DATS 6203
Final Project:
Fake Review Detection

Sean Pili
Pedro Uria

Motivation

- Online reviews are critical for the success of Restaurants and other businesses.
- Businesses buy fake reviews to help themselves and hurt their competitors.

- Online Review Platforms like Amazon & Yelp must filter out Fake Reviews
- Yelp displays “recommended” (real) reviews more prominently
  - Non-Recommended Reviews are less visible.
    - And aren’t factored into a company’s star-rating

Recommended Reviews

Your trust is our top concern, so businesses can't pay to alter or remove their reviews. Learn more.

Search within reviews

Sort by **Yelp Sort**

Start your review of **Starbucks**.

Laura F.
Centreville, VA
0 friends
7 reviews
2 photos

10/9/2019
1 check-in

This has to be THE happiest Starbucks location on earth. Everyone here on the morning shift is happy and friendly. They greet their regulars with enthusiasm--like so loudly that I don't dare take a conference call from here, but who wants to work anyway?? And they know how to make a flat white, many baristas just hand me an over-frothed latte but here you get the microbubbles It's also the CLEANEST location I've ever been in. Five stars for sure.

Useful | Funny | Cool

Keisha L.
Woodbridge, VA

2/5/2019

Adam J.
Washington, DC
1627 friends
257 reviews
790 photos

Share review
Embed review

10/6/2013
2 photos 1 check-in

This starbucks is pretty much good as it gets. The customer service was exceptional, as was the speed of service. The place was very clean. When I went one Saturday morning, there was a line at the door before they even opened. I say down and enjoyed my coffee and must have seen at least 30 customers in 30 minutes. Always a good sign. They had all of their desserts and items fully stocked and neatly presented.

Useful 1 | Funny 1 | Cool 1

10 other reviews that are not currently recommended

10 reviews for Starbucks that are not currently recommended

Note: The reviews below are not factored into the business's overall star rating.

Emilie H.
Woodbridge, VA
0 friends
2 reviews

12/7/2018

Went in for a simple coffee today. Stood in line forever and then realized there was only one barista. How do you plan on serving people efficiently with only on person working?

M S.
Alexandria, VA
5 friends
62 reviews
24 photos

5/17/2019

Great service! Clean establishment, very inviting atmosphere, and great staff! I give the gift of five stars to this business.....and they well deserve five stars.

I hope that the YELP algorithm leaves my five star review up. This business well deserves it.

Hey YELP, it sure would be great if this business could rate me as an individual. Other applications are doing that now. Who knows, I might not have been professional in my interactions with them. This process seems a little one sided.

Either way, I'm very happy to give this business five stars!

Albulbek A.
Burke, VA
0 friends
2 reviews

6/30/2018

luv u bbys thx fur makeing muy cuffe v3ry yumme tast3s liek my ded meow meow alm0st so thenkz!!!! *heart emoji* lmao yeet hahahah yeet yeet *dabz*

Problem Statement

- Yelp has not made their filtering algorithm Public, but…
  - Their “Recommended/Not Recommended” sections give DS practitioners labeled training data to train their own Fake Review classifier
- One group of researchers scraped over 100k Real and “Fake” Yelp Reviews of Hotels & Restaurants in Chicago to train their own fake review detector
  - They gave us their dataset
  - They extracted text and behavioural features
- Our goal is to **analyze** such **features** using **Bayesian Logistic Regression** and use this technique to **classify** a **review** as either real (negative) or fake (positive)

The Data

- 61,541 Reviews across 129 Restaurants in Chicago.
  - Filtered for Recurrent Reviewers (>1 Review)
    - 35,850 Real real, 2,086 were fake.
  - Contains Review Text, Star Rating (1-5), Review ID, ReviewerID, & Product ID (Restaurant Name)
  -
- 5 Behavioral Features
  - **MNR:** Most reviews a user posted in a day → Positive prior
  - **max_cosine** Max similarity among all reviews of a user → Positive prior
  - **avg_revL:** Average review length → Negative prior
  - **avg_posR:** Percent of 4-5 Star ratings: → Positive prior
  - **Reviewer_Deviation:** Abs Deviation from average restaurant rating → Positive prior

