WHICH WITCHER?

web scraping/nlp classification

Katherine Hickok

General Assembly DSI Project 03

January 18, 2022

THE WITCHER WILD HUNT ?

OR

NETFLIX THE WITCHER ?

https://www.gamesradar.com/the-witcher-4-release-date/

https://wccftech.com/the-witcher-netflix-geralt-ciri-yennefer/

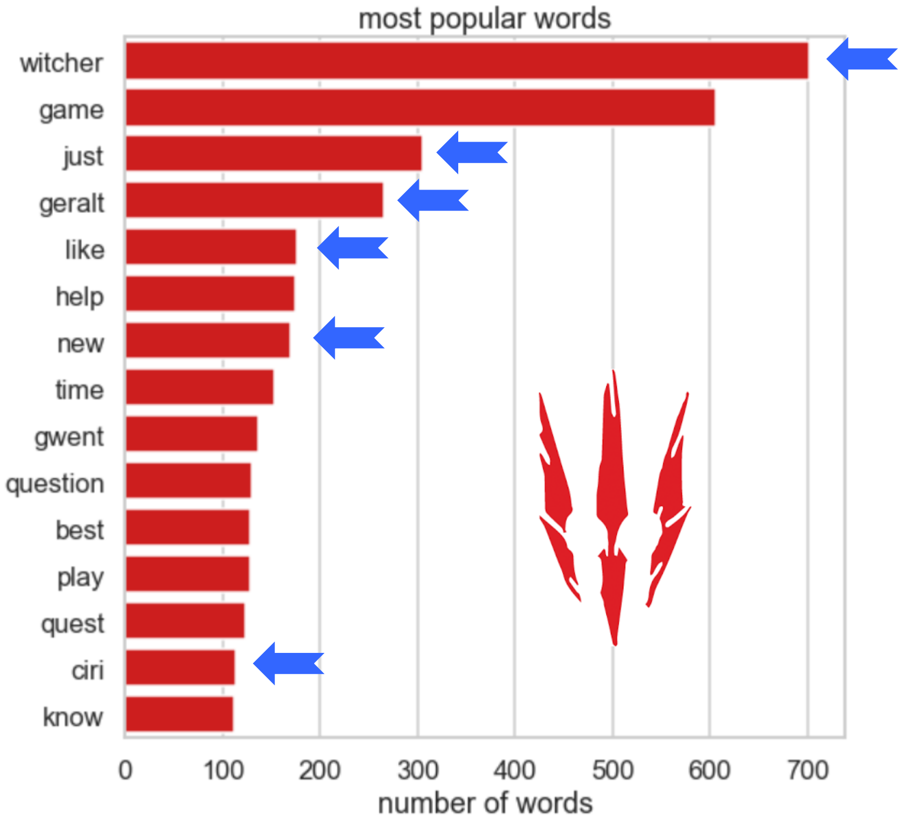
Can we predict whether a post would be in the *Witcher3* or *NetflixWitcher* subreddit better than the baseline accuracy?

