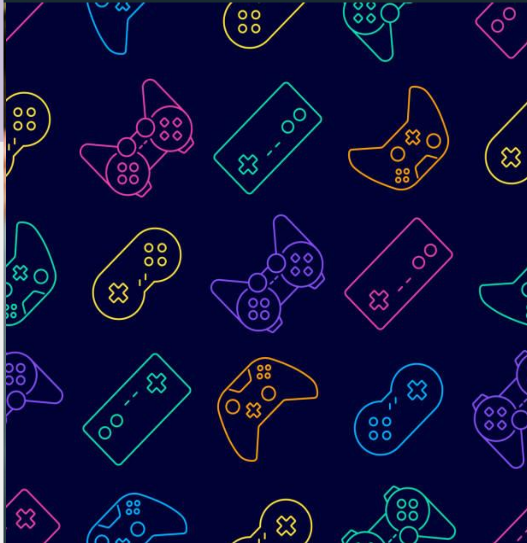
Steam Game Reviews Analysis

Ashley Bakaitus
LING2020
December 2, 2025



Contents

01 Steam Intro & Info

02 Data Gathering & Preparation

03 Data Information & Analysis

04 Classifier

05 Summary

Steam Info

- ## What is Steam?

App & Marketplace for video games

Hosts more than 100,000 games
- 18,000 added in 2024

- ## Steam Reviews

User feedback on games for other players to consult before shopping for themselves

- ## Game Rankings

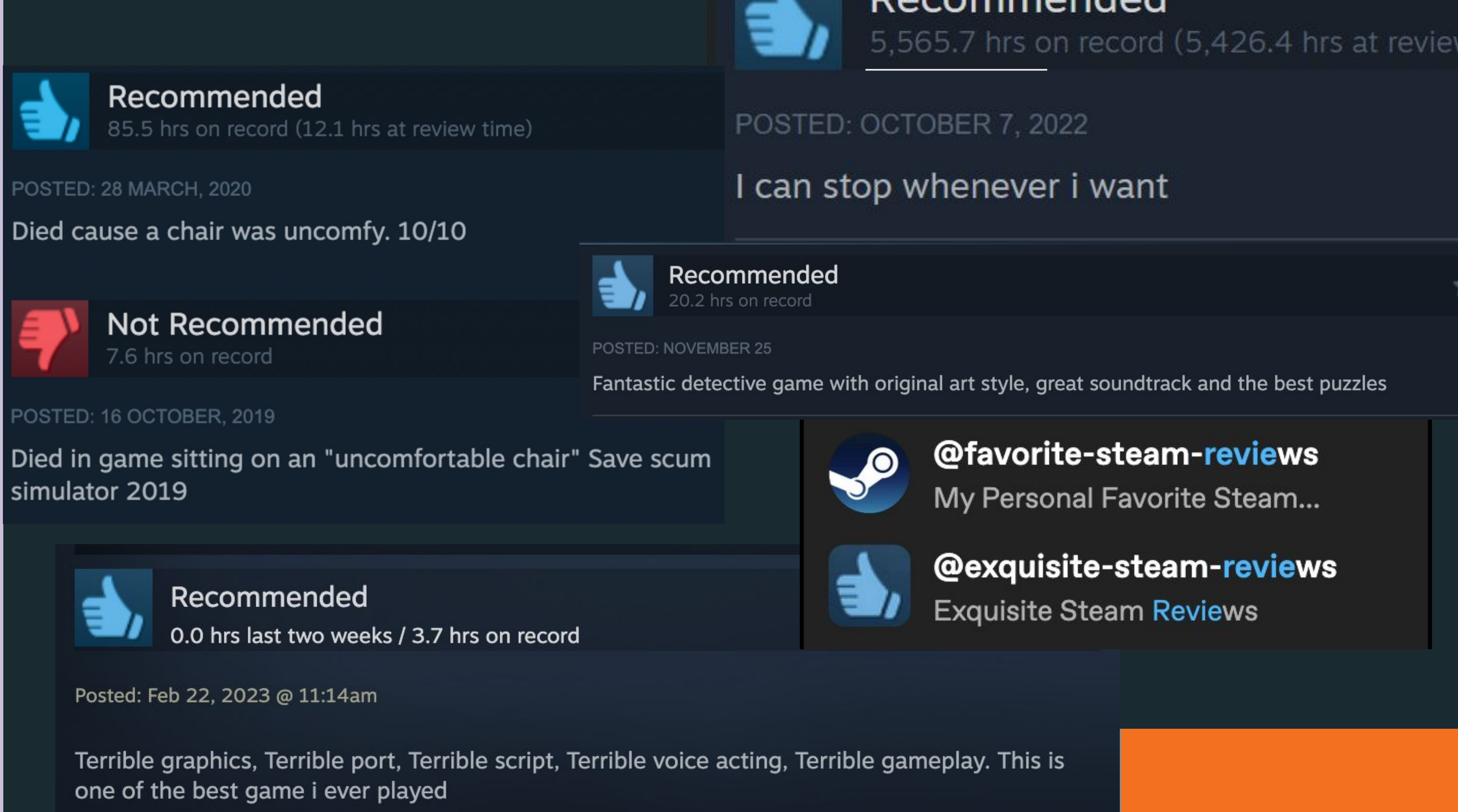
How well a game is being reviewed

Based on the ratio of positive/negative reviews and total reviews shared

Overwhelmingly Positive, Very Positive, Positive, Mostly Positive, Mixed, M. Negative, Negative, V. Negative, O. Negative

Ashley Bakaitus, cool, 1998.

Why Game Reviews?



Recommended
5,565.7 hrs on record (5,426.4 hrs at revie

POSTED: OCTOBER 7, 2022

I can stop whenever i want

Recommended
85.5 hrs on record (12.1 hrs at review time)

POSTED: 28 MARCH, 2020

Died cause a chair was uncomfy. 10/10

Recommended
20.2 hrs on record

POSTED: NOVEMBER 25

Fantastic detective game with original art style, great soundtrack and the best puzzles

Not Recommended
7.6 hrs on record

POSTED: 16 OCTOBER, 2019

Died in game sitting on an "uncomfortable chair" Save scum simulator 2019

@favorite-steam-reviews
My Personal Favorite Steam...

