Reddit NLP Analysis: Subreddit Prediction

Edward Lee

The problem: building a simple & robust classification model

Our Goal

Predict which subreddit a given post-title or comment came from:

- r/geopolitics
- r/worldnews

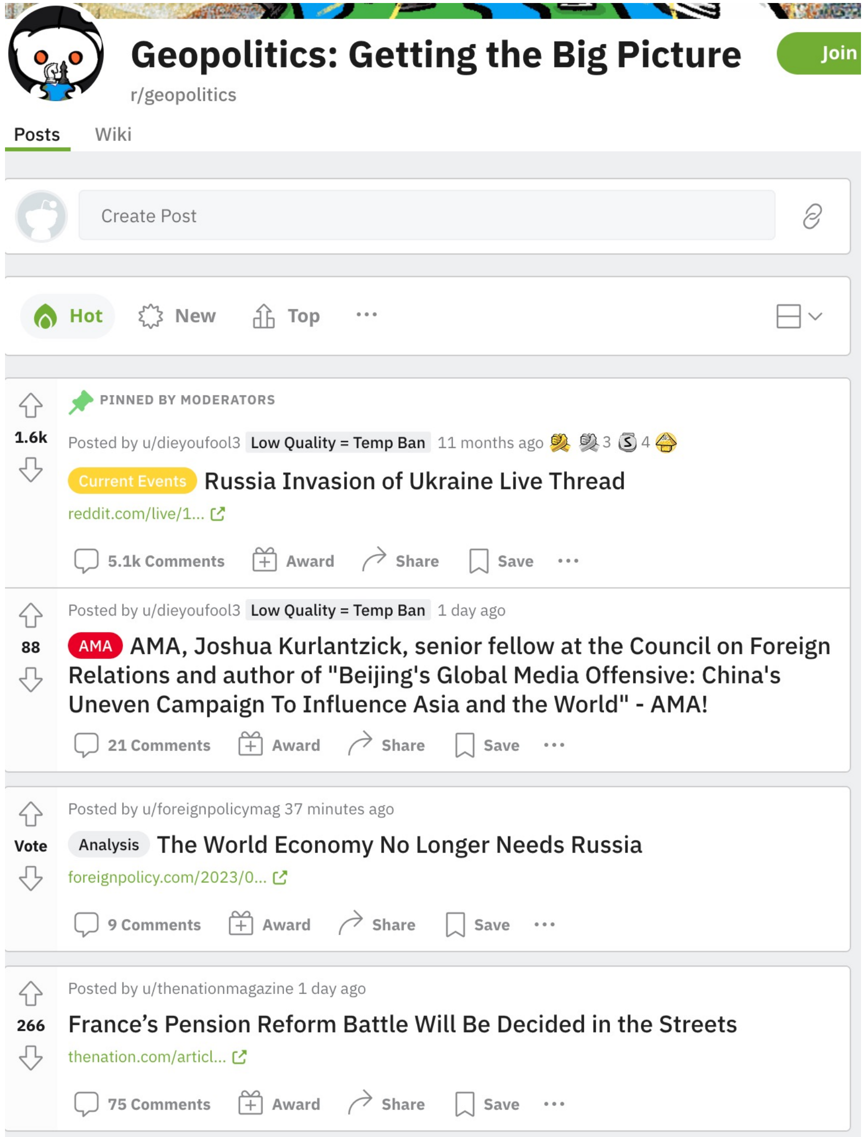
Model Use Case

- The purpose of this exercise is to simply build the most accurate and effective model.
- There are many real world applications this model can be applied to at a later stage

Audience

Semi-technical audience

Context: r/geopolitics vs. r/worldnews



Geopolitics: Getting the Big Picture

r/geopolitics

Join

Posts | Wiki

Create Post

Hot | New | Top | ···

PINNED BY MODERATORS

1.6k

Posted by u/dieyoufool3 Low Quality = Temp Ban 11 months ago 3 4

Current Events Russia Invasion of Ukraine Live Thread

reddit.com/live/1...

5.1k Comments | Award | Share | Save | ···

88

Posted by u/dieyoufool3 Low Quality = Temp Ban 1 day ago

AMA AMA, Joshua Kurlantzick, senior fellow at the Council on Foreign Relations and author of "Beijing's Global Media Offensive: China's Uneven Campaign To Influence Asia and the World" - AMA!

21 Comments | Award | Share | Save | ···

Vote

Posted by u/foreignpolicymag 37 minutes ago

Analysis The World Economy No Longer Needs Russia

foreignpolicy.com/2023/0...