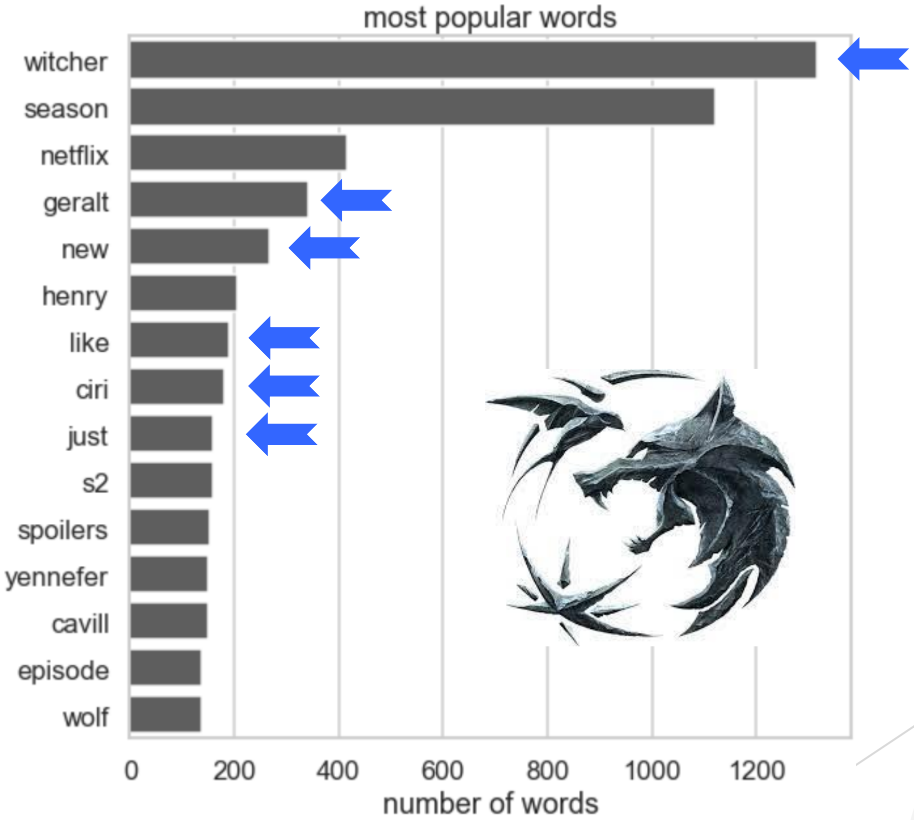
SUBREDDITS

- 5000 posts from each subreddit
- dropped nulls, [deleted], [removed]

