Topic Model Classification

via Yelp Reviews

Marc Kelechava

PROJECT GOAL

Can Topic Model distributions predict review sentiment on unseen text?

Train LDA* Model → **Get Topic Distribution Per Review** → **Feed to Supervised Classifier**

Use Cleaned & Bigram-ed Review Text

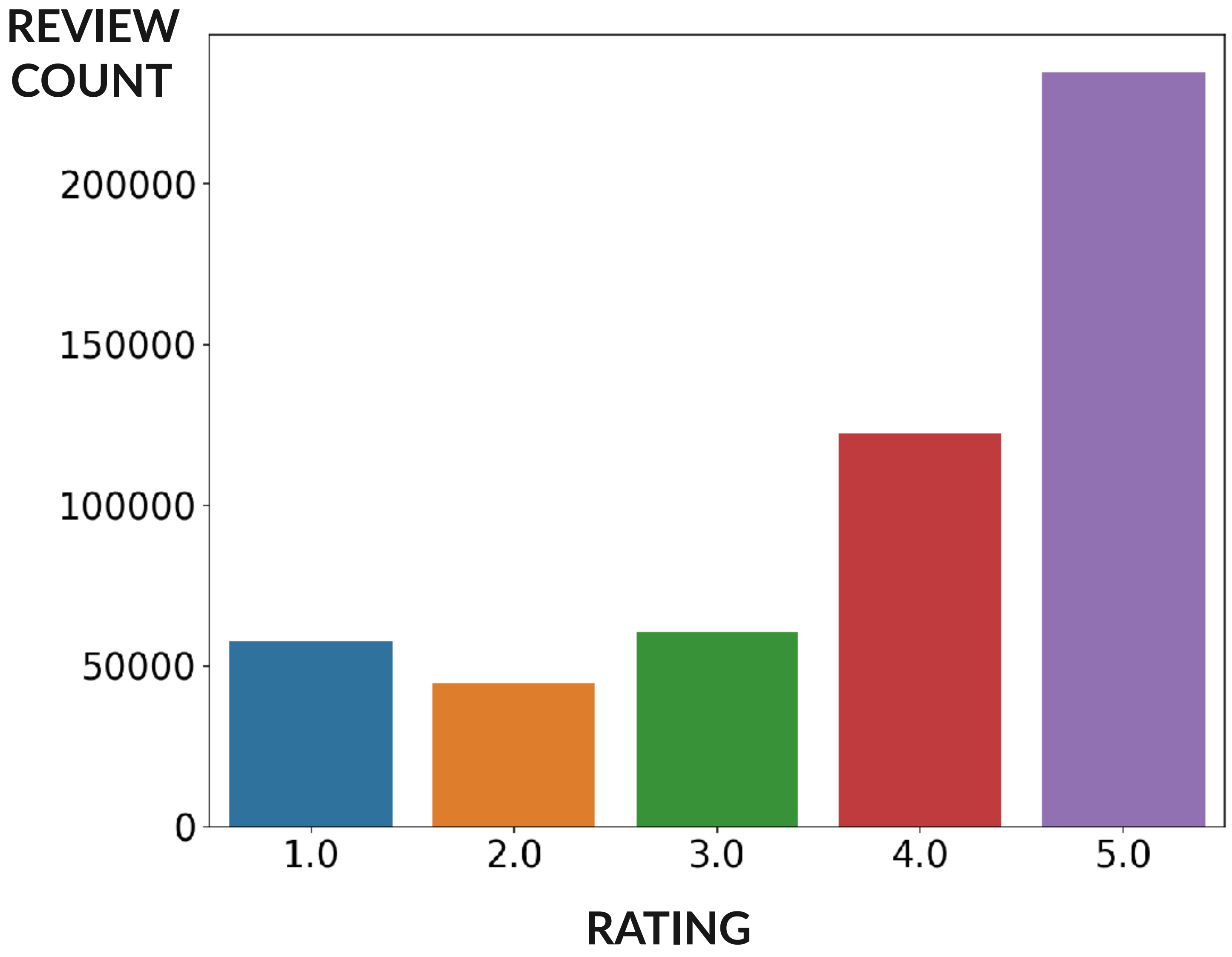
Feature Vector is a Probability Distribution

