Project 2

Jeremy Mednik, Jiayi Du, Karim Naous, Stephen Owen

Classifying authors based on word usage

We can classify authors fairly well based on just their word usage.

Classification rate on validation data set:

Unigrams: **68%**

Bigrams: **76%**

Trigrams: **78%**

Coefficient Analysis

Here are the most distinguishable lemmas between reviewers.

```
                blackaciddevil : alot; lil; Portable; admit; gripe; Gameboy; Naruto; Dragonball; Tak; &; '
                         Bryan : sophisticated; unintereste; significantly; haha; walk; NES; nonstop; theme; danger; think; appear; spike; remember; probably; because;
Playstation; platform; stage; around; way; music
                  Bullet Theory : 4.5/5; Negatives; native; 1080p; 1:1; firstly; Last; etc; PS4; Dishonored; Buy; enjoyable; positive
                         Cloud : Music; Sound
                        Deimos : origional
               Inspector Gadget : Commodore; Graphics; em; appeal; chuck; Gameplay; epileptic
                       Ishmael : didin't; absolutley; alot; intreste
                    Ivan Orozco : ^_^; sorry; true; universe
                   Jeff Johnson : downfall; Half; 100s
                      Lisa Shea : 8/10; cartooney; 7/10; 6/10; essence; gentle; :); reasonably; beneath; fund; multiplay; slay; XBox; highly; situation; quickly
                         M. King : unstead; r; i; basicly; u; allot; atari; becide
                  Michael Kerner : Enjoyment; convience; accessory; Convience; b; b+; Nintendo; d+; c; b-
                    Micheal Hunt : tho; simmilar; ok; then; ...; unlock; elimination; 1; option; would; coarse
                       N. Durham : noticable; Robert; Aerosmith; excursion; wonderfully; 2-d; flaw; horde; NHL; aside
                    NeuroSplicer : RECOMMENDED; CIVILIZATION; stutter; deliciously; DUNE; palatable; nVidia; STEAM; DRM; SecuROM; artificially; BALDUR; customer; hence; k;
(; -
Oreocokiemarshmallowkrispy 2018 : recomend; truley; lego; sweet
                        Reggie : menue
                  Richard Baker : rental; passive; Bad; Good; sexy; visual; ugly; story; use; feel; frustrating; just; mechanic; thank; weapon; extremely; great
         RJ the Great Cat Lover : terrific; neat; ammunition; Awesome
Ryan Sil. (Gamer &amp;amp; PC/Android indie dev) : Stages; aka; contain; 2.5d; platformer; Pac; Game; Kart; -
                     SleepyJD : Basicaly; basicaly
                  thomas henry : vita; gamestop; genesis; gameboy; i
                       Tsanche : +; also; begin; actually; will; however; .; not
                       TwistaG : A.I.; PlayStation; --; effect; cooperative; 's; Fighter; -
            Video Game History : recommendation; Broken; .....
```

The most likely gaming platforms used by some authors

Gameboy: **blackaciddevil, thomas henry**

NES: **Bryan**

Playstation: **Bullet Theory, TwistaG, Bryan**

XBox: **Lisa Shea**

Atari: **M. King**

Nintendo: **Michael Kerner**

PC: **NeuroSplicer**

```
TwistaG : two -; - player; PlayStation; --; effect; 's; . the
thomas henry : vita; gamestop; genesis; gameboy; i
```

Which authors used certain misspellings, abbreviations, and/or emoticons

blackaciddevil : -PRON- '; admit ,; alot; Gameboy Advance; . yet; but ,; gripe

Micheal Hunt : tho; what not; tho -PRON-; ok; then; this version; more then; this game; ...; unlock; , or; ... but; , then; or 3