@exquisite-steam-reviews
Exquisite Steam Reviews

Recommended
0.0 hrs last two weeks / 3.7 hrs on record

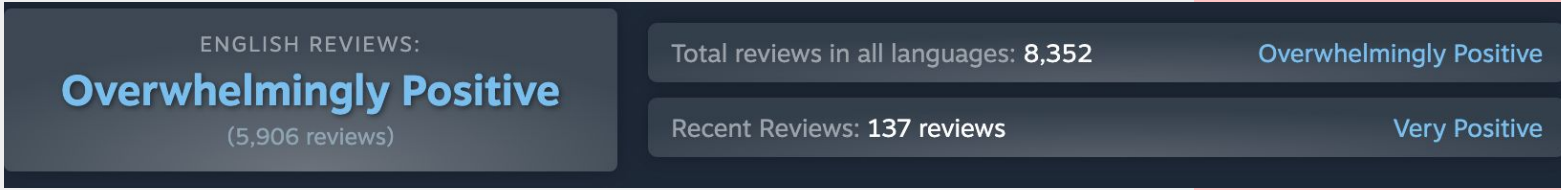
Posted: Feb 22, 2023 @ 11:14am

Terrible graphics, Terrible port, Terrible script, Terrible voice acting, Terrible gameplay. This is one of the best game i ever played

Gathering data

- Single-player games (primarily)
- older than 1 or 2 years (where possible)
- Not AAA studios (filtered roughly by selecting by price)
- Price range <$40 (not worried about minimum price but >$10 ideal)
- Has at least 1000 number of reviews (where possible)
- Game in English, only considering English reviews
- Full games only (no DLC), sequels OK
- No limitations on genre or game play style, single player only main limit
- only select games that the overall review ranking and the english review ranking are the same

Narrowed my search down to 48,121 game titles

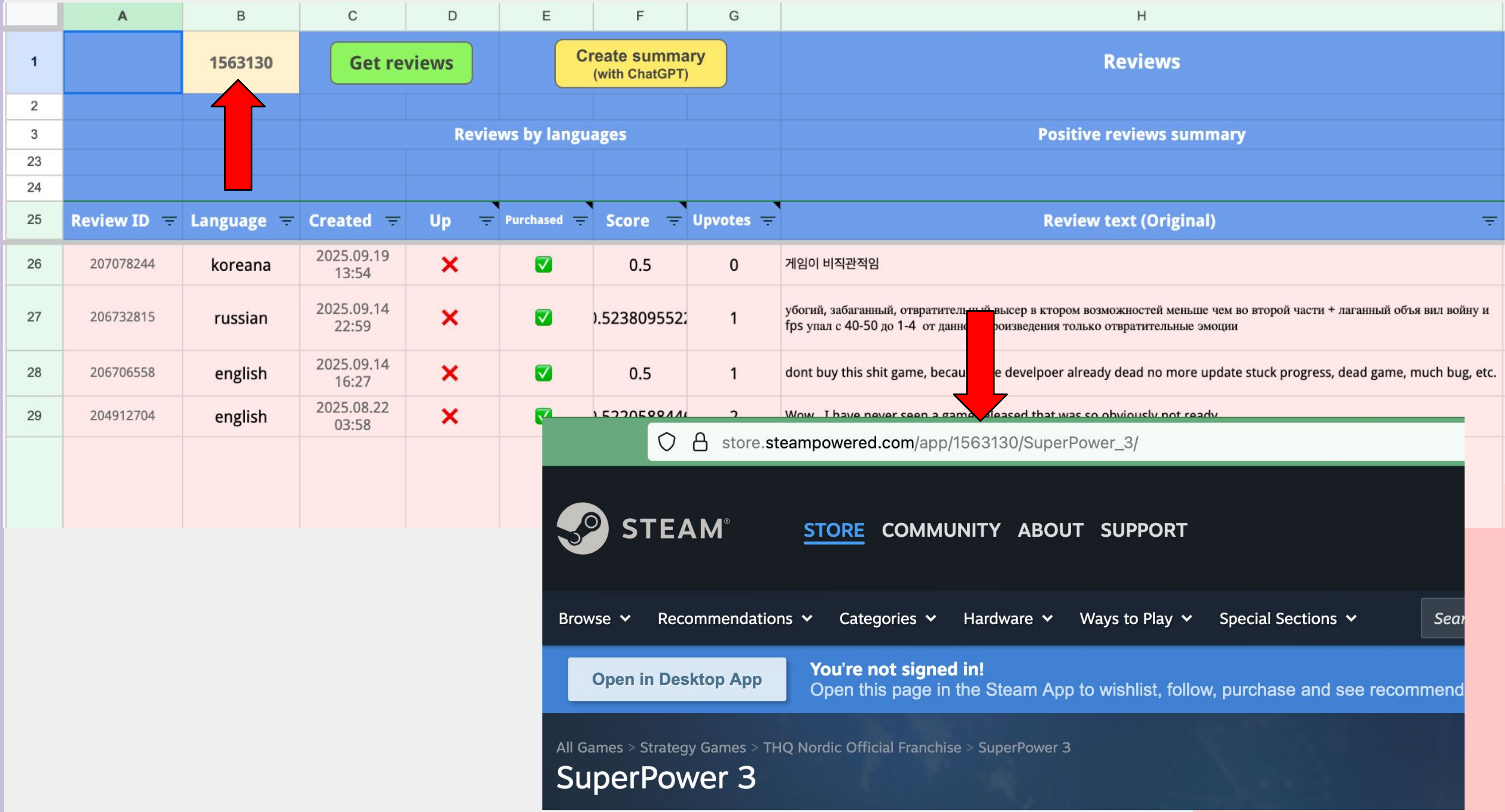


ENGLISH REVIEWS:
Overwhelmingly Positive
(5,906 reviews)

Total reviews in all languages: 8,352 Overwhelmingly Positive

Recent Reviews: 137 reviews Very Positive

Gathering data



	A	B	C	D	E	F	G	H
1		1563130	Get reviews		Create summary (with ChatGPT)			Reviews
2								
3			Reviews by languages					Positive reviews summary
23								
24								
25	Review ID	Language	Created	Up	Purchased	Score	Upvotes	Review text (Original)
26	207078244	koreana	2025.09.19 13:54	✗	✓	0.5	0	게임이 비직관적임
27	206732815	russian	2025.09.14 22:59	✗	✓	).523809552	1	убогий, забаганный, отвратитель высер в ктором возможностей меньше чем во второй части + лаганный объя вил войну и fps упал с 40-50 до 1-4 от данн роизведения только отвратительные эмоции
28	206706558	english	2025.09.14 16:27	✗	✓	0.5	1	dont buy this shit game, becau e develpoer already dead no more update stuck progress, dead game, much bug, etc.
29	204912704	english	2025.08.22 03:58	✗	✓	).522058844	2	Wow, I have never seen a game leased that was so obviously not ready

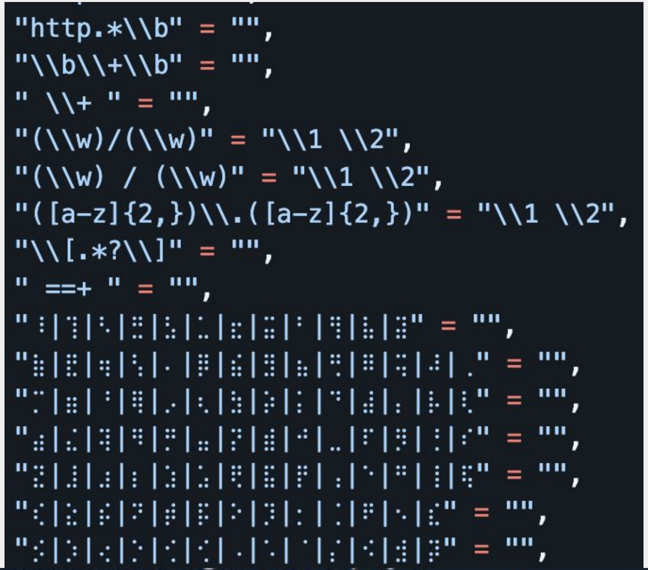
store.steampowered.com/app/1563130/SuperPower_3/