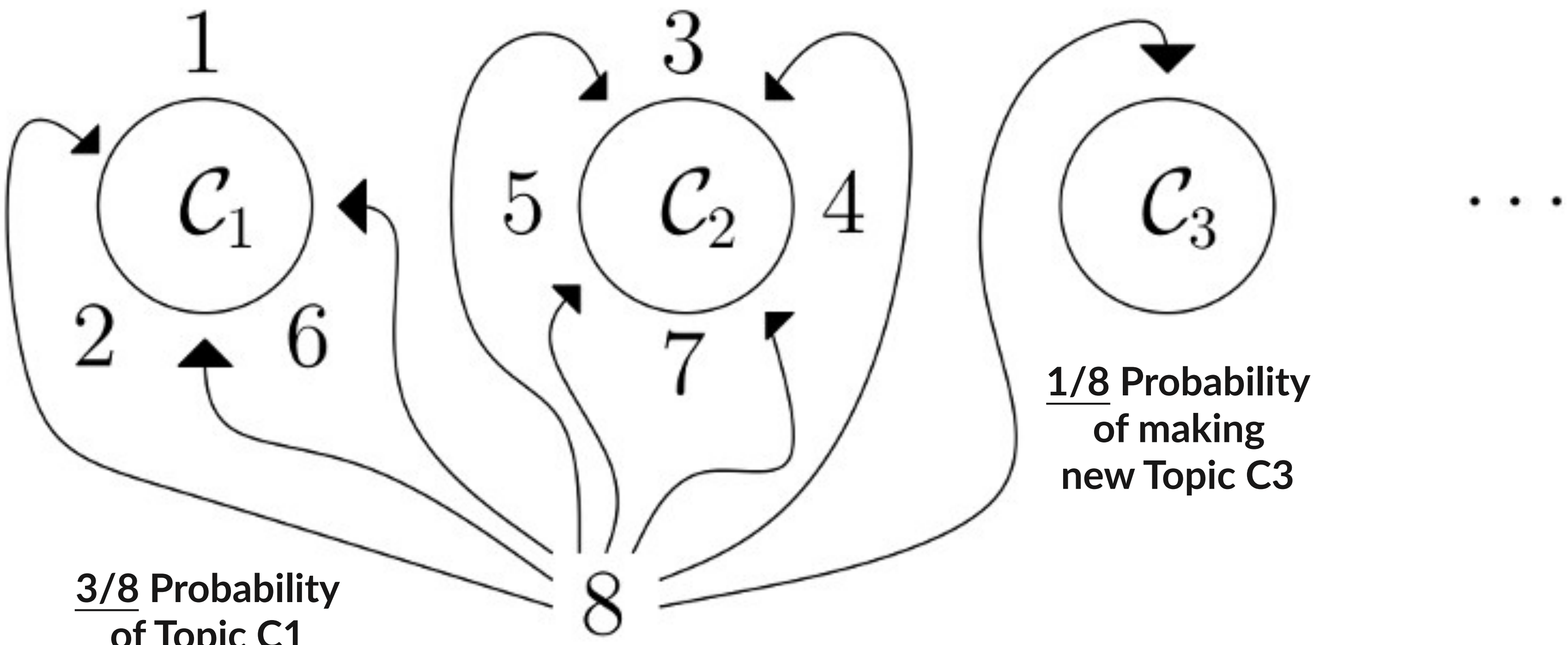
Predict Review Sentiment

*LDA = Latent Dirichlet Allocation

2016 Review Ratings

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





3/8 Probability of Topic C1