The Data : Language Features

- 3 Text Features:
  - Trained BERT (Neural Language Model) to classify the reviews
  - Added a linear layer to BERT's 768-dimensional output with 3 neurons, and a final output linear layer with 1 output neuron and a sigmoid output transfer function
  - Minimized a Binary Cross-Entropy loss, giving more weight to the fake reviews
  - Use trained BERT to extract 3 features for each review, by getting the output of the second to last linear layer
- BERT reported 88% recall but only 43 % accuracy
- There is only so much you can do with just the text, and we also forced BERT to provide higher recall in order to have more meaningful features towards identifying fake reviews

Modeling:

Model: Robust Bayesian Logistic Regression with Feature Selection

$$\beta_i \sim \text{ Normal distribution}, \quad \delta_i \sim \text{ Bernoulli distribution}, \quad \alpha \sim \text{ Beta distribution } (1,\, 9)$$

$$\uparrow$$

$$\mu = \frac{1}{2}\alpha + (1-\alpha)\frac{1}{1+e^{-\left(\delta_0\beta_0+\sum_i \delta_i\beta_i x_i\right)}}$$

$$\uparrow$$

$$y \sim \text{ bernoulli}(\mu)$$

Priors

BERT Features:

- **μ** = BERT's output layer's weight, **σ** = 1/4

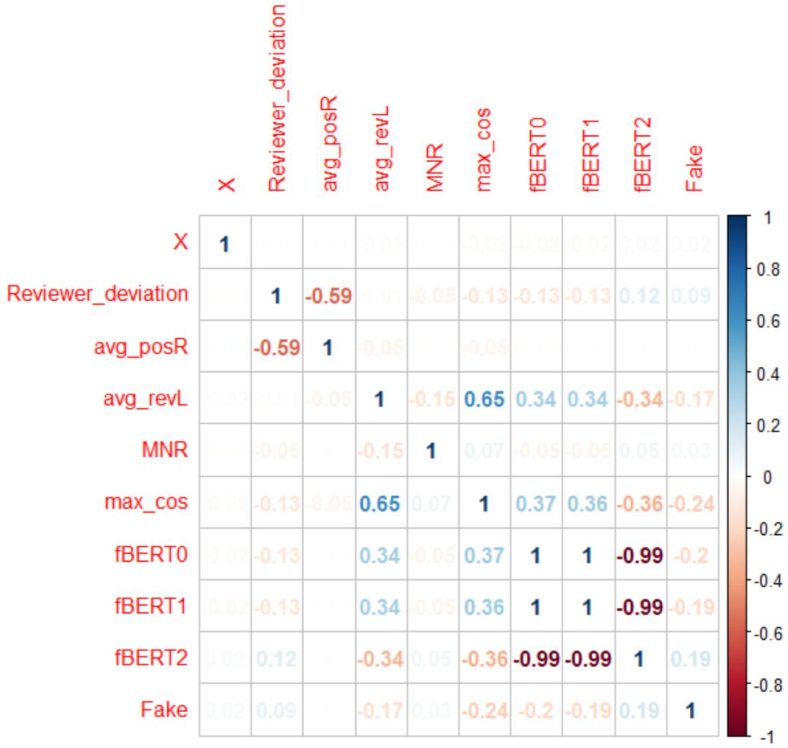
Intercept:

- **μ** = BERT's output layer's bias, **σ** = 1/4

Behavioral Features:

- MNR**:** **μ** = 1, **σ** = 1/4
- Max_Cosine**:** **μ** = 1, **σ** = 1/4
- avg_RevL**:** **μ** = -1, **σ** = 1/4
- avg_PosR**:** **μ** = 1, **σ** = 1/4
- Reviewer_deviation**:** **μ** = 1, **σ** = 1/4

Experiment # 1



- Included all features, ended in very bad convergence due to high correlation between many of them
- Therefore, no fair posterior analysis can be made
- Decided to drop max_cos because it was the feature with the most correlation with the rest of the features

