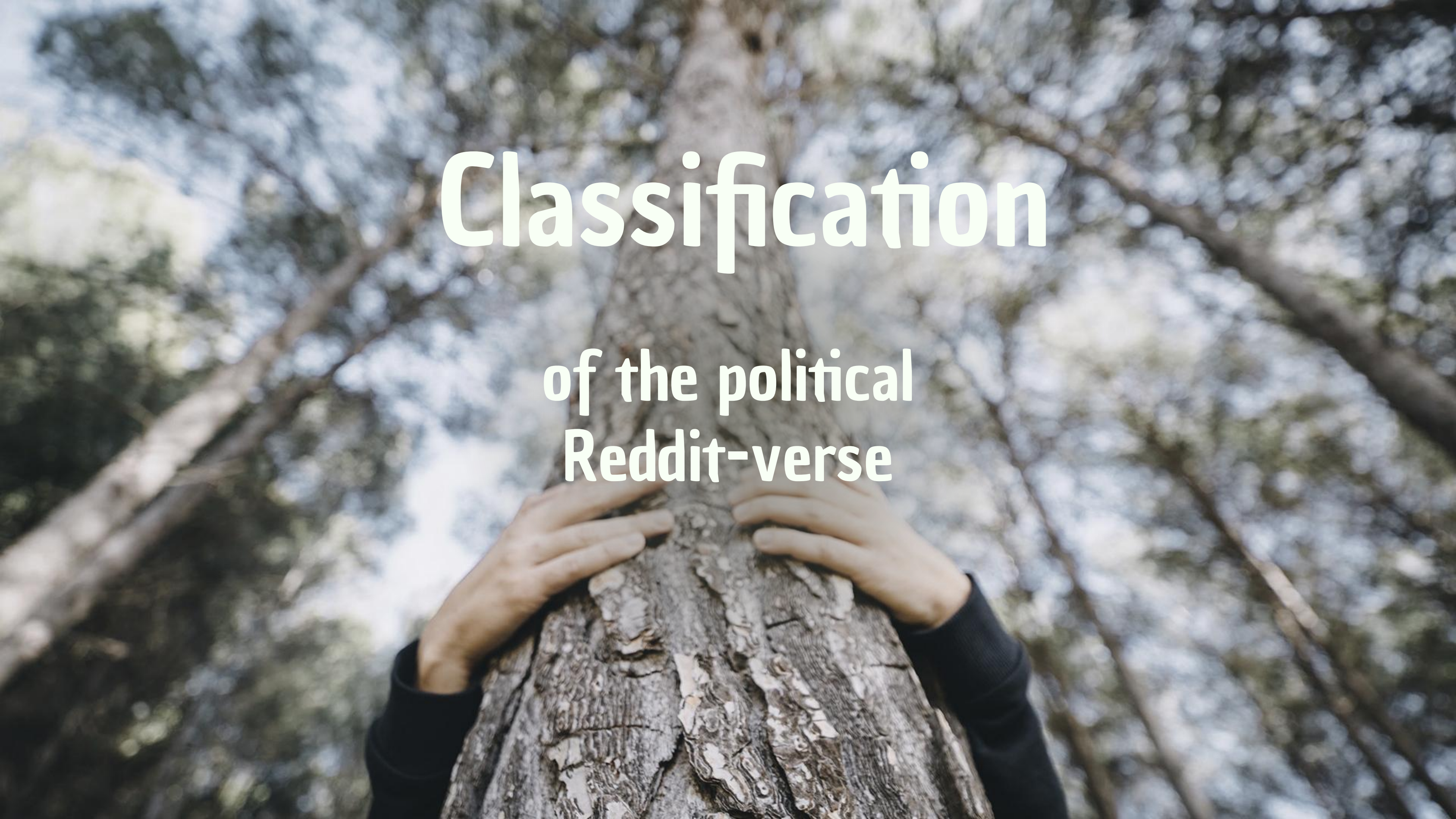
Classification

of the political Reddit-verse



AGENDA

01

Pete
BUTTIGIEG

02

Kamala
HARRIS

03

Bernie
SANDERS

04

Elizabeth
WARREN

Can we predict which candidate's Reddit a message was posted in?