lil: **blackaciddevil**

r, u: **M. King**

tho: **Michael Hunt, Ishmael**

ok: **Michael Hunt**

k: **NeuroSplicer**

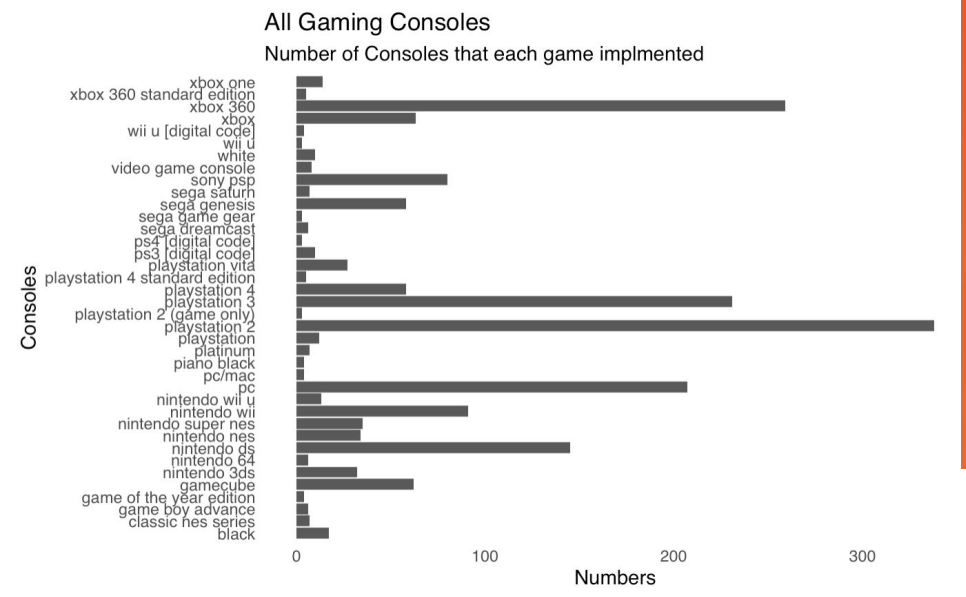
alot: **blackaciddevil**

Haha: **Bryan**

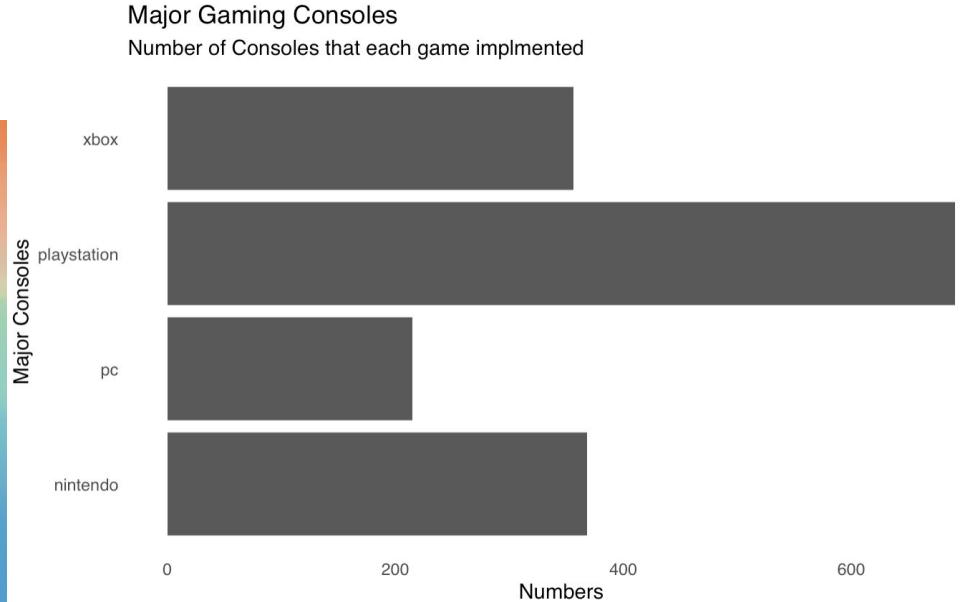
^_^: **Ivan Orozco**

(;: **NeuroSplicer**

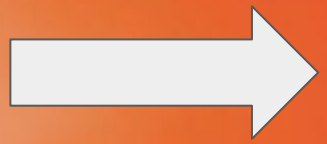
:): **Lisa Shea**



Consoles Covariate



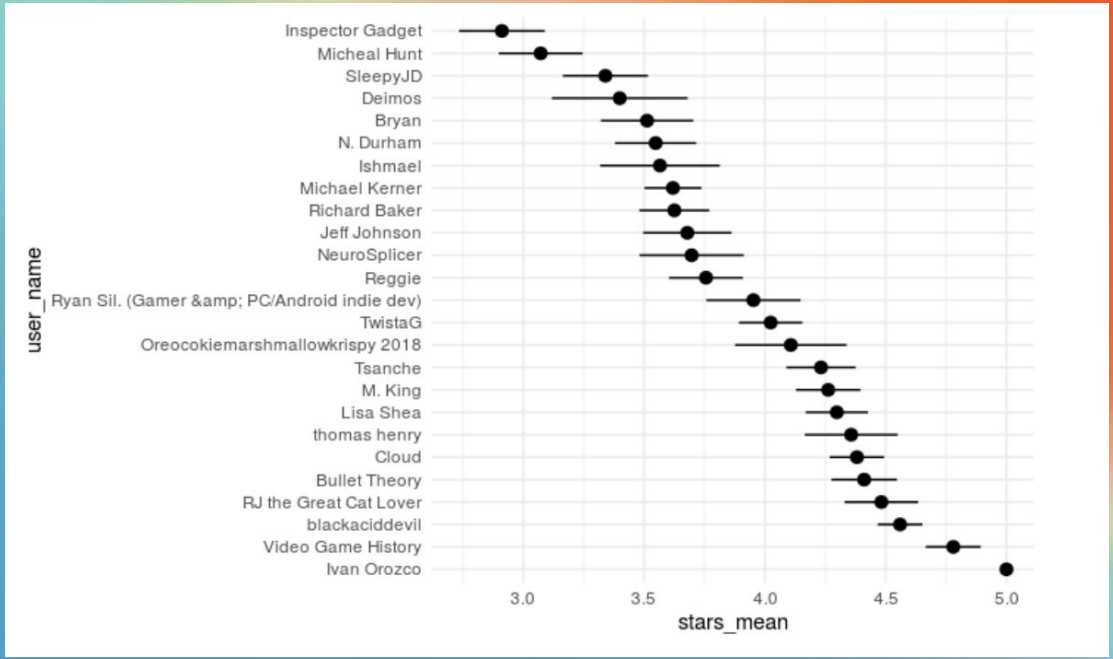
Train_id	Class Rate
train	.89
valid	.59



User_name	Train	Vaild
Deimos	0.6904762	0.3809524
Ryan Sil	0.9523810	0.4404762
Ishmael	0.8928571	0.4761905
Lisa Shea	0.9880952	0.4880952
.....		
NeuroSplicer	1.0000000	0.7261905
RJ the Great Cat Lover	0.9047619	0.7261905
Cloud	1.0000000	0.7380952

User_name	Train	Vaild	Consoles own	Confusion Matrix
Deimos	0.69	0.38	Diverse	31/84
RJ the Great Cat Lover	0.90	0.72	Major PlayStations	53/84

Stars Covariate



Reviewer with lowest average score: Inspector Gadget (2.9 Average)

Reviewer with highest average score: Ivan Orosco

-Typically only gives 5 star reviews