MOST COMMON WORDS

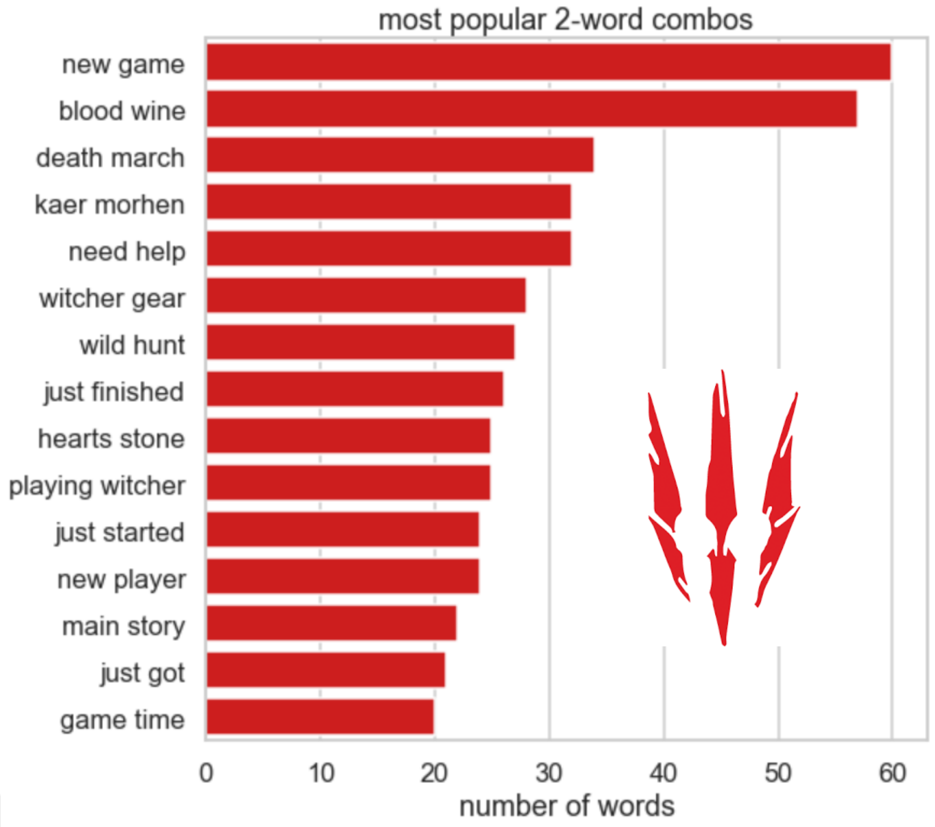
1-word

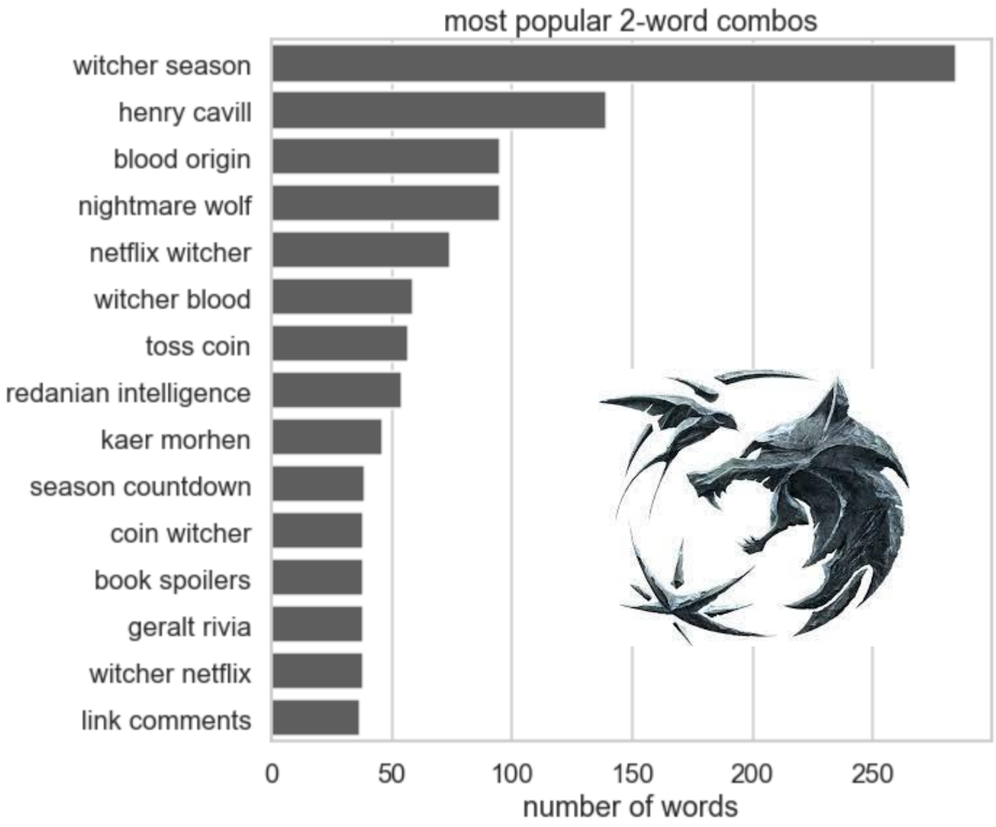




MOST COMMON WORDS

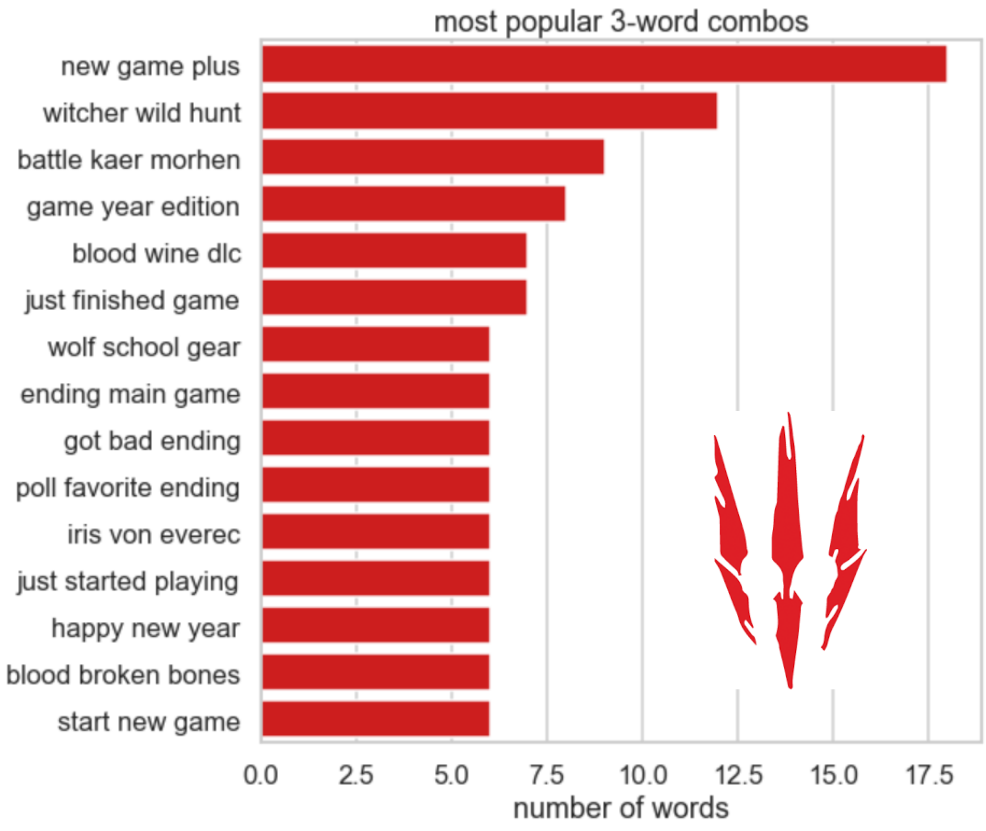
2-word combos

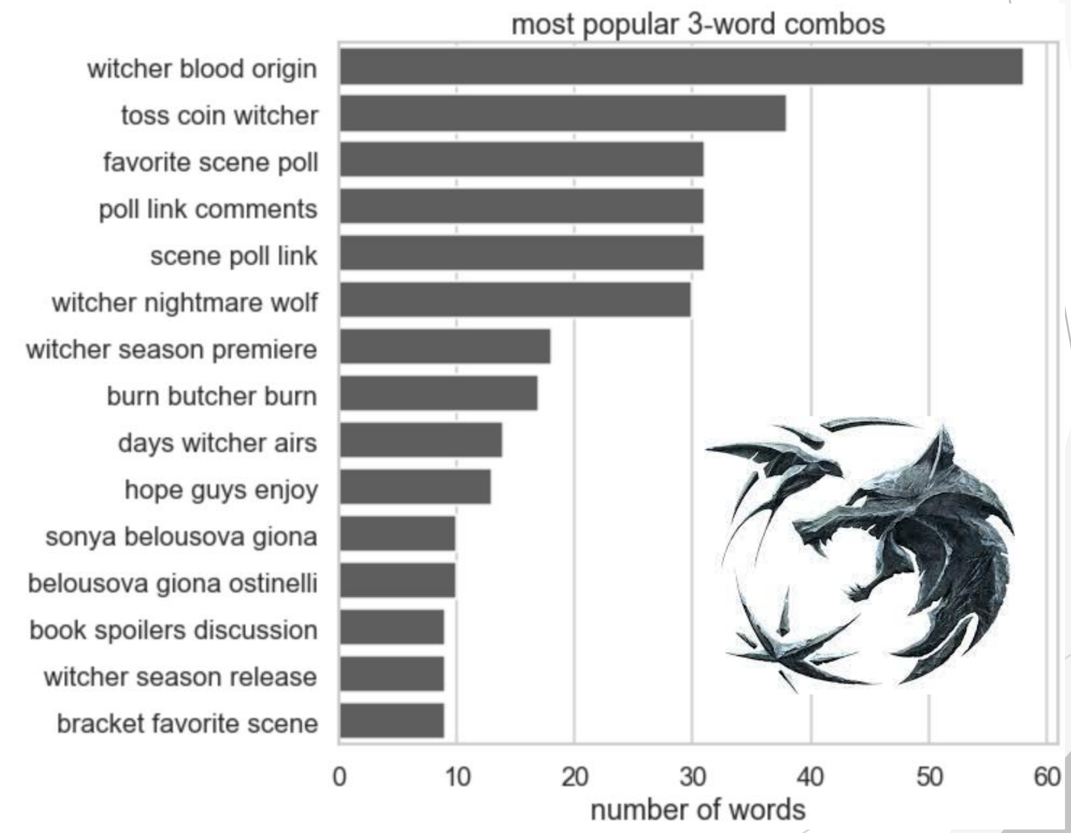




MOST COMMON WORDS

three-word combos





SENTIMENTAL INTENSITY ANALYSIS

averages

subreddit	score	number of comments	neg_%	neu_%	pos_%
	3.02	5.10	7.34	79.82	12.70
	10.51	10.59	4.28	85.96	9.69

- Very Positive: “Superior Feline gear bug. Help please”
- Neutral: “Here we go again (YEEEEEEES)”
- Very Negative: “I never thought that the toughest enemy on death march are gonna be ******* rats”

