REDDIT NLP

Web Scraping & NLP
Text Classification

Kade Higgins

PROBLEM STATEMENT

Did this subreddit text come from r/C_S_T or r/conspiracy?

LET'S PLAY A GAME!

"Am I missing something else going on?"

"Did time speed up after 9/11?"

"Is Our Whole Government Corrupt Or Is It Certain Networks?"

"Washington DC was staged just like Gorge Floyd"

METHODOLOGY

Scrape r/C_S_T & r/conspiracy reddit threads

Clean (remove HTML, Lemmatize)

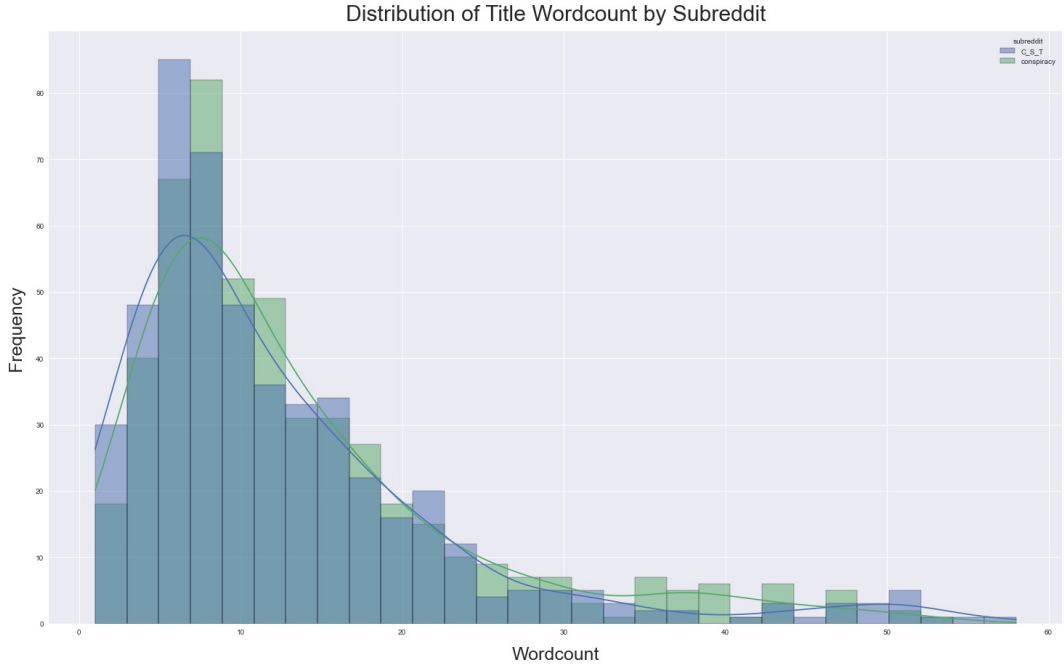
EDA & Visualization

Model - Naïve Bayes

Model – Logistic Regression

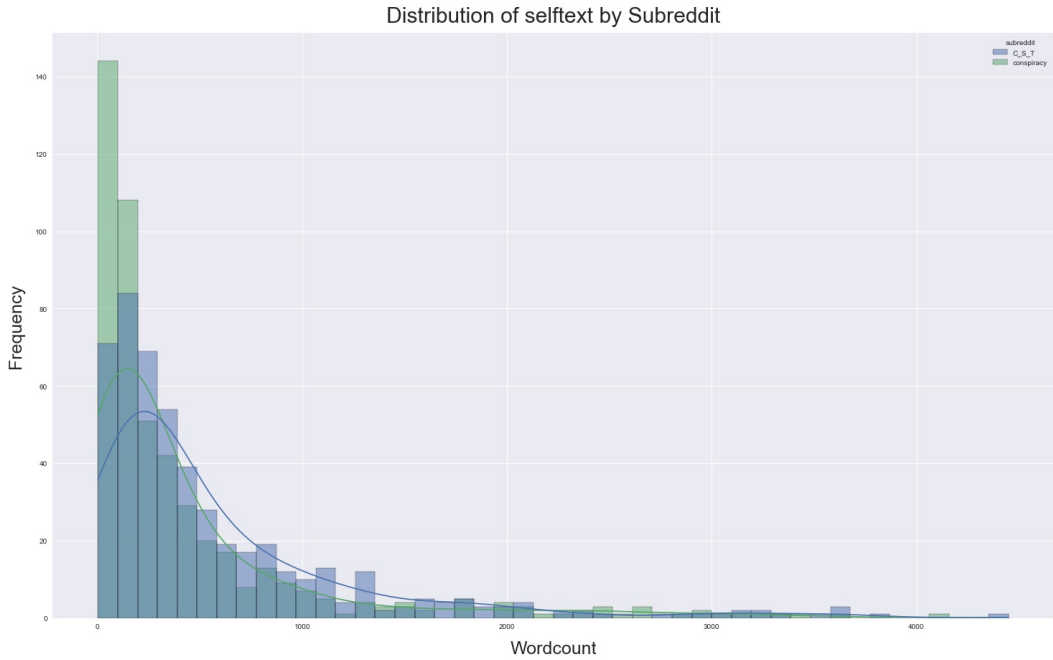