Can this covariate accurately predict the reviewer?

Stars Covariate

```r
amazon %>%
  select(-user_name) %>%
  left_join(uname, by = "user_id") %>%
  group_by(user_name) %>%
  summarize(sm_mean_ci_normal(stars)) %>%
  arrange(desc(stars_mean)) %>%
  mutate(user_name = fct_inorder(user_name)) %>%
  ggplot(aes(user_name, stars_mean)) +
    geom_pointrange(aes(ymin = stars_ci_min, ymax = stars_ci_max)) +
    coord_flip()

X_cov <- amazon %>%
  model.frame(user_id ~ stars, data = .) %>%
  model.matrix(attr(., "terms"), .)

X <- cbind(X_cov, X)

X_train <- X[amazon$train_id == "train", ]
y_train <- amazon$user_id[amazon$train_id == "train"]

model <- cv.glmnet(
  X_train,
  y_train,
  alpha = 0.9,
  family = "multinomial",
  nfolds = 3,
  trace.it = FALSE,
  relax = FALSE,
  lambda.min.ratio = 0.01,
  nlambda = 100
)

amazon %>%
  mutate(pred = as.vector(predict(model, newx = X, type = "class"))) %>%
  group_by(train_id) %>%
  summarize(class_rate = mean(user_id == pred))
```

