StarWars vs StarTrek

the battle for Comic Con dominance

By Christopher Villafuerte

Goal.

StarWars corporate (Disney) wants to have an overwhelming turnout of fans for the Comic Con 2021

How.

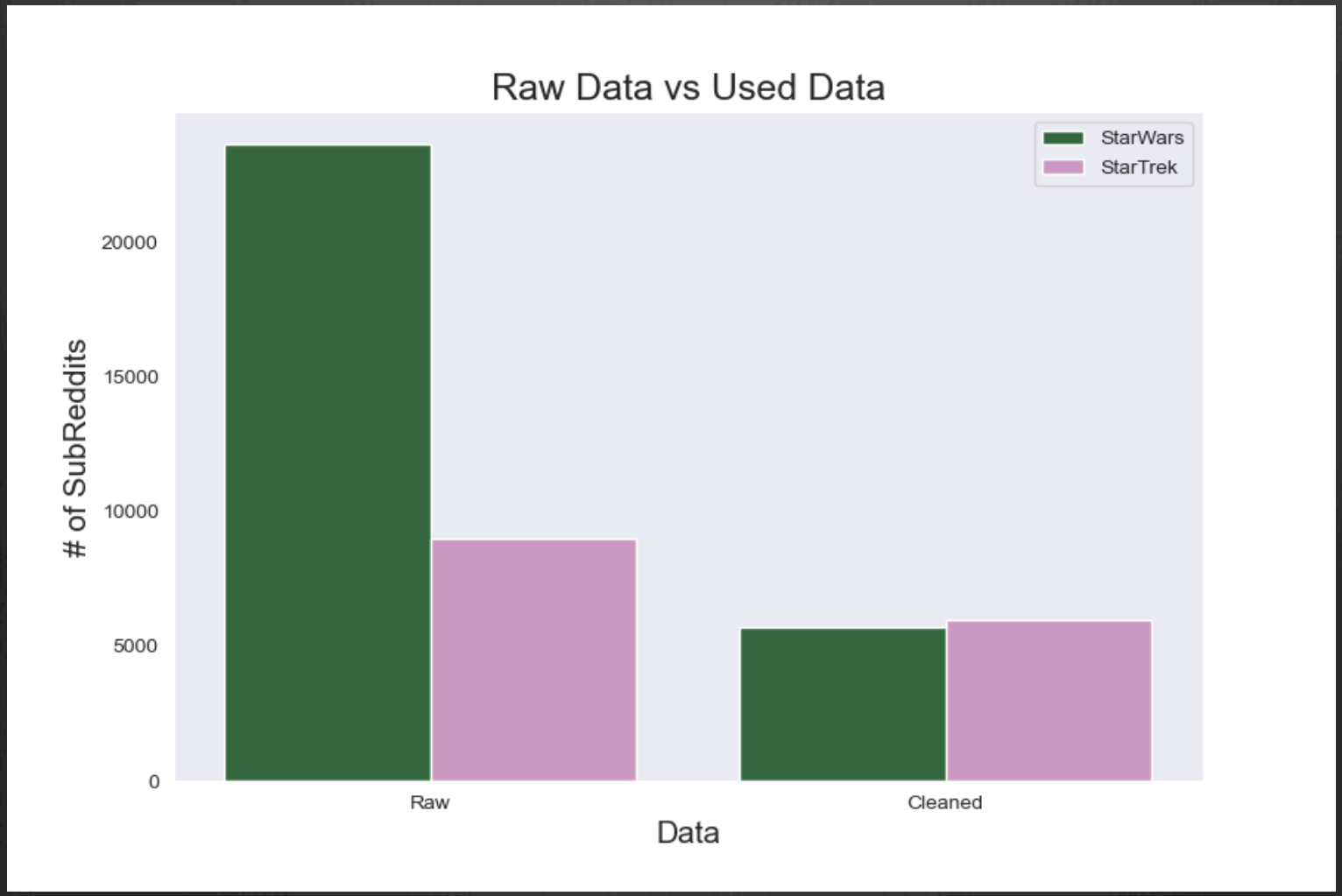
To achieve this Disney is sending reward their fans on the subbredit StarWars, offering a secret $15 discount off ComicCon 2021 entry fee.