EXPLORATORY DATA ANALYSIS

TITLE WORD COUNT

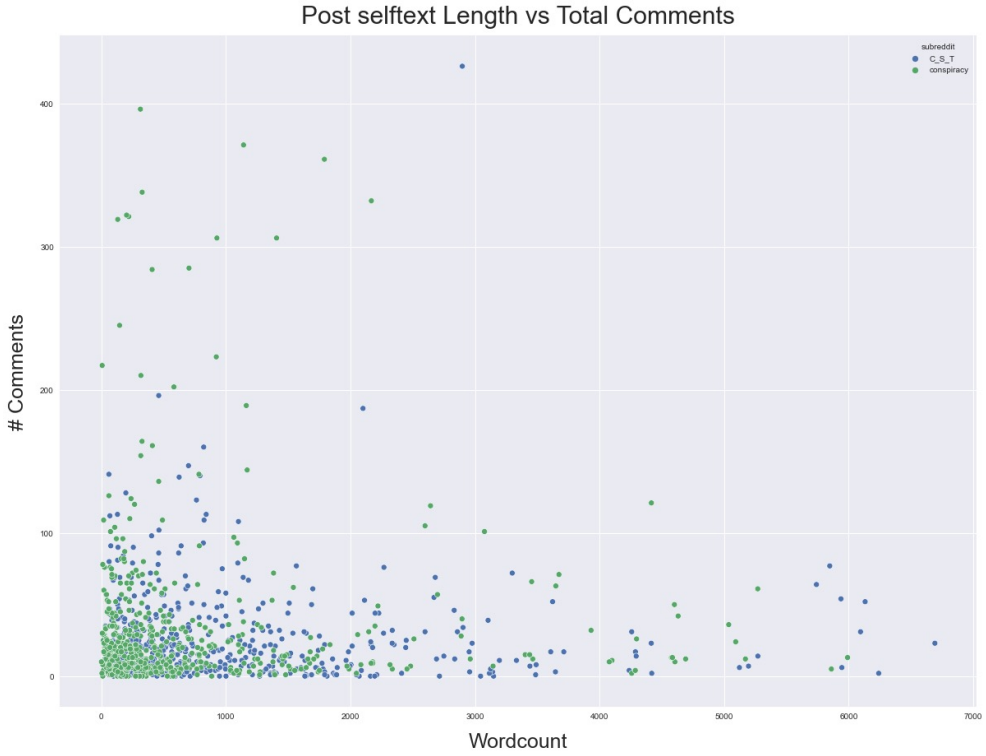


EXPLORATORY DATA ANALYSIS

SELFTEXT WORD COUNT

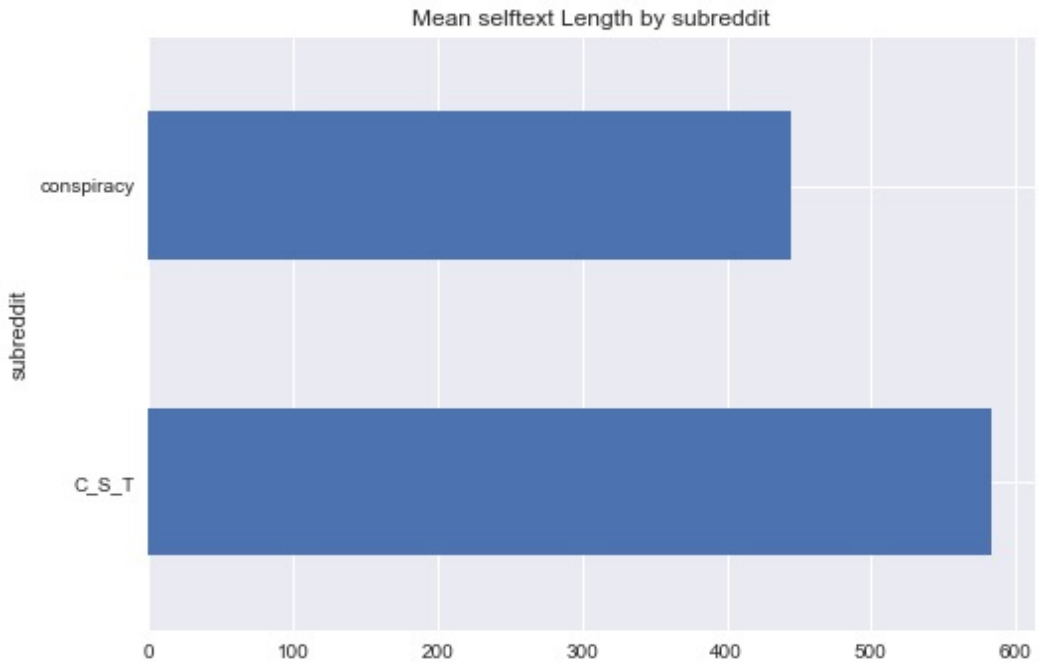


EDA
POST LENGTH VS. COMMENTS

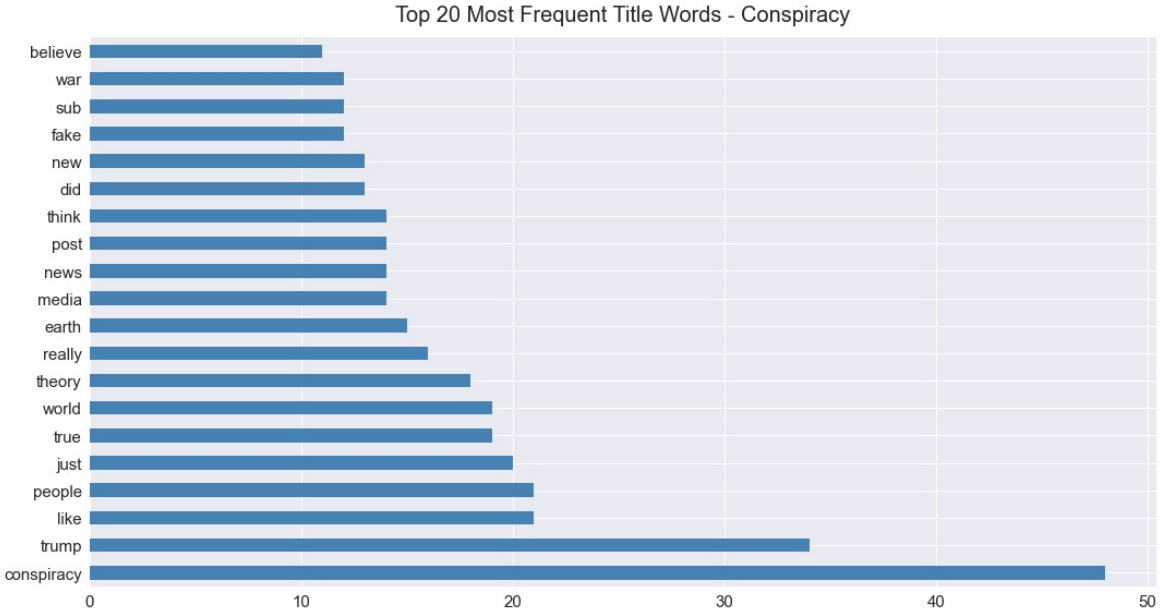


EDA

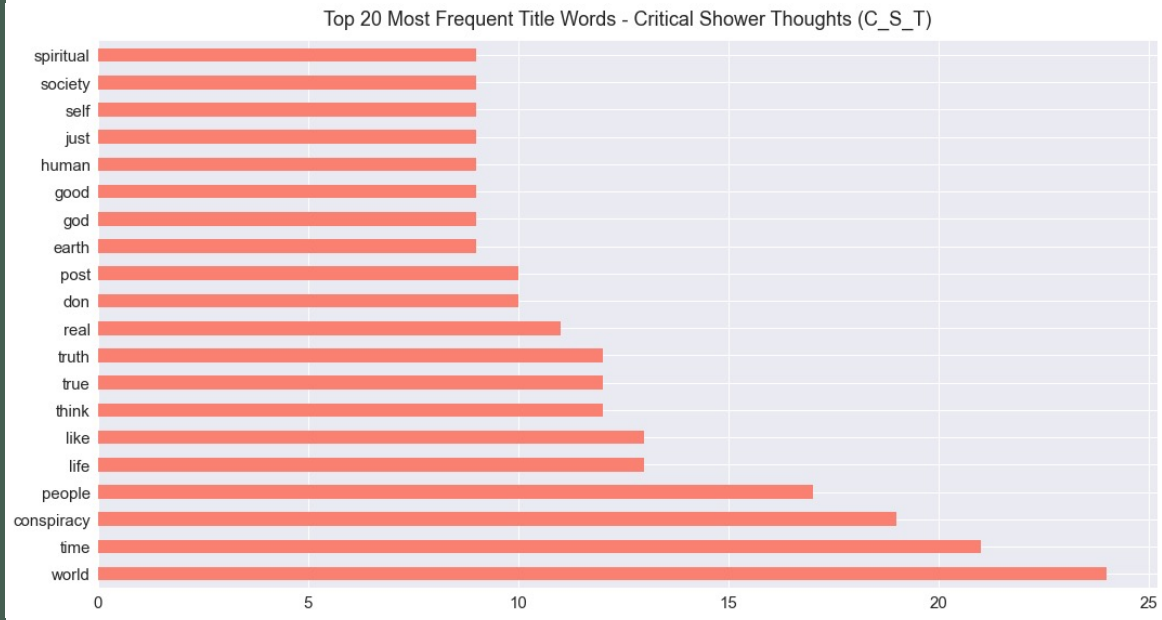
