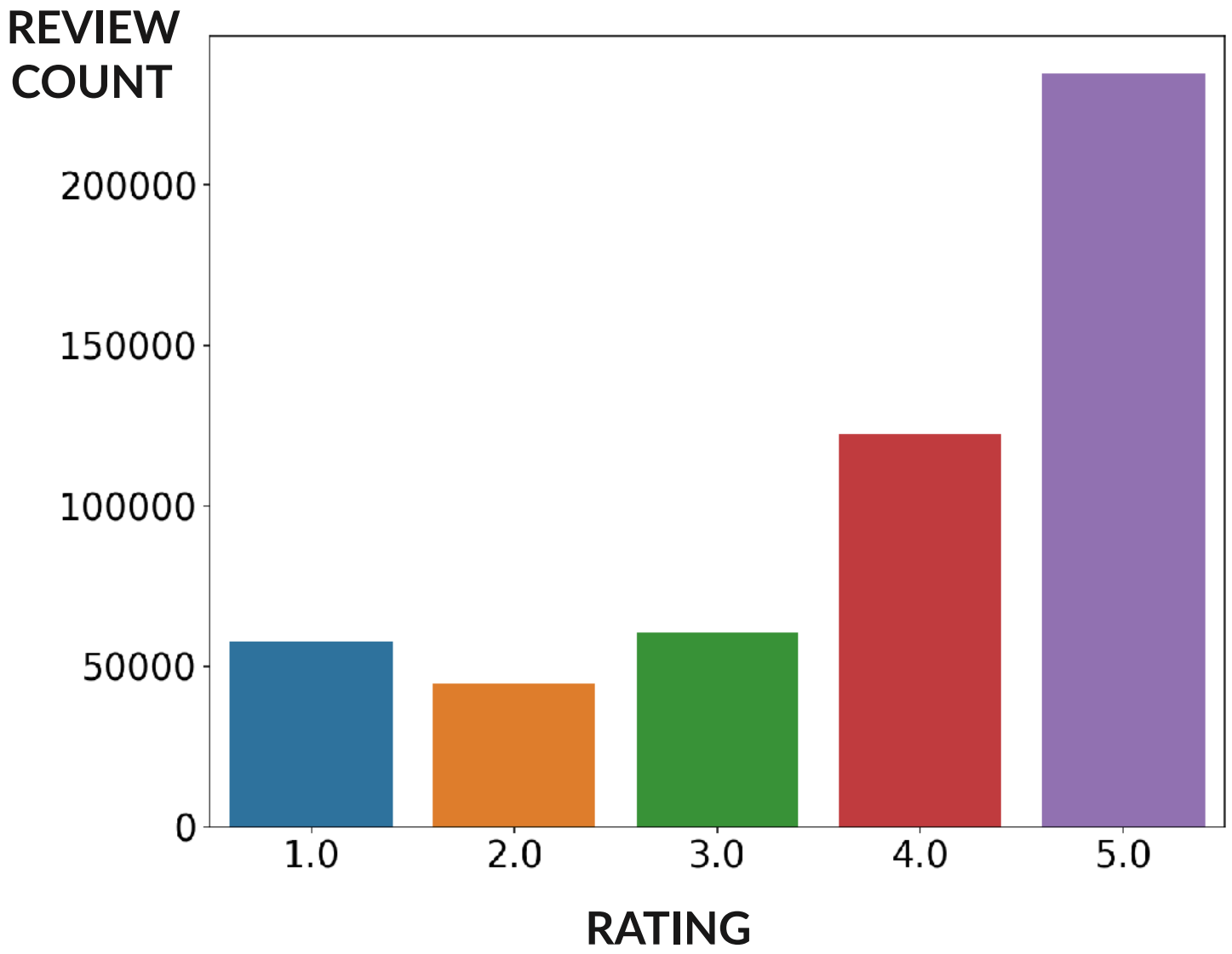
PROJECT GOAL



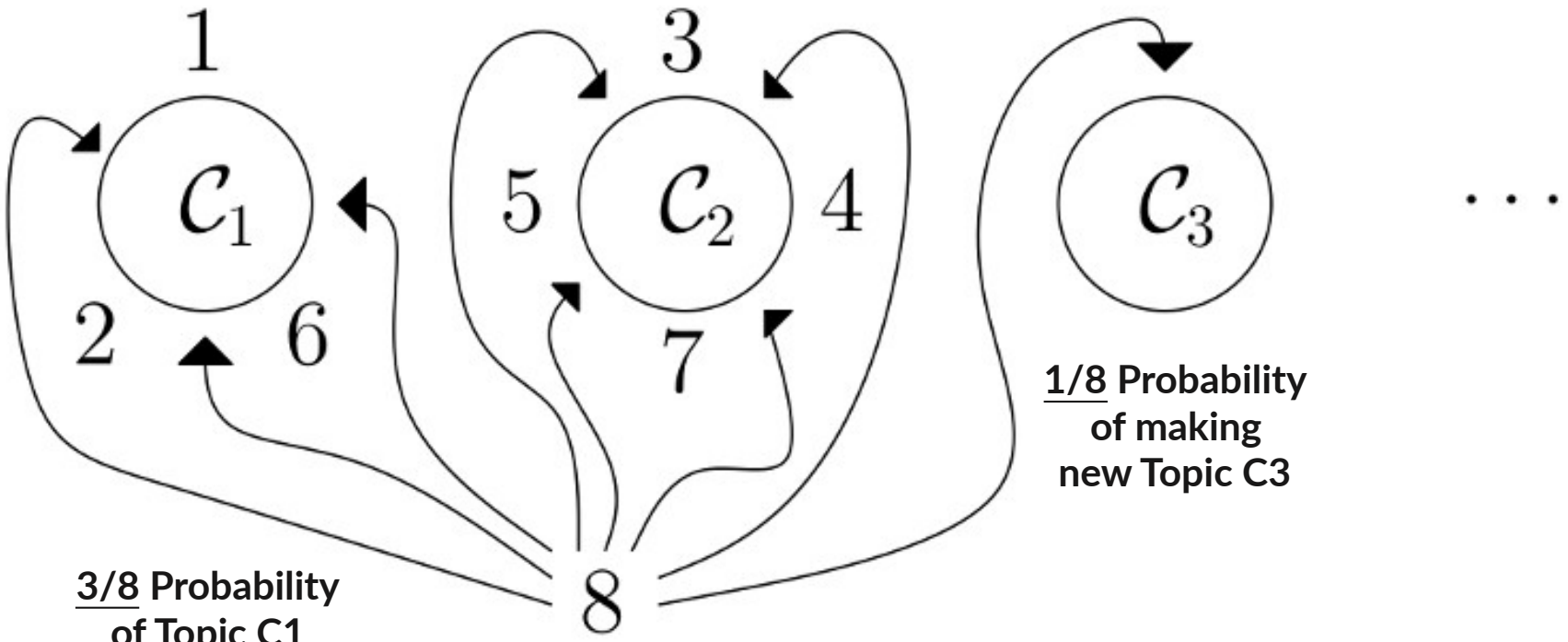
*LDA = Latent Dirichlet Allocation

2016 Review Ratings

- Heavily tilted toward 4s and 5s.
- Will correct for this imbalance when training classifiers.



How to choose the Number of Topics?



- Hierarchical Dirichlet Process does not assume fixed number of topics.
- Above is rough explanation of doc-topic grouping.

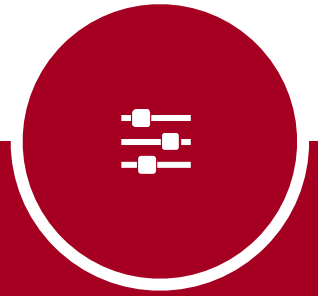
Source(s): http://gerin.perso.math.cnrs.fr, http://blog.echen.me

FIT HIERARCHICAL DIRICHLET PROCESS

Extract Topic Distributions with LDA

Train LDA Model on
Year 1 Reviews
w/ 20 Topics
(implied by HDP)

