DESTINATION: OHIO

CYNTHIA RODRIGUEZ
DSI-822

PROBLEM STATEMENT

I work for TourismOhio, where our mission statement is to “aggressively position Ohio as a relevant travel destination and support Ohio’s tourism industry to drive economic prosperity throughout the state.” We saw a 21% drop in visitor spending in 2020, but rebounded in 2021, and we are now trying to continue the growth momentum with more relevant ads and offers.

To do this, I’ve first been asked to scrape social media sites and find out what people are saying about our state and why they may or may not visit so that we can adapt our advertisements to our target audiences. I am starting on Reddit, with what I believe to be two relevant subreddit pages:

PROBLEM STATEMENT

I work for TourismOhio, where our mission statement is to “aggressively position Ohio as a relevant travel destination and support Ohio’s tourism industry to drive economic prosperity throughout the state.” We saw a 21% drop in visitor spending in 2020, but rebounded in 2021, and we are now trying to continue the growth momentum with more relevant ads and offers.

To do this, I’ve first been asked to scrape social media sites and find out what people are saying about our state and why they may or may not visit so that we can adapt our advertisements to our target audiences. I am starting on Reddit, with what I believe to be two relevant subreddit pages:

- Ohio: Created on October 4, 2008, with 342k members
- IHateOhio: Created on September 28, 2018, with 4.6k members

PROBLEM STATEMENT

I work for TourismOhio, where our mission statement is to “aggressively position Ohio as a relevant travel destination and support Ohio’s tourism industry to drive economic prosperity throughout the state.” We saw a 21% drop in visitor spending in 2020, but rebounded in 2021, and we are now trying to continue the growth momentum with more relevant ads and offers.

To do this, I’ve first been asked to scrape social media sites and find out what people are saying about our state and why they may or may not visit so that we can adapt our advertisements to our target audiences. I am starting on Reddit, with what I believe to be two relevant subreddit pages:

- Ohio: Created on October 4, 2008, with 342k members
- IHateOhio: Created on September 28, 2018, with 4.6k members

Before I can analyze the messages or make any recommendations, I need to be able to pull posts from each of these threads into an aggregated data frame. My goal is to build a classification model that can predict, with at least 85% accuracy, which subreddit each post is from.