You're not signed in!
Open this page in the Steam App to wishlist, follow, purchase and see recommend

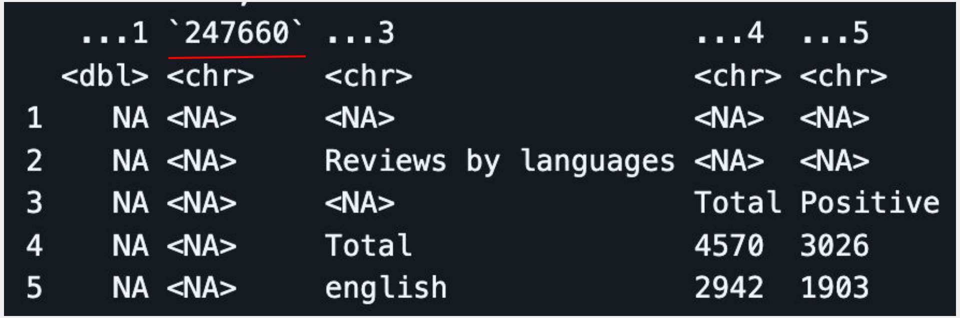
All Games > Strategy Games > THQ Nordic Official Franchise > SuperPower 3

SuperPower 3

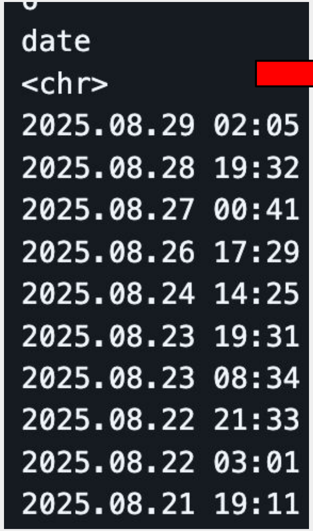
Cleaning data



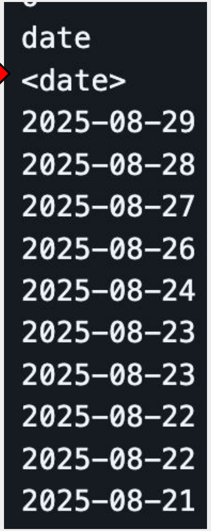
```
"http.*\\b" = "",
"\\b\\+\\b" = "",
" \\+ " = "",
"(\\w)/(\\w)" = "\\1 \\2",
"(\\w) / (\\w)" = "\\1 \\2",
"([a-z]{2,})\\.([a-z]{2,})" = "\\1 \\2",
"\\[.*?\\]" = "",
" ==+ " = "",
"⠸|⠹|⠱|⠿|⠵|⠤|⠮|⠿|⠁|⠻|⠷|⠿" = "",
"⠷|⠿|⠶|⠣|⠂|⠟|⠾|⠿|⠴|⠛|⠿|⠾|⠲|." = "",
"⠉|⠶|⠈|⠻|⠔|⠣|⠳|⠺|⠨|⠉|⠼|⠨|⠺|⠫" = "",
"⠴|⠦|⠹|⠛|⠏|⠤|⠏|⠾|⠊|⠤|⠟|⠻|⠸|⠌" = "",
"⠿|⠼|⠴|⠸|⠹|⠦|⠻|⠿|⠟|⠠|⠑|⠛|⠸|⠻" = "",
"⠪|⠵|⠮|⠋|⠾|⠟|⠪|⠹|⠨|⠰|⠟|⠢|⠪" = "",
"⠎|⠕|⠔|⠕|⠪|⠪|⠄|⠑|⠈|⠌|⠪|⠾|⠏" = "",
```



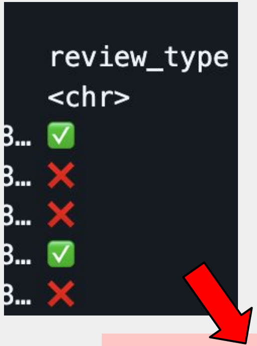
```
    ...1 `247660` ...3                 ...4  ...5
   <dbl> <chr>    <chr>                <chr> <chr>
1     NA <NA>     <NA>                 <NA>  <NA>
2     NA <NA>     Reviews by languages <NA>  <NA>
3     NA <NA>     <NA>                 Total Positive
4     NA <NA>     Total                4570  3026
5     NA <NA>     english              2942  1903
```



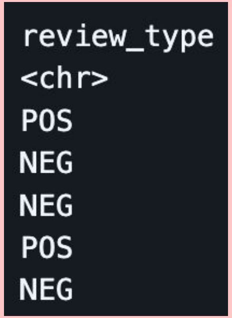
```
date
<chr>
2025.08.29 02:05
2025.08.28 19:32
2025.08.27 00:41
2025.08.26 17:29
2025.08.24 14:25
2025.08.23 19:31
2025.08.23 08:34
2025.08.22 21:33
2025.08.22 03:01
2025.08.21 19:11
```



```
date
<date>
2025-08-29
2025-08-28
2025-08-27
2025-08-26
2025-08-24
2025-08-23
2025-08-23
2025-08-22
2025-08-22
2025-08-21
```



```
   review_type
   <chr>
8… ✅
8… ❌
8… ❌
8… ✅
8… ❌
```



```
review_type
<chr>
POS
NEG
NEG
POS
NEG
```

