# CMPT 318 Project Results

ELLIOT HEISLER

**ACM Reference Format:**
Elliot Heisler. 2023. CMPT 318 Project Results.  1, 1 (December 2023), 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 HYPERPARAMETER STUDY

A grid search of 5 hyperparameters was conducted.

**selection** Use sentiment readings from either the beginning, middle, or end of the tokenized steam reviews. Has values `('left', 'middle', 'right')` in the code, respectiely

**k** number of columns to use in `SelKBest()`. `(5, 2)` were checked for k as well as using all of the 10 relevant columns.

**C, gamma, and kernel** These are used directly as parameters to the grid search of SVC. Values `(1, 10, 0.1)`, `('scale', 'auto')`, and `('linear', 'rbf', 'sigmoid')` were checked, respectively.

## 2 MAIN RESULTS

With the precision varying only between 0.87 and 0.88, I cannot say that I achieved a high precision, but I can discuss what parameters affected results in the grid search the most. As far as `selection`, selecting the rightmost tokens performed slightly better than middle on average, and left performed significantly worse. The case was similar for kernel selection. `rbf` performed slightly better than `linear`, and `sigmoid` much worse than both. The `C` parameter seemed to increase potential performance slightly linearly for my chosen values of 0.1, 1, and 10. It makes sense that a larger C would allow better performance, if the correlation between sentiments and `['user_suggestion']` is smooth and not noisy. Finally, the `k` in `SelKBest` did not affect performance much. This was to be expected since the non-sentiment columns are hardly even correlated with `[user_suggestion]`. In the end, the most accurate and precise classifiers used an `rbf` kernel and sentiment readings from the final 512 tokens.

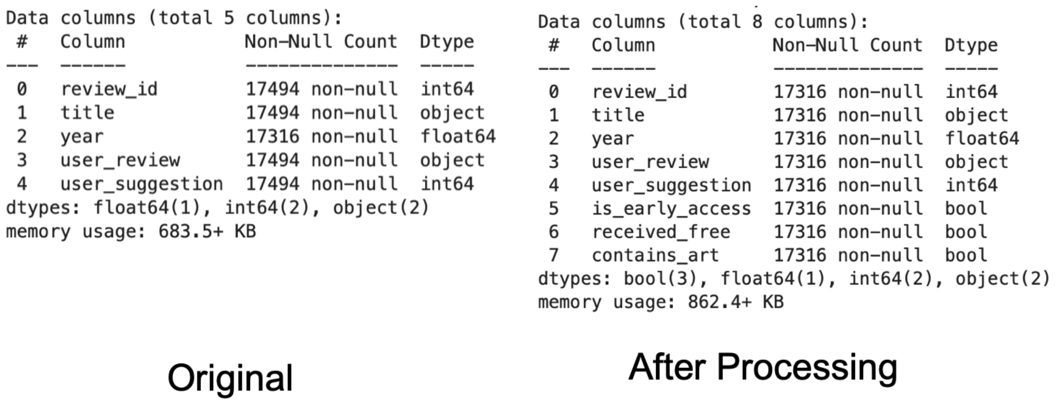
## 3 ROADBLOCKS

There is one main roadblock worth mentioning. While running sentiment analysis, I discovered that the twitter transformer and most huggingface transformers only support 512-token inputs, too small for roughly half of the reviews. After some research, I found out that selected only a substring of tokens is a valid solution and that selecting the beginning, middle, or end can have different results. Supposedly for long paragraphs or essay-like documents, selecting the end tokens yields more accurate sentiment readings since that is where the concluding remarks are found. This seemed to prove true as is shown in the figures.

## 4

Author’s address: Elliot Heisler.

2023. XXXX-XXXX/2023/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



```
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   review_id        17494 non-null  int64
 1   title            17494 non-null  object
 2   year             17316 non-null  float64
 3   user_review      17494 non-null  object
 4   user_suggestion  17494 non-null  int64
dtypes: float64(1), int64(2), object(2)
memory usage: 683.5+ KB
```

```
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   review_id        17316 non-null  int64
 1   title            17316 non-null  object
 2   year             17316 non-null  float64
 3   user_review      17316 non-null  object
 4   user_suggestion  17316 non-null  int64
 5   is_early_access  17316 non-null  bool
 6   received_free    17316 non-null  bool
 7   contains_art     17316 non-null  bool
dtypes: bool(3), float64(1), int64(2), object(2)
memory usage: 862.4+ KB
```

Fig. 1. Dataset: The dataset was unchanged.