OhiO find it here. ohio.org

PREPROCESSING

1. Undersampling

PREPROCESSING

1. Undersampling
2. Null replacement

PREPROCESSING

1. Undersampling
2. Null replacement
3. URL removal

PREPROCESSING

1. Undersampling
2. Null replacement
3. URL removal
4. Features created

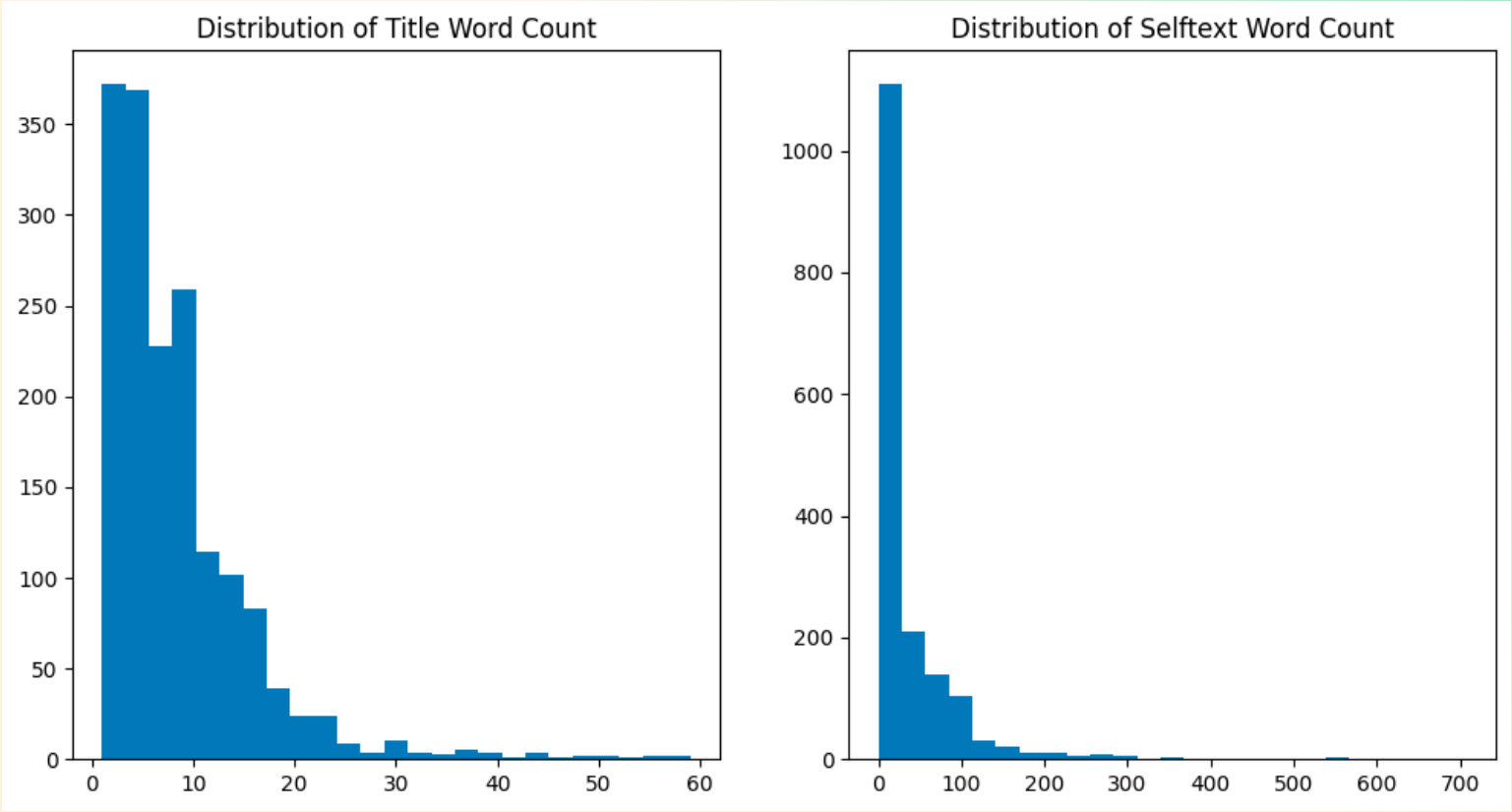
PREPROCESSING

1. Undersampling
2. Null replacement
3. URL removal
4. Features created
5. Stop words added

PREPROCESSING

1. Undersampling
