Using Natural Language Processing to Sort Reddit Posts Based on Text-Based Content

Source: https://www.reddit.com/, https://www.digitaltrends.com/web/what-is-reddit

Overview

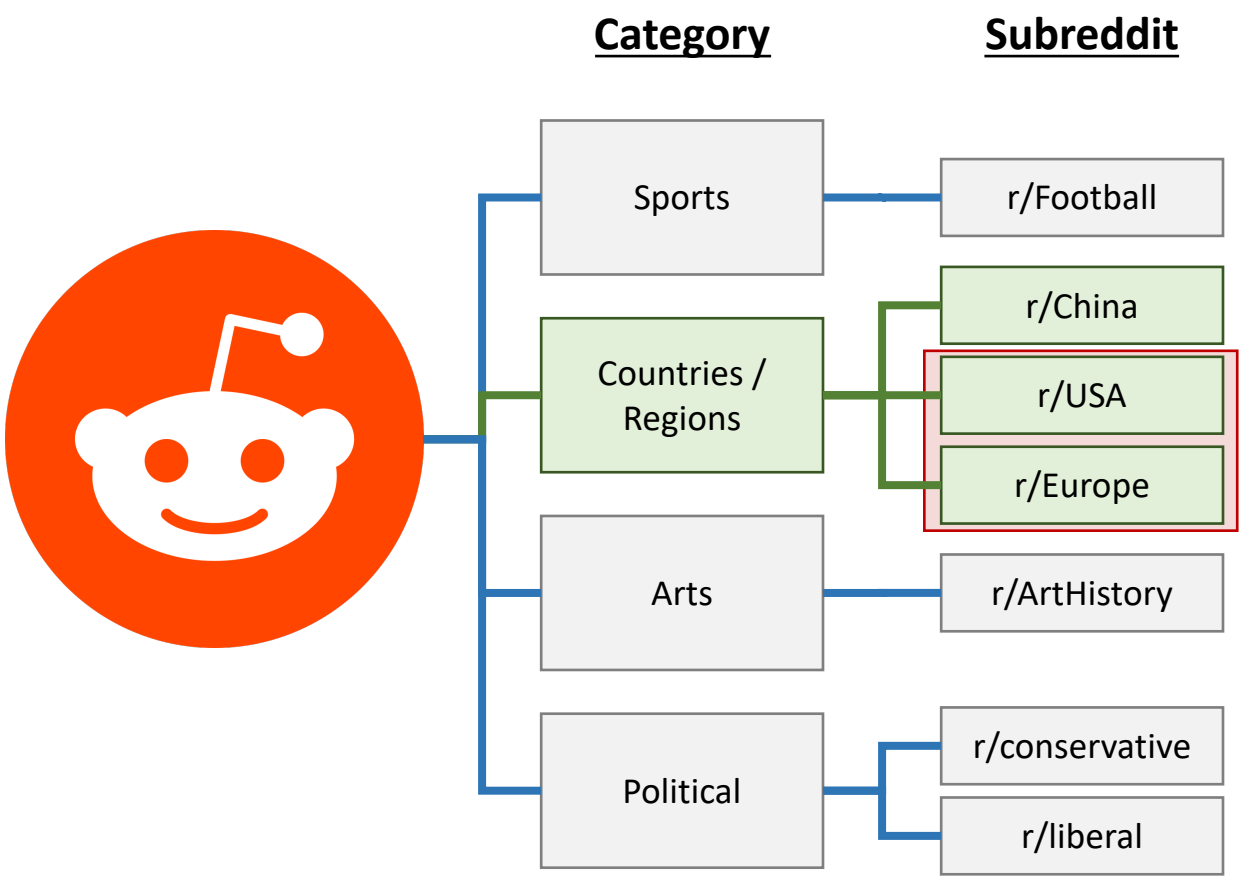
What is Reddit? What are Subreddits? How are they related to each other?

Reddit

Reddit is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members.