AVG POST LENGTH



CONSPIRACY WORDS

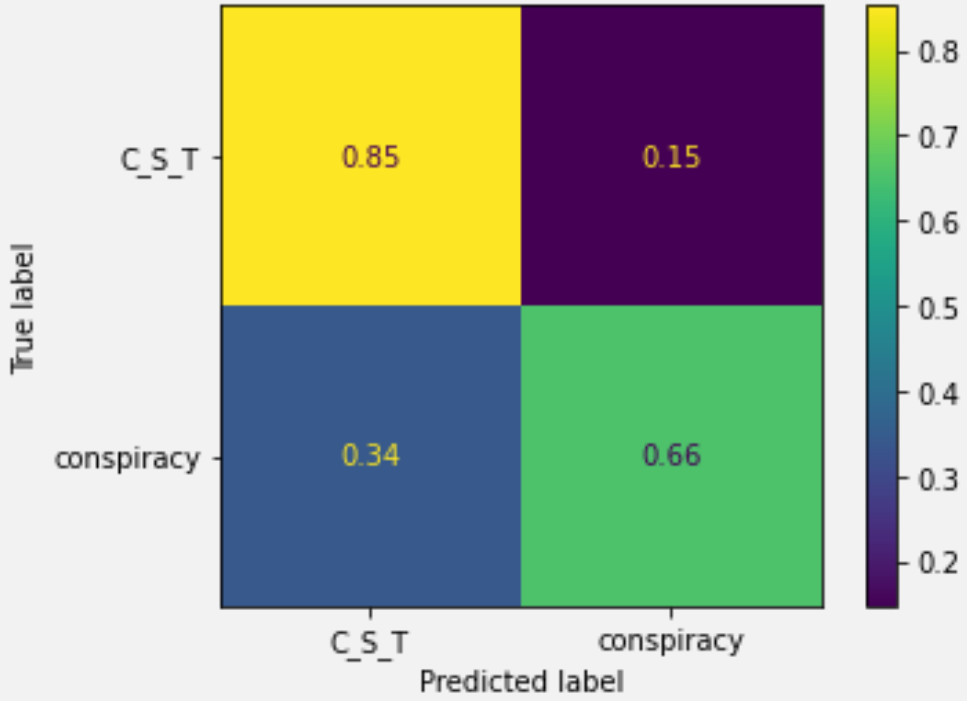


C_S_T WORDS


Top 20 Most Frequent Title Words - Critical Shower Thoughts (C_S_T)
spiritual
society
self
just
human
good
god
earth
post
don
real
truth
true
think
like
life
people
conspiracy
time
world
0
5
10
15
20
25