9 Comments | Award | Share | Save | ···

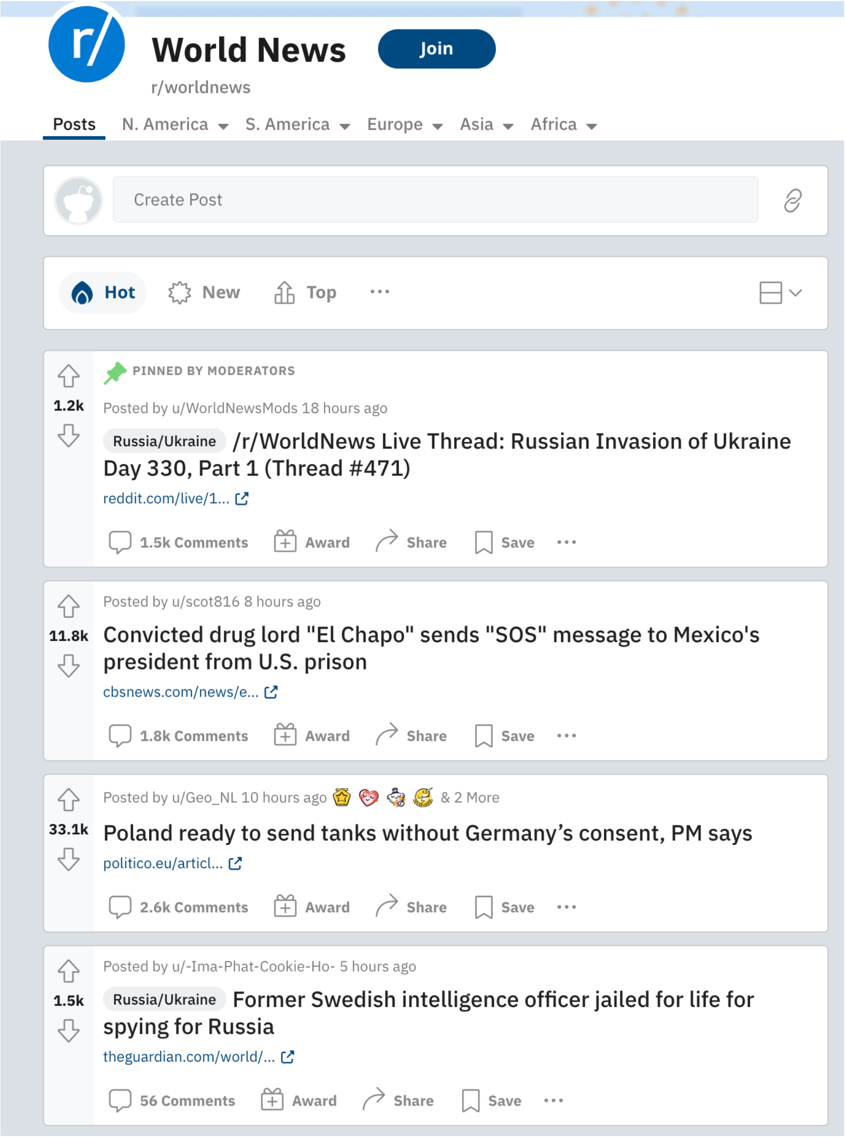
266

Posted by u/thenationmagazine 1 day ago

France’s Pension Reform Battle Will Be Decided in the Streets

thenation.com/articl...

75 Comments | Award | Share | Save | ···



World News

r/worldnews

Join

Posts | N. America | S. America | Europe | Asia | Africa

Create Post

Hot | New | Top | ···

PINNED BY MODERATORS

1.2k

Posted by u/WorldNewsMods 18 hours ago

Russia/Ukraine /r/WorldNews Live Thread: Russian Invasion of Ukraine Day 330, Part 1 (Thread #471)

reddit.com/live/1...

1.5k Comments | Award | Share | Save | ···

11.8k

Posted by u/scot816 8 hours ago

Convicted drug lord "El Chapo" sends "SOS" message to Mexico's president from U.S. prison

cbsnews.com/news/e...

1.8k Comments | Award | Share | Save | ···

33.1k

Posted by u/Geo_NL 10 hours ago & 2 More

Poland ready to send tanks without Germany’s consent, PM says

politico.eu/articl...

