HISTORY OR CONSPIRACY?

A SUBREDDIT CLASSIFICATION PROBLEM

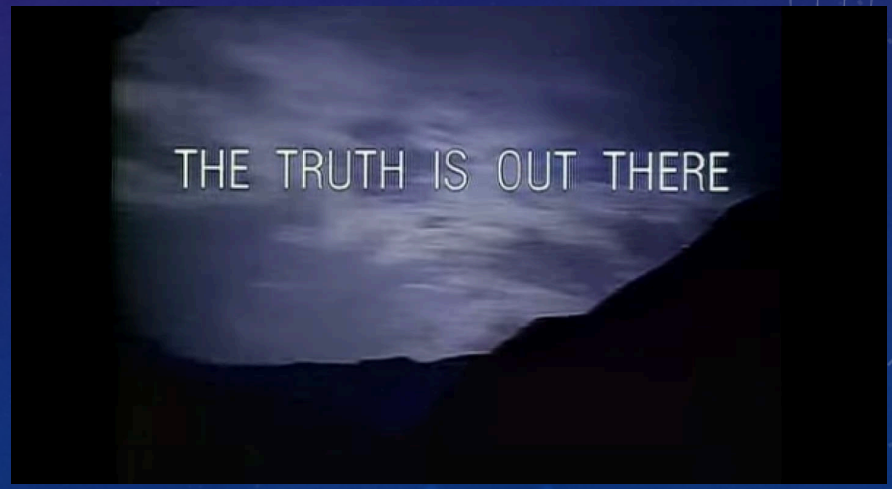
DYLAN BLOUGH

GA DSI-DC

PRESENTED 01/31/20

THE PROBLEM

- Awareness of widespread inaccurate or fake information on the internet
- Rapid rise of senior citizens online: 12% to 67% since 2000
- Conspiracy theories as a central part of mainstream American culture.
  - Research indicates that 50% of Americans believe some form of conspiracy theory
  - Propagated through movies, tv, and even government officials
- How do we vet information we receive online, especially on social media sites like Reddit and Twitte before we decide to believe it



Source: Pew Research, *https://www.pewresearch.org/internet/2017/05/17/tech-adoption-climbs-among-older-adults/*
Source: Washington Post, https://www.washingtonpost.com/news/monkey-cage/wp/2015/02/19/fifty-percent-of-americans-believe-in-some-conspiracy-theory-heres-why/

