Who does r/Politics hate the most?

By Oliver Green
Project number: 130

Abstract

Name Entity Recognition and Sentiment Analysis were used to analyze posts from Reddit's r/Politics. The ultimate takeaway was that Reddit seriously does not like Donald Trump. He was talked about the most by far and all of it was very negative.

Introduction

It is often hard to survey people voluntarily about their political opinions, but on forums such as r/Politics people share them freely. This presents a gold mine of free political data. With this in mind, I decided to use a combination of Name Entity Recognition (NER) and Sentiment Analysis to gather political opinions from Reddit's r/Politics.

Methods

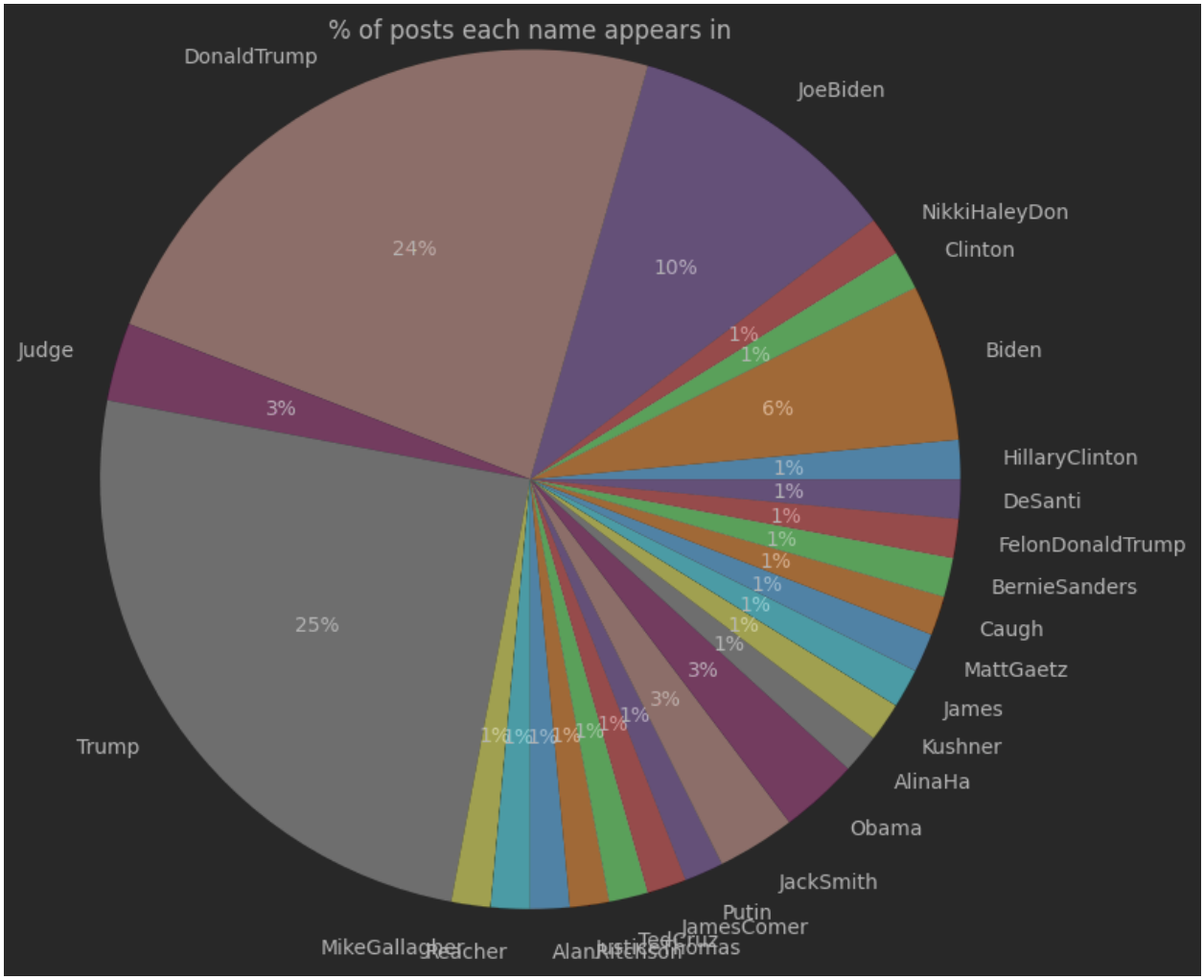
As mentioned above, I used NER and Sentiment Analysis to discover and aggregate the opinions of r/Politics. I performed NER on the titles of various posts to extract the names of the politicians being discussed. I then performed Sentiment Analysis on the comments underneath those posts to get a general idea of how people felt about the politicians that were mentioned in the post's title.