Experiment #2

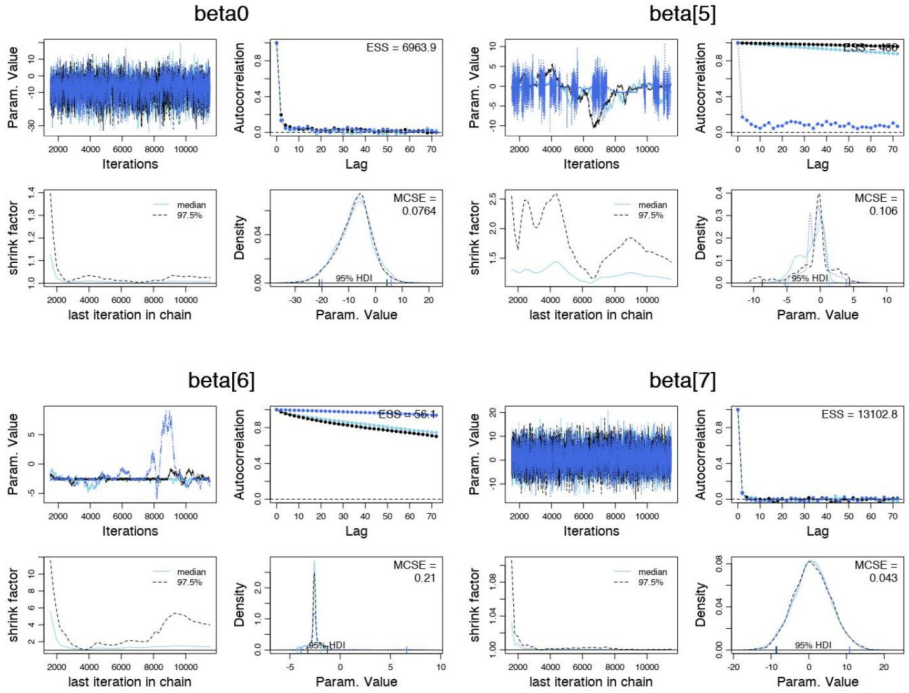
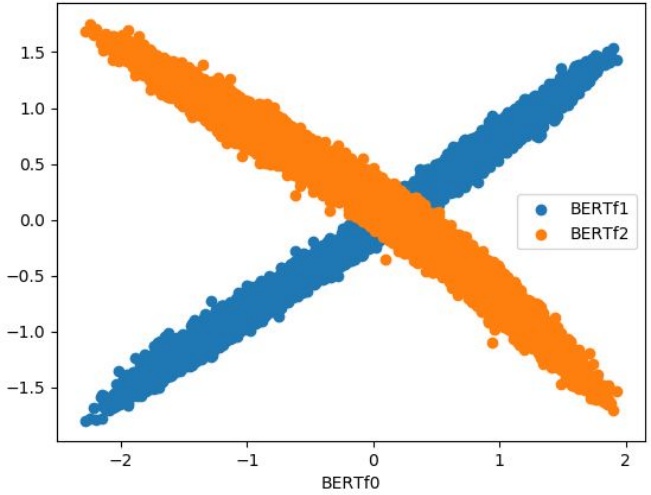