> “”Politicians are the same all over. They promise to build a bridge even where there is no river.”
>
> —Nikita Khruschev

Google

Bernie

Politician since 1991.
Senator from Vermont since ‘07
Longest serving Independent
Running for president in 2020

Pete

Politician.
Mayor of South Bend, In
Former military officer
Running for president in 2020

Liz

Politician & Academic.
Senator from Massachussets
Former law school professor
Running for president in 2020

Kamala

Lawyer & Politician.
Senator from California
Former district attorney
Running for president in 2020

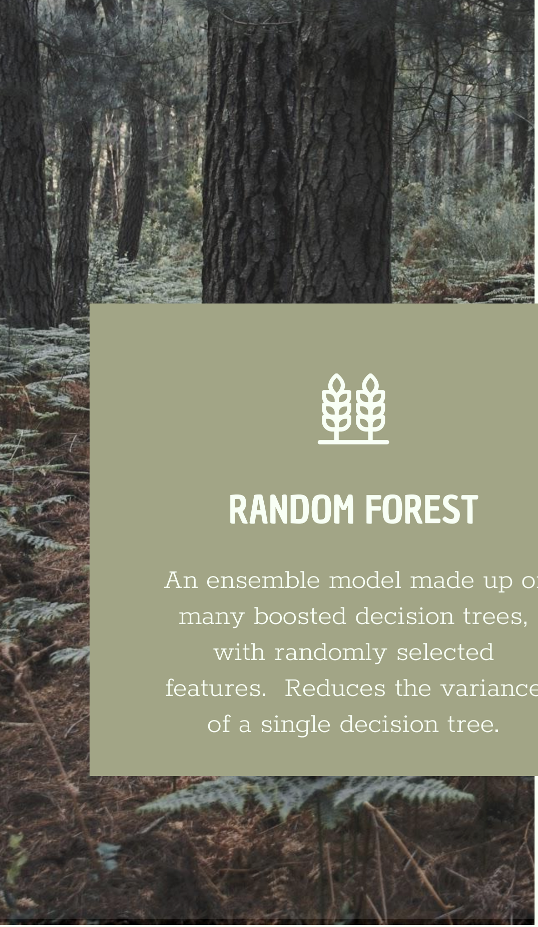
Says

Determine the Baseline

01

Slightly Unbalanced Class

PETE	26%
KAMALA	26%
BERNIE	26%
LIZ	21%



Three Classification Models were chosen to test

RANDOM FOREST

An ensemble model made up of many boosted decision trees, with randomly selected features. Reduces the variance of a single decision tree.

LOGISTIC REGRESSION

Aims to find the relationship between features, and the probability of a particular outcome

GRADIENT BOOSTING

Another decision tree based ensemble model, using regression trees instead of decision trees.

01

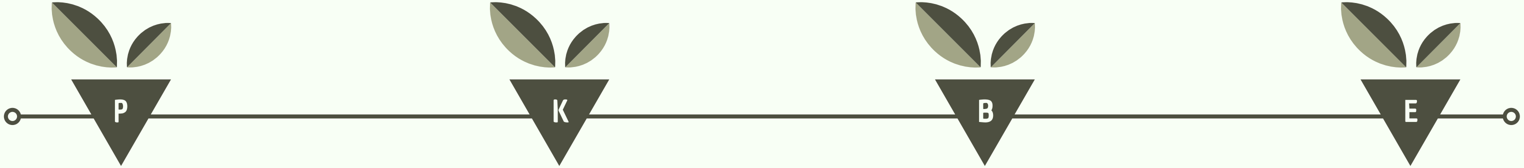
In the end, the simplest model proved to be the best! (but still not so great)

… and now for some performance stats!

	RANDOM FOREST	NAIVE BAYES	LOGISTIC REGRESSION
ENGLISH STOP WORDS	Train: 99.1% Test: **85.2%**	Train: 96.1% Test: **79.1%**	Train: 97.5% Test: **87.5%**
SPEED TO RUN	MEDIUM	FAST	FAST

WOOT!!

