Harry Skywalker

and the Vectorizer of Random Forests

Project 3

Starring Juston Dea

October 20th, 2023


Turn up Volume
and enjoy

Problem Statement

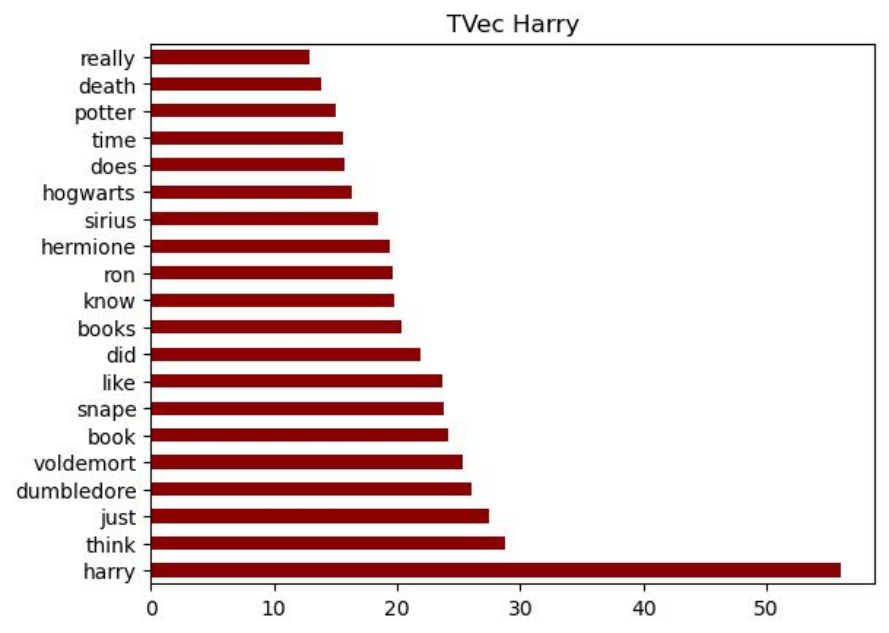
A Star Wars movie and Harry Potter movie are both being released this Christmas and my marketing team wants to use Reddit to identify the fans of each movie. We want to send targeted advertisements of the *other* movie to Reddit users in hopes to promote both movies to non-fanbase Reddit users. The issue is the hype is so real that the team is collecting posts at mass volume and cannot distinguish which subreddit it originated from.

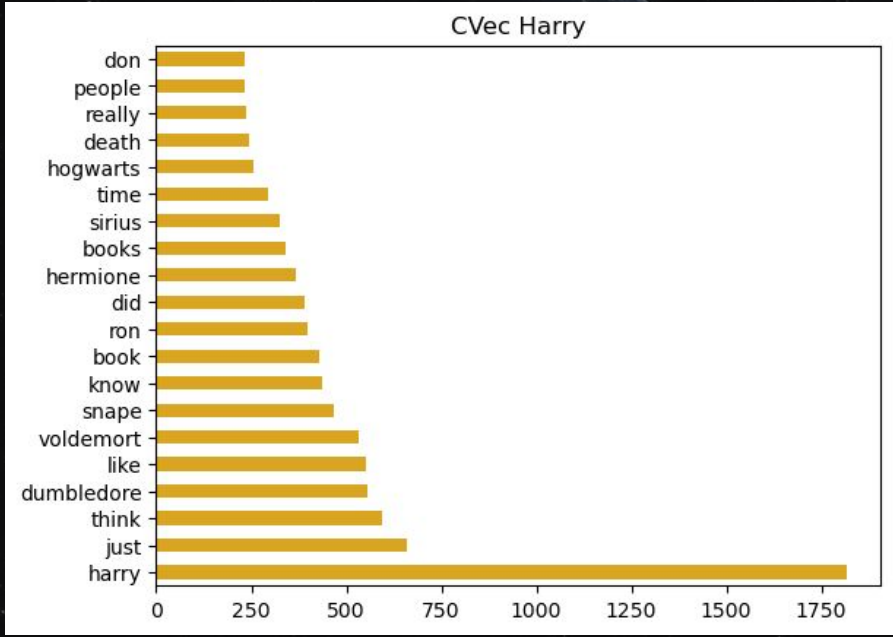
In order to predict which subreddit the post came from I have created three binary classification models which include combinations of Logistic Regression and Random Forest.

