

Political discourse on Reddit

Classifying posts and comments from
r/politics and **r/conservative**

Friday, May 5th 2023

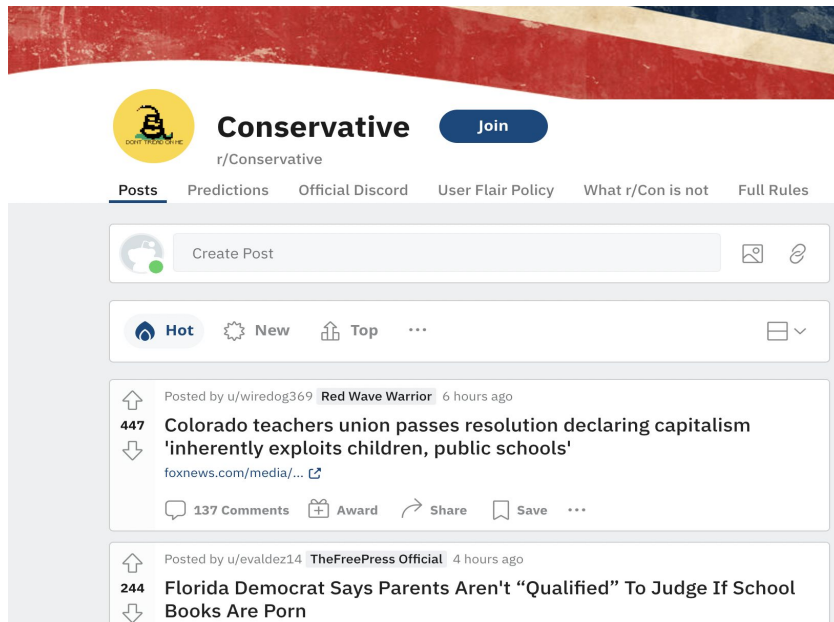
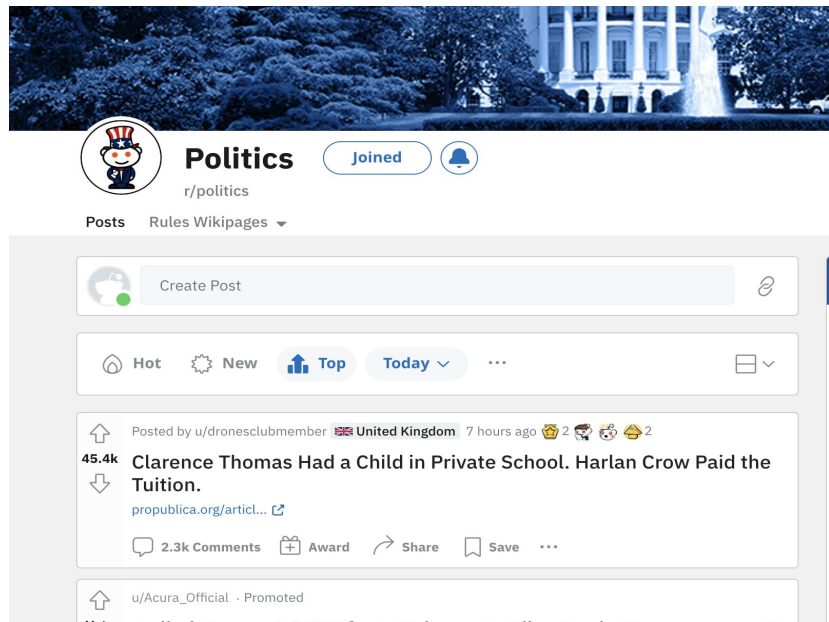
Martijn de Vries

General Assembly

DSIR-0320 Project 3



Problem Statement



Classify posts and comments from October 2022
Evaluate performance with Accuracy

Data Collection

Collected with Pushshift API

October 6th - November 5th, 2022

1 data request for each day (1000 maximum)