4/8 Probability of Topic C2

1/8 Probability of making new Topic C3

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

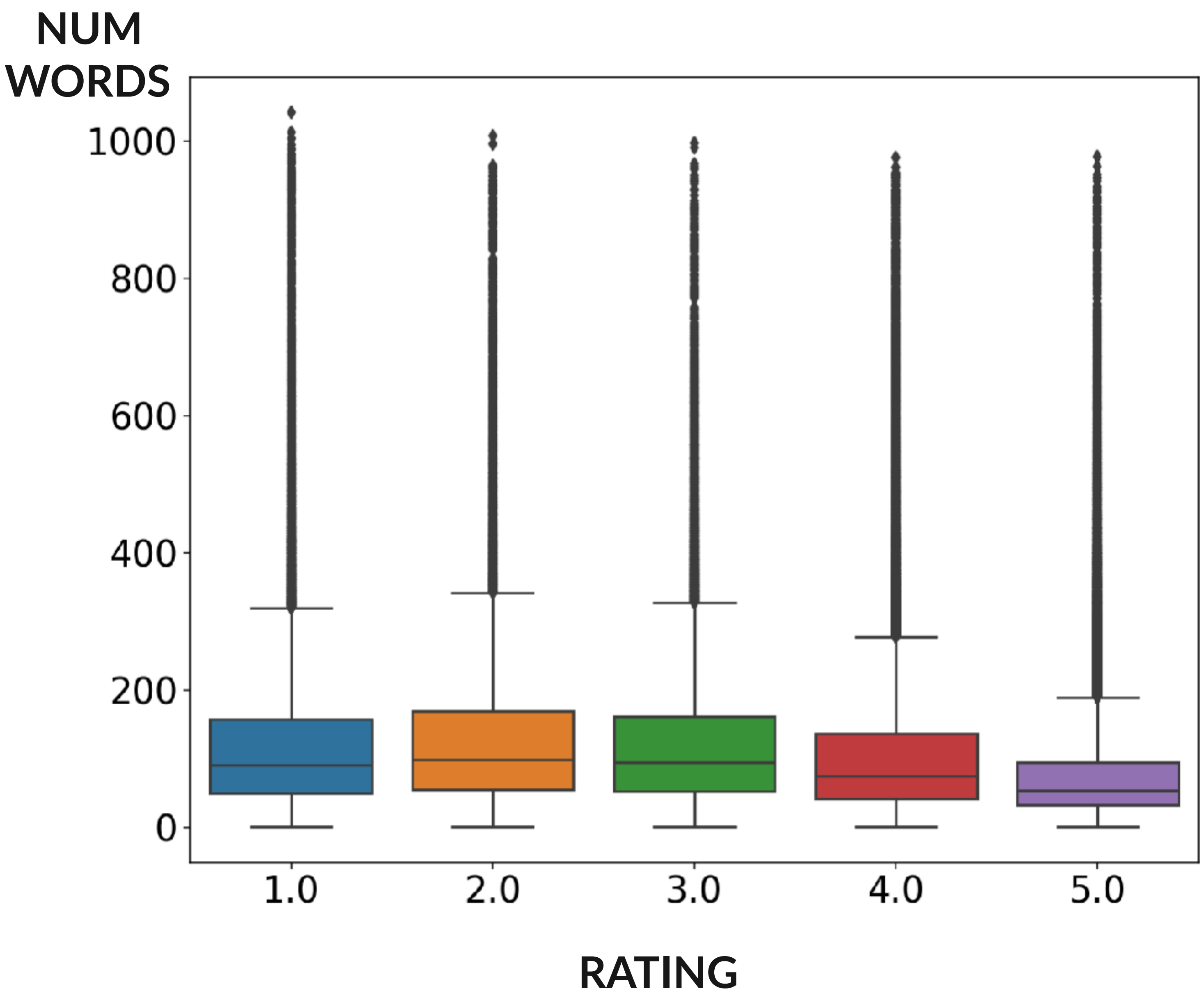
Train LDA Model on
Year 1 Reviews
w/ 20 Topics
(implied by HDP)