2. Null replacement
3. URL removal
4. Features created
5. Stop words added
6. Lemmatization

WORD COUNT DISTRIBUTION



QUICK STATS

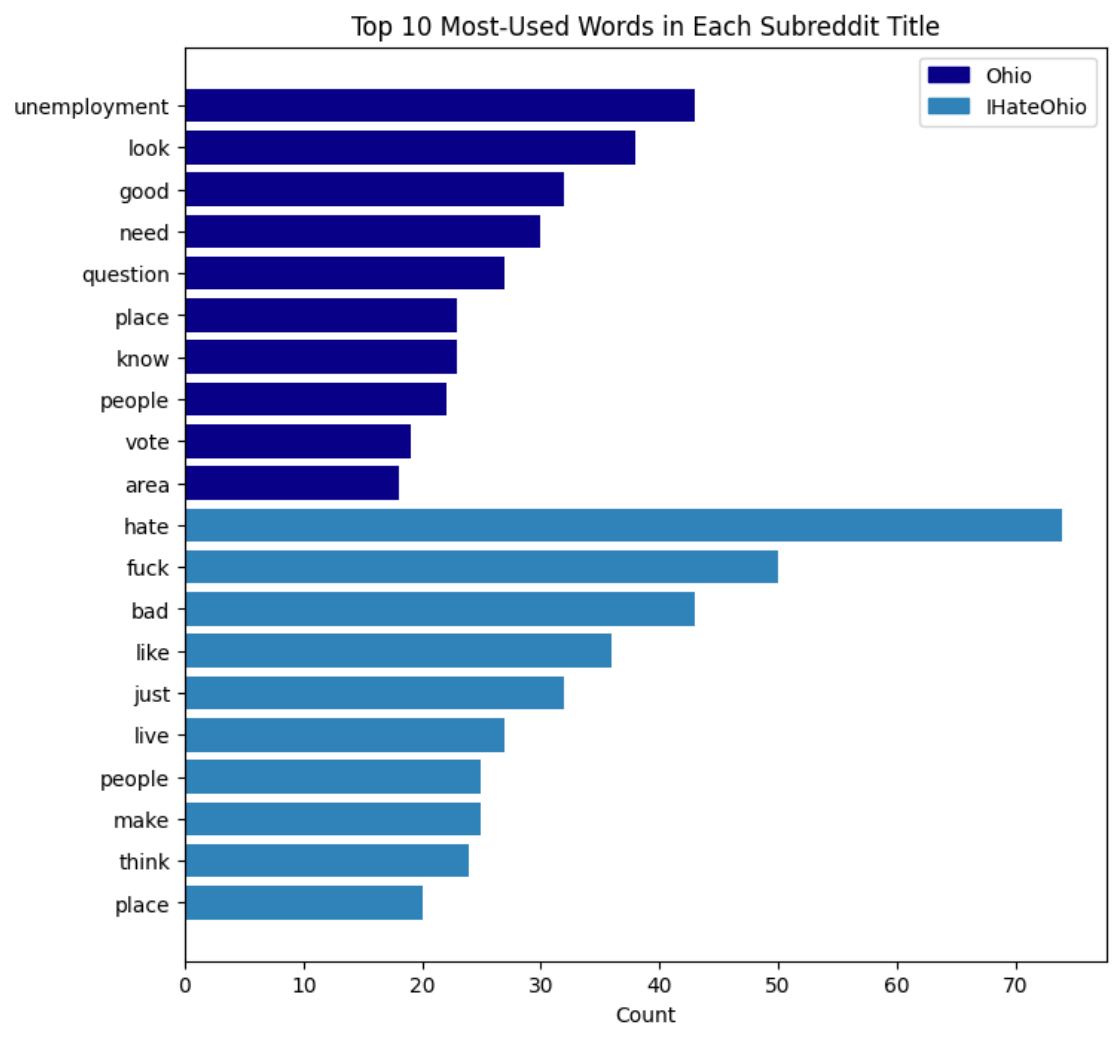
1. Mean r/Ohio title word count: 8.75
2. Mean r/IHateOhio title word count: 8.06

QUICK STATS

1. Mean r/Ohio title word count: 8.75
2. Mean r/IHateOhio title word count: 8.06
3. Mean r/Ohio selftext word count: 58.37
4. Mean r/IHateOhio selftext word count: 9.96