SENTIMENTAL INTENSITY ANALYSIS

ratioed and controversial

- Being ‘ratioed’ defines controversial posts on Twitter.
- Reddit has controversial posts too.
- 52 posts: 7 in Witcher3, 45 in NetflixWitcher

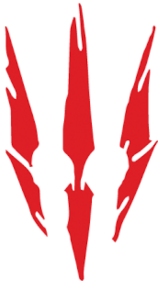
upvote ratio: 0.36, comments: 12, score: 0, sia rating: negative

> “Does anyone else hate the Netflix casting choice for Yennefer? I feel so alone in this opinion. I don't hate the incorporation of POC into white character roles, but this actor just doesn't suit IMO”

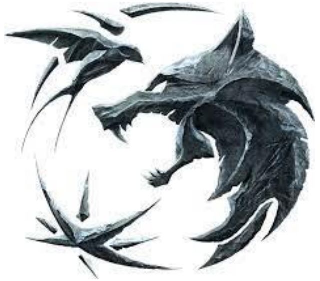
MODELS

baseline

- after model set up, the baseline accuracy is…



0.508



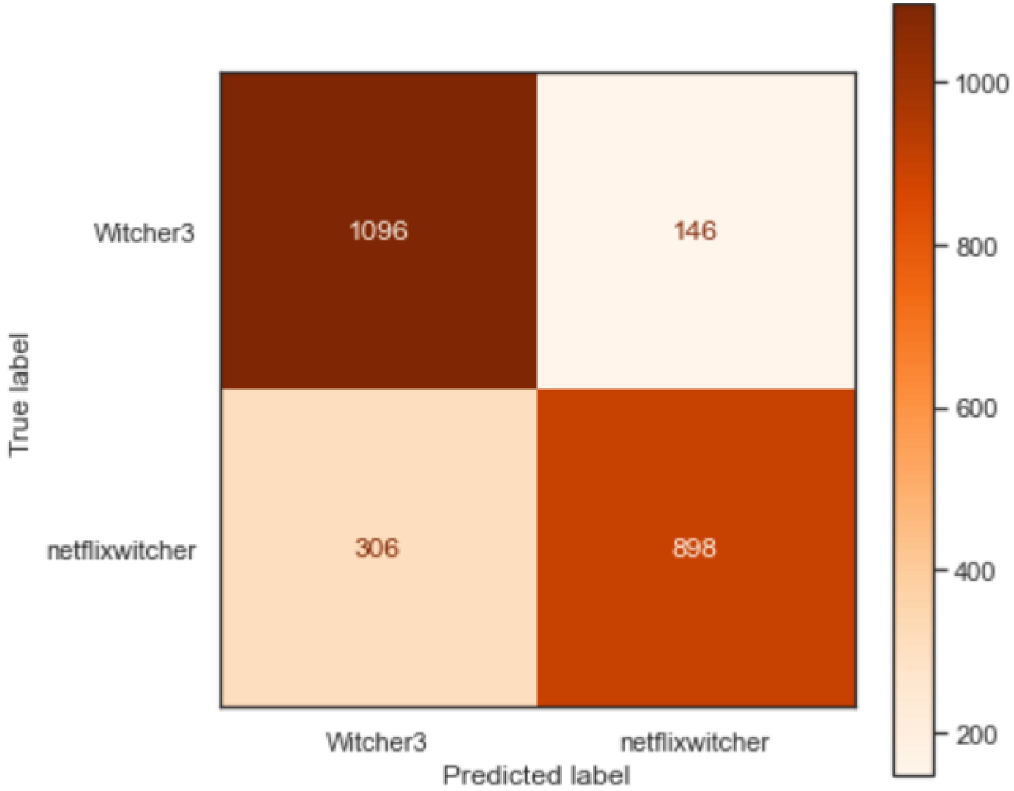
0.492

MODELS

cv/log

- Pipeline/Gridsearch
  - CountVectorizer
  - LogisticRegression

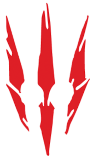
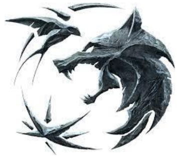
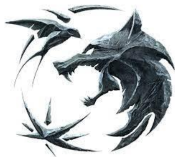
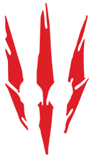
- train R^2 score= 0.855
- test R^2 score= 0.815