Data Collection & EDA

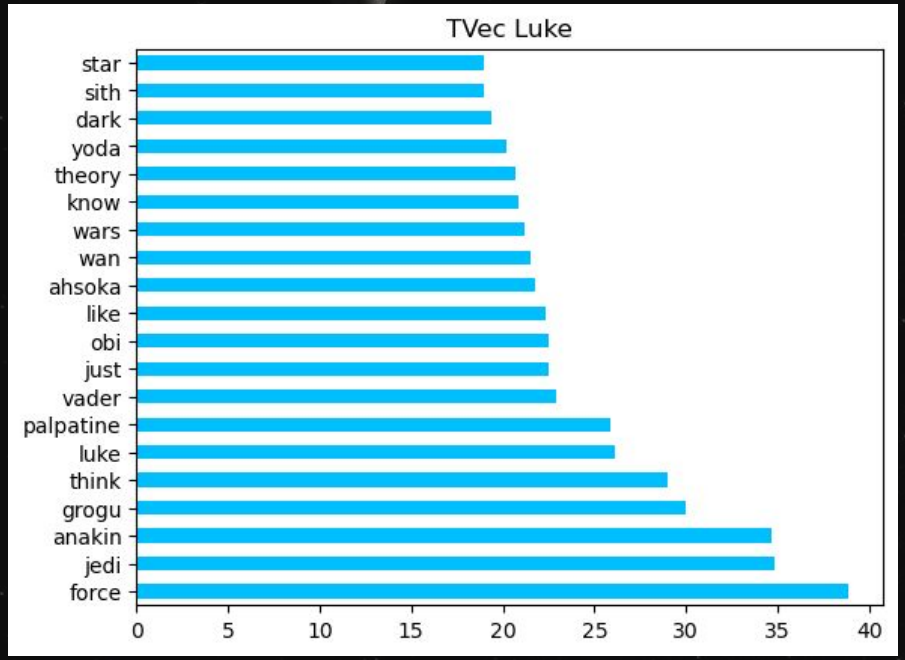
- Subreddits used - r/HarryPotterBooks, r/StarWarsTheories
- **PRAW** to pull posts from each subreddit, create into a DataFrame
- Deleted any duplicate posts, input a blank at any missings
- Combined the **“title”** and **“self_text”** columns to make one text column
- Binarized the subreddit column **Harry : 0, Star Wars : 1**
- Models used:
  1) Term Frequency-Inverse Document Frequency Vectorizer (**Tvec**) with Logistic Regression
  2) Count Vectorizer (**CVec**) with Random Forest
  3) TVec with Random Forest
- Noticeable differences from TVec to Cvec - CVec simply **counts words** and puts them into a matrix. TVec does the same but it takes **frequency** into consideration to determine **how important** a word is.