2.6k Comments | Award | Share | Save | ···

1.5k

Posted by u/-Ima-Phat-Cookie-Ho- 5 hours ago

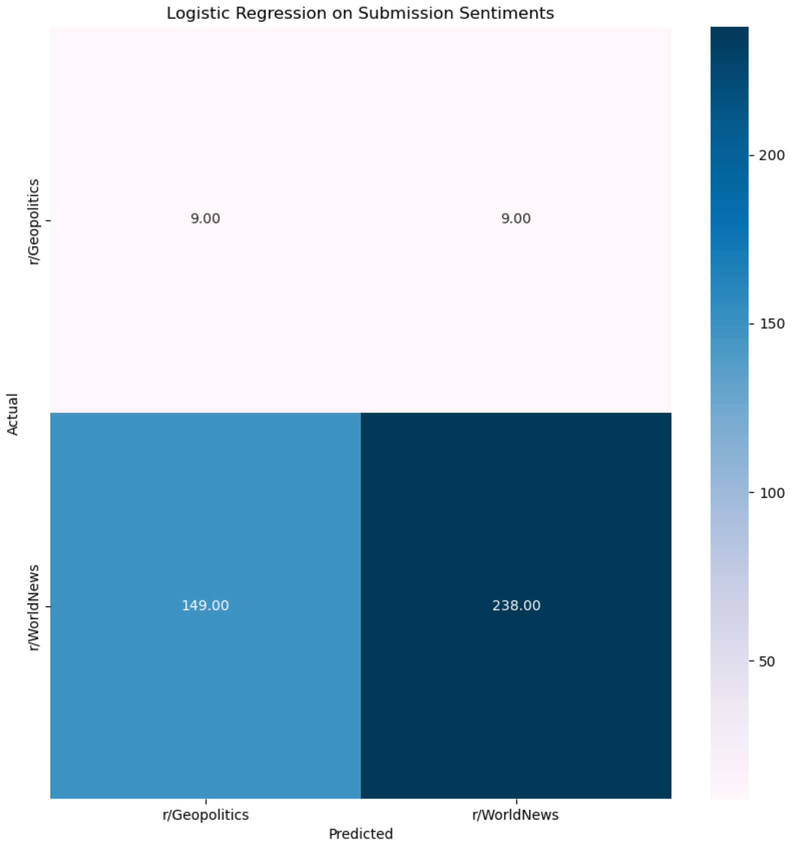
Russia/Ukraine Former Swedish intelligence officer jailed for life for spying for Russia

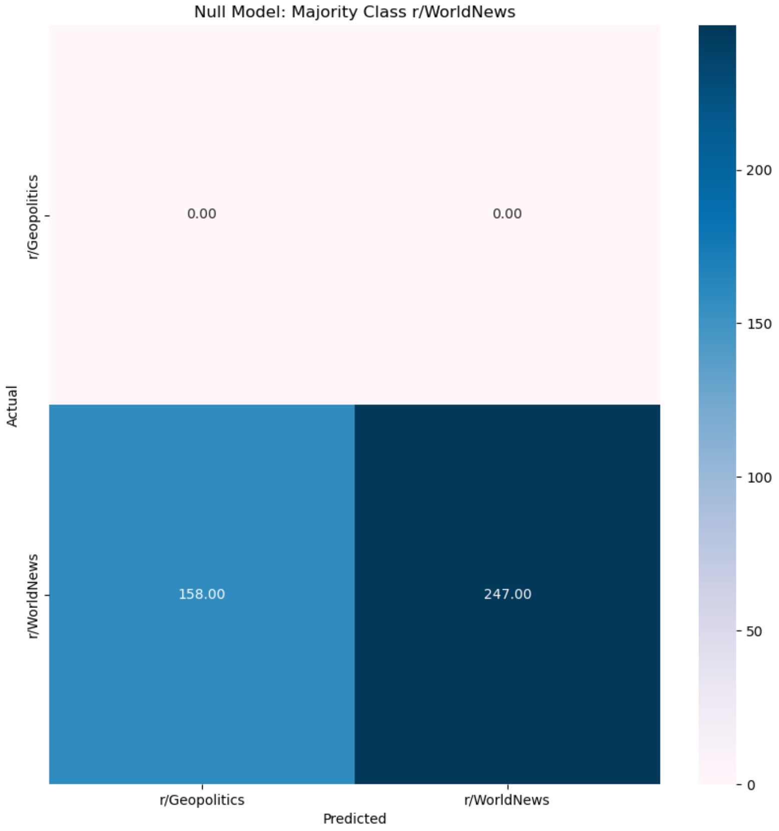
theguardian.com/world/...