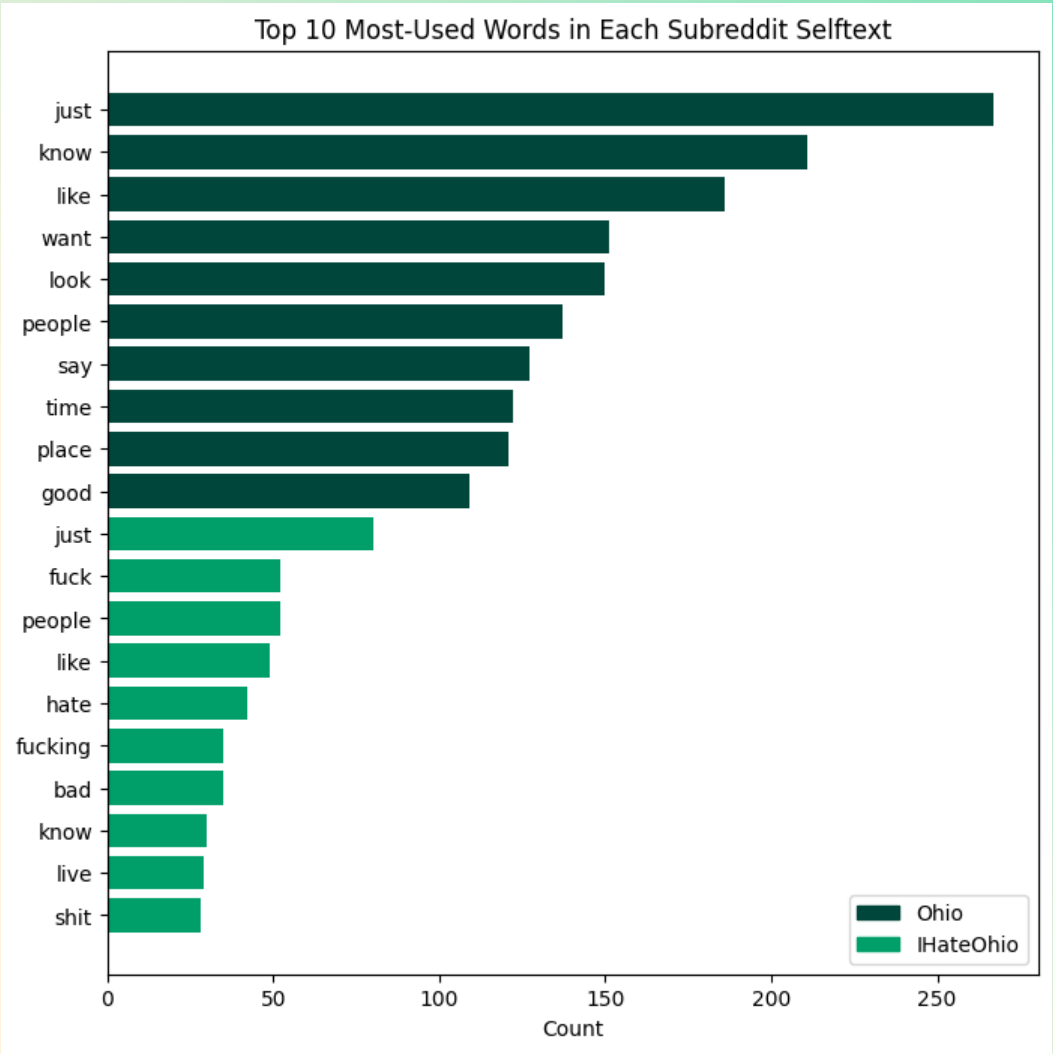
QUICK STATS

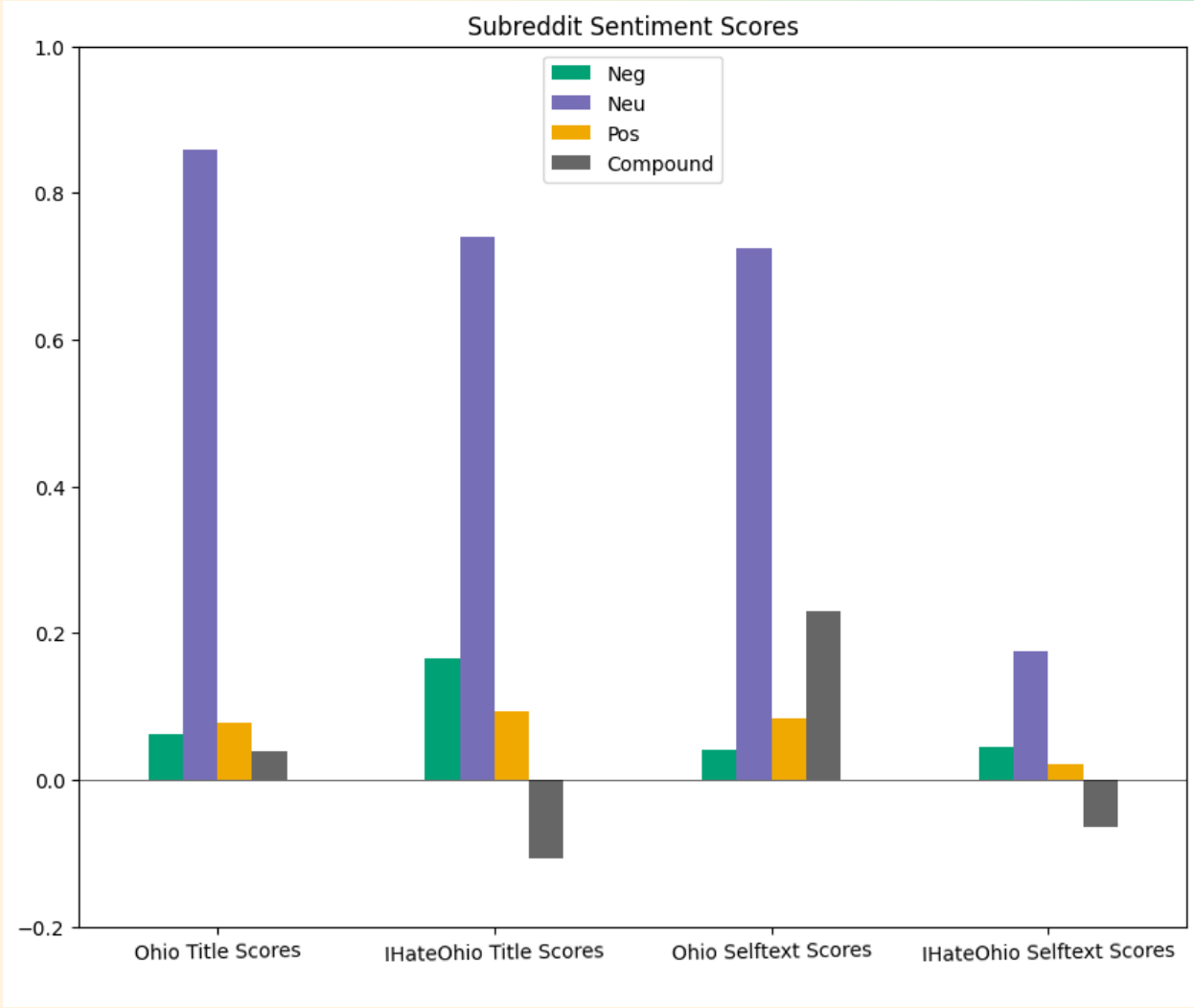
1. Mean r/Ohio title word count: 8.75
2. Mean r/IHateOhio title word count: 8.06
3. Mean r/Ohio selftext word count: 58.37
4. Mean r/IHateOhio selftext word count: 9.96
5. Users who posted in both subreddits: 1