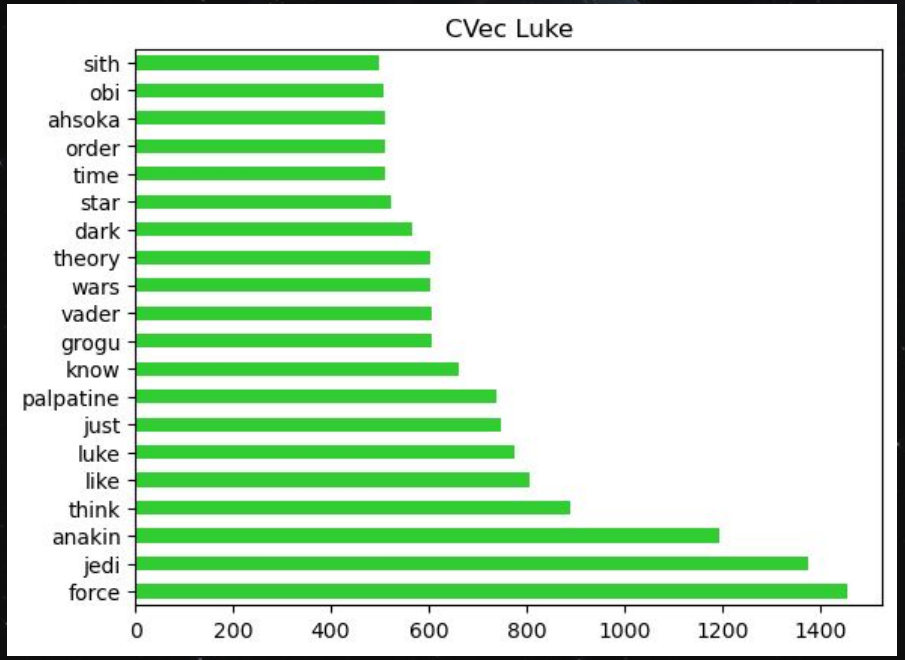
Harry TVEC vs CVEC





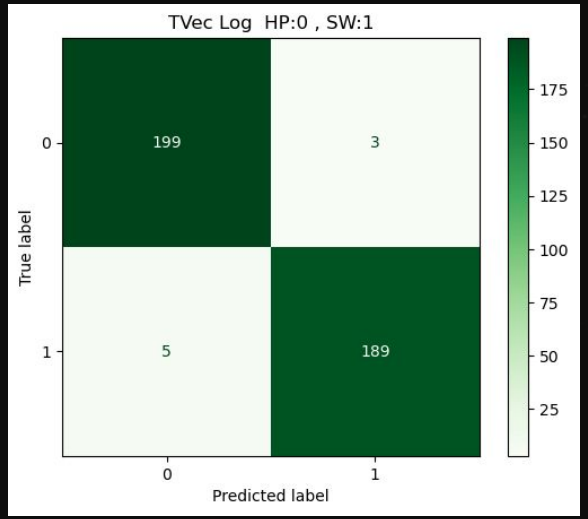
Luke TVEC vs CVEC



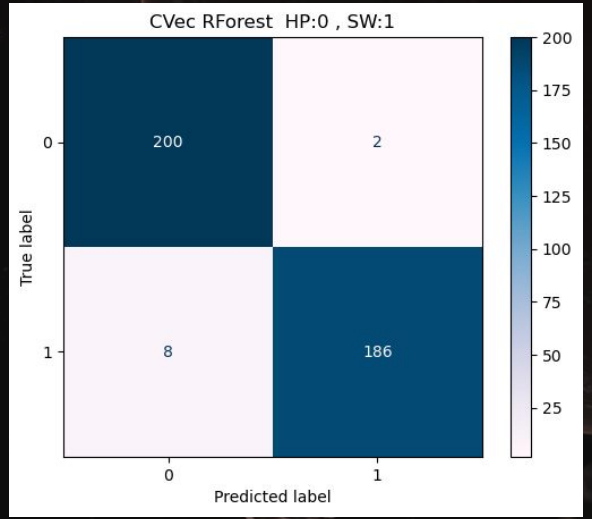


Models

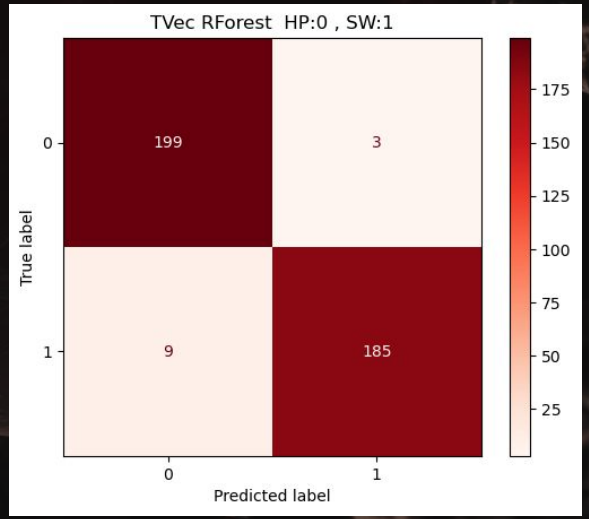
TVec/Log Regression
- **97.9%** test score



