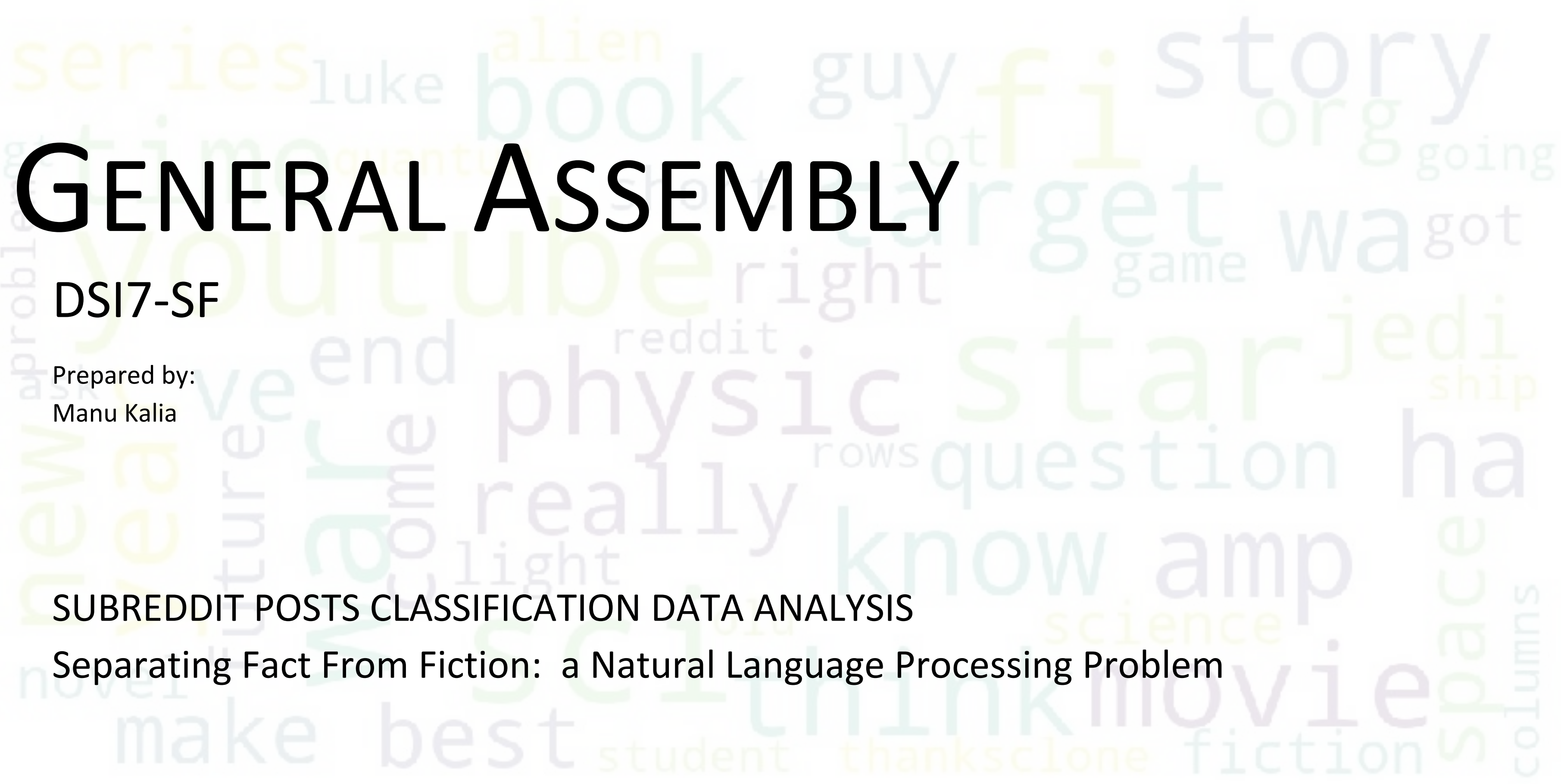


# General Assembly

## DSI7-SF

Prepared by:
Manu Kalia

SUBREDDIT POSTS CLASSIFICATION DATA ANALYSIS

Separating Fact From Fiction: a Natural Language Processing Problem

# PROBLEM STATEMENT

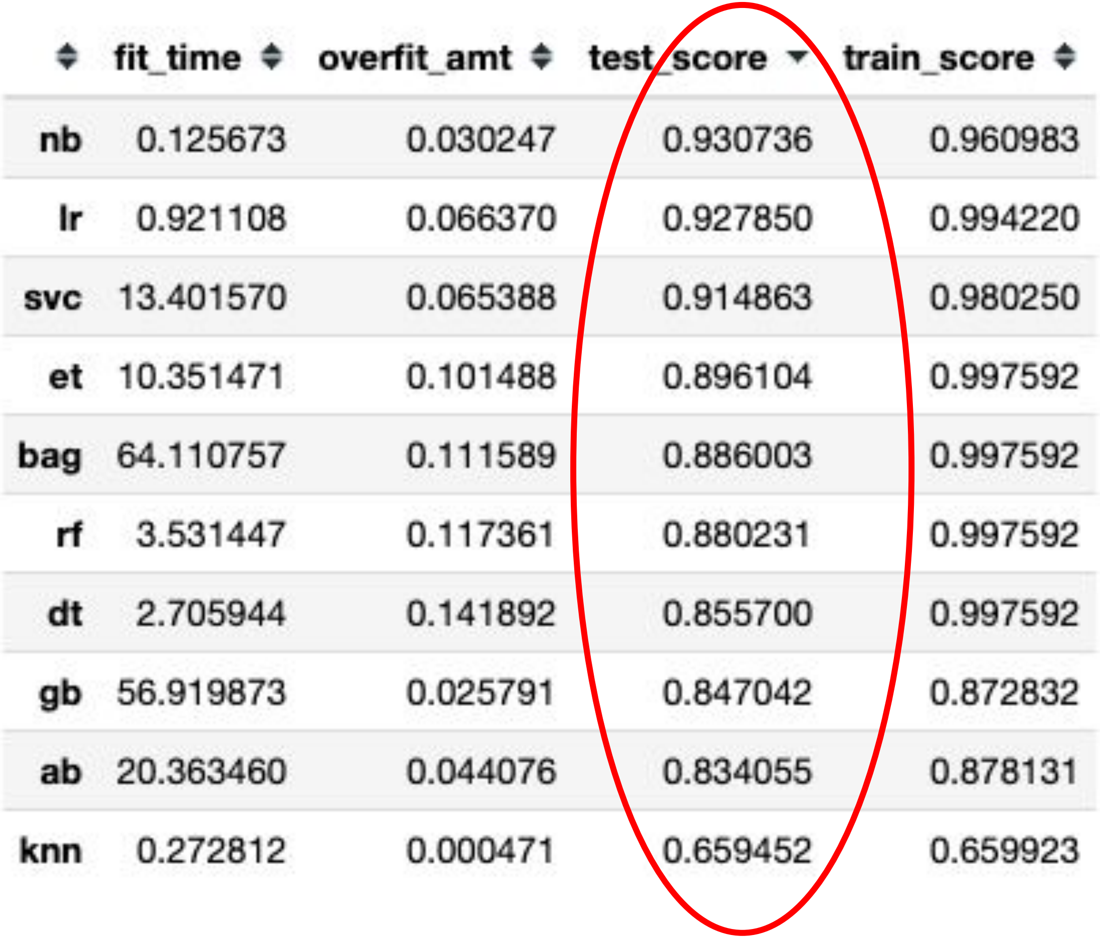
***Natural Language Processing*** has a number of very high-utility application areas: classification, machine translation, sentiment analysis, chat bots/ cust service, marketing message, targeting, etc.

***The question here is***: can a classification model successfully classify subreddit posts into one of two categories, even if the subreddits are related in subject matter? If so, which estimator performs best in terms of accuracy and compute resources?

# THE ANSWER

Yes!

Naive- Bayes achieved 93% of accuracy, followed closely by Logistic Regression



|  | fit_time | overfit_amt | test_score | train_score |
|---|---|---|---|---|
| **nb** | 0.125673 | 0.030247 | 0.930736 | 0.960983 |
| **lr** | 0.921108 | 0.066370 | 0.927850 | 0.994220 |
| **svc** | 13.401570 | 0.065388 | 0.914863 | 0.980250 |
| **et** | 10.351471 | 0.101488 | 0.896104 | 0.997592 |
| **bag** | 64.110757 | 0.111589 | 0.886003 | 0.997592 |
| **rf** | 3.531447 | 0.117361 | 0.880231 | 0.997592 |
| **dt** | 2.705944 | 0.141892 | 0.855700 | 0.997592 |
| **gb** | 56.919873 | 0.025791 | 0.847042 | 0.872832 |
| **ab** | 20.363460 | 0.044076 | 0.834055 | 0.878131 |
| **knn** | 0.272812 | 0.000471 | 0.659452 | 0.659923 |

# WHAT DOES THE DATA LOOK LIKE?

Combined 'Title' and 'Selftext' into 'POSTS'

Originally selected two subreddits:

- Scifi
- Physics

***- but -***

… both categories seemed to have a great many duplicates...