- English and custom stop words removed
- Ohio words are more information-focused and specific
- IHateOhio words are more opinion-based
- One shared word: people

- Ohio words are similar to those seen in title chart
- IHateOhio words are more negative
- Four shared words: just, know, like, people





- All sections rank highest for neutrality
- IHateOhio Title compound score: −0.107
- Ohio Selftext compound score: 0.229

MODEL TRIALS

1. Baseline score: 0.522

MODEL TRIALS

1. Baseline score: 0.522
2. No hyperparameters added

MODEL TRIALS

1. Baseline score: 0.522
2. No hyperparameters added
3. 18 models built

MODEL TRIALS

1. Baseline score: 0.522
2. No hyperparameters added
3. 18 models built
4. Selftext outscored title in all models except one

MODEL TRIALS

1. Baseline score: 0.522
2. No hyperparameters added
3. 18 models built
4. Selftext outscored title in all models except one
5. Average train score: 0.8647

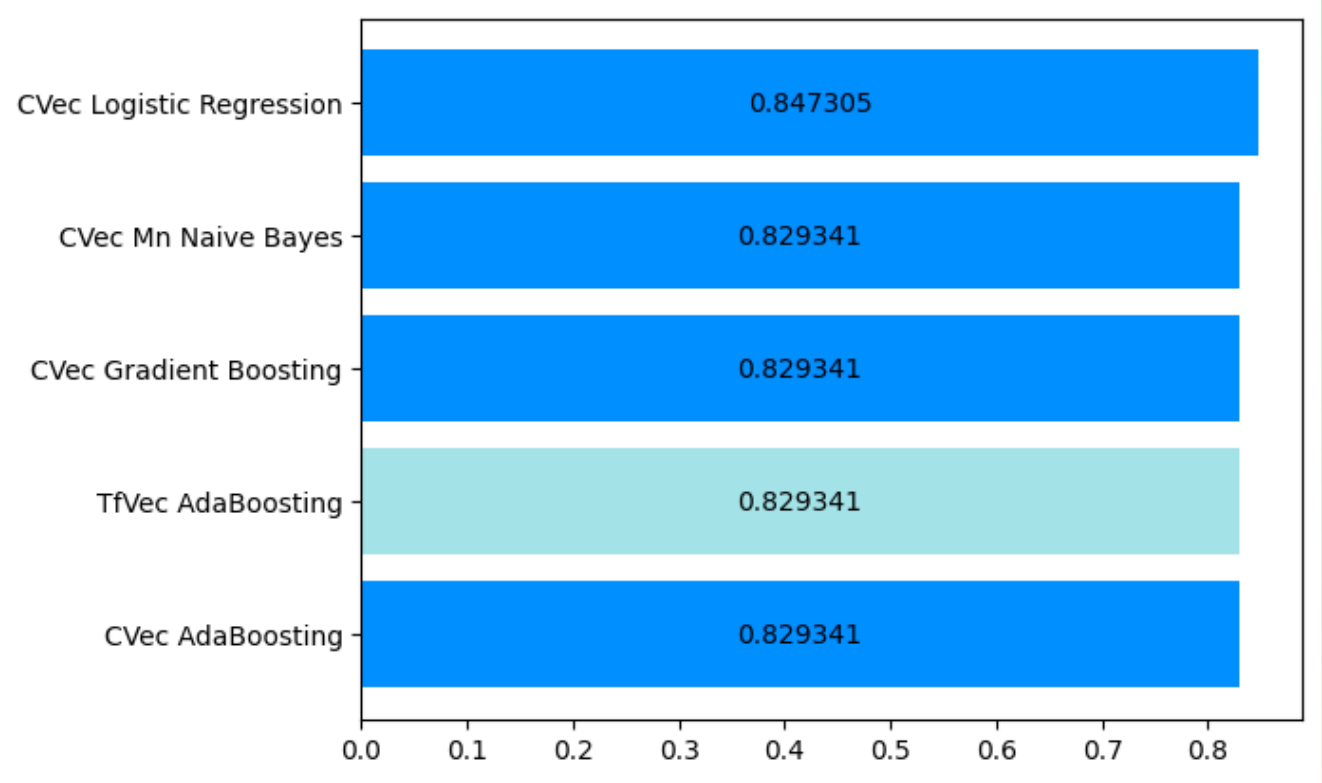
MODEL TRIALS

1. Baseline score: 0.522
2. No hyperparameters added
3. 18 models built
4. Selftext outscored title in all models except one
5. Average train score: 0.8647
6. Average test score: 0.7500

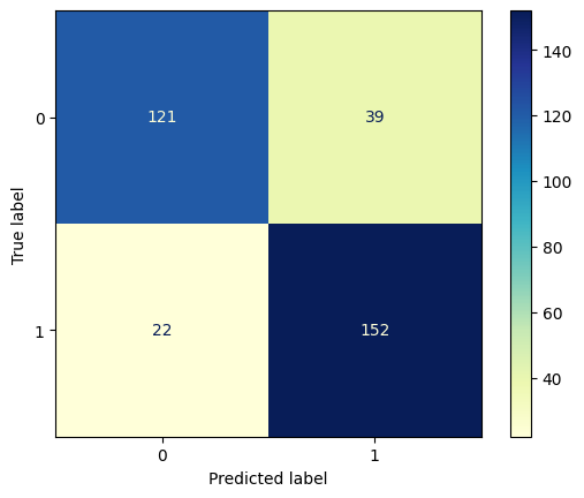
MODEL TRIALS

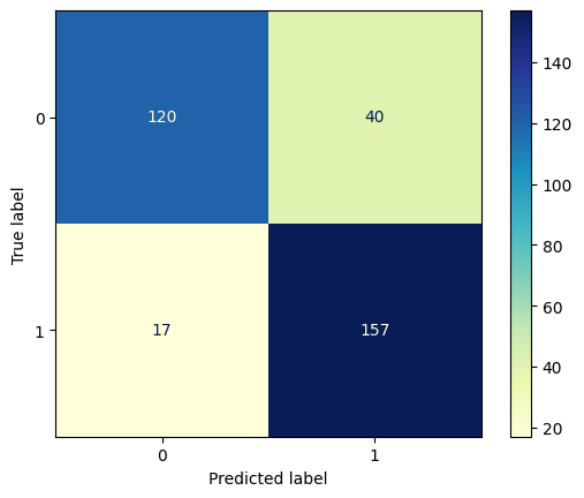
1. Baseline score: 0.522
2. No hyperparameters added
3. 18 models built
4. Selftext outscored title in all models except one
5. Average train score: 0.8647
6. Average test score: 0.7500
7. Minimum test score: 0.522