Almost no selftext



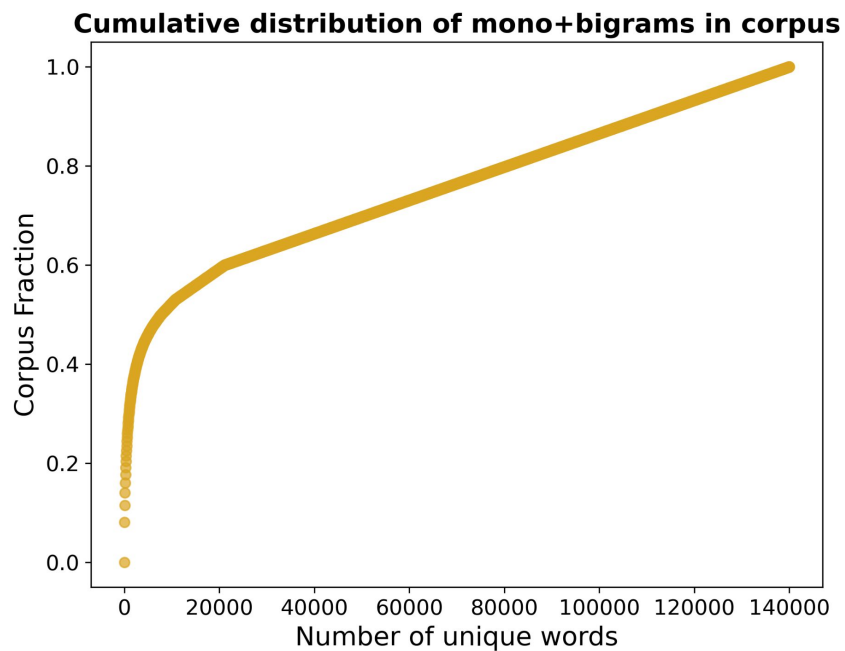
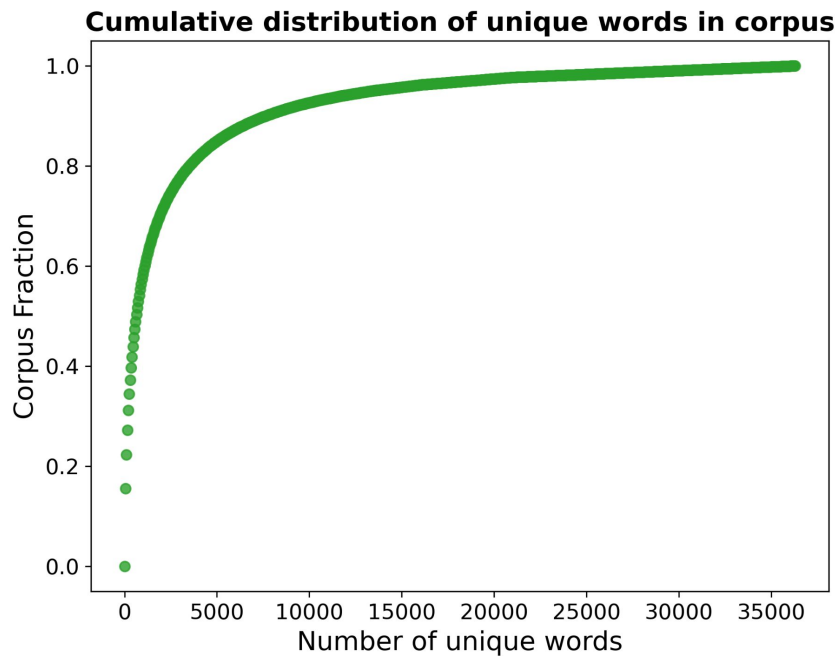
After Cleaning:

	r/conservative	r/politics	Class balance
Posts	9974	8962	53%/47%
Comments	22097	25874	46%/54%

Cleaning and EDA

Removed:

- 1) Duplicate/deleted entries
- 2) Entries with subreddit info or by mods/bots
- 3) Non-english posts



Model features

Posts:

- 1) Post title (text - vectorized with TF-IDF)
 - 2) Number of comments (numerical)
 - 3) Domain name (categorical)
-