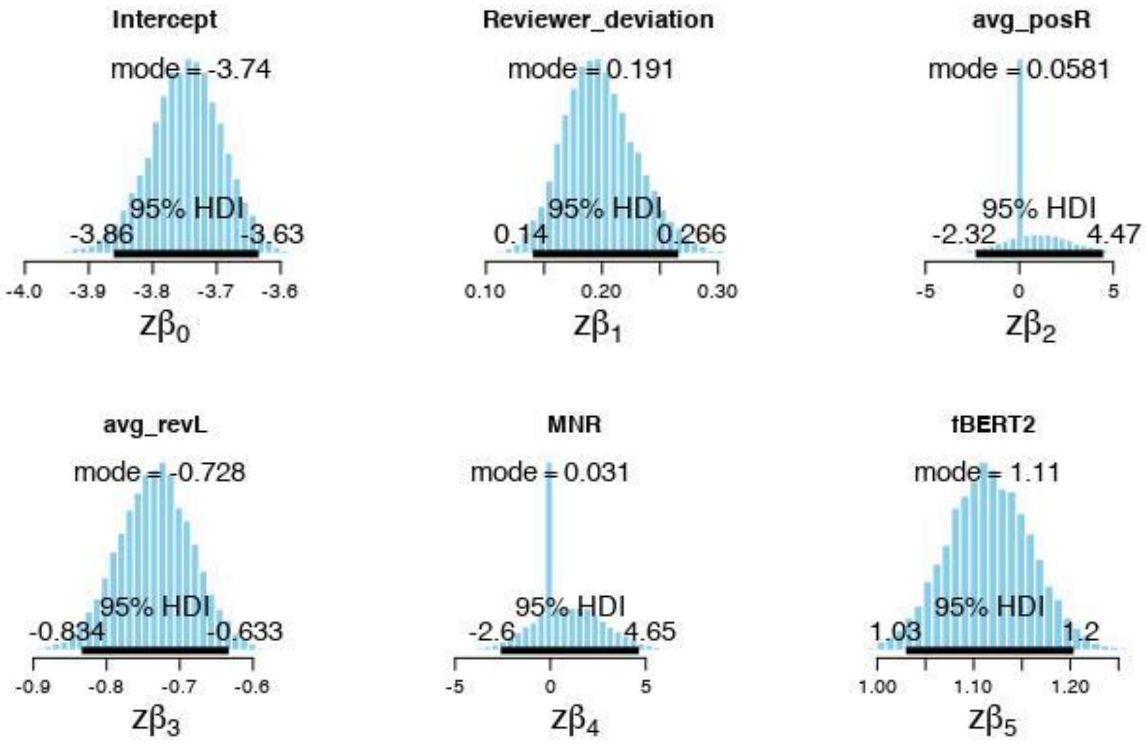
Figure 1. Intercept (beta0) and BERT's features MCMC convergence

- Every featured converged well except for the first and second BERT features. This was also expected given the high correlation among these features.



- Dropped the first two BERT features

Experiment #3



- No signs of bad convergence for any of the features
- avg_posR and MNR have very wide HDIs and their mode is very close to 0
- We can conclude that this features are not very significant, which was expected given our EDA and the disagreement with the literature. We exclude them from our next experiment