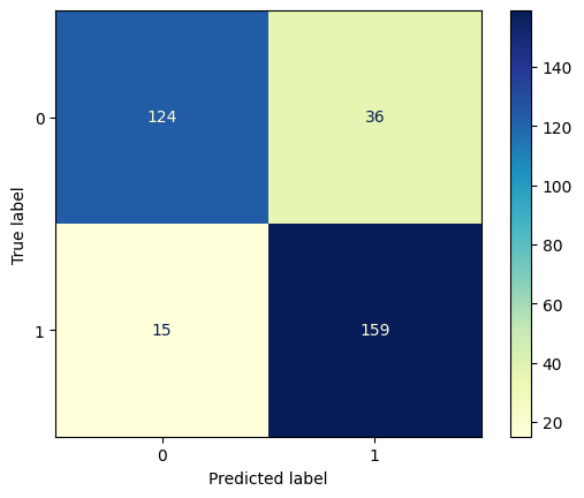
TOP-SCORING MODELS



MISCLASSIFICATIONS







MISCLASSIFICATIONS

1. Empty strings

MISCLASSIFICATIONS

1. Empty strings
2. Positive words out of context

Post text	Pred class	True class
for the love of god someone save me the allergens are killing me and it was snowy until april	Ohio	IHateOhio
my wife is from ohio, i'm from eastern pa. multiple times a year i have to drive into that hell hole and pretend that terrible pizza is good pizza. f**k ohio pizza, f**k ohio.	Ohio	IHateOhio

MISCLASSIFICATIONS

1. Empty strings
2. Positive words out of context

Post text	Pred class	True class
for the *love* of god someone save me the allergens are killing me and it was snowy until april	Ohio	IHateOhio
my wife is from ohio, i'm from eastern pa. multiple times a year i have to drive into that hell hole and pretend that terrible pizza is *good* pizza. f**k ohio pizza, f**k ohio.	Ohio	IHateOhio

MISCLASSIFICATIONS

1. Empty strings
2. Positive words out of context
3. Locations

Post text	Pred class	True class
hey i’m born and raised in ohio its f**king awful here we have nothing but corn and cleveland	Ohio	IHateOhio
i absolutely hate ohio. that stupid state. it's shaped like a f**king deformed square, doesn't even have anything special about it compared to chad states like texas	Ohio	IHateOhio

MISCLASSIFICATIONS

1. Empty strings
2. Positive words out of context
3. Locations

Post text	Pred class	True class
hey i’m born and raised in ohio its f**king awful here we have nothing but corn and *cleveland*	Ohio	IHateOhio
i absolutely hate ohio. that stupid state. it's shaped like a f**king deformed square, doesn't even have anything special about it compared to chad states like *texas*	Ohio	IHateOhio

MODEL TUNING SUMMARY

1. GridSearchCV used to test hyperparameters

MODEL TUNING SUMMARY

1. GridSearchCV used to test hyperparameters
2. 26 trial and 3 stacked models built

MODEL TUNING SUMMARY

1. GridSearchCV used to test hyperparameters
2. 26 trial and 3 stacked models built
3. Average train score: 0.89367

MODEL TUNING SUMMARY

1. GridSearchCV used to test hyperparameters
2. 26 trial and 3 stacked models built
3. Average train score: 0.89367
4. Average test score: 0.82381

MODEL TUNING SUMMARY

1. GridSearchCV used to test hyperparameters
2. 26 trial and 3 stacked models built
3. Average train score: 0.89367
4. Average test score: 0.82381
5. Best-scoring model: CountVectorizer Multinomial Naïve Bayes, 0.86228 test score

MODEL CONCLUSIONS

- Minimum threshold score of 0.85 was met
- Feature importance did not provide any valuable information
- Reddit is not useful for identifying places or experiences in Ohio that an advertisement team could build on
- Overall, full project statement goal was not achieved

NEXT STEPS

1. Pull data from websites like TripAdvisor or Yelp

NEXT STEPS

1. Pull data from websites like TripAdvisor or Yelp
2. Add in more custom stop words to filter the data

NEXT STEPS

1. Pull data from websites like TripAdvisor or Yelp
2. Add in more custom stop words to filter the data
3. Further model trials with a focus on decreasing variance

NEXT STEPS

1. Pull data from websites like TripAdvisor or Yelp
2. Add in more custom stop words to filter the data
3. Further model trials with a focus on decreasing variance
4. Implementation of best-scoring model

NEXT STEPS

1. Pull data from websites like TripAdvisor or Yelp
2. Add in more custom stop words to filter the data
3. Further model trials with a focus on decreasing variance
4. Implementation of best-scoring model
5. Advise the marketing team