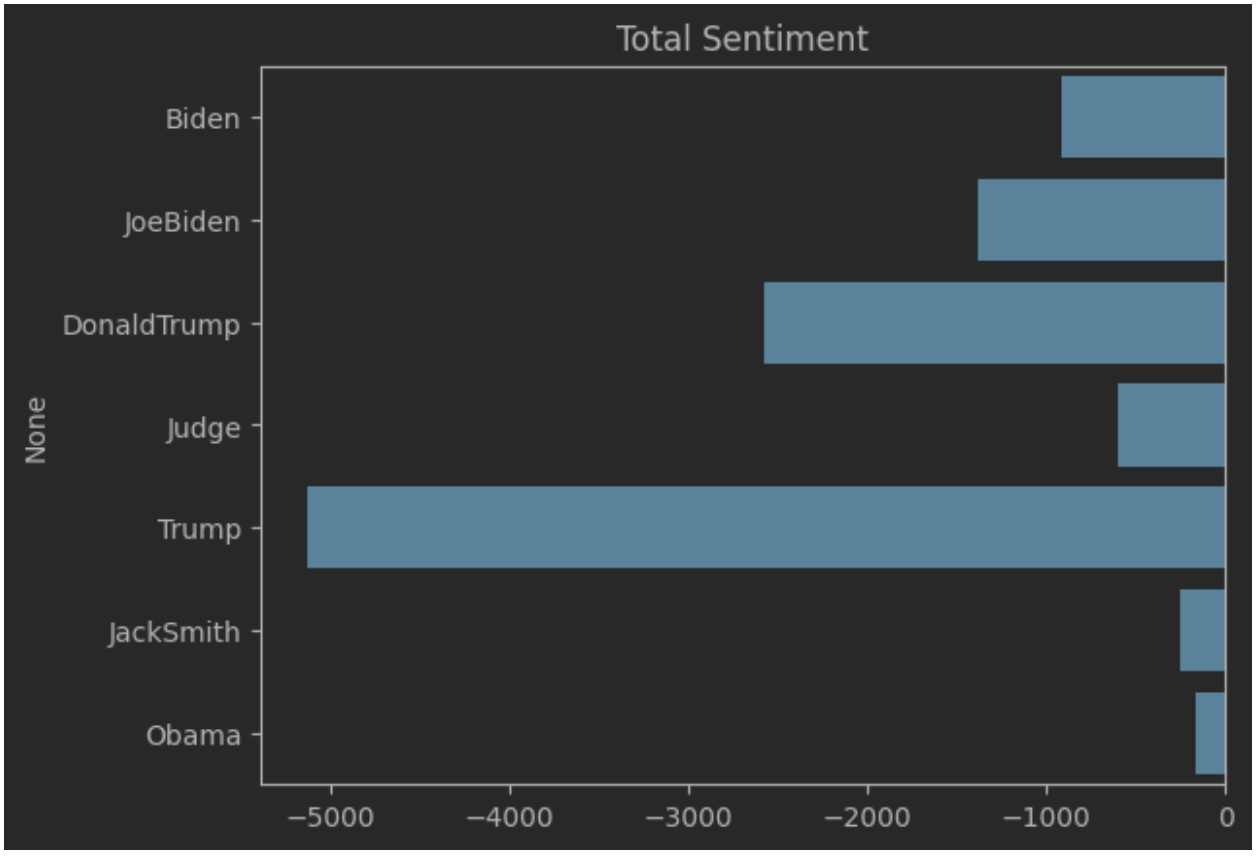
Data

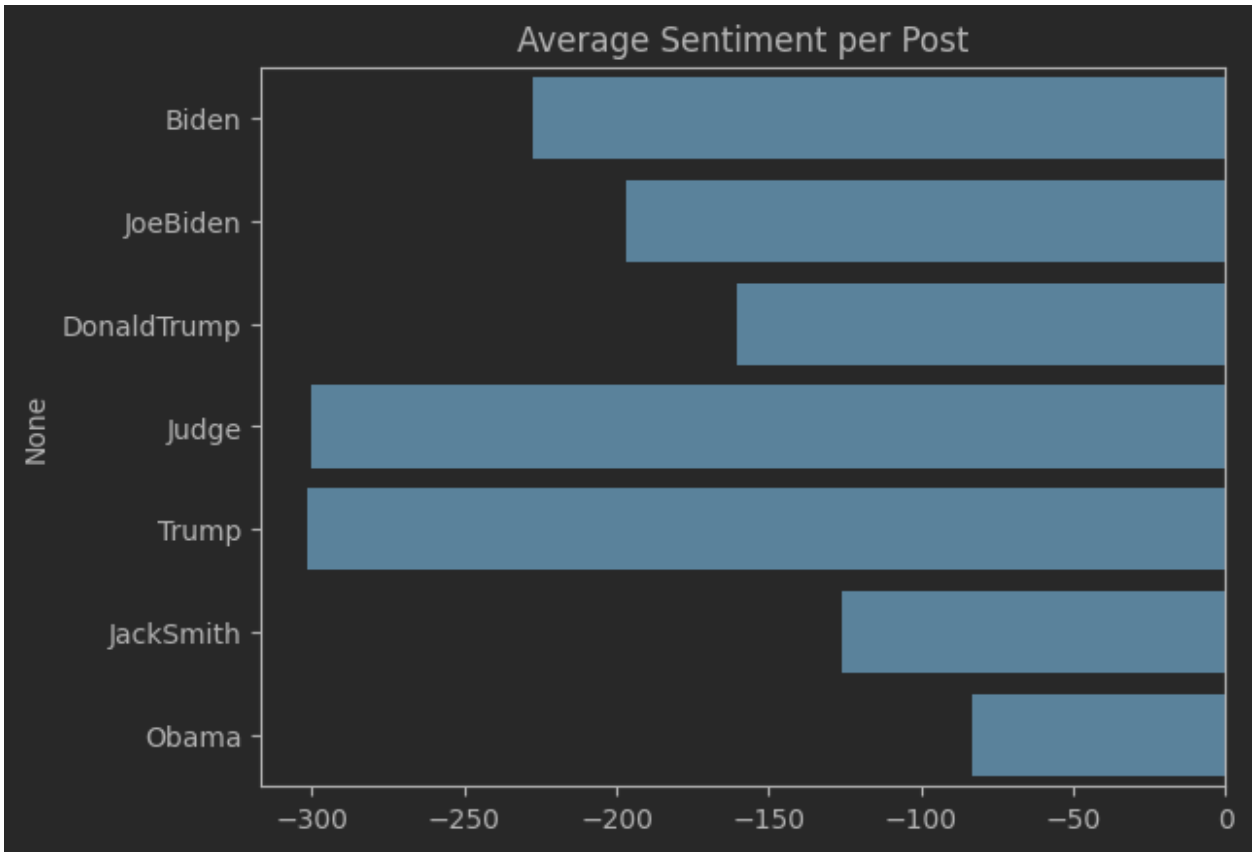
The raw data was gathered using PRAW, a Python wrapper for Reddit's API. I then stored this data in a dictionary and later dumped it into a .json file for long-term storage. I was able to gather data from about 50 posts before I reached Reddit's API limit.