specificity:0.882
sensitivity: 0.746

MISCLASSIFICATIONS

cv/log

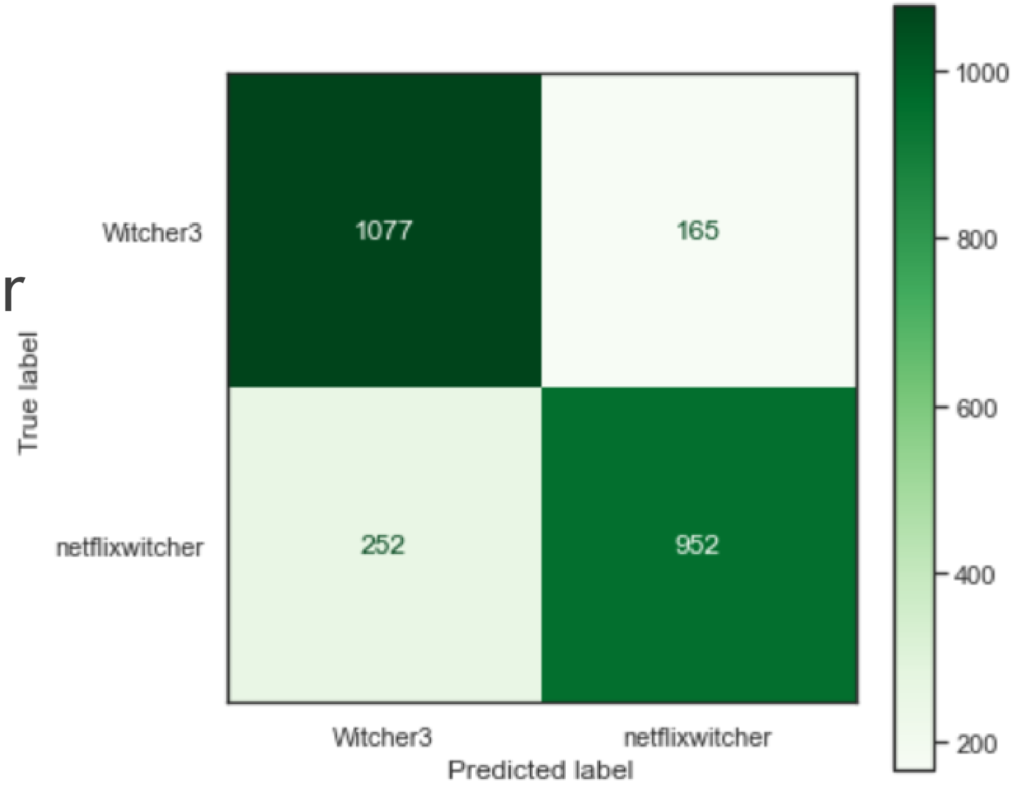
Quote	Predicted Class	True Class
“Henry Cavill Would Like The Witcher Show To Go to Toussaint”		
“Started my 4th playthrough after watching season 2 and how I've missed my Ciri, the game Ciri is my canon Ciri. I do not adore Freya Allen.”		

MODELS

random forest

- CountVectorizer
- Gridsearch
  - RandomForestClassifier

- train R^2 score= 0.960
- test R^2 score= 0.830



specificity:0.867
sensitivity: 0.791

MISCLASSIFICATIONS

random forest

Quote	Predicted Class	True Class
“I feel like it's weird they make Ciri look like an adult but still talk to her like a child. It's jarring and breaks immersion every time they address her like the 12 year old she is”		
“Best and Worst Actor in the Show”		

CONCLUSIONS and FURTHER RESEARCH

- NLP and associated classification modeling did a good job at beating the baseline model
  - Overall, better at predicting Witcher3 than NetflixWitcher
  - RandomForestClassifier performed the best

- Sentiment Intensity Analysis was faulty but was a good analog to look at controversial/ratioed posts

- Further research to investigate misclassifications…
  - some were surprising and others deceptive

WORKS CITED

- https://www.reddit.com/r/Witcher3/
- https://www.reddit.com/r/netflixwitcher/
- https://www.dictionary.com/browse/ratio#:~:text=On%20the%20social%20media%20platform,and%20considering%20its%20content%20bad
- https://www.reddit.com/r/explainlikeimfive/comments/1rqjwp/eli5_what_does_top_new_hot_controversial_old_mean/