***- so -***

… combined *pairs* of subreddits:

- Scifi + StarWars = Fiction (neg class)
- Physics + Astronomy = Factual (pos class)

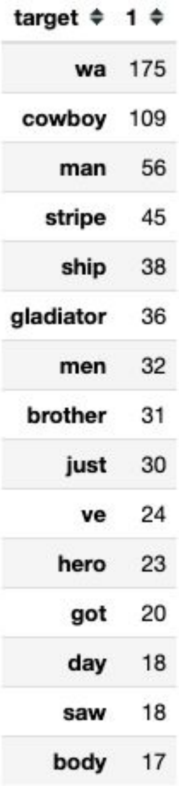
| SUBREDDIT | TOTAL DOWNLOADS | UNIQUE POSTS | CLASS |
|---|---|---|---|
| SciFi | 991 | 789 | 0 |
| StarWars | 987 | 759 | 0 |
| Fiction | 1,978 | 1,548 | 0 |
| | | | |
| Physics | 990 | 712 | 1 |
| Astronomy | 984 | 509 | 1 |
| Factual | 1,974 | 1,221 | 1 |
| | | | |
| **TOTAL** | **3,952** | **2,769** | |

# DATA CLEANING & VECTORIZATION

1. Employed a cleaning + lemmatization function to:
   - remove line-breaks
   - remove non-letter characters
   - tokenize by splitting on spaces
   - lemmatize words
   - re-combine into a post string

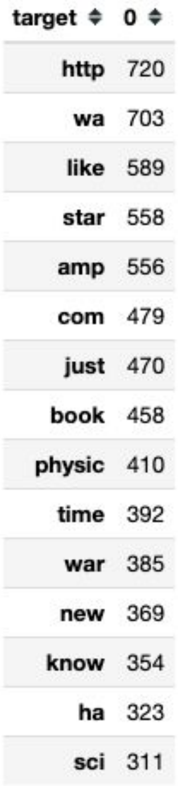
2. Vectorize
   - max_features = 5,000
   - ngram_range = (1, 2)

Top 15 Words (by Freq) in Factual Posts



| target | 1 |
|---|---|
| wa | 175 |
| cowboy | 109 |
| man | 56 |
| stripe | 45 |
| ship | 38 |
| gladiator | 36 |
| men | 32 |
| brother | 31 |
| just | 30 |
| ve | 24 |
| hero | 23 |
| got | 20 |
| day | 18 |
| saw | 18 |
| body | 17 |

Top 15 Words (by Freq) in Fiction Posts

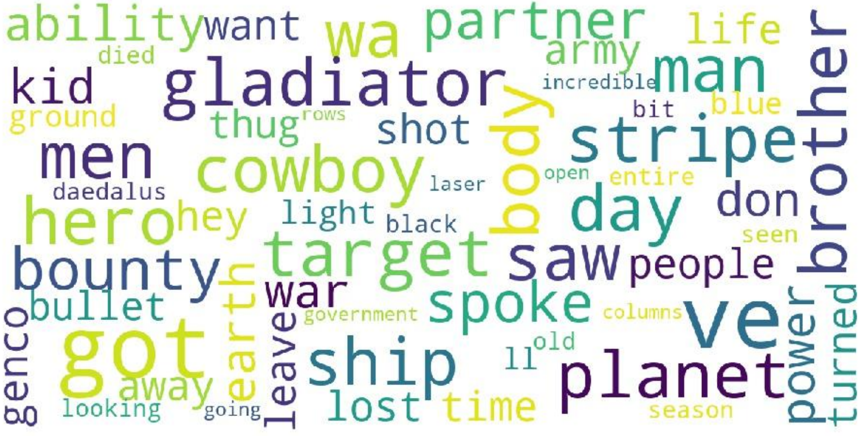


| target | 0 |
|---|---|
| http | 720 |
| wa | 703 |
| like | 589 |
| star | 558 |
| amp | 556 |
| com | 479 |
| just | 470 |
| book | 458 |
| physic | 410 |
| time | 392 |
| war | 385 |
| new | 369 |
| know | 354 |
| ha | 323 |
| sci | 311 |