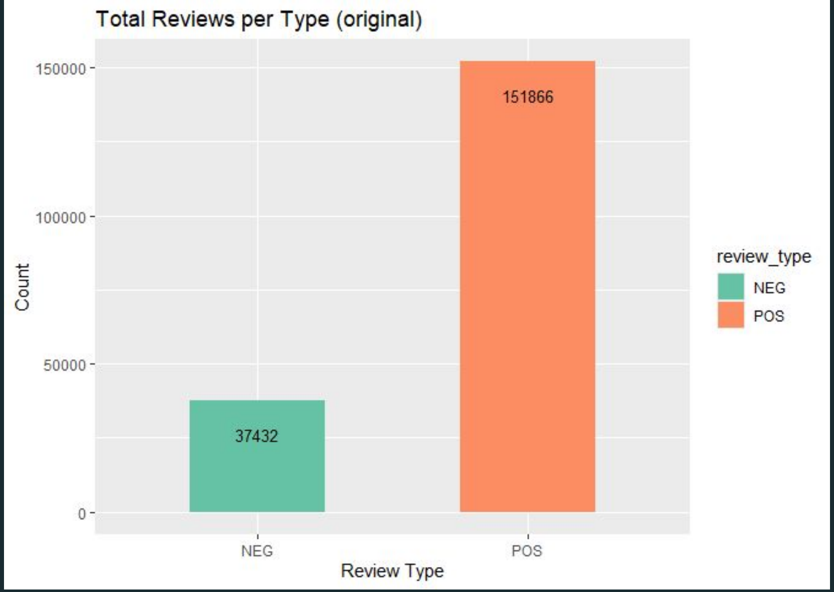
The Data

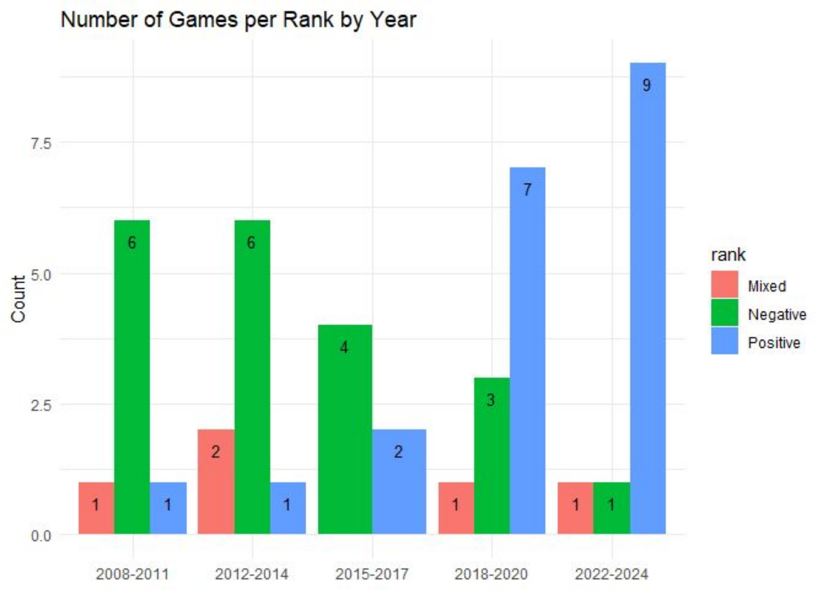
45 games

- 5 per Steam ranking

2008-2024 (skipping 2021)

189,298 reviews





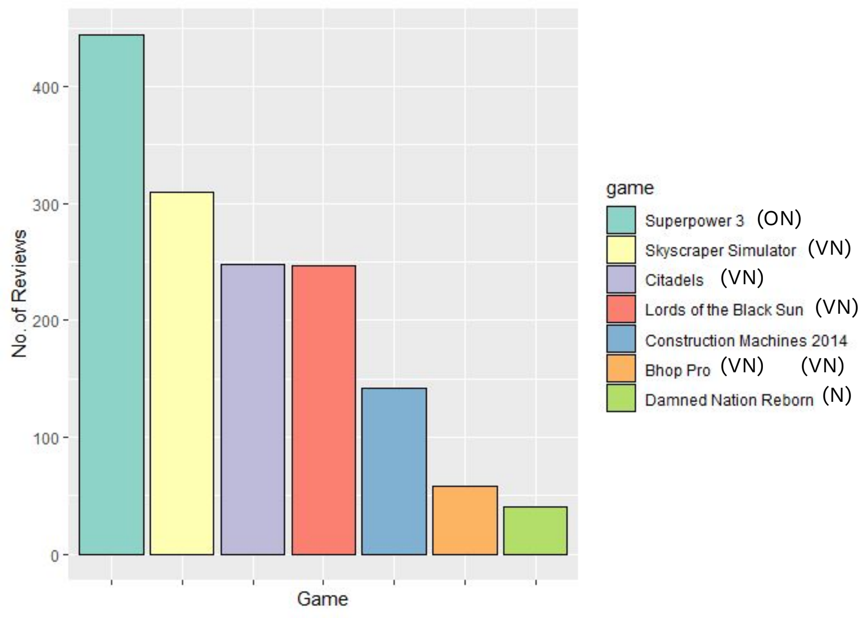
But!

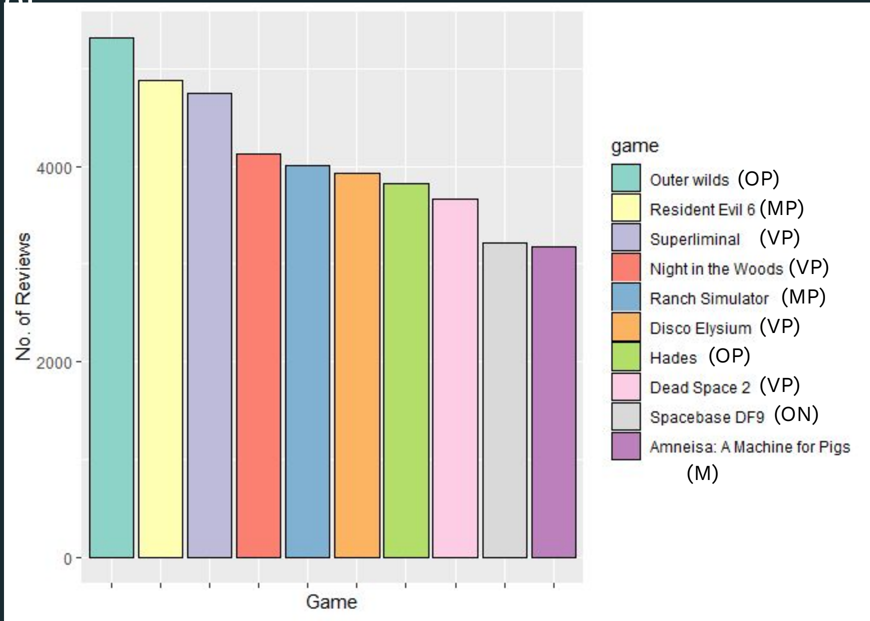
Steam Reviews Analysis

The Data

Positively reviewed games get a significant amount of reviews more than negatively reviewed games*

This caused some problems with running comparisons



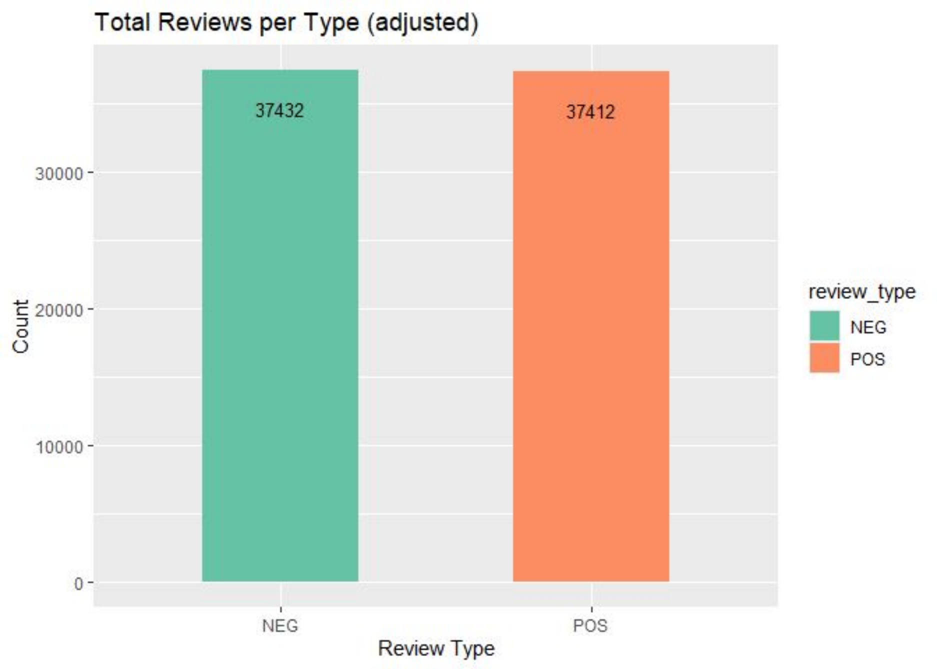


*one exception for the games in the plain Negative and Positive rankings, which are only populated by games with ≤ 50 reviews