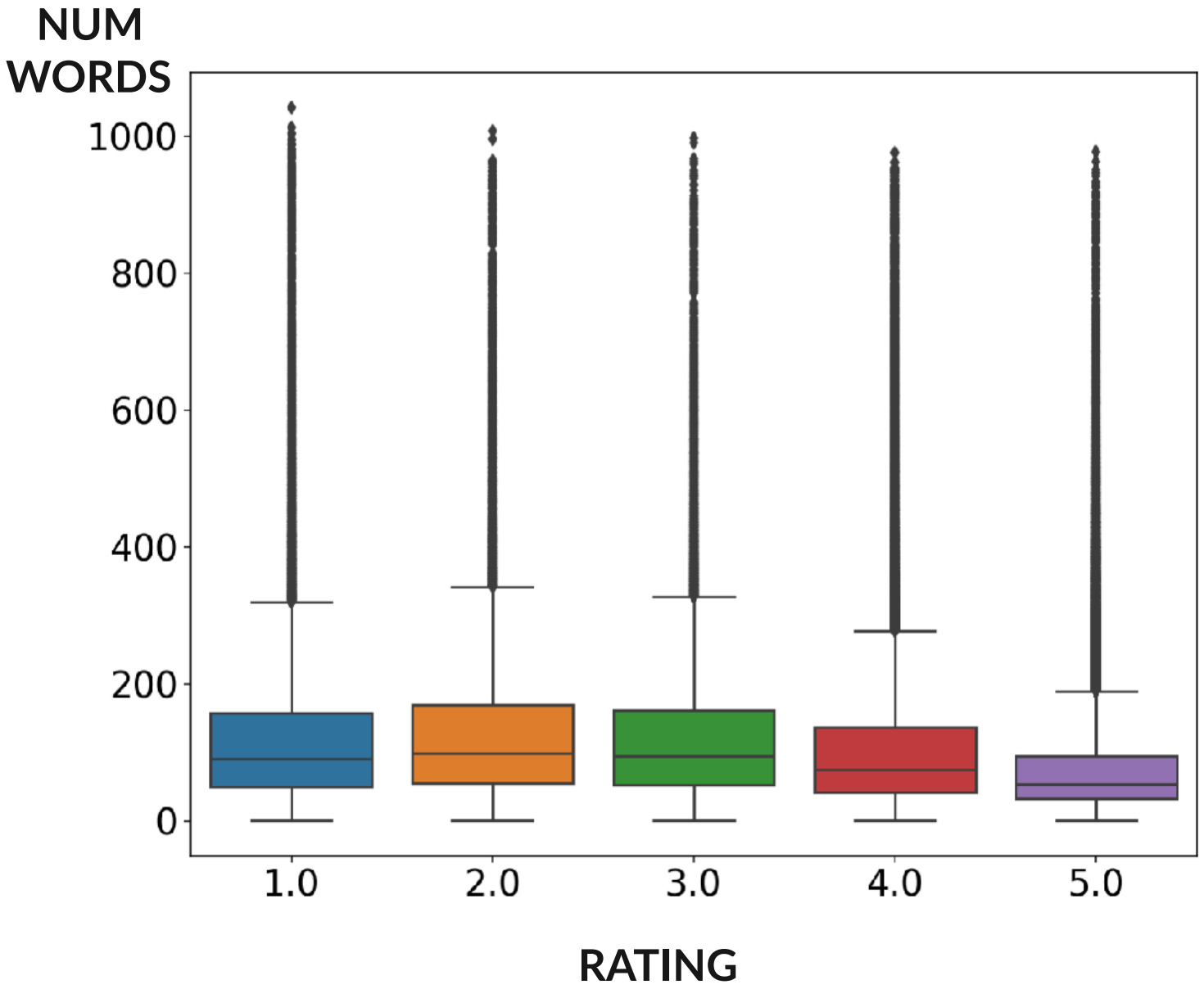
Apply Year 1 Reviews to
trained LDA Model
and get Features
(Probability Distribution
over TOPICS
for each Review)



