Flatiron Capstone Project

Analysis of former reddit the_donald

Angelo Turri

Stakeholder

- A social media communications team for the Republican Party wants me to extract meaningful insights from Reddit data
    - Goal: have a better grasp of the sentiment within their voter base
- Reddit: r/the_donald (no longer exists)

Data

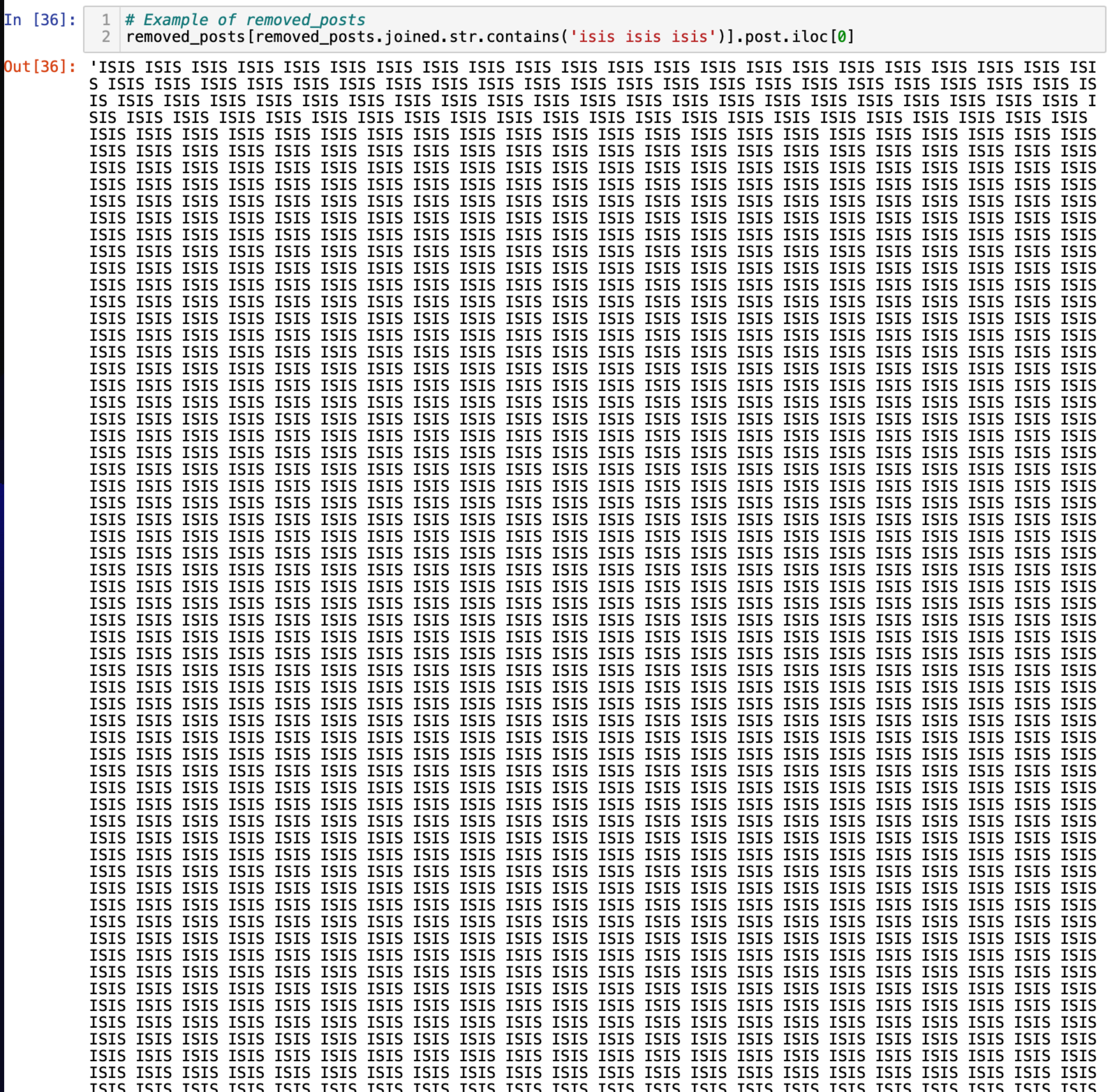
- Origin: https://the-eye.eu/redarcs/
- Dataset properties
    - Date range: 2015–2020
    - 178,308 unique authors
    - 48 million posts
    - Reduced to 2.1 million due to excessively large dataset size