Downsampling

Proportionally removing positive reviews per game until they're a bit more equal to the amount of negative reviews.

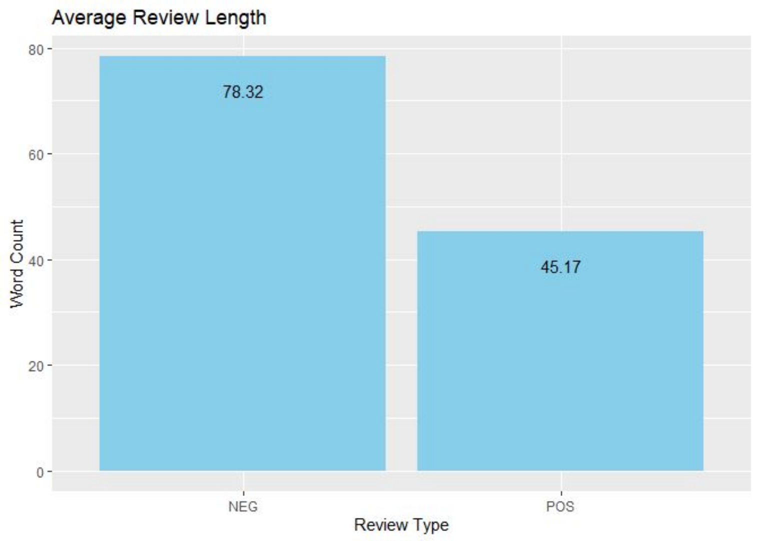
This way the elements of the reviews to look at, word count, TTR, tf-idf, can be more easily comparable.

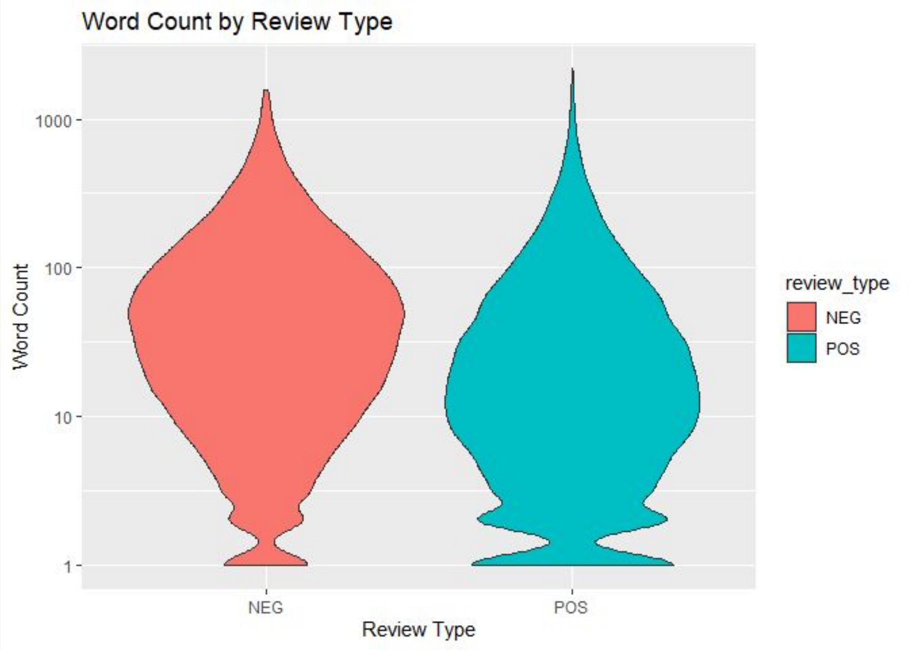


Review Length

Overall avg review length = ~62 words (median 22)

Negative reviews are a bit longer on average!