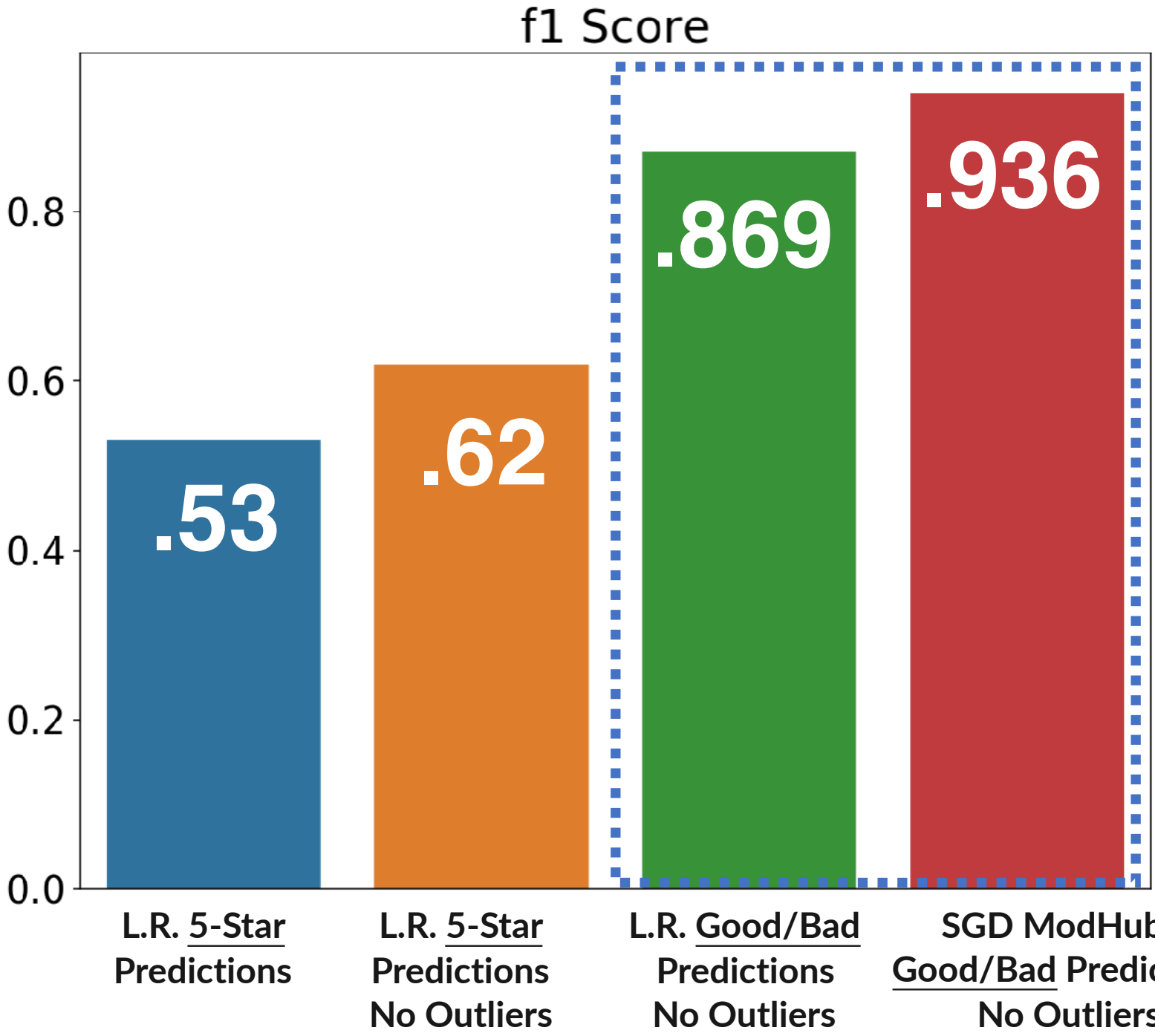
Add some hand-engineered
features and pass to
supervised classifier
CV-Loop

2016 Review Word Count

- Large outlier range.
- Removing these outliers helped immensely with classifier training.