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

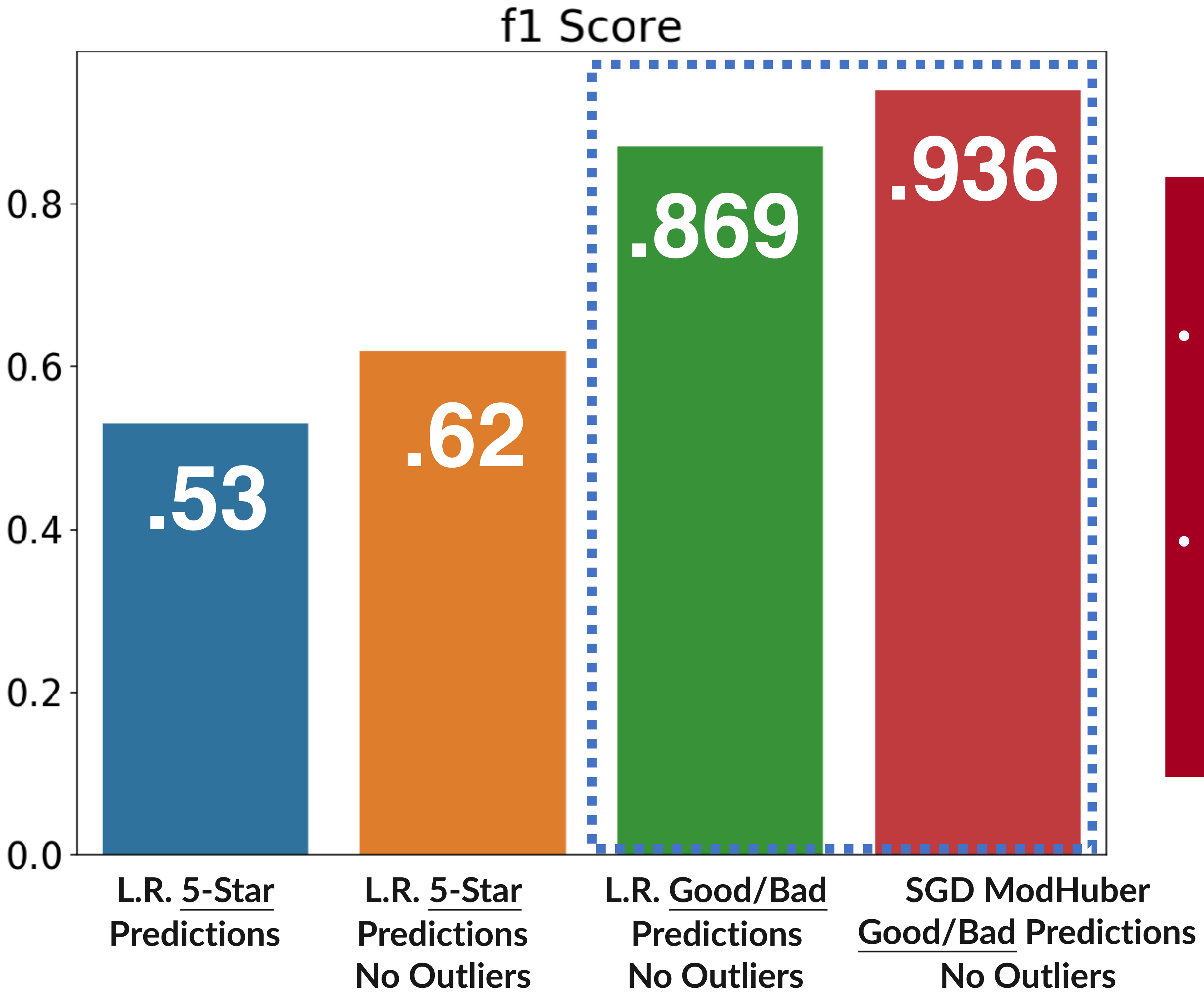
Add some hand-engineered
features and pass to
supervised classifier
CV-Loop