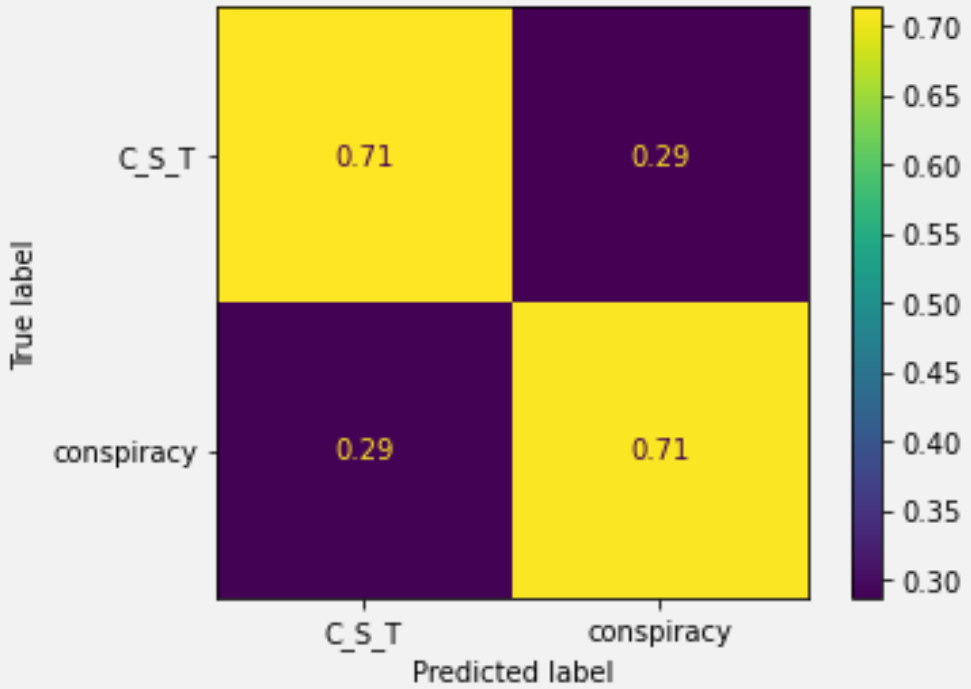
MODEL 1: COMPLEMENT NB

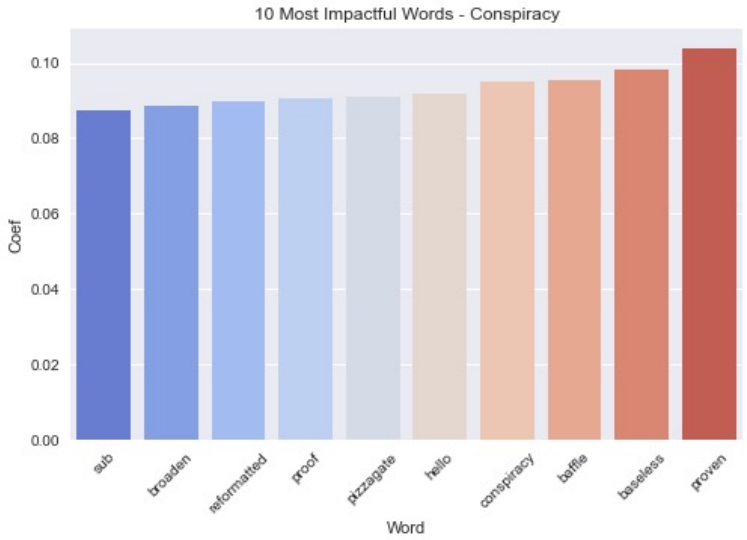
- CountVectorizer & Complement Naïve Bayes on selftext
- ***0.94*** *Train /* ***0.76*** *Test*
- *Balanced Accuracy Score:* ***0.76***
- **our baseline = 0.50

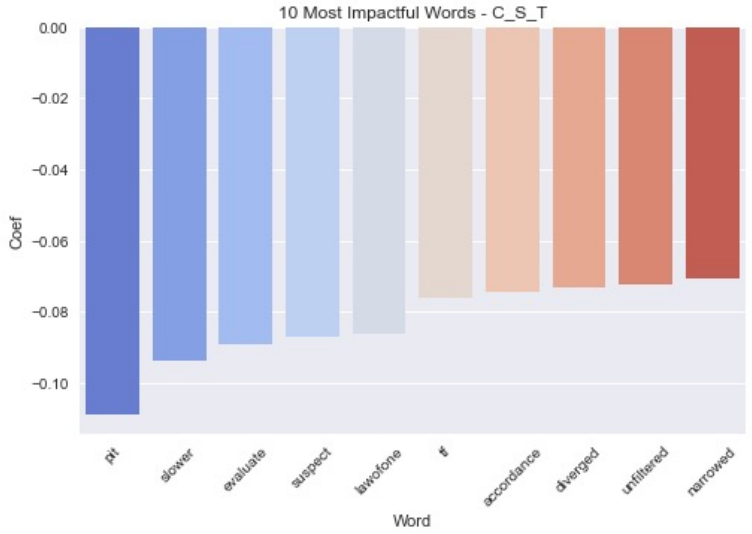


LOGISTIC REGRESSION

- Using TfidfVectorizer & Logistic Regression, the results were:
- ***Best Params*** *(max_iter=5_000, penalty='elasticnet', C=1.0, solver='saga', l1_ratio=0.65)*
- ***0.77*** *Train /* ***0.72*** *Test*
- *Balanced Accuracy Score:* ***0.70***




10 Most Impactful Words - Conspiracy
Coef
0.10
0.08
0.06
0.04
0.02
0.00
sub
broaden
reformatted
proof
pizzagate
hello
conspiracy
baffle
baseless
proven
Word


10 Most Impactful Words - C_S_T
Coef
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
pit
slower
evaluate
suspect
iswofone
ff
accordance
diverged
unfiltered
narrowed
Word

CONCLUSIONS

Naïve Bayes Complement Wins!

With a slightly higher train/test, F1 & Baseline Accuracy scores

- However, Logistic Regression might win out on an easier dataset. LR does typically have much more interpretable coefs

Next Time…

- Combine variables like num_comments & score to see if they can increase the overall accuracy of the model.

THANK YOU FOR LISTENING

I'm happy to answer any questions

Kade Higgins