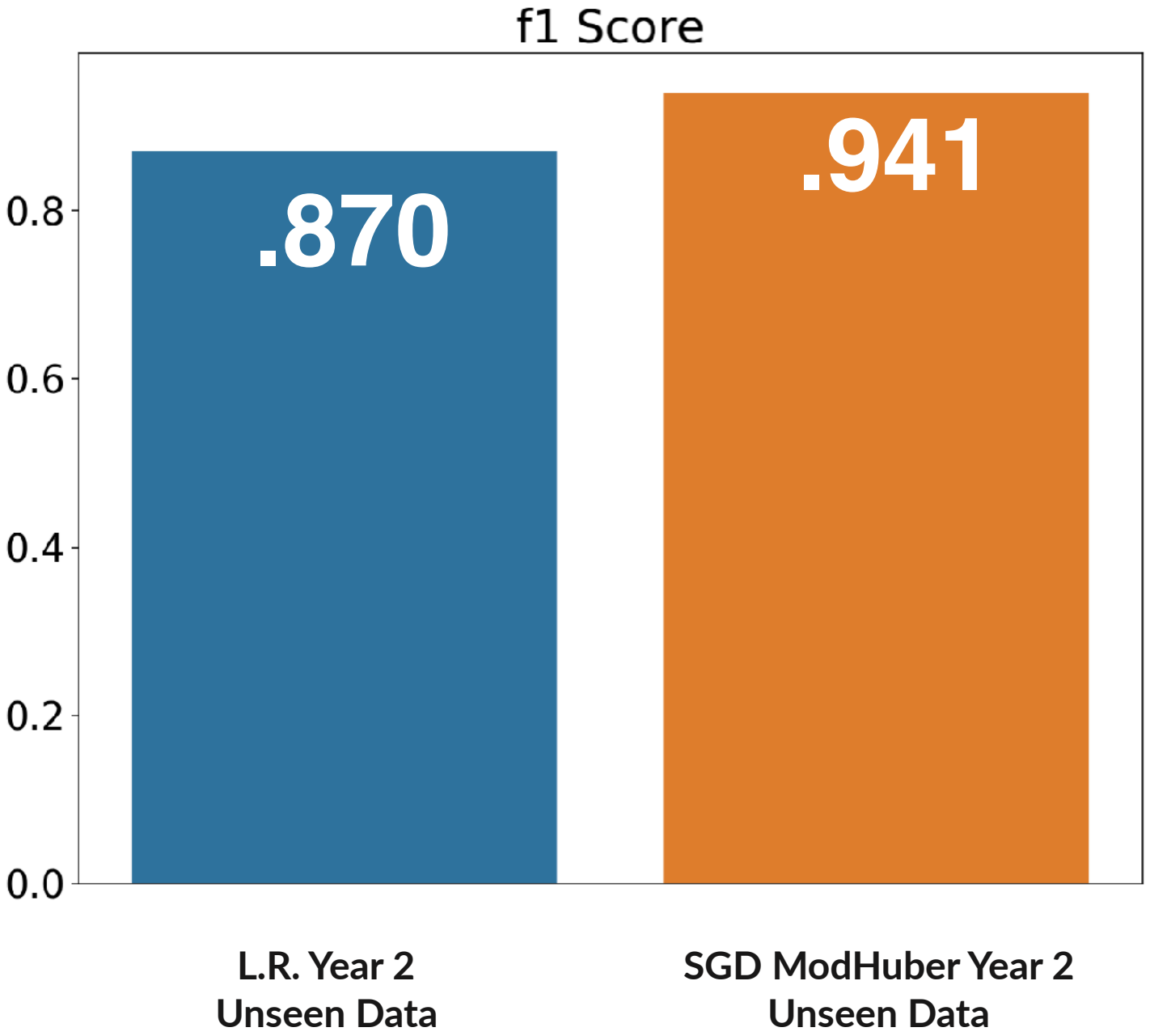
TRAINING RESULTS



5-Fold CV

- Balanced Method on Logit Regression approximates undersampling
- Stochastic Gradient Classifier jumped up when using Modified Huber loss