Cleaning

- Removing:
  - Bot posts
  - Spam (any post containing numerous consecutive duplicate words or phrases)
  - Stop-words
    - Commonly used words with little meaning that make very small contributions at best to models

Spam example

D:



In [36]:
```
# Example of removed_posts
removed_posts[removed_posts.joined.str.contains('isis isis isis')].post.iloc[0]
```

Out[36]:
```
'ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISI
S ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS IS
IS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS I
SIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS ISIS
```

Feature Engineering (Bag of Words)

- Three levels of analysis
  - Unigrams (single words)
  - Bigrams (two-word phrases)
  - Trigrams (three-word phrases)
- Limited vocabulary to top-100 unigrams, bigrams, or trigrams
- Every sentence including one or more of these words/phrases was used for modeling

How is a sentence represented?

Vocabulary: 5 words - [‘Donald’, ‘Trump’, ‘became’, ‘President’, ‘apple’]

“Donald Trump became President”

is_donald	is_trump	is_became	is_president	is_apple
1	1	1	1	0

All Target Variables

- Post score (upvotes - downvotes)
- VADER sentiment score (engineered)
  - “Valence Aware Dictionary sEntiment Reasoner”

🤦‍♂️

- These are different ways of measuring sentiment, we will choose whichever one allows for a better model and more interpretable results.

Type of model used: Linear Regression

- Highly interpretable model: it provides the values of coefficients
  - Lets us know which words the model thought negatively impacted sentiment and which ones positively impacted it
- I will create a separate linear regression model for each set of features, so that unigrams, bigrams and trigrams can be analyzed separately