Subreddits

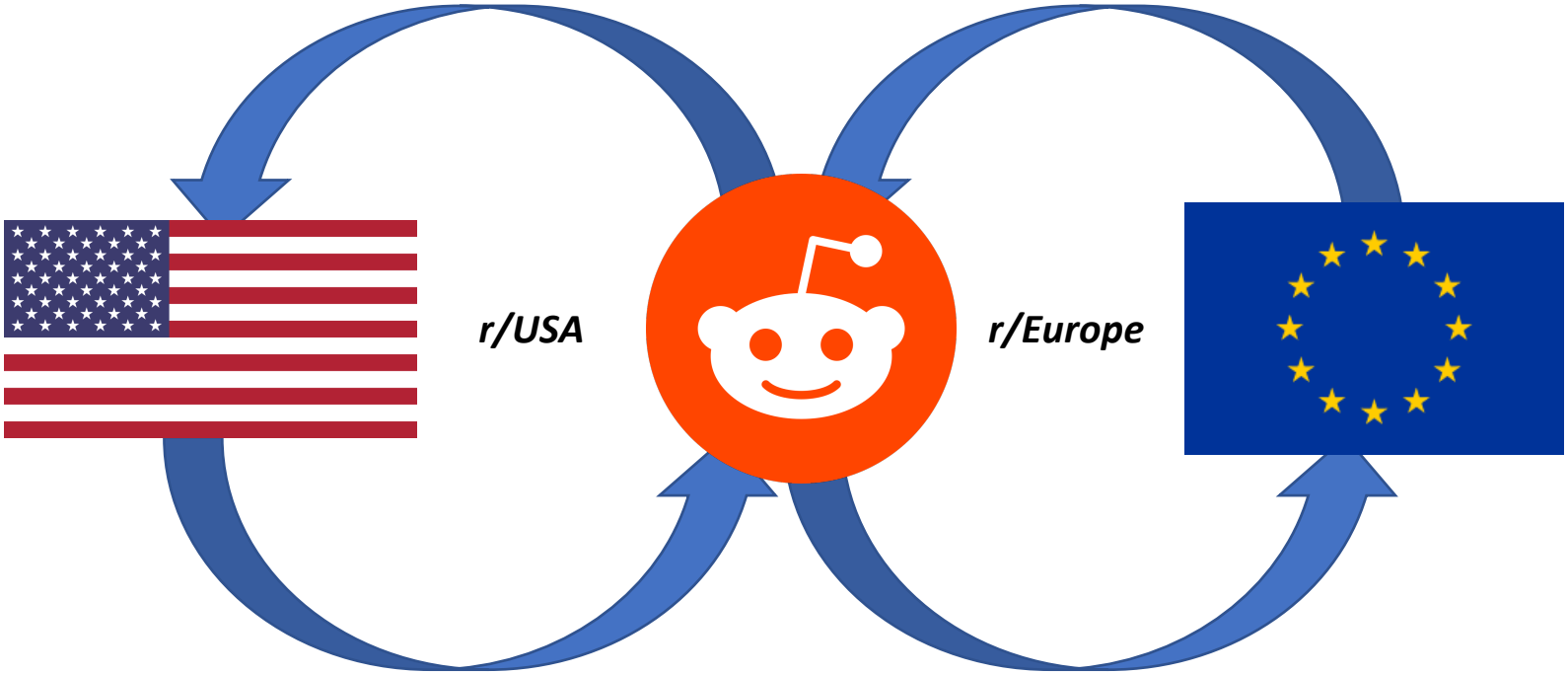
Reddit is broken into more than a million communities known as “subreddits”, each of which covers a different topic.



Source: https://www.reddit.com/, https://www.digitaltrends.com/web/what-is-reddit

Problem Statement

Establishing the Problem Statement and the two component subreddits.



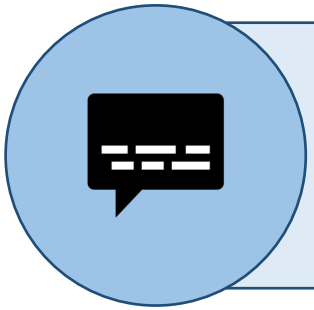
Problem Statement

Using Natural Language Processing, is it possible to create a model that can accurately sort Reddit posts based on their subreddit of origin?