Experiment #3

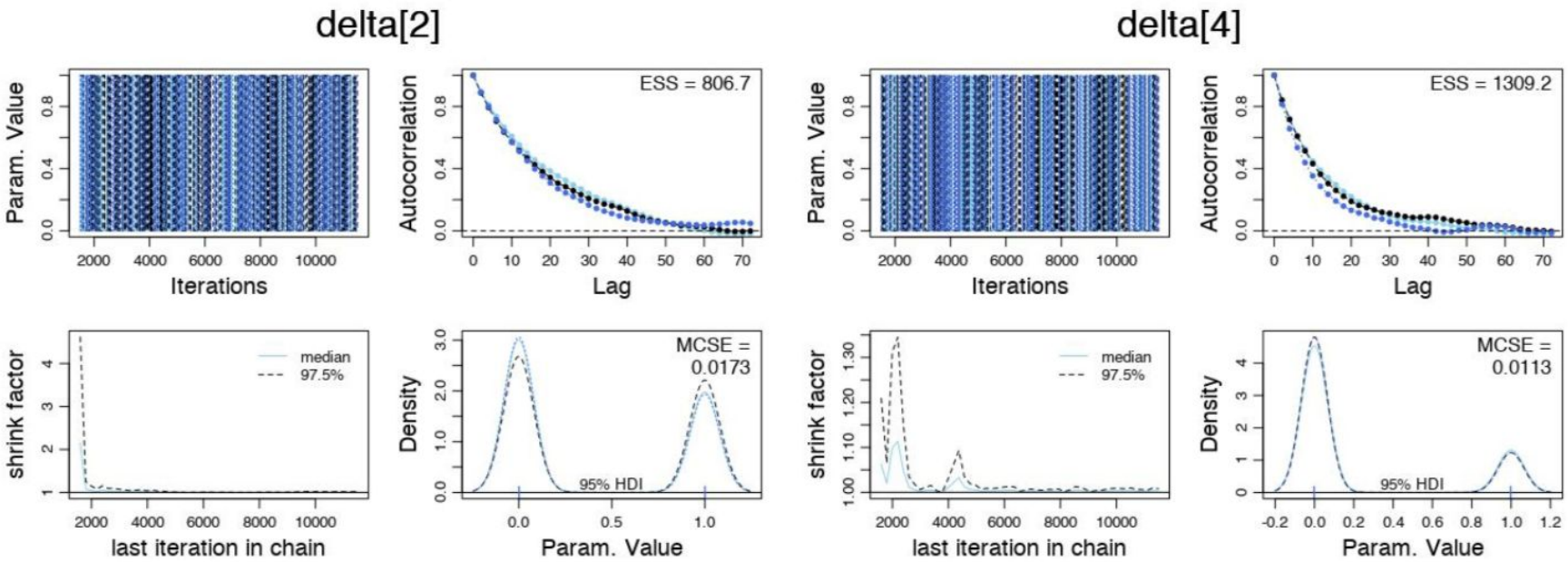


Figure 4. PRP (left) and MNR (right) Deltas from Experiment #3

Experiment #4

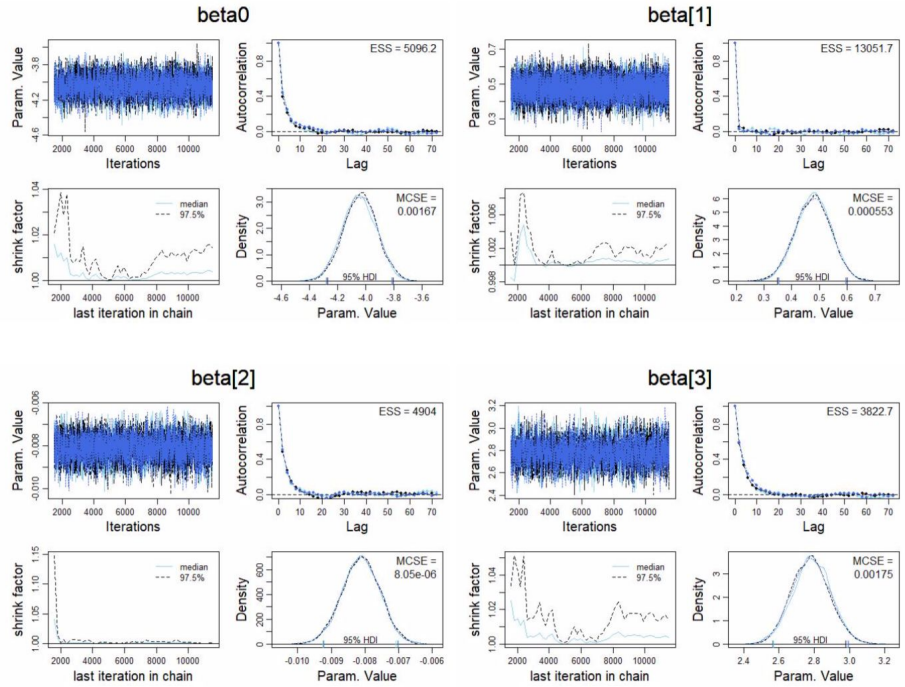
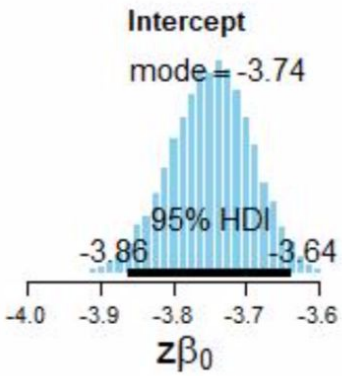
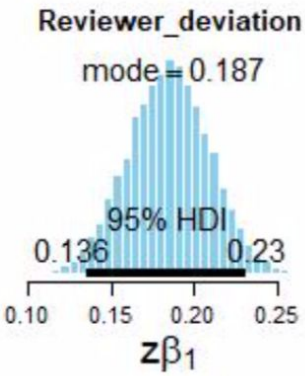