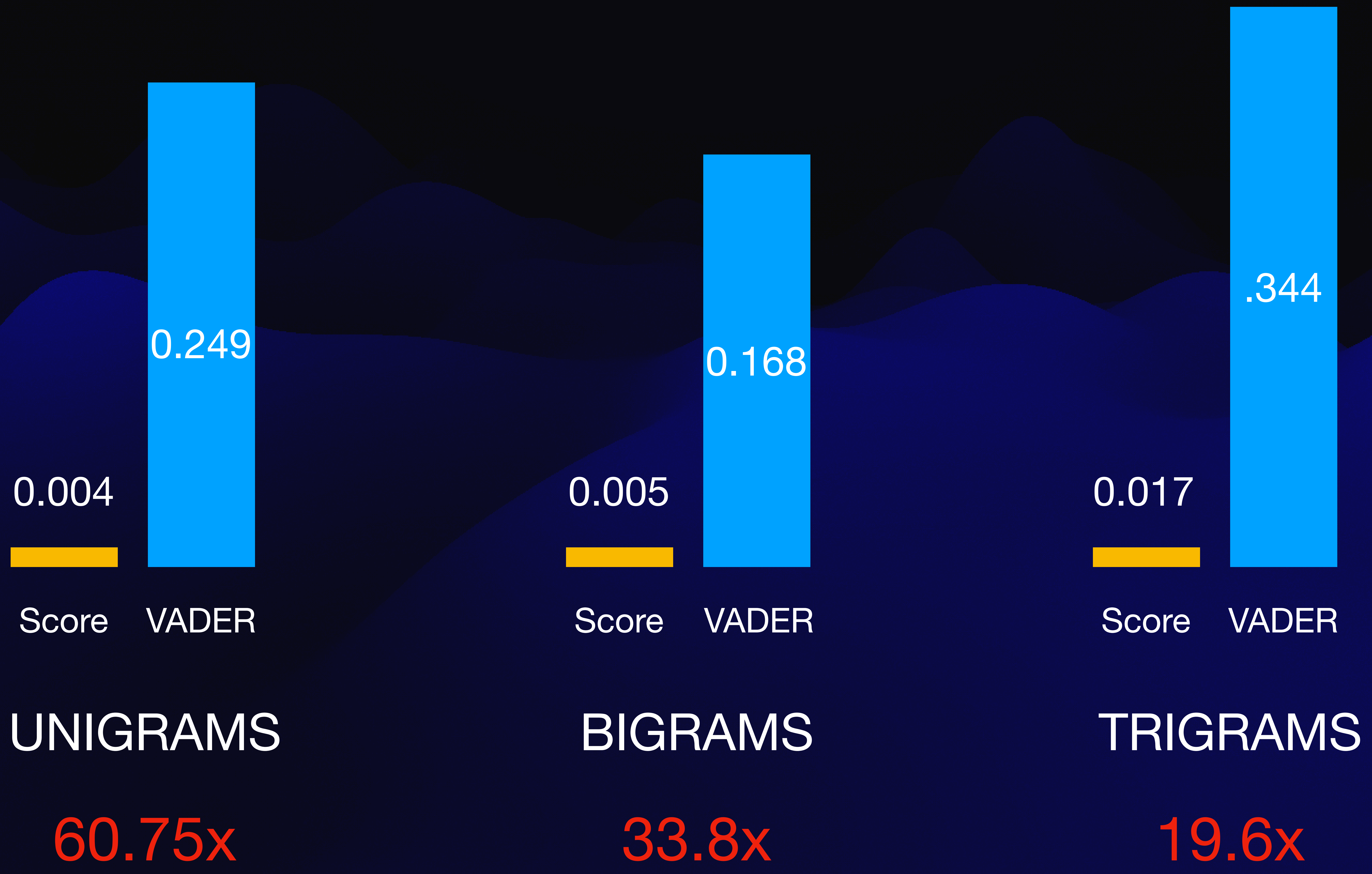
Target Variable: Score

- Abysmal results across the board
- R-squared < 0.1 for unigrams, bigrams and trigrams models
  - R-squared ranges from 0 to 1
- Models categorized negative words as positive
- Models categorized positive words as negative
- Off by an average of 7-8 upvotes where the average score was 7-8 upvotes

Target Variable: VADER sentiment

- Much better results
- R-squared of 0.243, 0.169, and 0.333 for unigrams, bigrams and trigrams respectively
- The broad sentiment of the top positive words and negative words was what you'd expect.
- Off on average by 0.33, 0.433, and 0.363 for unigrams, bigrams and trigrams respectively
  - VADER sentiment ranges from -1 to 1