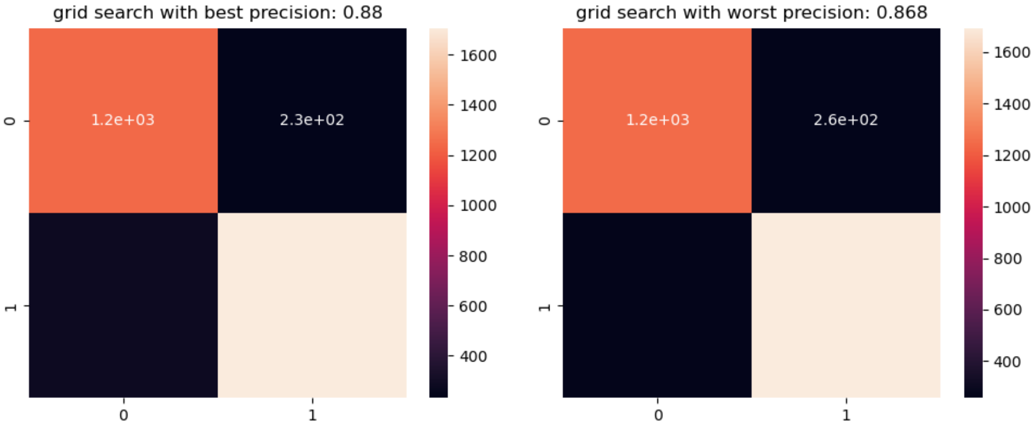


Fig. 2. Results: Confusion matrices of classifiers with the best and worst precisions

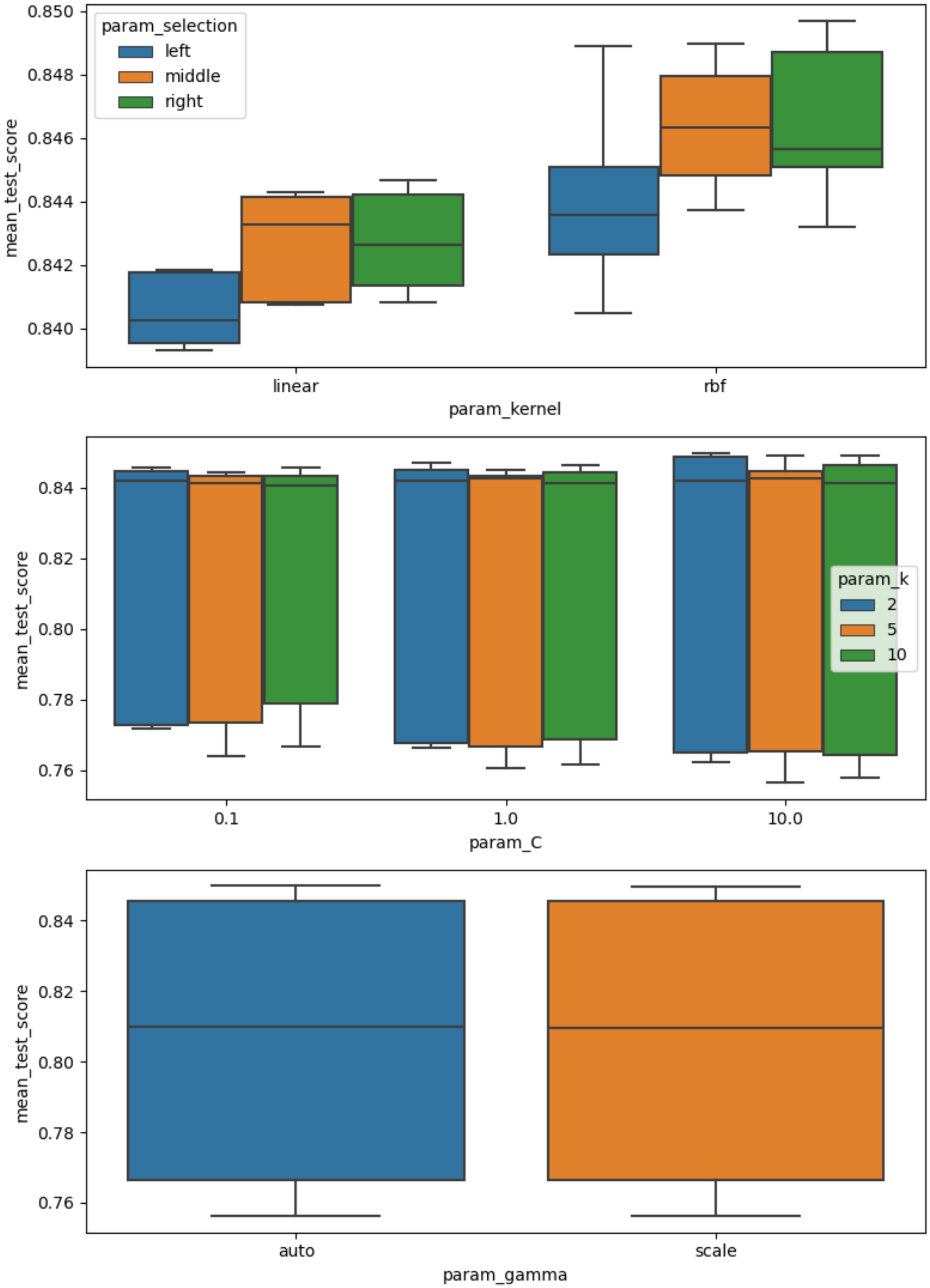


Fig. 3. Additional Results: Boxplots showing means of classifier accuracies accross the different hyperparameters