Problem:

The promotion is only for participants on the StarWars subreddit, therefore corporate want to make sure that only StarWars fans receive this promotions. (and not Trekies)

Process is to
create a model that yields a high accuracy and precision score in a binary classification between the subreddits
StarWars and StarTrek?

Data:


Raw Data vs Used Data
StarWars
StarTrek
20000
15000
10000
5000
0
# of SubReddits
Raw
Cleaned
Data

BaseLine

51%

Models

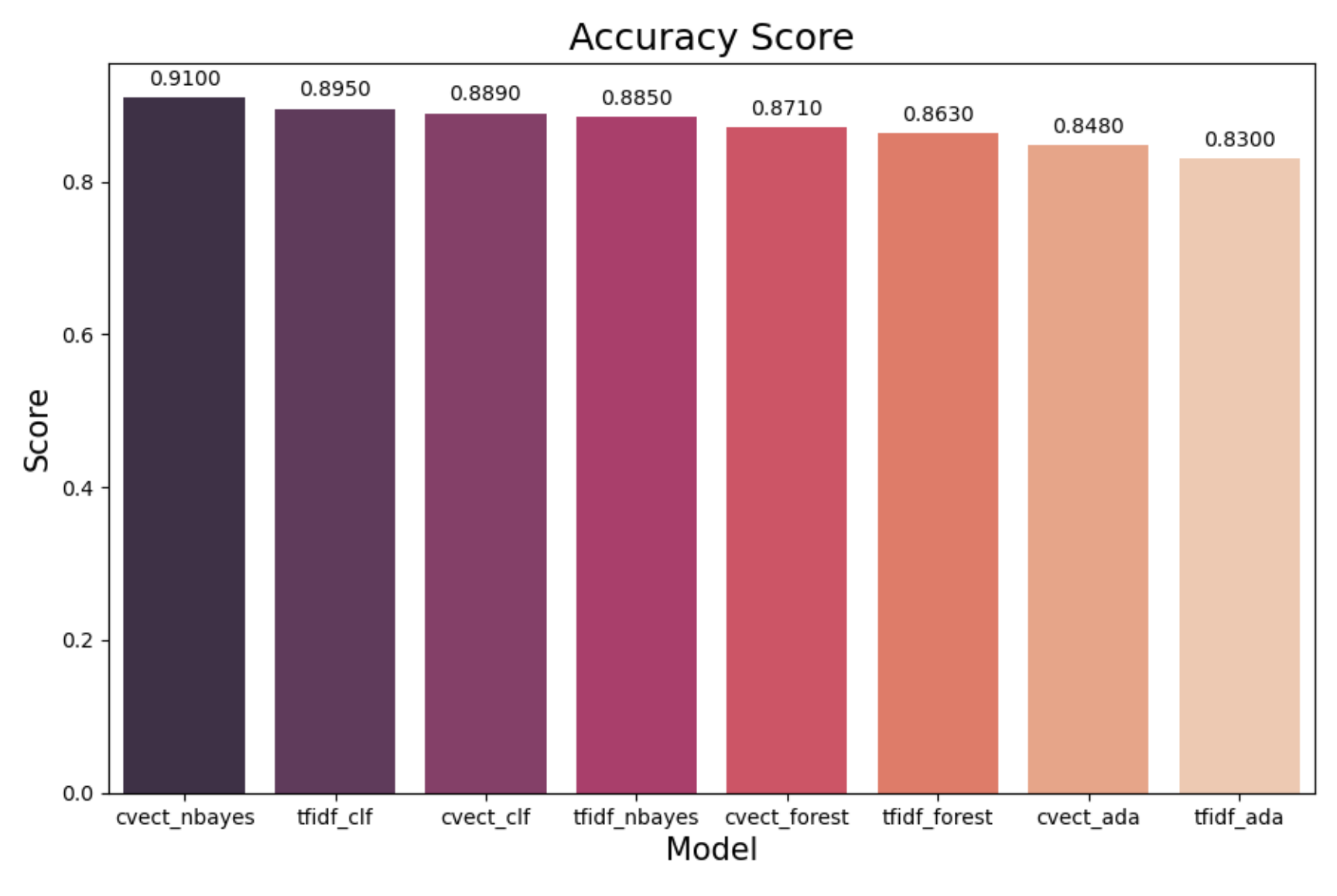
Naïve Bayes,
Logistic Regression,
Random Forest
Adaboost
classifier models