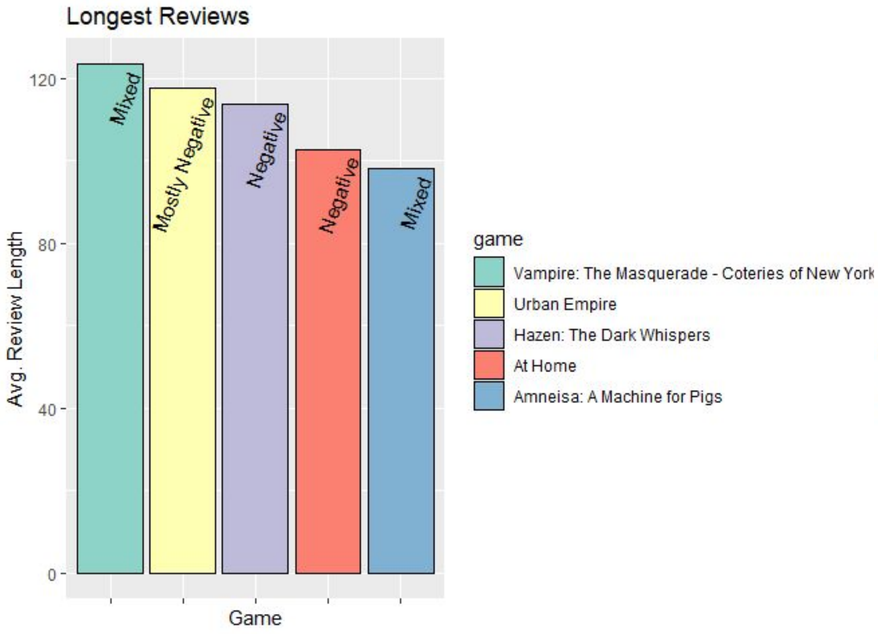


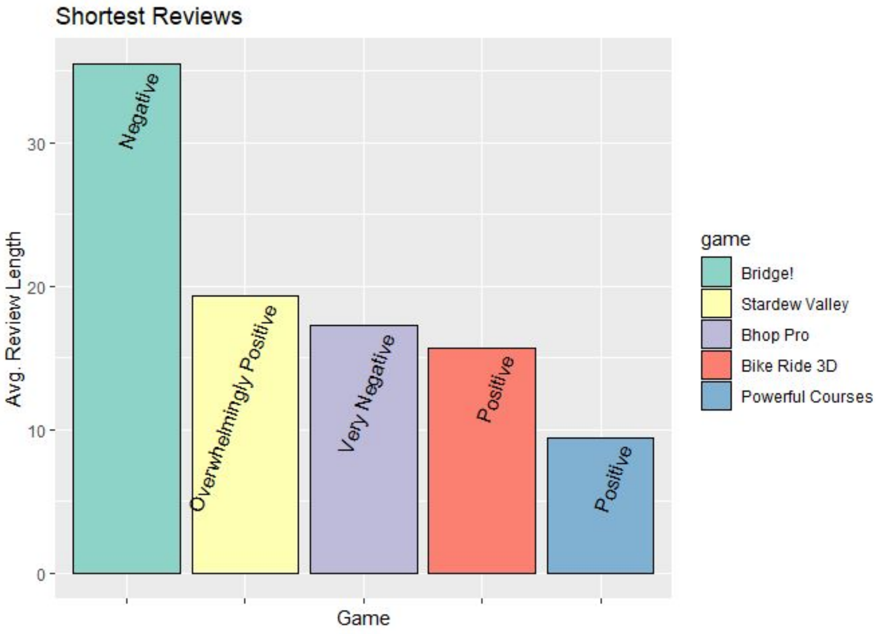
Positive reviews rule the extremes.

27,493 Positive reviews are 5 words or shorter (14,475 Neg)

17 reviews > 1700 words long are all Positive

Review Length





Steam Reviews Analysis

Review Length

TTR

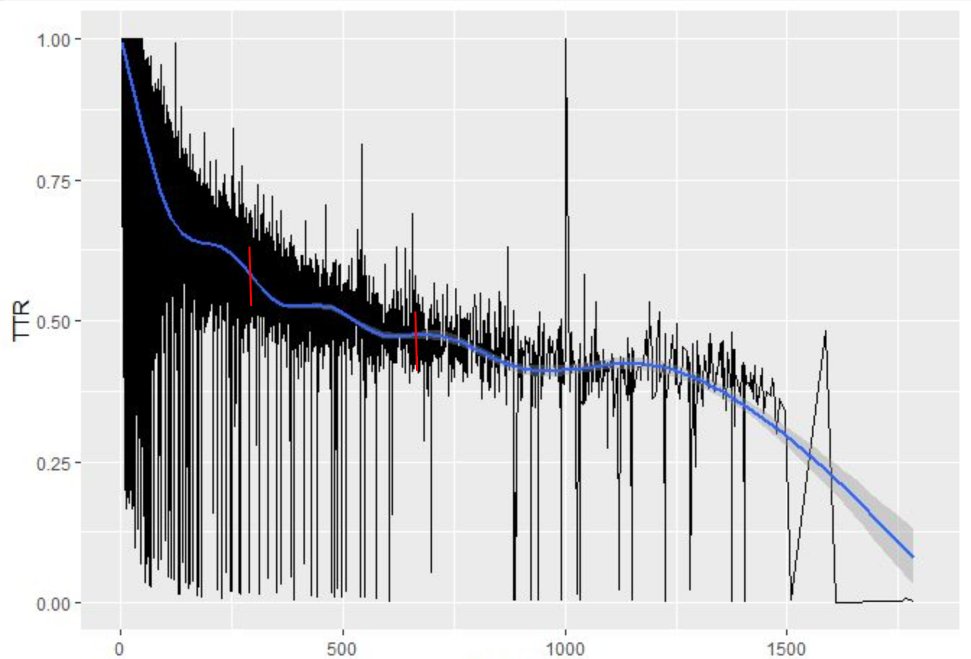
Type-Token Ratio

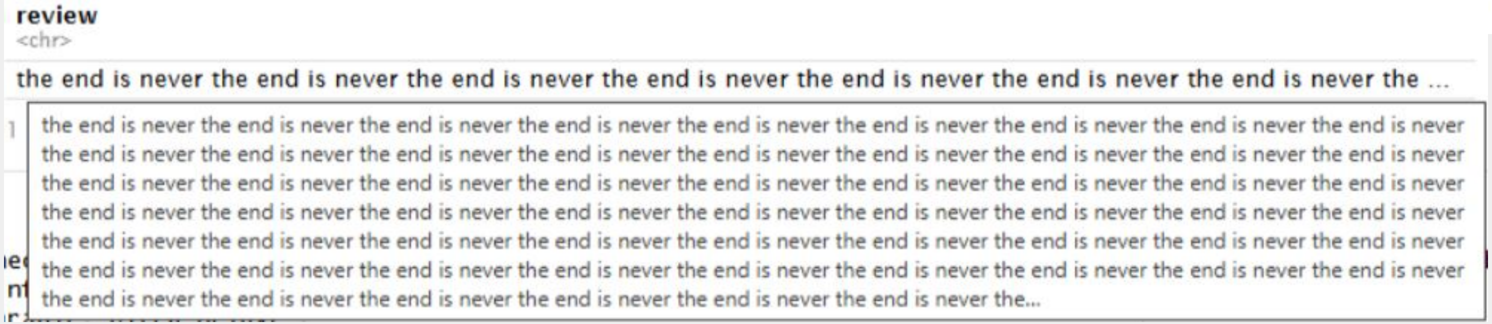
A measurement of language variety of a document

Dividing word types (unique words) by word tokens (all words used).

Closer to 0 -> repetitive document
Closer to 1 -> varied document

Very sensitive to document length!

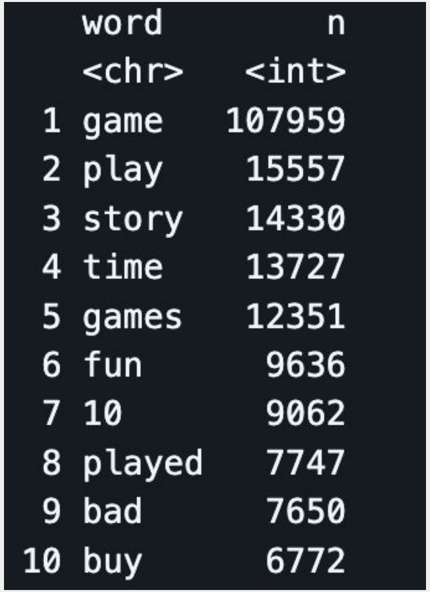



review
<chr>

the end is never the end is never the end is never the end is never the end is never the end is never the end is never the ...

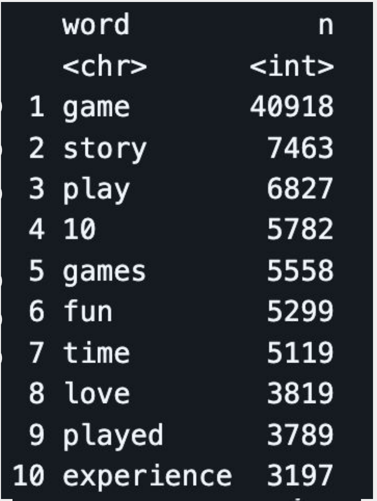
1 the end is never the...