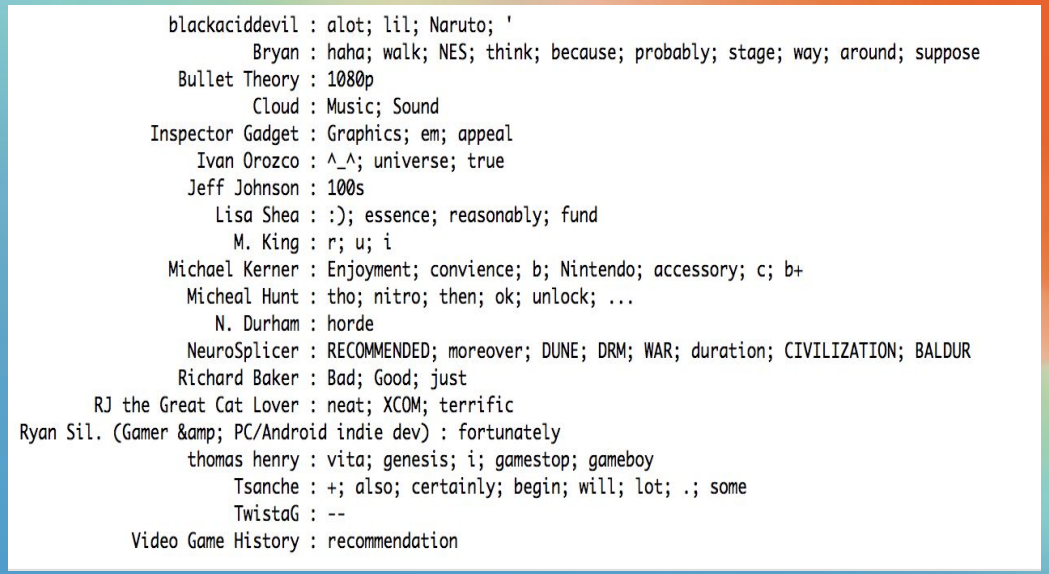
NO IT CANT!

High accuracy for the training data, but very low for the valid data

Train_id	Class Rate
train	.93
valid	.62

Stars Covariate

Is there particular products that each person reviews is the main driver of our model? Can we remove this factor?



As you can see, the particular products each person reviews is a driving factor for a model based on the what each user commonly said in each of their reviews.

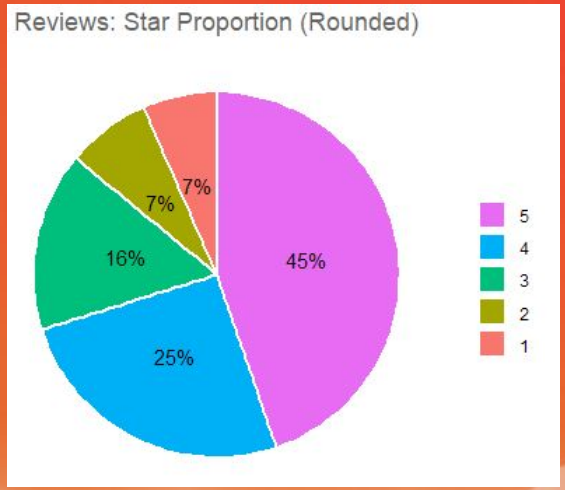
-"Dune","Baldur", "Nintendo"

How to remove?

Create a multinomial model and filter out proper nouns to not included.

Star Bias

- 70% of reviews are 4 or 5 stars
- Only 589 negative reviews
- *Possible* voluntary response bias
  - More negative reviews
    - More varied language
      - More accurate predictions



Other models

- Neural networks proved difficult
  - Required large layers due to input size
    - Smaller networks would barely “train”
  - Large layers -> longer to train
  - 6 hours to train 30 epochs on GPU
    - Only was able to go from 4% -> 12% accuracy.
- More advanced models could potentially work
  - 1D Convolution
    - Take advantage of spatial locality
  - RNNs / Transformers
    - Tend to excel at NLP type tasks.

Conclusions

- Decent prediction metrics as is
- Higher training scores compared to validation
  - Overfitting
- Additions / modifications to dataset could help
  - Additional negative reviews / more data in general