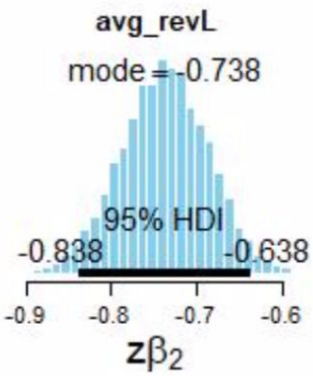
Figure 6. MCMCs from Experiment #4. Intercept (top left), Reviewer Deviation (top right), avg_RevL (bottom left) and fBERT2 (bottom right)

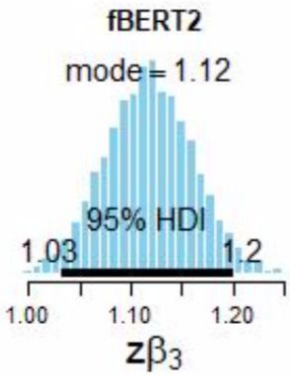
- Very good convergence
- Although some of the features are still somewhat correlated, we can draw some conclusions by analyzing their posteriors

Experiment #4









- All features are significant.
- Sign of Reviewer Deviation and BERT are **positive**
- Sign of avg_revL and intercept **negative**

Experiment #5

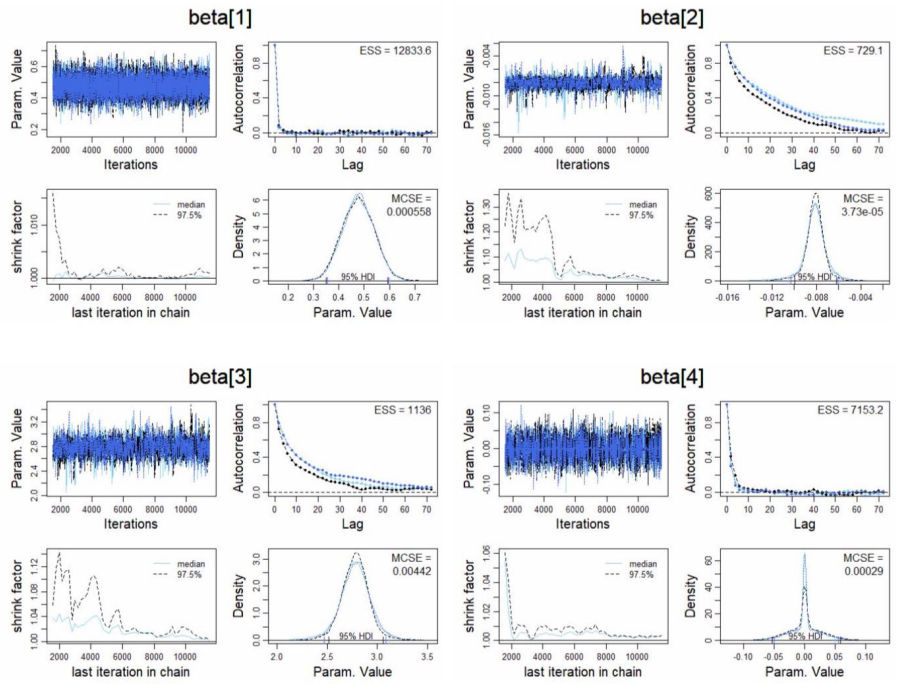
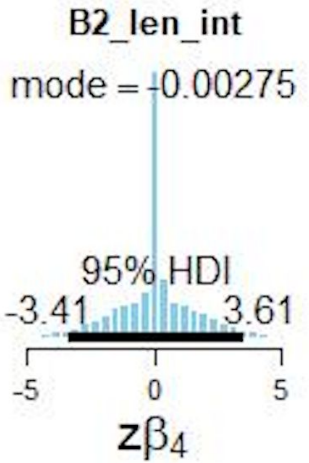


Figure 8. MCMCs from Experiment #5. Reviewer Deviation (top left), avg_RevL (top right), fBERT2 (bottom left) and avg_RevL x fBERT2 interaction (bottom right)

- Added interaction between BERT and avg_revL
- Interaction term does not appear to be not relevant…
- And its presence seems to affect the convergence of the chains negatively