56 Comments | Award | Share | Save | ···

Simple Sentiment Analysis

- Sentiment Scores calculated (Vader SentimentIntensity Analyzer)
- Logistic regression using Sentiment Intensity scores as only feature





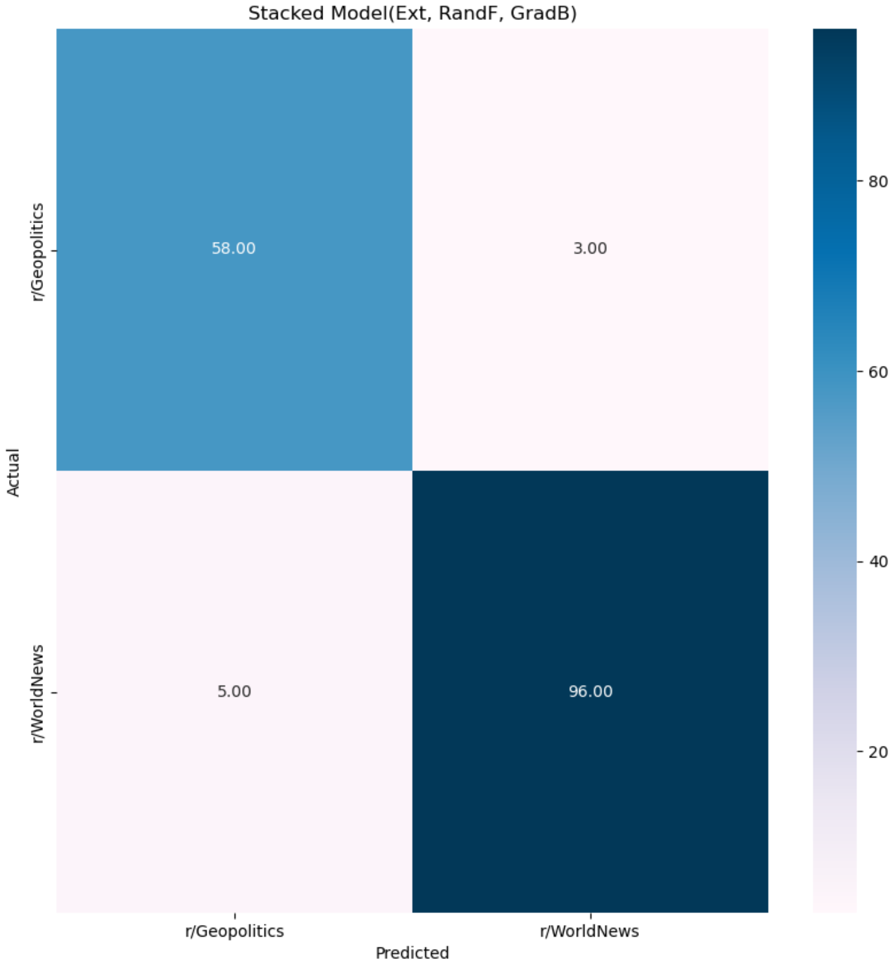
Submissions Data Models

Baseline : 0.611 : r/worldnews

Model	Training Score	Testing Score	Final Testing Score
Logistic Regression (CVEC)	0.993	0.797	0.759
Logistic Regression (TFIDF)	0.909	0.775	0.765
Random Forest (CVEC)	1.0	0.821	**0.796**
Random Forest (TFIDF)	1.0	0.813	0.778
Extra Trees (CVEC)	1.0	0.84	**0.783**
Extra Trees (TFIDF)	1.0	0.808	0.759
Gradient Boost (CVEC)	0.995	0.780	0.771
Gradient Boost (TFIDF)	0.994	0.783	**0.778**

Simple Stacking Model

Model	Training Score	Testing Score	Final Score
Random Forest (CVEC)	1.0	0.821	0.796
Extra Trees (CVEC)	1.0	0.84	0.783
Gradient Boost (TFIDF)	0.994	0.783	0.778
Stacked Model	**1.0**	**0.800**	**0.864**



A look at top unigrams and bigrams from false positives

- russia
- power
- caspian
- war
- counts
- atlas
- rape
- germany
- books
- poland
- police
- days
- end
- batteries
- nato
- articles
- biden

- end war
- deploy patriot
- officer admits
- collapsing war
- ukraine atlas
- war yemen
- war ukraine
- coming days
- met police
- books articles
- articles west
- atlas report
- biden promised
- csto crisis
- fares better
- promised end
- crisis Russia
- germany deploy
- patriot batteries
- doing opposite
- police deny

Recommendations

- Re-train models after adjusting for imbalanced classes

- Increase observations to make model more robust

- Further exploration of sentiment score variants

- Train and test model on comments data and aggregated comments-posts data