TOP 5 WORDS BY CANDIDATE SUBREDDIT

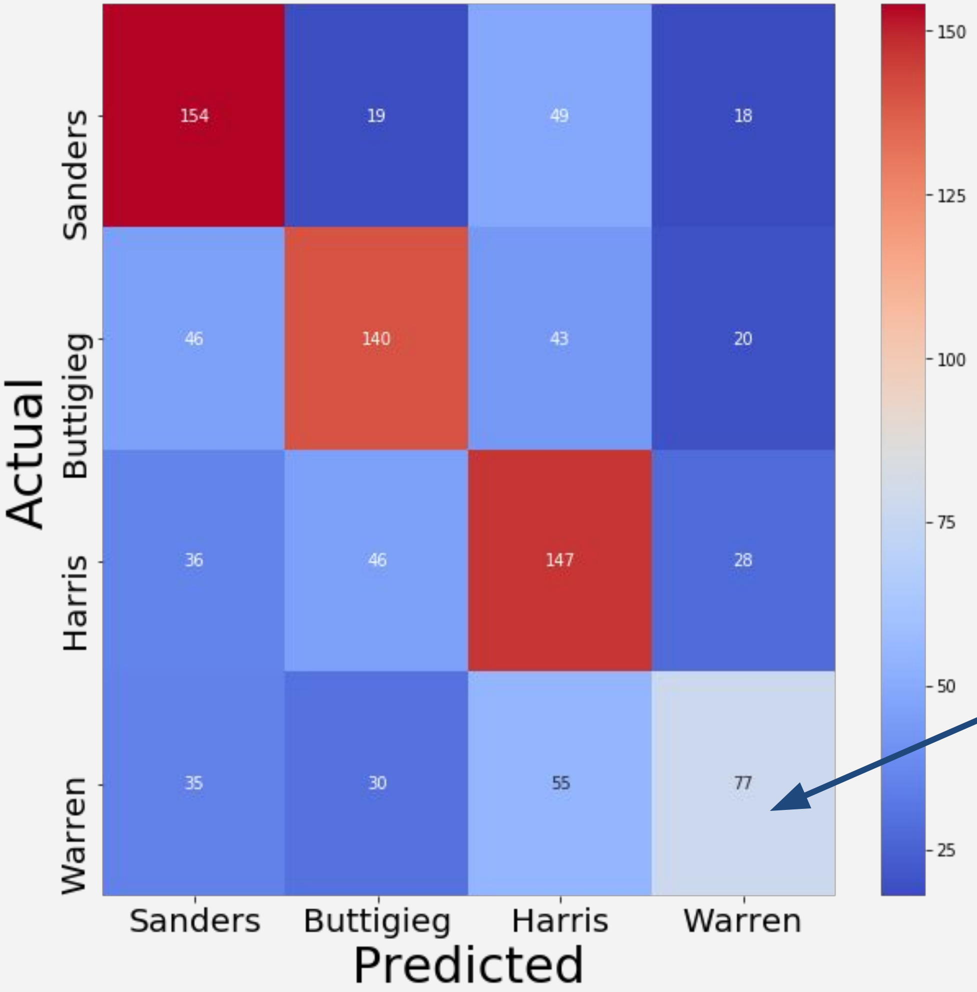


BUTTIGIEG	HARRIS	SANDERS	WARREN
he/him	her/she	he/him	her/she
pete	kamala	bernie	warren
buttigieg	harris	sanders	elizabeth
petebuttigieg	kamalaharris	bernie sanders	elizabeth warren
mayor	barr	medicare	plan
chasten	women	deadline	liz

... and now for some USEFUL stats!

	RANDOM FOREST	NAIVE BAYES	LOGISTIC REGRESSION
ENGLISH STOP WORDS	Train: 99.1% Test: 85.2%	Train: 96.1% Test: 79.1%	Train: 97.5% Test: 87.4%
REMOVED CANDIDATE NAMES	Train: 97.7% Test: 52.3%	Train: 86.4% Test: 54.7%	Train: 90.6% Test: 54.9%
SPEED TO RUN	MEDIUM	FAST	FAST