Experiment #6

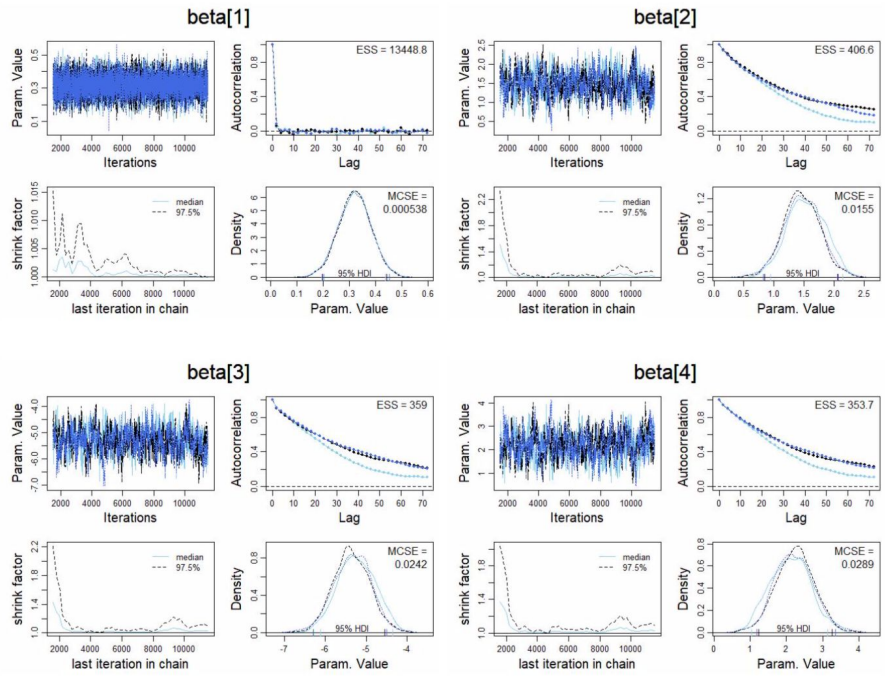
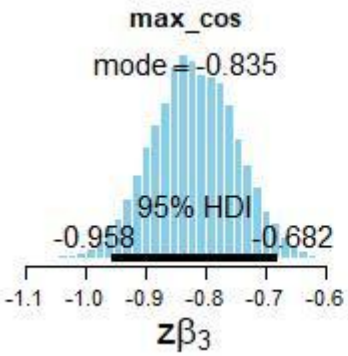
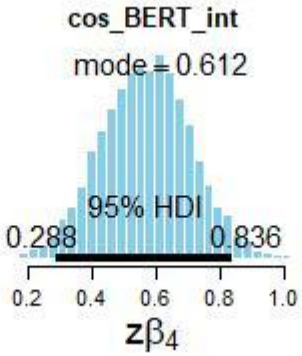


Figure 9. MCMCs from Experiment #6. Reviewer Deviation (top left), fBERT2 (top right), max_cos (bottom left) and fBERT2 x max_cos interaction (bottom right)

- Replaced avg_revL with max_cos and added interaction with BERT
- In this case, both terms appear to be significant





Final Results

Experiment	**Accuracy**	**Recall**	**AUC**
Experiment 4	69.24%	74.9%	.786
Experiment 5	73.43%	63.09%	.751
Experiment 6	***71.98%***	***77.31%***	***.807***
Sklearn Logistic Regression (All Features)	**71.72%**	**75.24%**	**.510**
Literature (SVC on Behavior Features)	***82.8%***	***87.9%***	*N/A*

Conclusions: Feature Importance

- There was ***no significant difference*** between using Bayesian Logistic Regression v. Traditional Logistic Regression
- And we could not meet the Baseline Metrics from the Literature Review
- ***However, the relationships present in our data did not match those in the literature review, which struck us as odd***
    - **We tried 'vanilla' priors and ones that fit the data, but achieved similar results.**

- **BERT: *Most Important*,** had the highest value in most of the experiments and was usually significant.
- **Reviewer Deviation: *Important:*** was significant in most experiments
- **Average Review Length**: ***Important:*** was significant in most experiments.
- **PRP: *Not Important.*** Converged in in Experiment 2 and 3, but never significant.
- **MNR: *Not Important.*** Converged in in Experiment 2 and 3, but never significant.