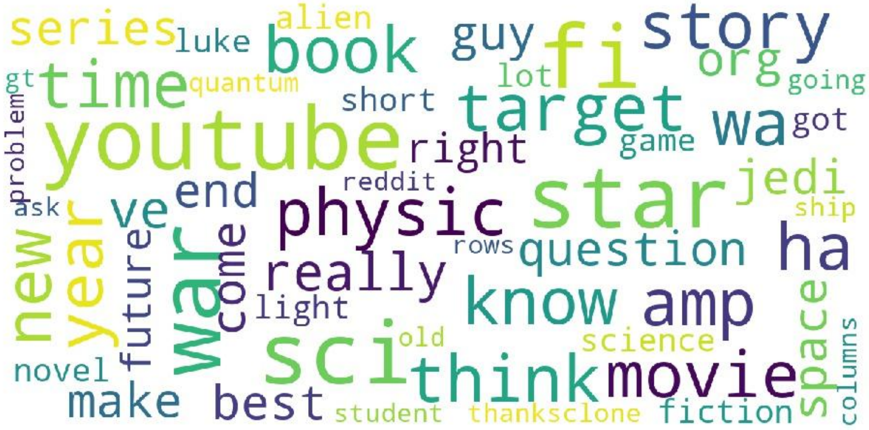
# WORD CLOUDS & COUNTS

Top 50 words in the "factual" group only overlap the top 50 words in the "fiction" group by 20%

TOP PHYSICS & ASTRONOMY SUBREDDIT WORDS



TOP SCIFI & STARWARS SUBREDDIT WORDS



20% OVERLAP

# MODELING STRATEGY

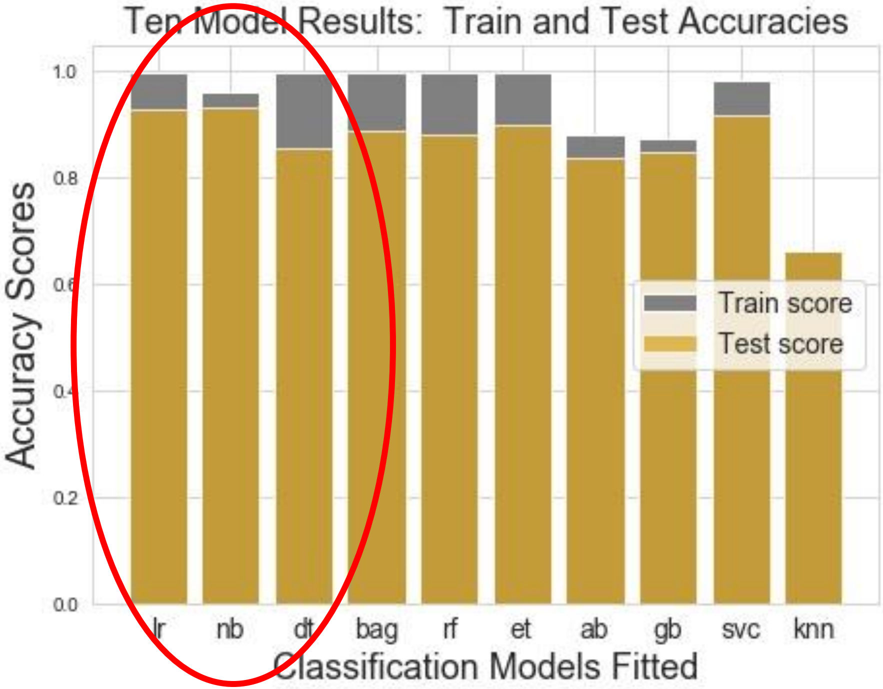
1. Run ten different models, adjusting 1-2 hyperparameters to see how each method performs, tracking fit times, train scores, and test scores (use a function to automate this process...