TF-IDF



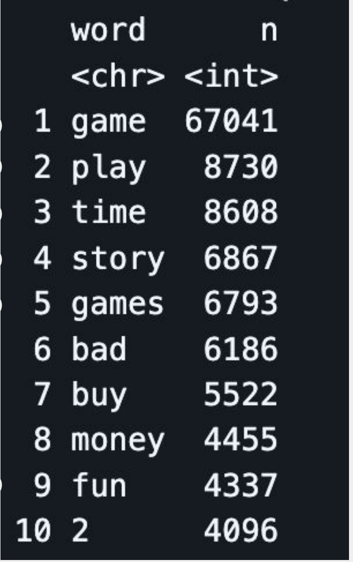
```
     word         n
     <chr>    <int>
   1 game    107959
   2 play     15557
   3 story    14330
   4 time     13727
   5 games    12351
   6 fun       9636
   7 10        9062
   8 played    7747
   9 bad       7650
  10 buy       6772
```

Positive



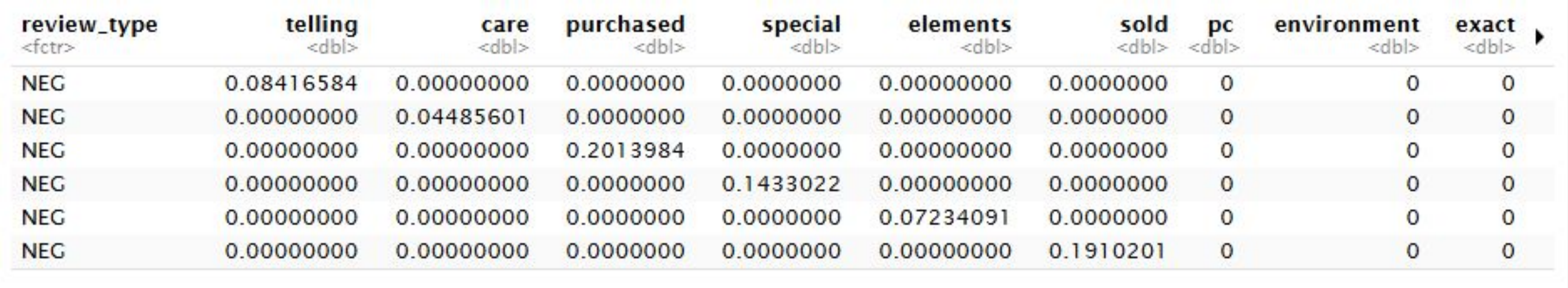
```
     word             n
     <chr>        <int>
●  1 game         40918
●  2 story         7463
●  3 play          6827
   4 10            5782
●  5 games         5558
●  6 fun           5299
●  7 time          5119
   8 love          3819
   9 played        3789
  10 experience    3197
```

Negative



```
     word      n
     <chr> <int>
●  1 game  67041
●  2 play   8730
●  3 time   8608
●  4 story  6867
●  5 games  6793
   6 bad    6186
   7 buy    5522
   8 money  4455
●  9 fun    4337
  10 2      4096
```

The Classifier



review_type	telling	care	purchased	special	elements	sold	pc	environment	exact
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
NEG	0.08416584	0.00000000	0.0000000	0.0000000	0.00000000	0.0000000	0	0	0
NEG	0.00000000	0.04485601	0.0000000	0.0000000	0.00000000	0.0000000	0	0	0
NEG	0.00000000	0.00000000	0.2013984	0.0000000	0.00000000	0.0000000	0	0	0
NEG	0.00000000	0.00000000	0.0000000	0.1433022	0.00000000	0.0000000	0	0	0
NEG	0.00000000	0.00000000	0.0000000	0.0000000	0.07234091	0.0000000	0	0	0
NEG	0.00000000	0.00000000	0.0000000	0.0000000	0.00000000	0.1910201	0	0	0

After TF-IDF and stop word removal, the data has a bit more NEG reviews than POS

Remember those overlapping high frequency words from before?

Things had to get a little tricky...

The Classifier

Some additional treatment

- Drop any column that has only 1 TF-IDF value
- Remove highly common overlapping words
  - Anti_join stop words
  - Filter by TF-IDF value
- Select an equal amount of POS and NEG vals

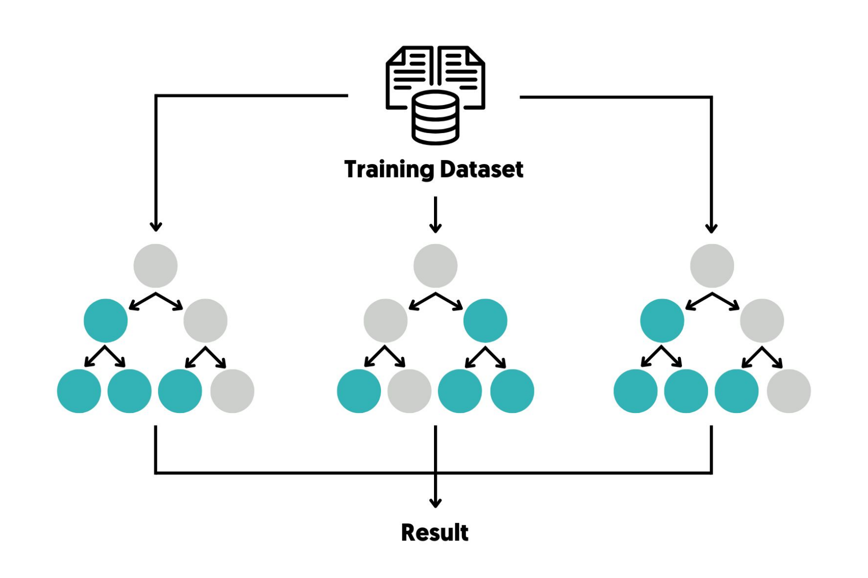
Sample Data

- 300,000 rows from the total dataframe

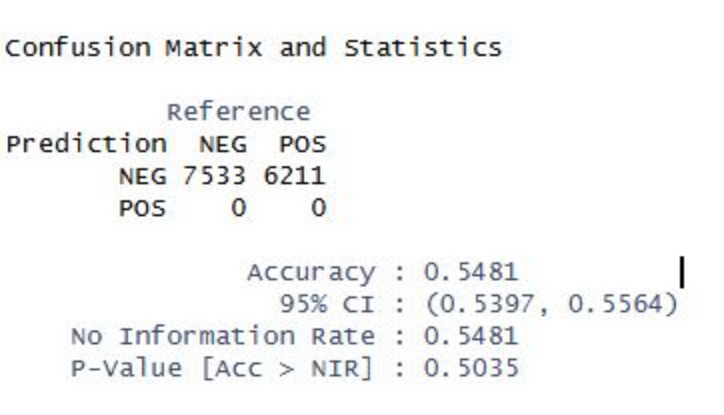
Training Data:

- 40,226 rows
- 14,123 columns (words)