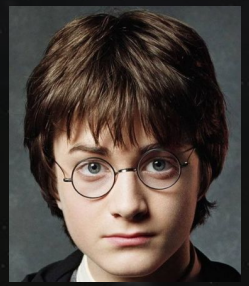
CVec/Random Forest
- **97.4%** test score



TVec/Random Forest
- **96.9%** test score



Conclusion



In conclusion by using binary classification models, we found a way to predict the subreddit from given words in a post. Of the three models I tried I think **TVec with Logistic Regression** had the best prediction outcomes with the least mismatched errors. This is a good start, but more testing and modeling will need be to done to find a better predictor so that the marketing team can optimize their advertisement strategy.

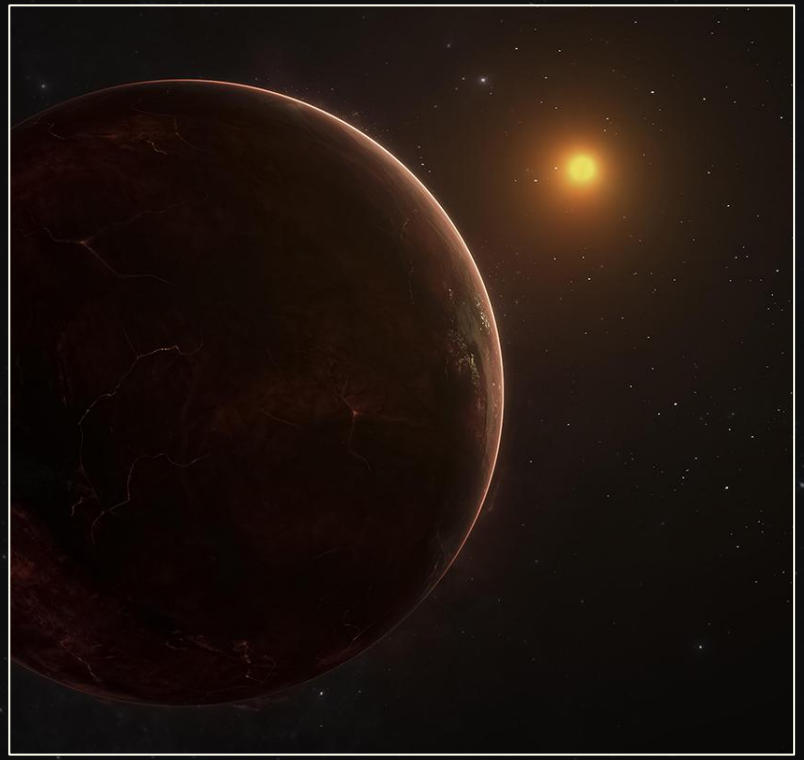
VADER

Of all the text the mean polarity score was **-0.011.**

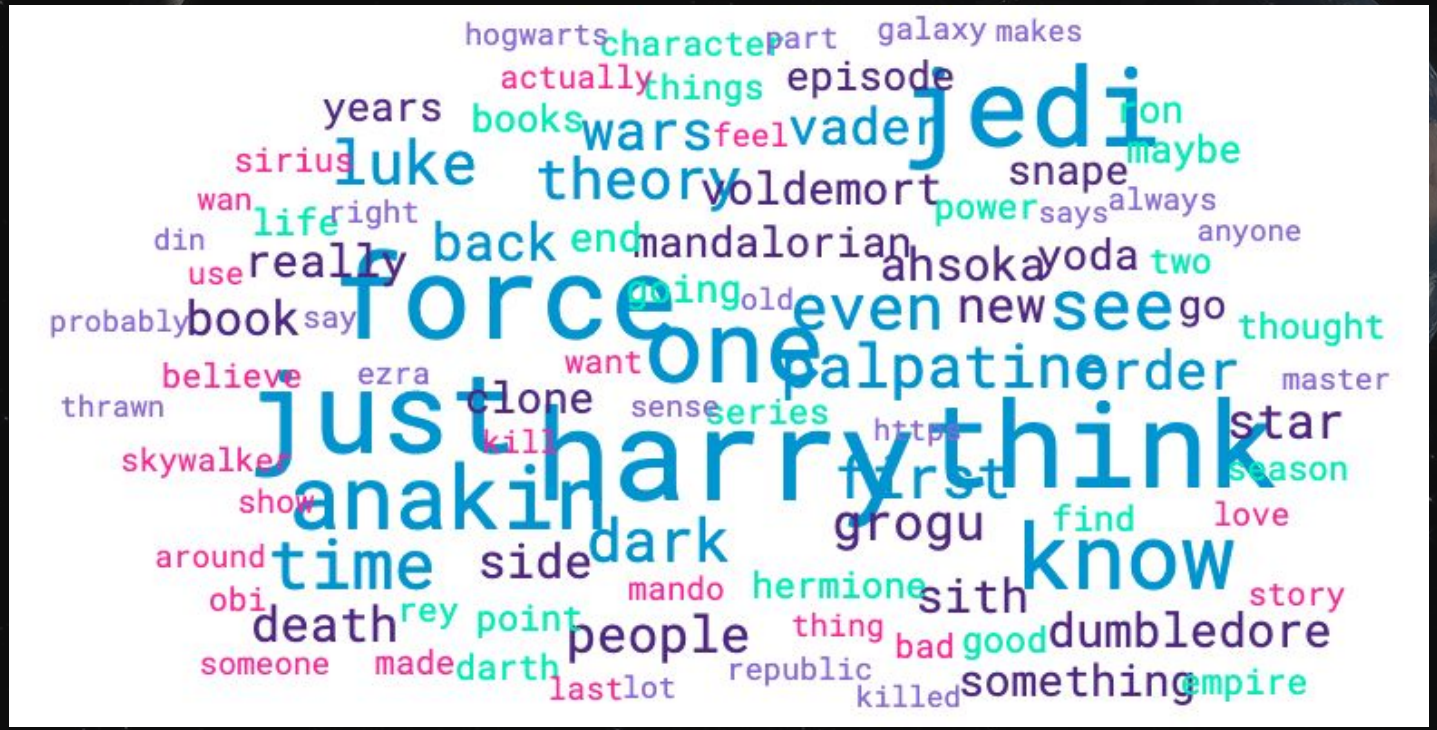
Interpretation:
These subreddits are pretty neutral but also conflicted to the dark-side almost like Anakin..who becomes... omg... VADER!!

Resources

- **Project 3 Breakout Room!**
- Office hours- Rowan and Emily
- PRAW
- Lesson 504: NLP I
- Lesson 505: NLP II
- Lesson 603: Random Forest
- Chat GPT
- Video production https://starwarsintrocreator.kassellabs.io/#!/ENh8-2hlmmDeKqr81k1G/edit



For Tim


hogwarts
character
part
galaxy
makes
actually
things
episode
jedi
years
books
wars
feel
vader
ron
sirius
luke
theory
voldemort
snape
maybe
wan
life
right
power
says
always
din
back
end
mandalorian
anyone
use
really
force
ahsoka
yoda
two
going
old
even
new
see
go
probably
book
say
one
palpatine
thought
believe
ezra
want
order
master
thrawn
clone
sense
series
star
just
harry
think
skywalker
kill
https
first
season
show
anakin
grogu
find
love
around
time
side
dark
know
obi
mando
hermione
sith
story
death
rey
point
people
thing
bad
good
dumbledore
someone
made
darth
republic
something
lastlot
killed
empire

The End

Questions?