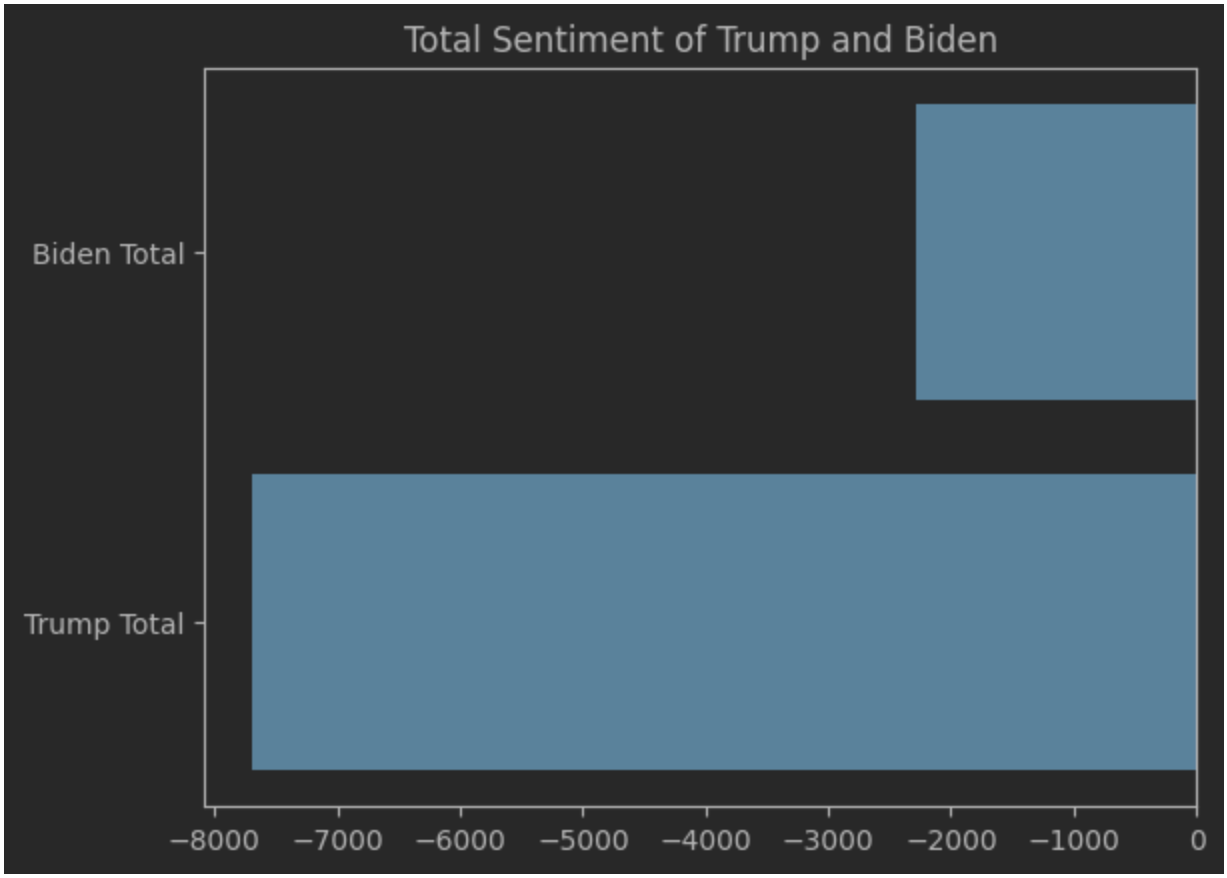
Results

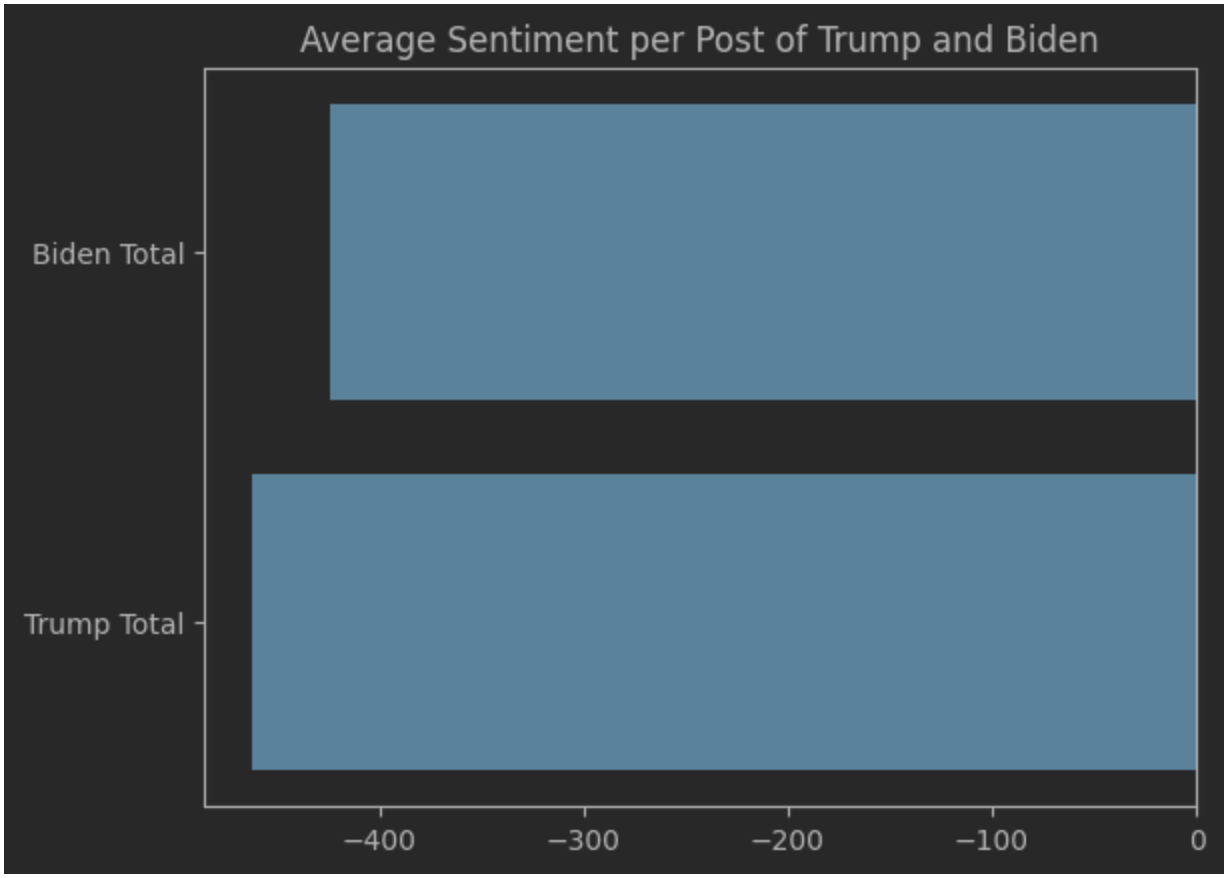
The following visualizations show the percentage of posts in which each politician was mentioned, the total sentiment for politicians' who were mentioned more than twice, the average sentiment per post of politicians mentioned more than twice, the total sentiment for just Trump and Biden, and finally the average sentiment per post of just Trump and Biden.