2016 Review Word Count

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



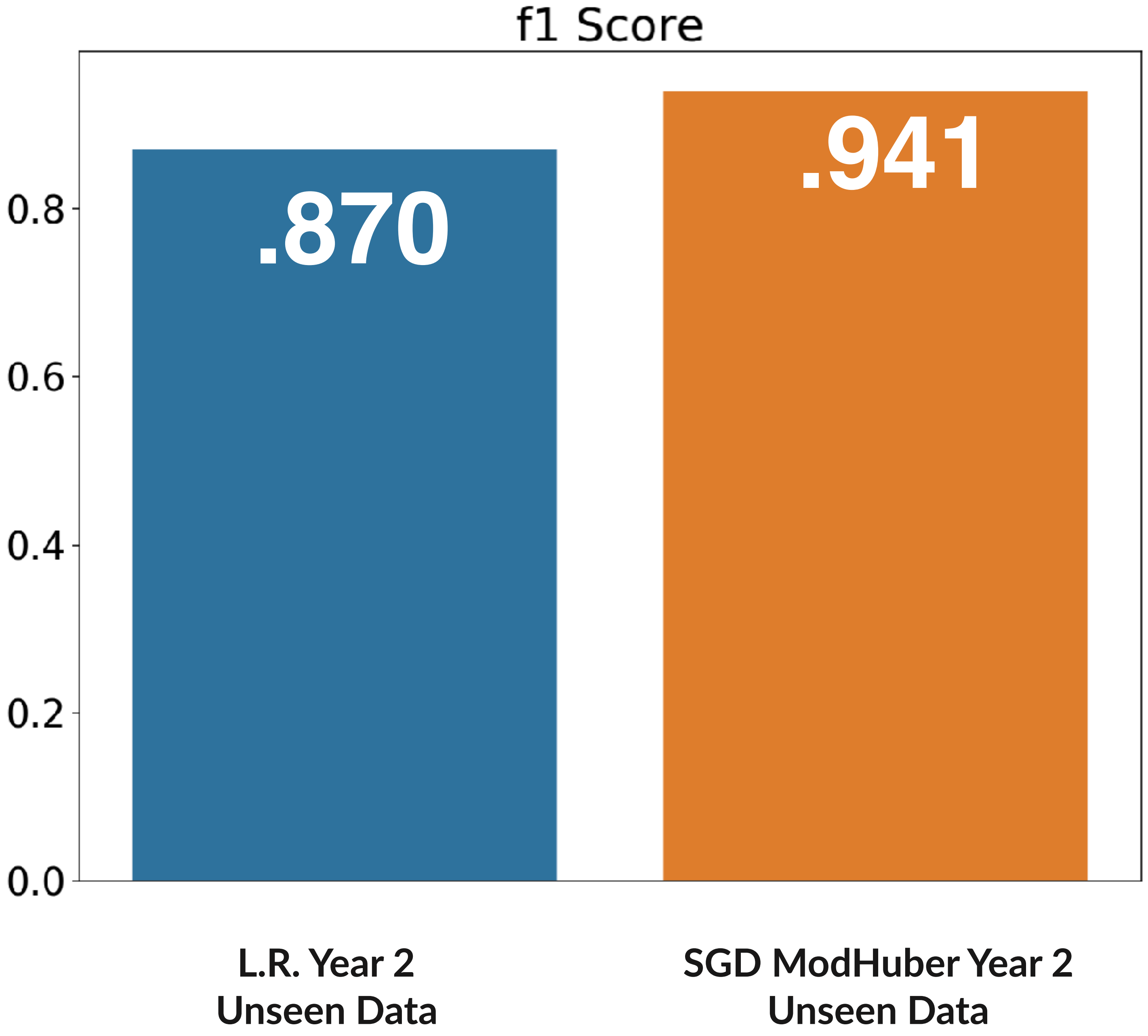
TRAINING RESULTS



5-Fold CV

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

UNSEEN DATA TEST RESULTS

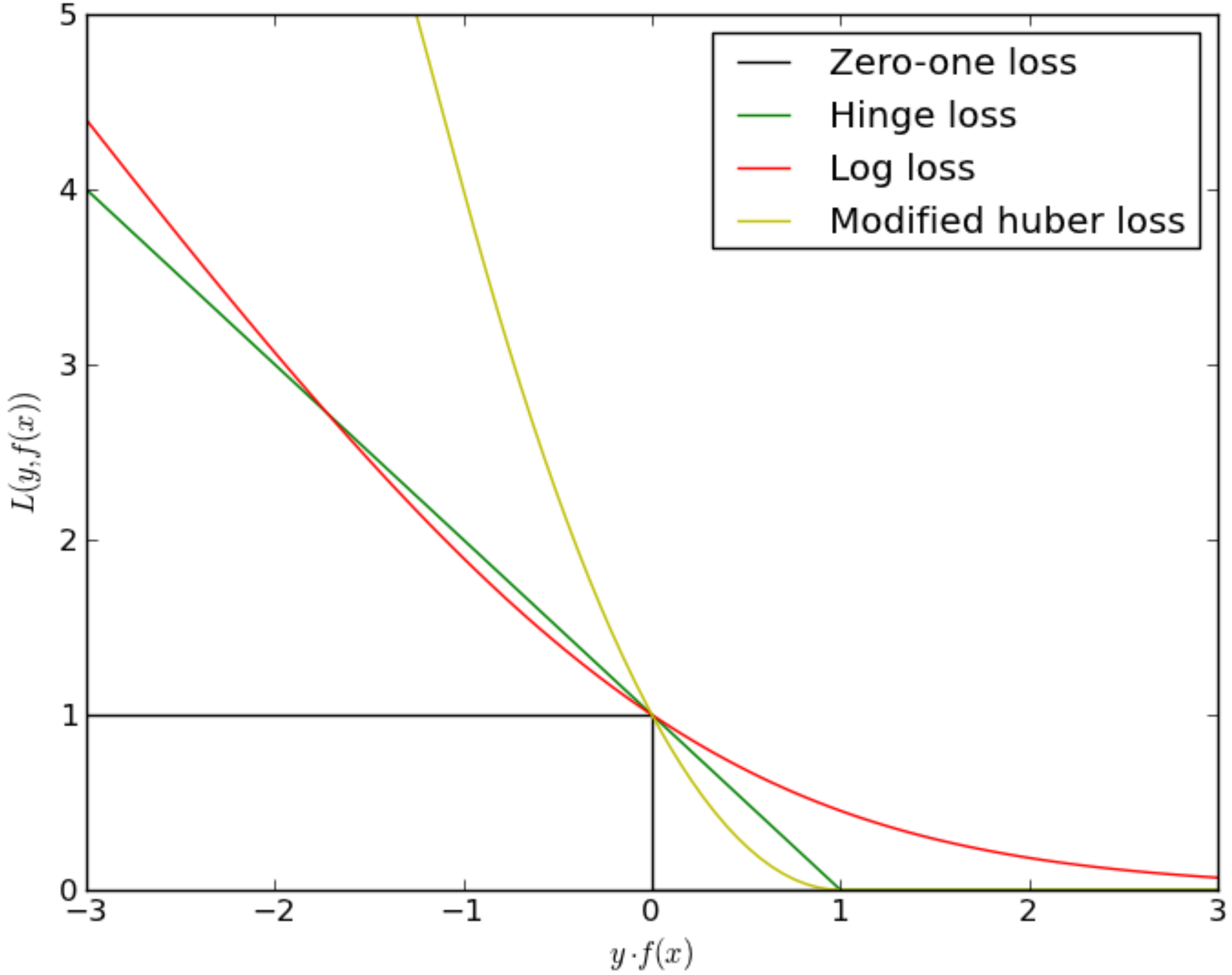


Hypothesis Testing

- 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?

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



- 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

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!

Topic Examples

```
 (6,
  '0.082*"friendly" + 0.078*"staff" + 0.049*"nice" + 0.034*"super" + 0.029*"amazing" + 0.026*"awesome" + 0.026*"serve
r" + 0.023*"delicious" + 0.022*"good" + 0.020*"definitely_back" + 0.018*"back" + 0.018*"attentive" + 0.018*"atmospher
e" + 0.017*"clean" + 0.017*"excellent"'),
 (7,
  '0.048*"tacos" + 0.042*"wait" + 0.034*"good" + 0.020*"back" + 0.019*"long" + 0.018*"delicious" + 0.016*"amazing" +
0.015*"taco" + 0.014*"burrito" + 0.014*"margaritas" + 0.013*"awesome" + 0.012*"check" + 0.011*"hour" + 0.011*"free" +
0.011*"okay"'),
 (8,
  '0.261*"best" + 0.076*"ever" + 0.059*"one" + 0.038*"town" + 0.029*"amazing" + 0.018*"vegas" + 0.016*"bbq" + 0.014
*"around" + 0.014*"las_vegas" + 0.013*"favorite" + 0.012*"thank" + 0.010*"must" + 0.010*"indian" + 0.010*"donuts" +
0.010*"party"'),
 (9,
  '0.044*"still" + 0.036*"say" + 0.024*"inside" + 0.021*"need" + 0.020*"anything" + 0.019*"home" + 0.018*"know" + 0.0
18*"outside" + 0.016*"something" + 0.014*"long_time" + 0.013*"drive" + 0.012*"tonight" + 0.011*"eat" + 0.010*"never"
+ 0.009*"seems"')]
```

UMAP Viz Gone Awry