Model R-squared improvement

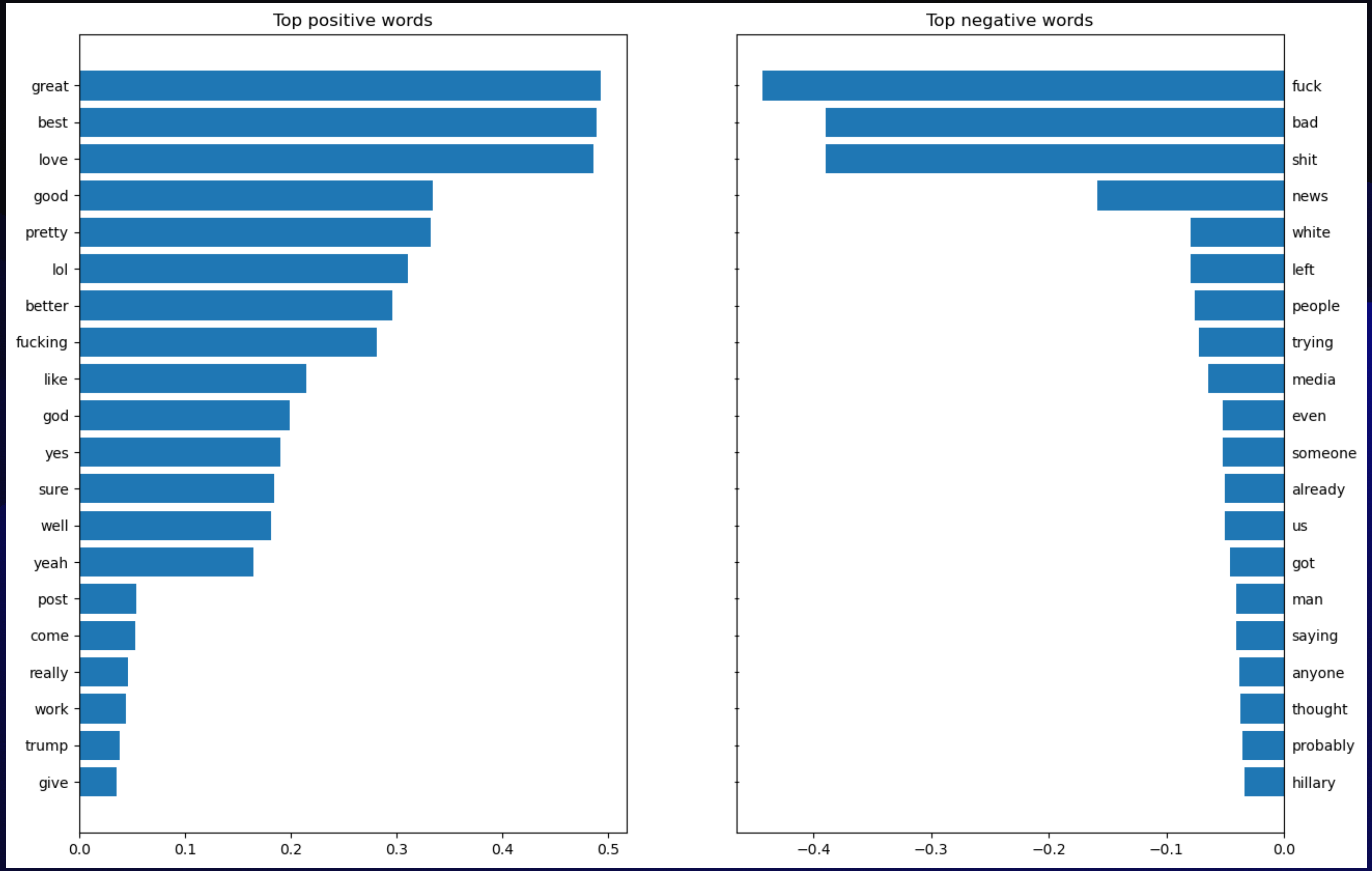


Top unigrams after tuning

- The longer the bar, the more emotionally charged that word/phrase

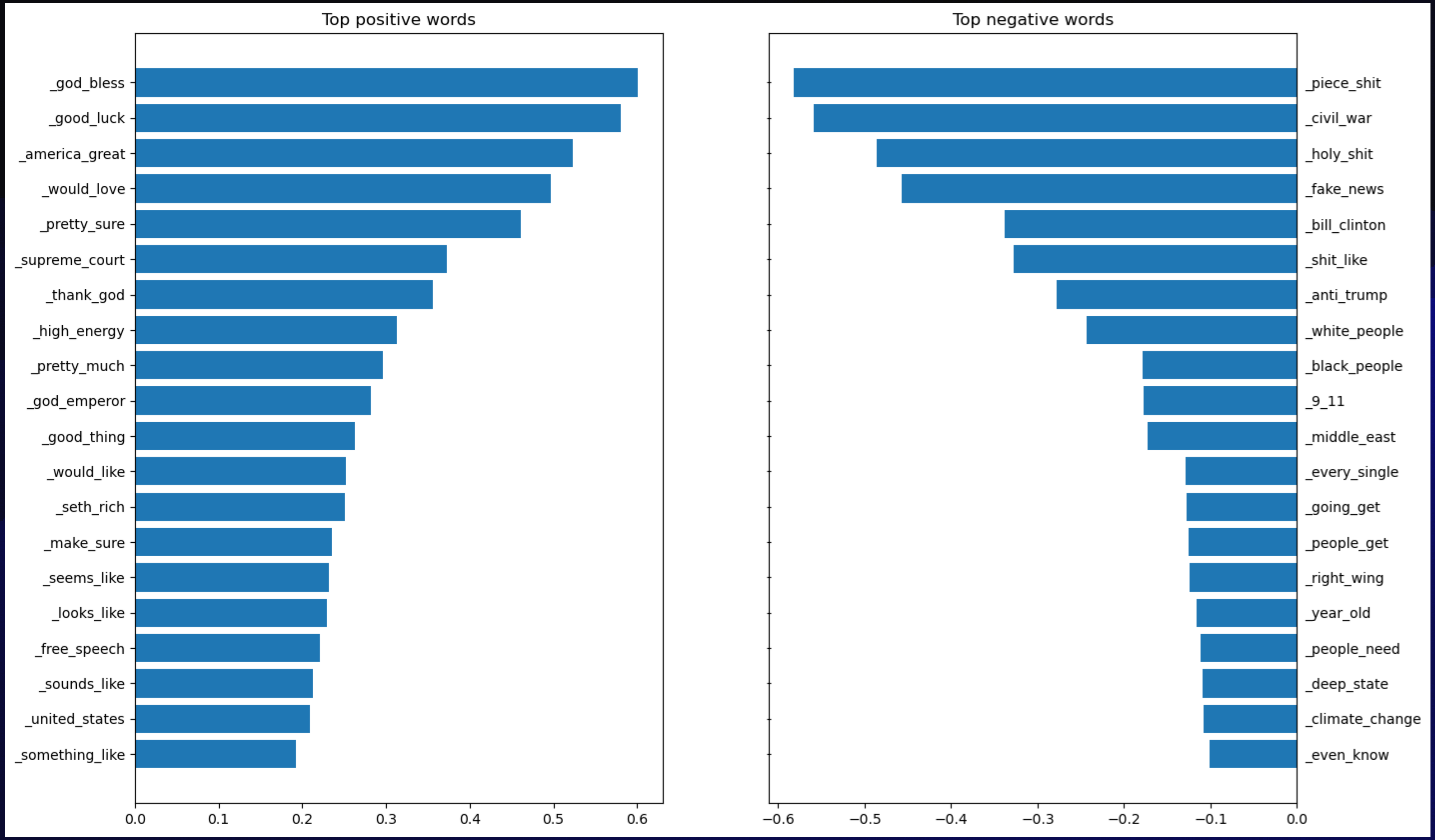
- Positive column: Best, God, Trump

- Negative column: News, media, Hillary



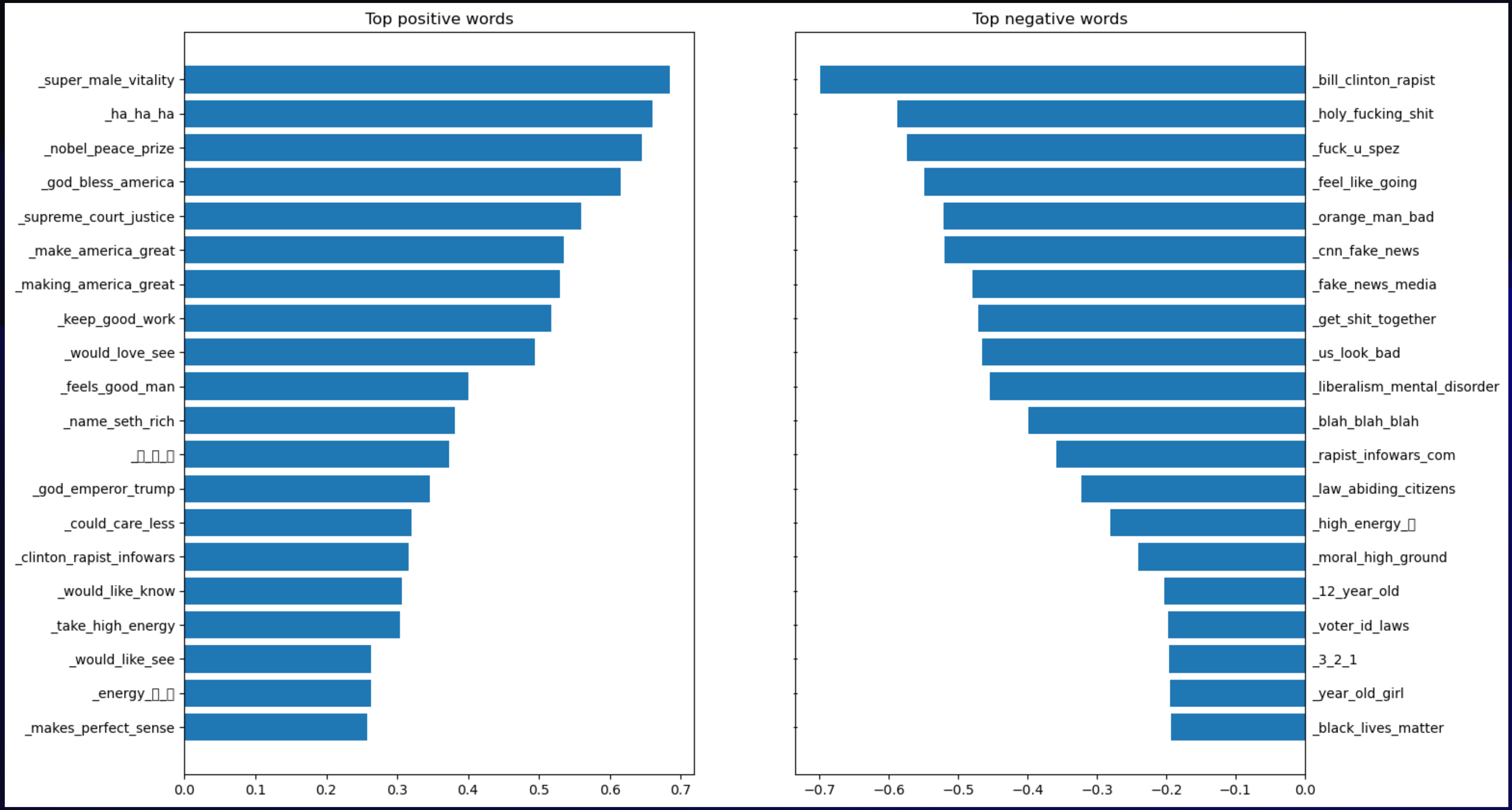
Top bigrams after tuning

- The longer the bar, the more emotionally charged that word/phrase
- Positive column: God bless, America great, thank God, god emperor, United States, free speech
- Negative column: civil war, fake news, Bill Clinton, anti-Trump, Middle East, Deep State



Top trigrams after tuning

- The longer the bar, the more emotionally charged that word/phrase
- Positive column: God bless America, make America great, God emperor Trump