% of posts each name appears in
DonaldTrump
24%
JoeBiden
10%
NikkiHaleyDon
1%
Clinton
1%
Biden
6%
Judge
3%
HillaryClinton
1%
DeSanti
1%
FelonDonaldTrump
1%
BernieSanders
1%
Caugh
1%
MattGaetz
1%
James
1%
Kushner
1%
AlinaHa
3%
Obama
3%
JackSmith
1%
Putin
1%
JamesComer
1%
TedCruz
1%
JustinThomas
1%
AlanAitkenson
1%
Reacher
1%
MikeGallagher
Trump
25%


Total Sentiment
Biden
JoeBiden
DonaldTrump
Judge
Trump
JackSmith
Obama
None
−5000
−4000
−3000
−2000
−1000
0


Average Sentiment per Post
Biden
JoeBiden
DonaldTrump
Judge
Trump
JackSmith
Obama
None
−300
−250
−200
−150
−100
−50
0


Total Sentiment of Trump and Biden
Biden Total
Trump Total
−8000
−7000
−6000
−5000
−4000
−3000
−2000
−1000
0



Discussion

I had originally planned to call this project, "What does Reddit think about Politicians?" I assumed some of the sentiments would be positive around more popular presidents like Obama, however, I soon discovered that r/Politics hates everyone. It's just a question of who they hate more. As the visualizations show, Donald Trump was mentioned the most by far of any politician in post titles. His total sentiment exceeded everyone else's by a large margin although his average sentiment per post was only nominally greater than Biden's.

References

- r/Politics
- BERT

Appendix:

The code is structured as a Jupyter Notebook that first connects to the Reddit API using a secret API key. Then it gathers the data from Reddit and performs NER on it simultaneously. Then it combines the name chunks that BERT spits out (eg. B, ##iden) into full names. Then it performs sentiment analysis on the comments underneath each post. Then it aggregates those sentiments into totals and averages. Finally, it produces visualizations.

The libraries I used were PRAW for wrapping the Reddit API, JSON for storing dictionary form data, tqdm was used to create loading bars, Seaborn and matplotlib were used for visualizations and finally, transformers was used for BERT NER and Sentiment Analysis.

Function Descriptions

- NER(text)
    - Performs NER on post titles using BERT
- extractNames(textNER)
    - Combines name chunks from bert (eg. B, ##iden) into full names with no spaces (eg. JoeBiden)
- sentimentAnalysis(post)
    - Performs Sentiment Analysis on the comments of a given post and returns either POSITIVE, NEGATIVE, or AMBIGUOUS

User Manual

To run the full Jupyter Notebook, you must create a Reddit account and request an API key. Then store that information in a file named secrets.json in the format:

```
{"id": "API_ID",
  "pass": "REDDIT_PASSWORD",
  "user": "REDDIT_USERNAME",
  "secret": "API_SECRET_KEY"}
```

Finally, run all of the cells as normal.

Demo Code

To run the demo code open demo.py then customize what you want your toy post titles and comments to be. These are shown at the top of the code under the comment header #TOY DATA. Then run the file as normal and observe the visualizations it depicts. They will not be incredibly detailed as this toy model only uses two posts with three comments per post. It does give you a general idea of how the code works though.