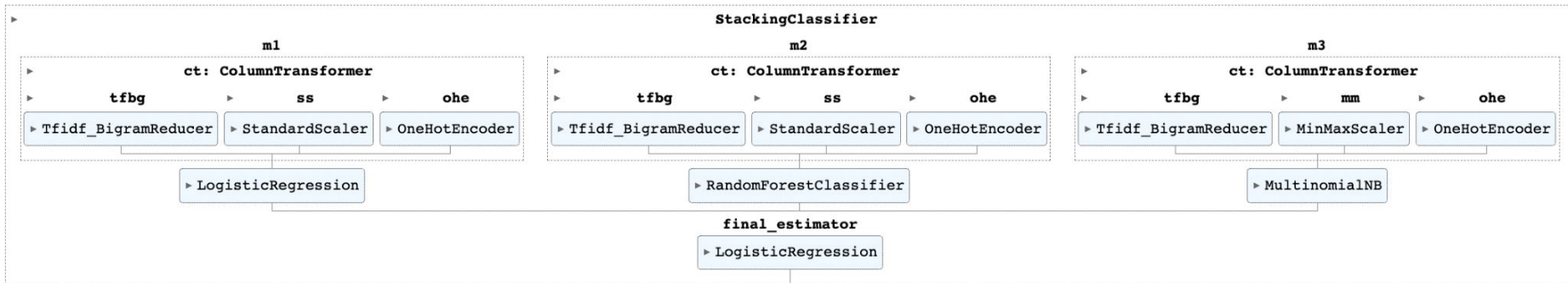
Comments:

- 1) Comment (text - vectorized with TF-IDF)
- 2) Score (numerical)
- 3) Word length (numerical)
- 4) Frequent poster yes/no (categorical)
- 5) Sentiment label (RoBERTa) (categorical)



Credit: XKCD, edited by
https://www.reddit.com/r/ProgrammerHumor/comments/9cu51a/shamelessly_stolen_from_xkcd_credit_where_is_due/

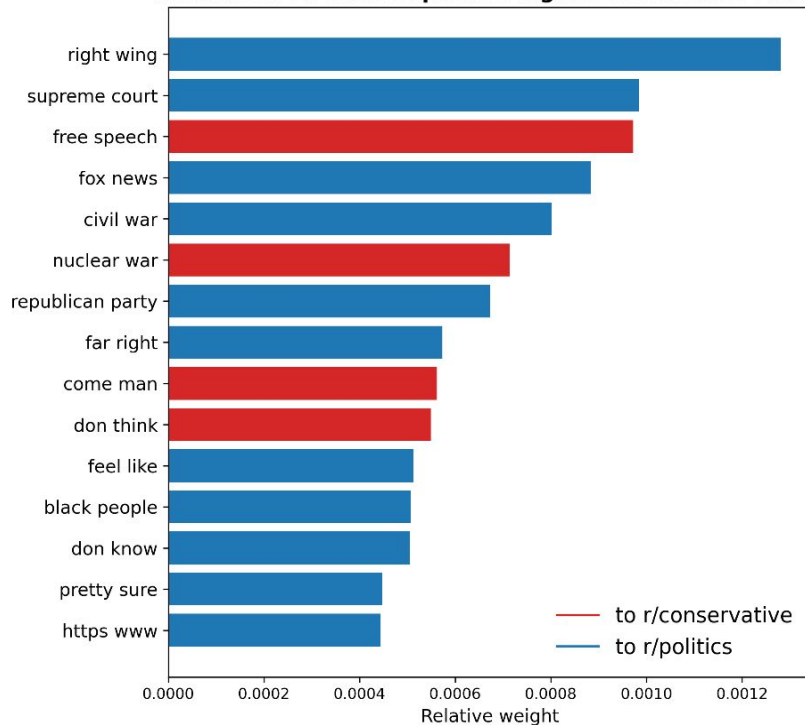
Best Model



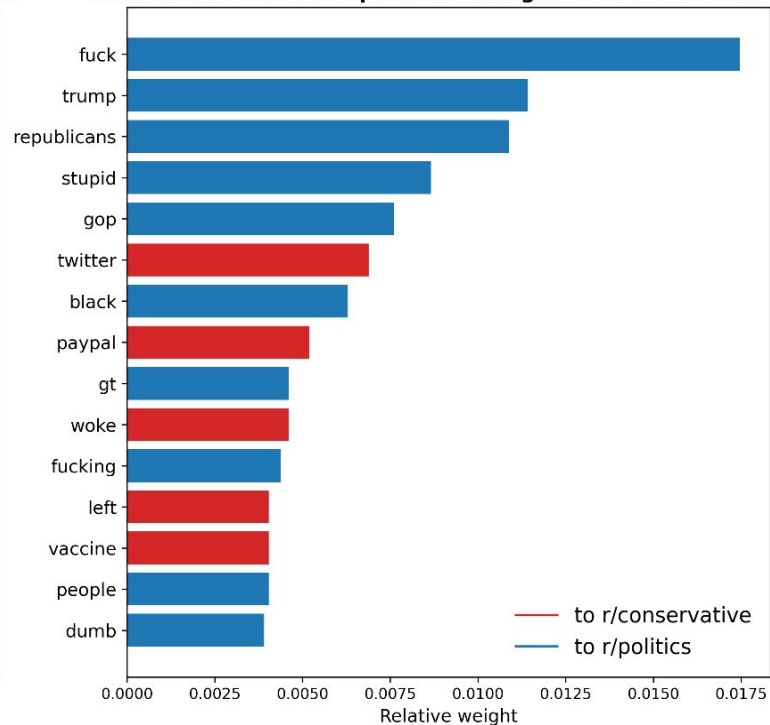
	Training Accuracy	Testing Accuracy
Posts Model	99.1%	88.6%
Comments Model	85.3%	65.4%