- Negative column: Bill Clinton rapist, orange man bad, CNN fake news, liberalism mental disorder



Insight #1

- The base is passionate about their representative
  - "Best" is listed in the most positive words
  - The word "Trump" is listed in the most positive words
  - The phrase "god emperor Trump" is listed in the most positive trigrams
  - Phrases insulting their representative such as "orange man bad" and "anti trump" are listed in the most negative trigrams

Insight #2

- The base is passionately against opposing politicians and organizations
  - “News”, “media”, “fake news”, “fake news media”, “CNN fake news” and “deep state” appear in the most negative phrases
  - “Hillary”, “Bill Clinton,” “Bill Clinton rapist”, “left”, and “liberalism mental disorder” appear in the most negative phrases

Insight #3

- The base is highly religious
  - “God”, “God bless,” “thank God”, and “God bless America” appear in the most positive phrases.

Insight #4

- The base is very passionate about the relevant issues
    - Swear words such as the f-word and “shit” commonly appear in both the most positive and negative phrases column.

Insight #5

- The base is very fond of their country
    - Country-oriented phrases such as “America great”, “United States”, “God bless America,” “make America great,” and “making America great” appear in the positive column.
    - Phrases such as “civil war” and “middle east” appear in the negative column.

Recommendations

- Try to match the base’s enthusiasm with your own
- Be aware that they are angry against certain people and organizations and want immediate change

Thank You!

- All questions are welcome.
- You may contact me here:
  - angelo.turri@gmail.com