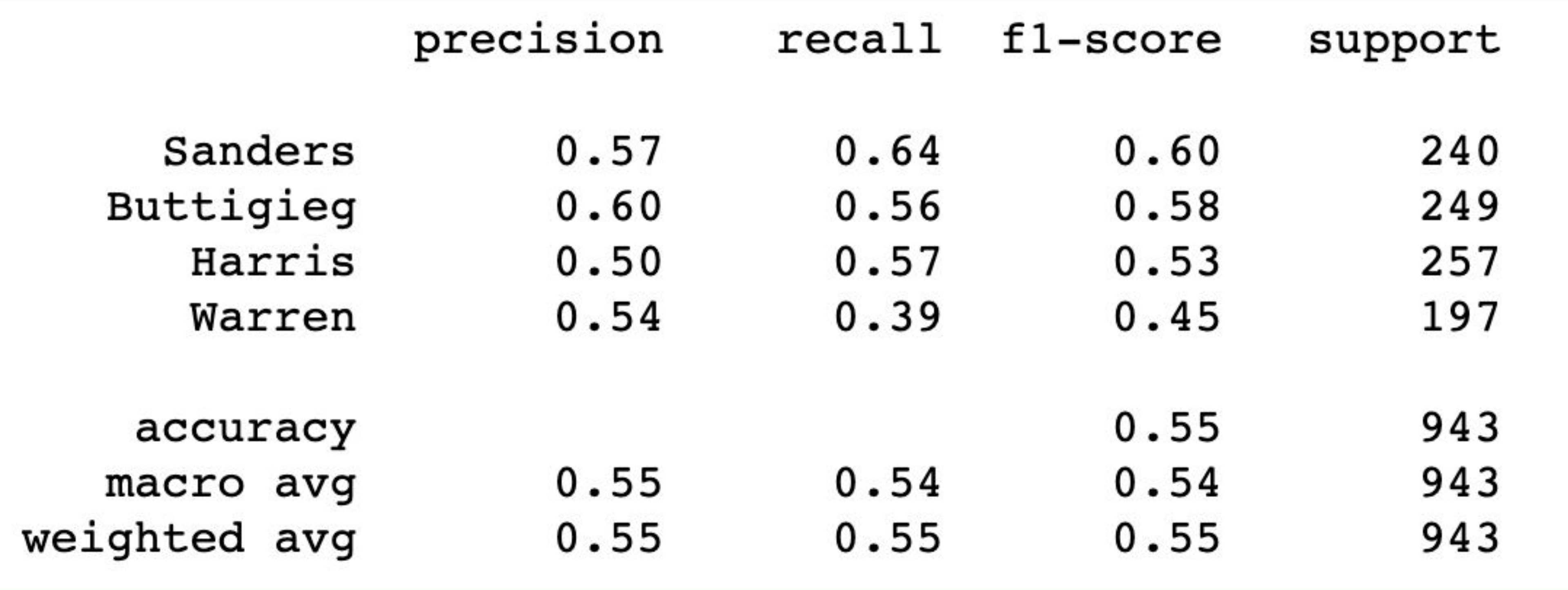
Heatmap of the **Confusion Matrix**



Actual \ Predicted	Sanders	Buttigieg	Harris	Warren
Sanders	154	19	49	18
Buttigieg	46	140	43	20
Harris	36	46	147	28
Warren	35	30	55	77

was there enough data?

Classification Report

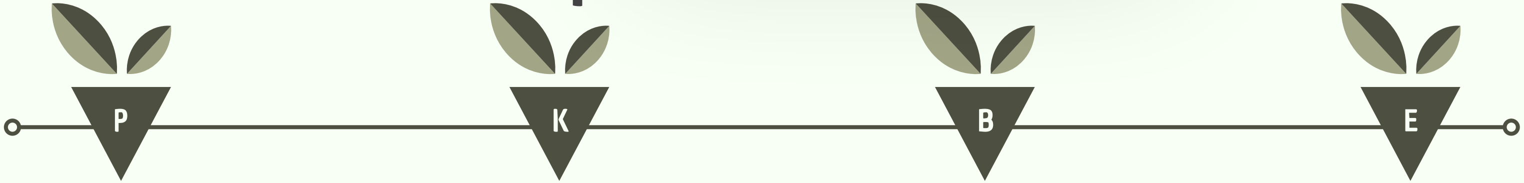


```
              precision    recall  f1-score   support

     Sanders       0.57      0.64      0.60       240
   Buttigieg       0.60      0.56      0.58       249
      Harris       0.50      0.57      0.53       257
      Warren       0.54      0.39      0.45       197

    accuracy                           0.55       943
   macro avg       0.55      0.54      0.54       943
weighted avg       0.55      0.55      0.55       943
```

TOP 5 WORDS BY CANDIDATE SUBREDDIT
after name removals



BUTTIGIEG	HARRIS	SANDERS	WARREN
chasten **(12.2x)**	barr **(7.31x)**	photo **(10x)**	plan **(14.58x)**
south bend	california	insurance **(5.7x)**	policy **(7.7x)**
police	women	medicare	theory
service	busing	deadline	rise
tomorrow	sen	donated	leads
florida	iowa	health	capitalism
shooting	2020	donation	debt
team	senator	today	ask
hire	trump	revolution	winning

Google

Bernie

Politician since 1991.
Senator from Vermont since ‘07
Longest serving Independent
Running for president in 2020

Pete

Politician.
Mayor of South Bend, In
Former military officer
Running for president in 2020

Liz

Politician & Academic.
Senator from Massachussets
Former law school professor
Running for president in 2020

Kamala

Lawyer & Politician.
Senator from California
Former district attorney
Running for president in 2020

Says

THIS IS JUST THE START!

OPPORTUNITIES

Get More Data

and categorized by time posted

Deeper Analysis

How does each candidate's subreddit correlate to policy subject threads

Advanced Modeling

Try to generate better predictions by using advance NLP modeling techniques like Spacy, Amazon Comprehend, Google

CAN YOU SEE THE POSSIBILITIES?!

Does anyone have any questions?

git.generalassmeb.ly/DSTrichter