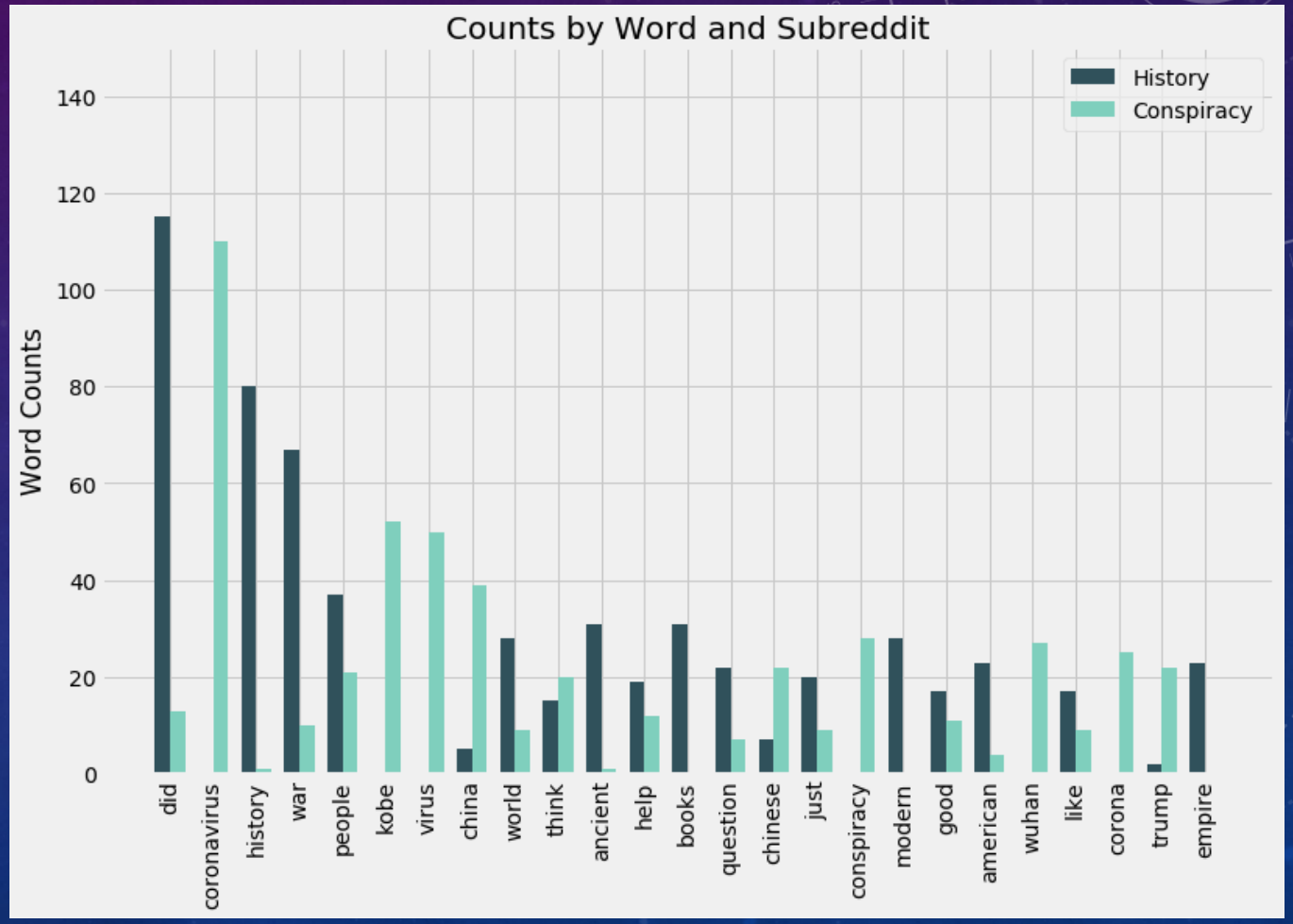
ABOUT THE SUBREDDITS

- /r/History: 14.5mil members, strict rules governing what content it allowed
  - No current politics (within 20 years)
  - No historical negationism or denialism
- /r/Conspiracy: 1.1 mil members, lax rules regarding submissions (it IS a conspiracy board after all!)

Source: reddit.com/r/history/index#wiki_rules
Source: redditcome/r/conspiracy/index#wiki_rules