UNSEEN DATA TEST RESULTS



Hypothesis Testing

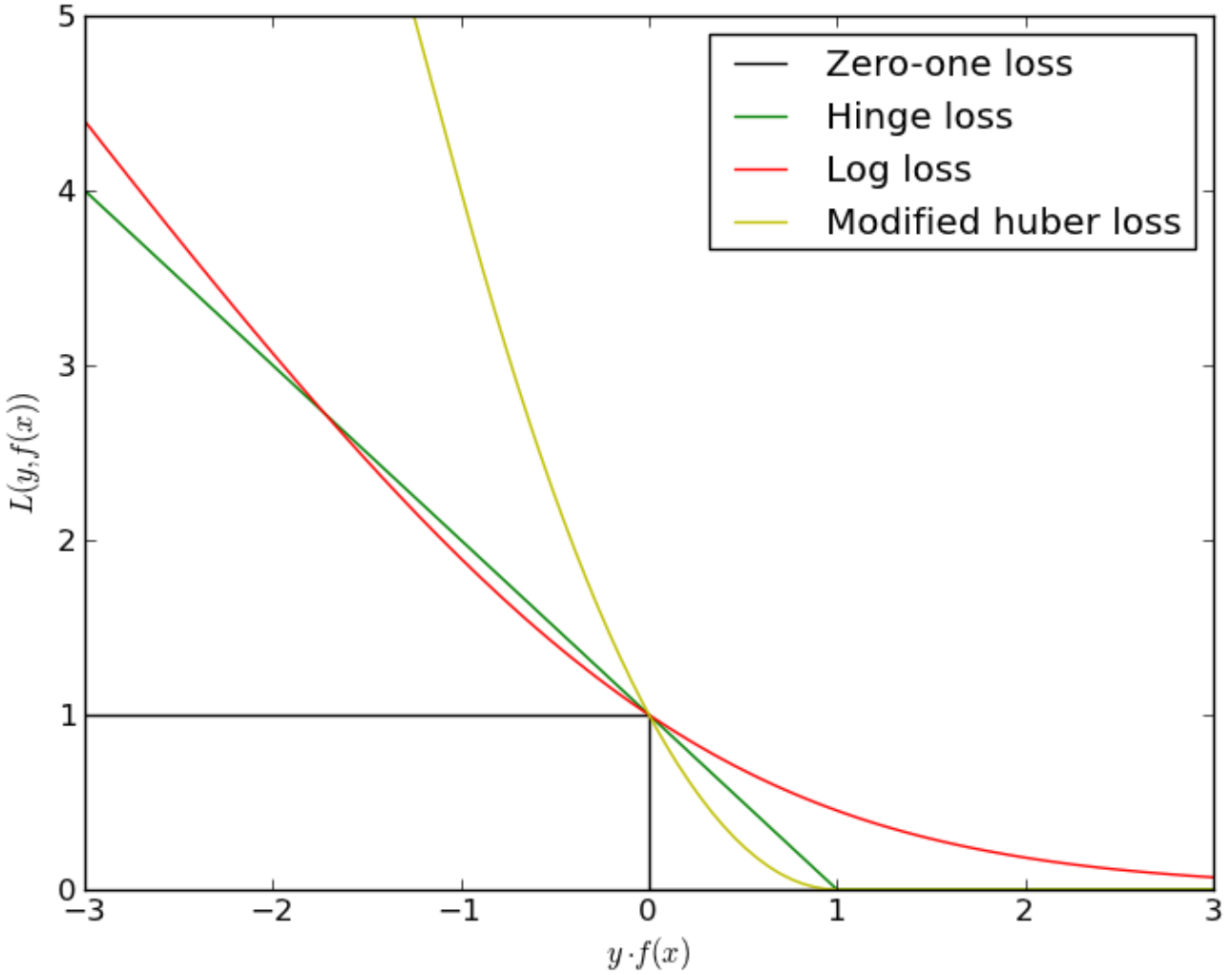
- Ran a McNemar Test (Chi-Squared for models) to see if scores are truly different
- Near 0 p-value on chi-square indicates models are truly different and SGD Mod. Huber wins.

What is Modified Huber Loss Exactly?

- Note Modified Huber punishes you more on outliers

- Suggests to me that that with one-by-one gradient descent, it learns faster to avoid misclassifications

Source: http://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/auto_examples/linear_model/plot_sgd_loss_functions.html



Future Work

Very, very slow LDA training. Would like to explore with trigrams and different lemmatization options.

More by-hand feature engineering on things like if restaurants offer takeout or not.

Extend to predict Neutral reviews as well, and also find high-dim visual projections to 2-D

Thank You!