Insights

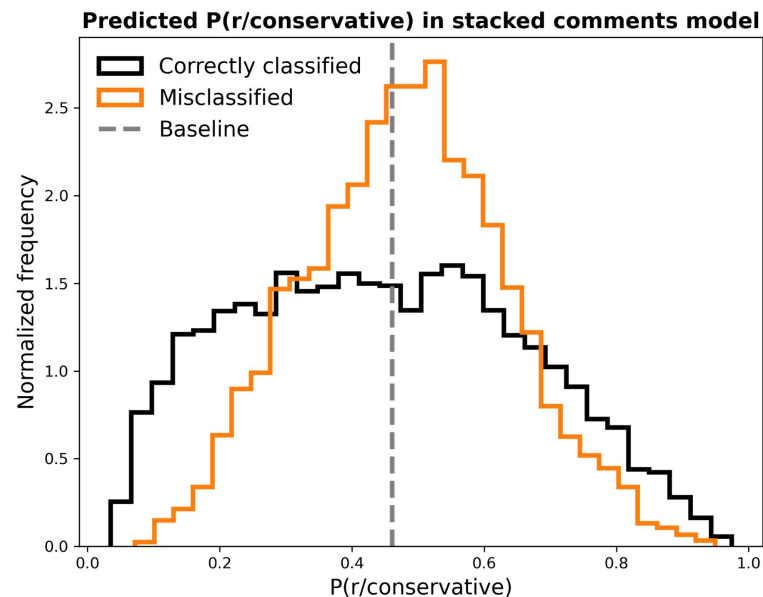
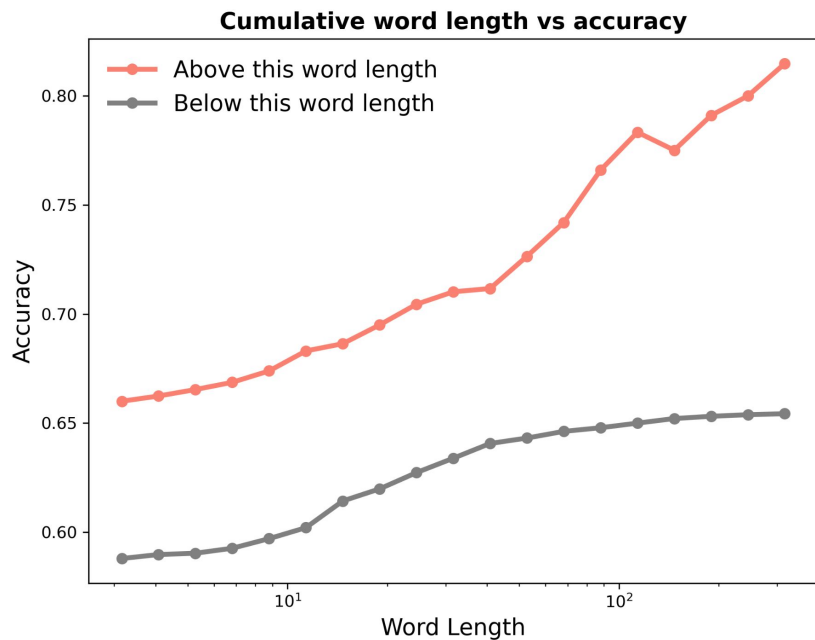
Comments: 15 Most important bigrams in Random Forest



Comments: 15 Most important monograms in Random Forest



Misclassified Comments



Conclusions

Classified **posts** and **comments** from **r/politics** and **r/conservative**

Accuracy: **88.6%** for posts, **65.4%** for comments

Posts model does well domain names

Classifying comments is difficult because there is little info

Lots of interesting information about political use of language