Paired with:

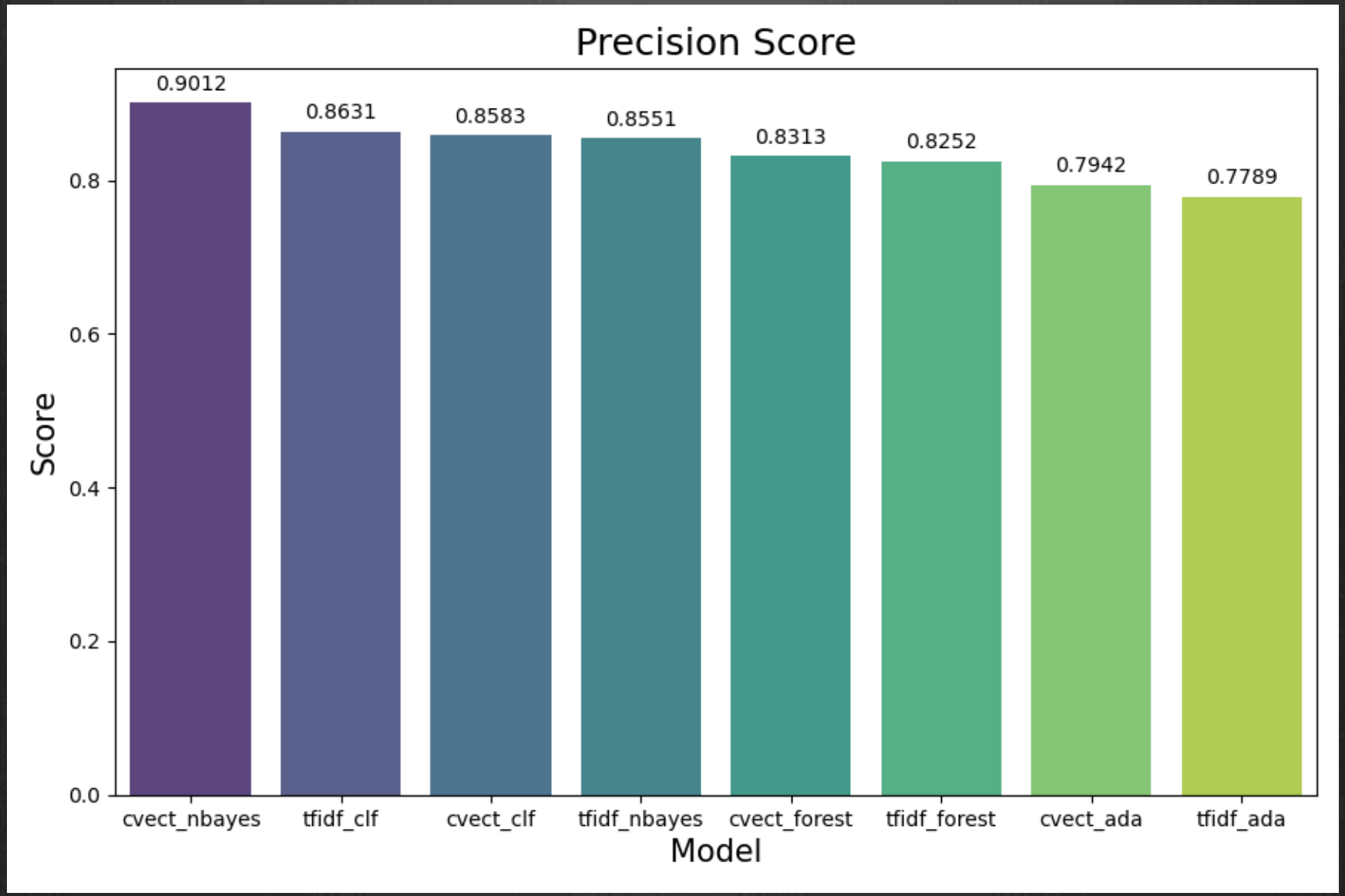
TfidfVectorizer and CountVectorizer

both with standard stop words.