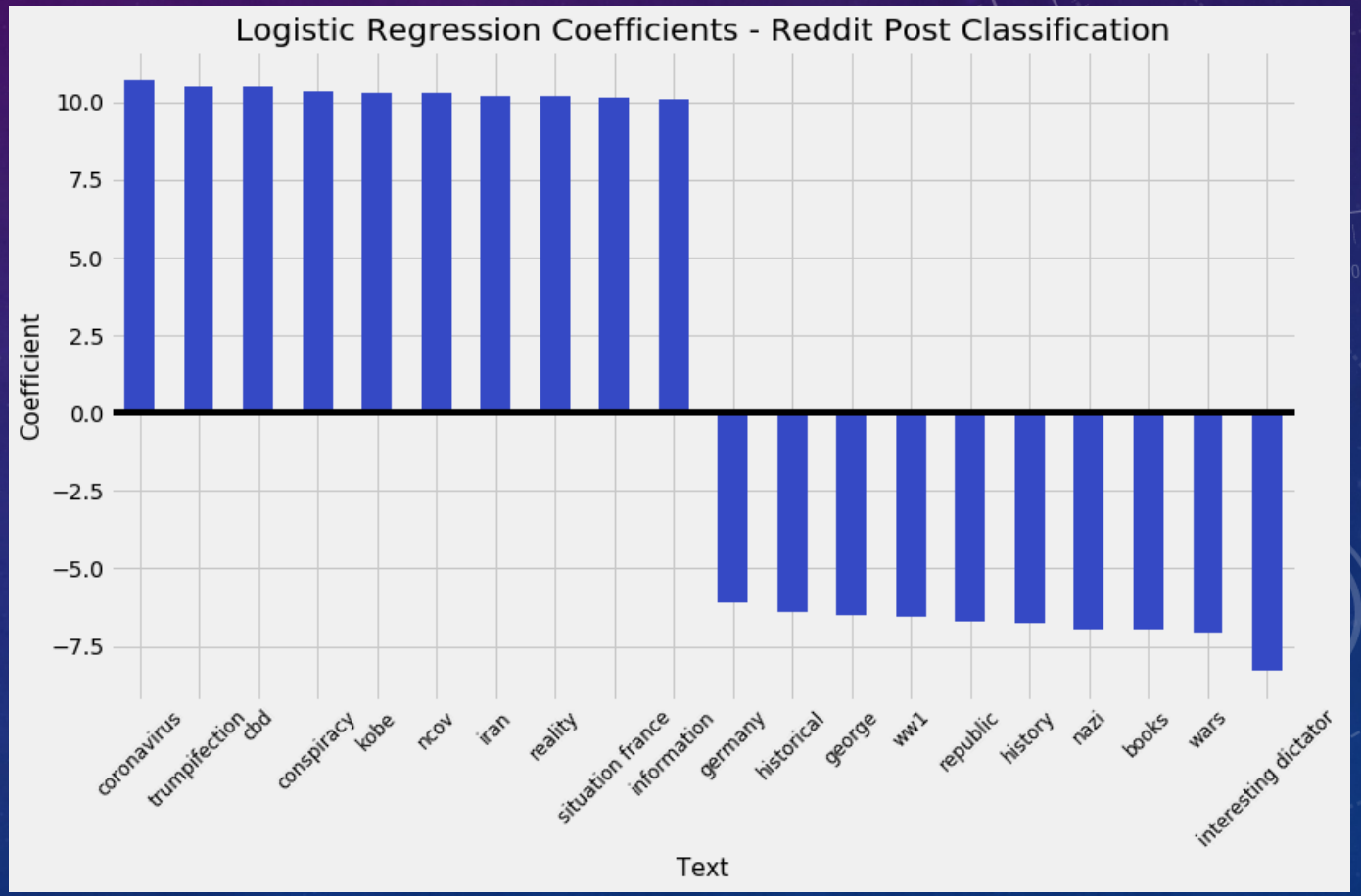
MOST COMMON WORDS

- Heavily influenced by current events (all submissions scraped were from the past month
- Focus on American & ancient history



STRONGEST LOGICAL REGRESSION COEFFICIENTS

- Strongest conspiracy coefficients are all directly tied in to current events
- Strongest history coefficient tied in to large areas of historical research: the World Wars, and especially Nazi Germany



HOW COUNT VECTORIZED LOGICAL REGRESSION STACKS UP

Model Type	Train Score	Test Score	ROC AUC Score
CountVectorizer Naive Bayes	.9821	.93224	.9325
CountVectorizer Logistic Regression	1.0	.9525	.952
TFIDF Logistic Regression	.9940	.9339	.9337
TFIDF Bagging	.9918	.92284	.9337
TFIDF Random Forest	.9599	.8857	.8821
TFIDF Bagging Random Forest	.9873	.9332	.931
TFIDF SVM	.9146	.8264	.8188

CONCLUSIONS & FUTURE WORK

- Since the history subreddit does not allow posts debating events less than 20 years ago, current events heavily weight the word counts in conspiracy posts.
- Larger data set & older posts
- Sentiment analysis