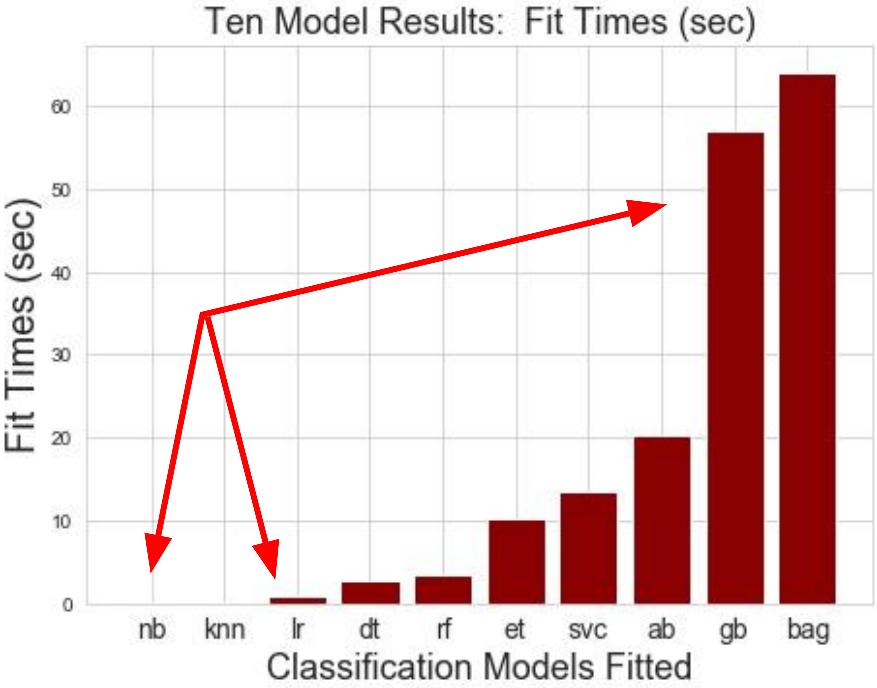
   Logistic Regression
   Naive-Bayes Multinomial
   Decision Tree Classifier
   Bagging Classifier
   Random Forest Classifier
   Extra Random Trees Classifier
   Ada Boost Classifier
   Gradient Boost Classifier
   Support Vector Classifier
   K Nearest Neighbor

2. Use GridSearchCV to further tune hyperparameters on the best-performing 2-3 models above.

# RESULTS

Logistic Regression and Naive-Bayes models delivered the highest test scores, the lowest overfit, and took the least compute time to fit.

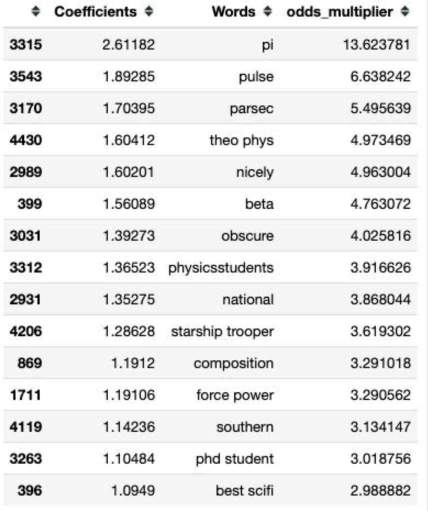




# INTERPRETATION

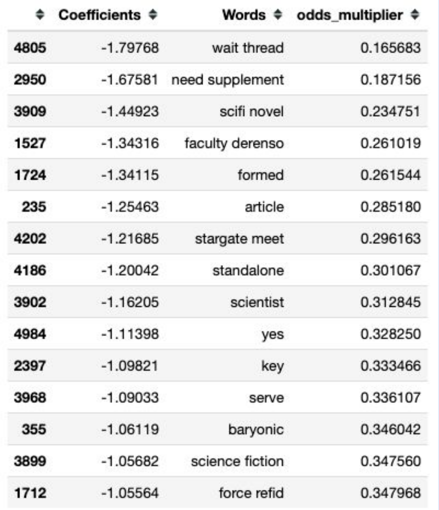
Below are pairs of rank-ordered lists (most-positive and most-negative) of coefficients and words for both the LR and NB models. The models achieve similar results, but use different words to do the classification… only 18% of top 50 words appear in both models' top-50 coefficients lists ['pi', 'pulse', 'theo phys', 'beta', 'starship trooper', 'vague', 'local', 'agreement', 'unlike'].