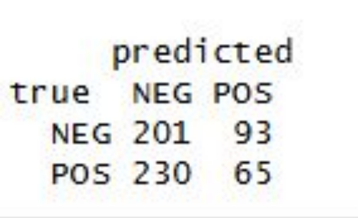
Random Forest Method



The Classifier



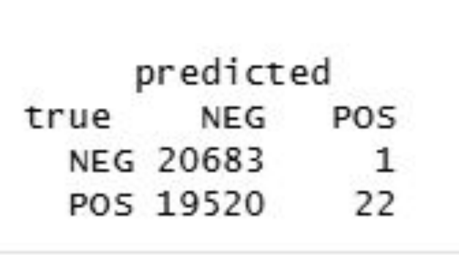
```
Confusion Matrix and Statistics

          Reference
Prediction  NEG  POS
       NEG 7533 6211
       POS    0    0

               Accuracy : 0.5481
                 95% CI : (0.5397, 0.5564)
    No Information Rate : 0.5481
    P-Value [Acc > NIR] : 0.5035
```



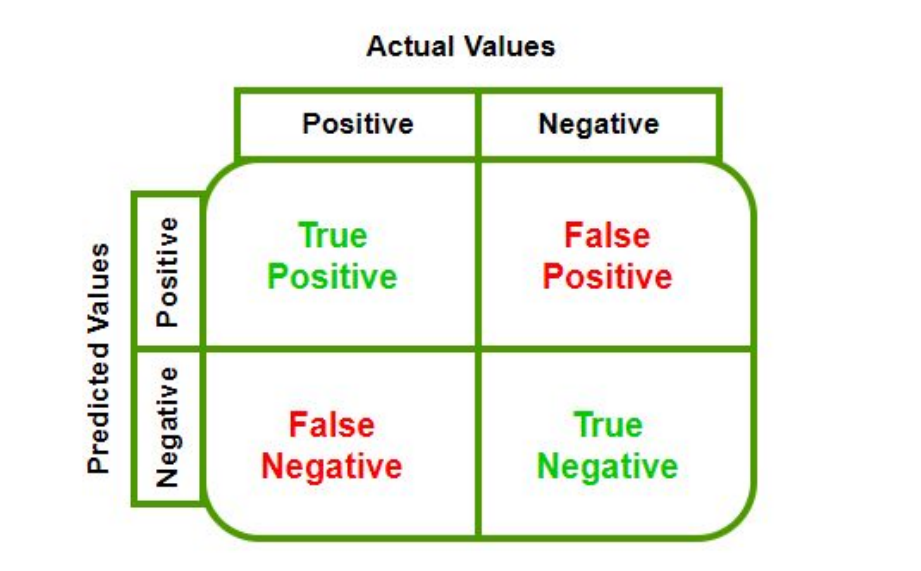
```
     predicted
true   NEG POS
   NEG 201  93
   POS 230  65
```

<- tiny sample

full sample ->



```
       predicted
true        NEG    POS
   NEG 20683      1
   POS 19520     22
```



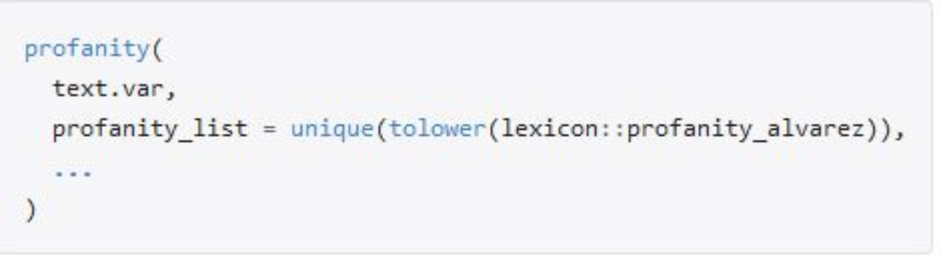
Actual Values

		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

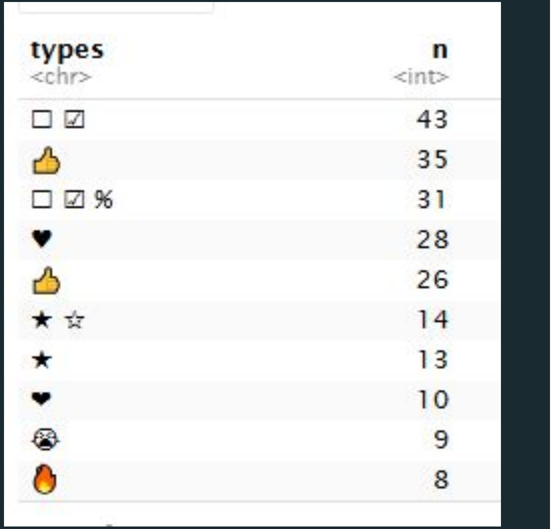
Over-predicting negative classification

Additional Goals

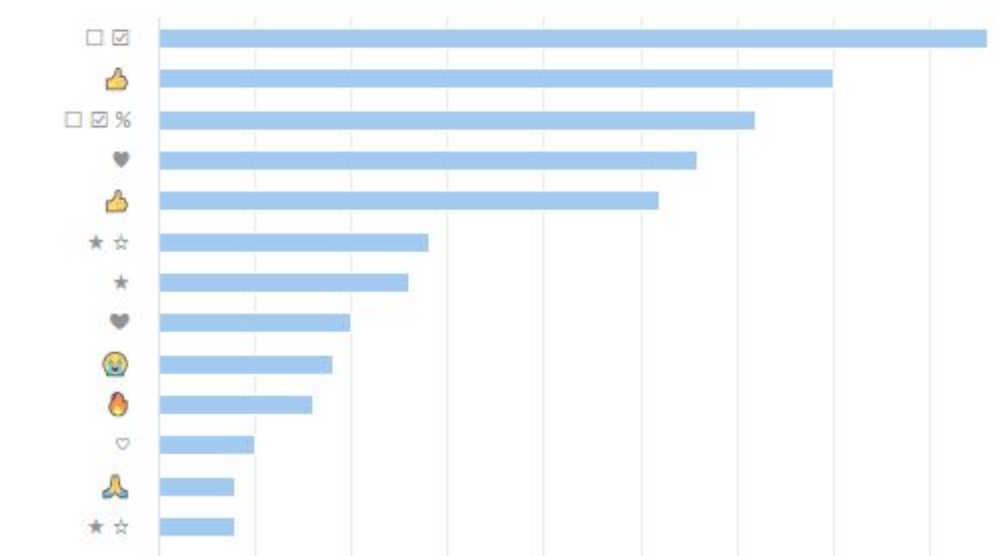
1. Emoji usage
2. Profanity usage
3. Number of reviews by date



```
profanity(
  text.var,
  profanity_list = unique(tolower(lexicon::profanity_alvarez)),
  ...
)
```



types <chr>	n <int>
□ ☑	43
👍	35
□ ☑ %	31
♥	28
👍	26
★ ☆	14
★	13
❤	10
😭	9
🔥	8



THANK YOU!