Accuracy Score



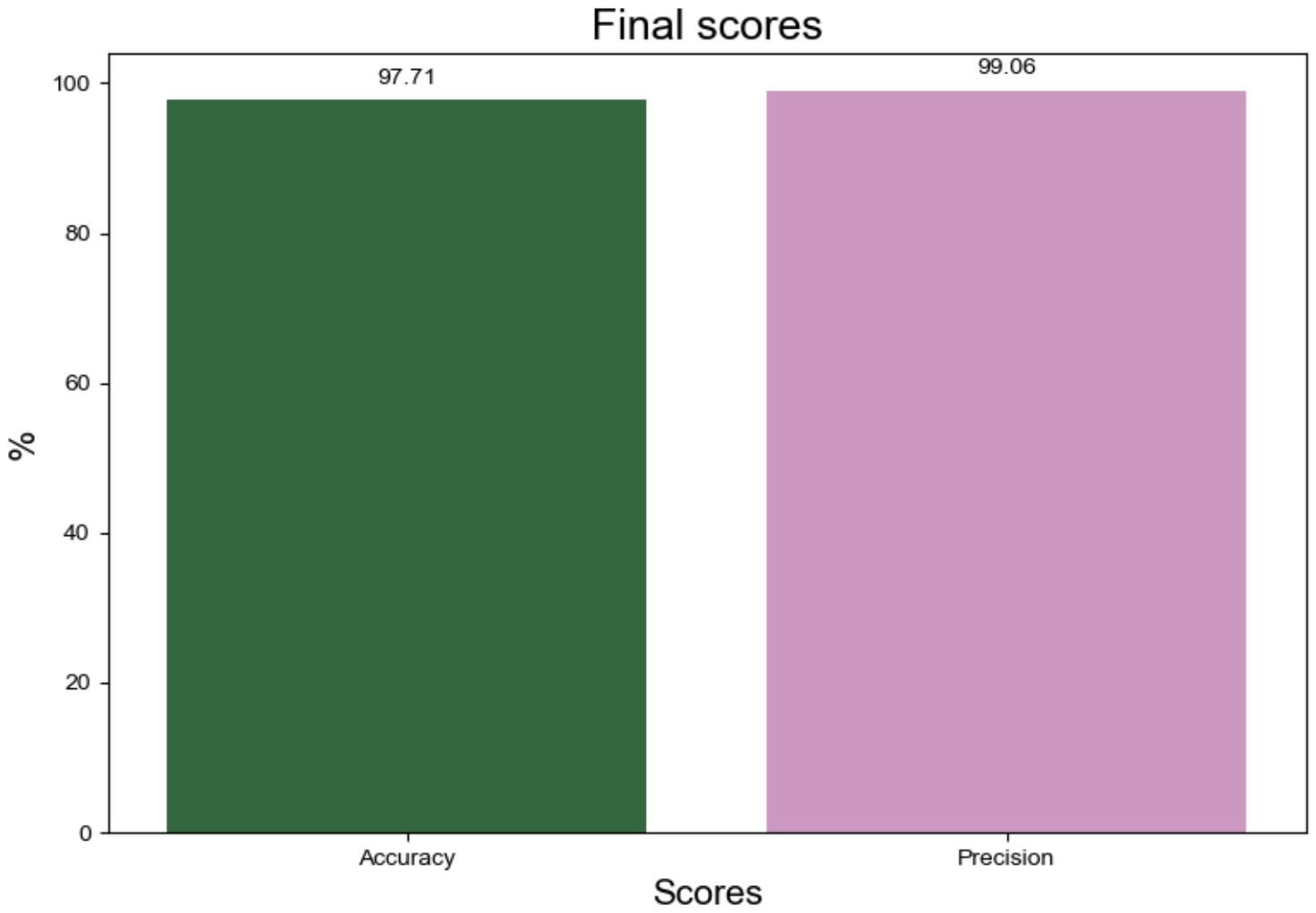
Precision Score



Final Model

Naive Bayes
&
CountVectorizer

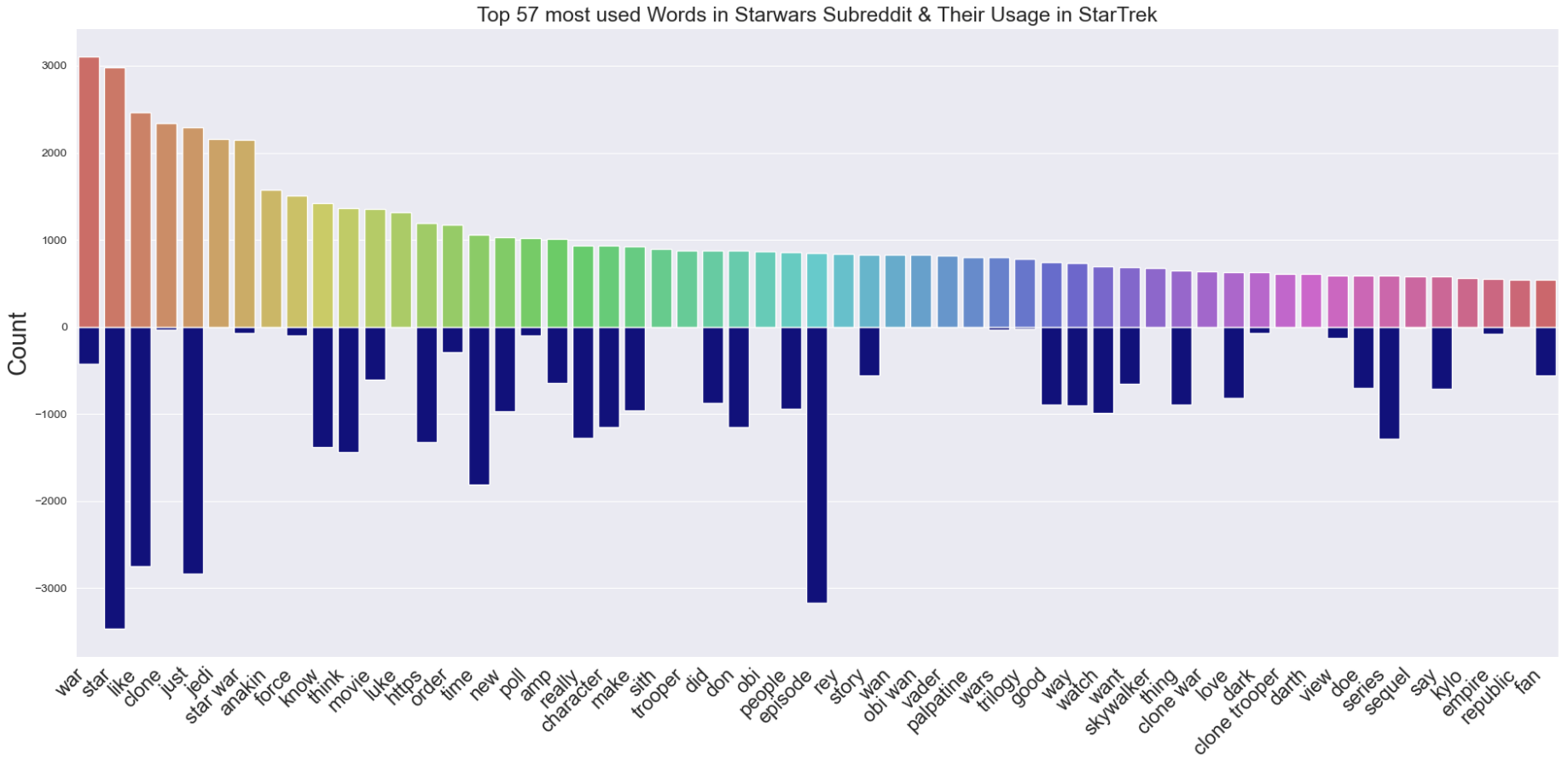
Final Scores



Tuning factors

-Heavy data cleaning
- Lemmatizing
- Custom stop words
(1-2) n-grams
5000 features

57 Most Used Words
in Starwars subReddit



With this model we can confidently infer the difference between StarWars and StarTrek Reddits