## Top LR Coefficients & Words



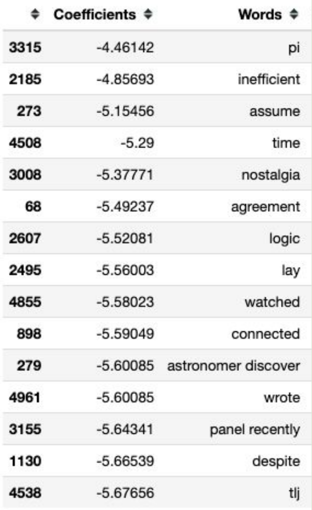
| | Coefficients | Words | odds_multiplier |
|---|---|---|---|
| 3315 | 2.61182 | pi | 13.623781 |
| 3543 | 1.89285 | pulse | 6.638242 |
| 3170 | 1.70395 | parsec | 5.495639 |
| 4430 | 1.60412 | theo phys | 4.973469 |
| 2989 | 1.60201 | nicely | 4.963004 |
| 399 | 1.56089 | beta | 4.763072 |
| 3031 | 1.39273 | obscure | 4.025816 |
| 3312 | 1.36523 | physicsstudents | 3.916626 |
| 2931 | 1.35275 | national | 3.868044 |
| 4206 | 1.28628 | starship trooper | 3.619302 |
| 869 | 1.1912 | composition | 3.291018 |
| 1711 | 1.19106 | force power | 3.290562 |
| 4119 | 1.14236 | southern | 3.134147 |
| 3263 | 1.10484 | phd student | 3.018756 |
| 396 | 1.0949 | best scifi | 2.988882 |

## Bottom LR Coefficients & Words



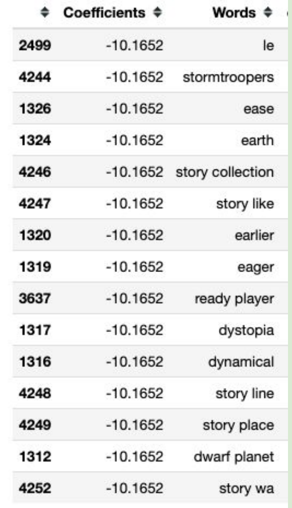
| | Coefficients | Words | odds_multiplier |
|---|---|---|---|
| 4805 | -1.79768 | wait thread | 0.165683 |
| 2950 | -1.67581 | need supplement | 0.187156 |
| 3909 | -1.44923 | scifi novel | 0.234751 |
| 1527 | -1.34316 | faculty derenso | 0.261019 |
| 1724 | -1.34115 | formed | 0.261544 |
| 235 | -1.25463 | article | 0.285180 |
| 4202 | -1.21685 | stargate meet | 0.296163 |
| 4186 | -1.20042 | standalone | 0.301067 |
| 3902 | -1.16205 | scientist | 0.312845 |
| 4984 | -1.11398 | yes | 0.328250 |
| 2397 | -1.09821 | key | 0.333466 |
| 3968 | -1.09033 | serve | 0.336107 |
| 355 | -1.06119 | baryonic | 0.346042 |
| 3899 | -1.05682 | science fiction | 0.347560 |
| 1712 | -1.05564 | force refid | 0.347968 |

## Top NB Coefficients & Words



| | Coefficients | Words |
|---|---|---|
| 3315 | -4.46142 | pi |
| 2185 | -4.85693 | inefficient |
| 273 | -5.15456 | assume |
| 4508 | -5.29 | time |
| 3008 | -5.37771 | nostalgia |
| 68 | -5.49237 | agreement |
| 2607 | -5.52081 | logic |
| 2495 | -5.56003 | lay |
| 4855 | -5.58023 | watched |
| 898 | -5.59049 | connected |
| 279 | -5.60085 | astronomer discover |
| 4961 | -5.60085 | wrote |
| 3155 | -5.64341 | panel recently |
| 1130 | -5.66539 | despite |
| 4538 | -5.67656 | tlj |

## Bottom NB Coefficients & Words



| | Coefficients | Words |
|---|---|---|
| 2499 | -10.1652 | le |
| 4244 | -10.1652 | stormtroopers |
| 1326 | -10.1652 | ease |
| 1324 | -10.1652 | earth |
| 4246 | -10.1652 | story collection |
| 4247 | -10.1652 | story like |
| 1320 | -10.1652 | earlier |
| 1319 | -10.1652 | eager |
| 3637 | -10.1652 | ready player |
| 1317 | -10.1652 | dystopia |
| 1316 | -10.1652 | dynamical |
| 4248 | -10.1652 | story line |
| 4249 | -10.1652 | story place |
| 1312 | -10.1652 | dwarf planet |
| 4252 | -10.1652 | story wa |

# CONCLUSION

Even though we might imagine that SciFi/StarWars posts in Reddit might have a great deal of similarity to posts in Physics/Astronomy posts, NLP estimators do a good job in classifying these posts… 93% accuracy.

In the cases of Naive-Bayes (multinomial) and Logistic Regression estimators, we also have very computationally-efficient models, compared to ensemble techniques like Random Forests, or ExtraRandom Trees.

Logistic Regression also generates coefficients that are interpretable as odds (when coeff is exponentiated). All in all, a very powerful tool for processing text.