Source: *https://www.reddit.com/*

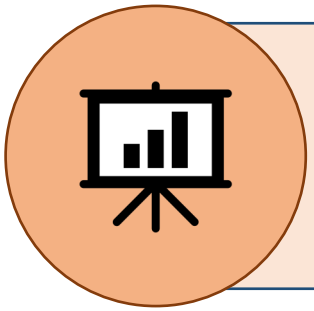
Web Scrapping

Reddit posts were initially gathered using the Reddit PRAW API.



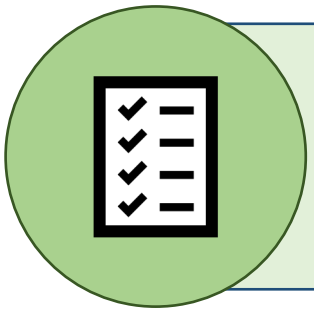
Web Scrapping

Web Scrapping was done using the Reddit PRAW API.



Metrics Pulled

- **ID**
- **Subreddit**
- **URL**
- **Title**
- **Downvotes**
- **Comments**
- **Upvotes**
- **Body (Text)**
- **Date Created**



Number of Posts

200 posts were pulled for each subreddit.

Source: *https://www.reddit.com/*

Model Interpretation - Correlations

Certain words have stronger correlations with different subreddits.

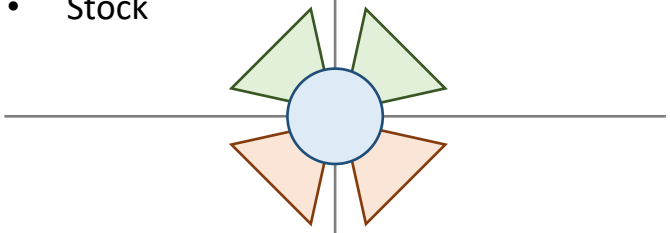
Correlations between words and subreddits

	Words	CVEC Coef	TVEC Coef
51	america	1.32	1.32
1729	us	1.25	1.38
1549	stock	1.22	1.11
1730	usa	1.14	1.08
1691	trump	1.12	1.06
...	...	...	...
660	germany	-0.83	-0.66
517	europe	-0.86	-0.60
1515	spain	-0.86	-0.74
1704	uk	-1.08	-1.05
514	eu	-1.61	-1.53

Certain words are more strongly correlated with the **USA** subreddit:

- America
- US
- Stock

The appearance of the word "Stock" is somewhat surprising. The other 4 words are largely expected.



Certain words are more strongly correlated with the **Europe** subreddit:

- Germany
- Europe
- Spain

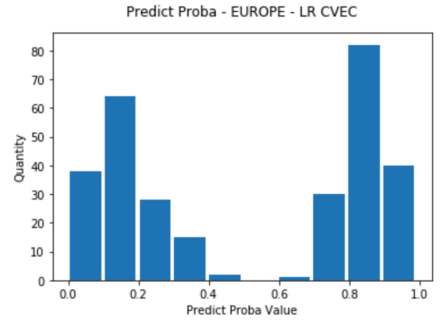
Germany, Europe, and Spain are all words you would expect to see in a Europe focused subreddit.

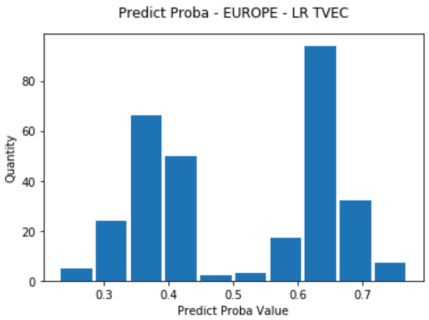
Source: *https://www.reddit.com/*

Model Interpretation - Probabilities

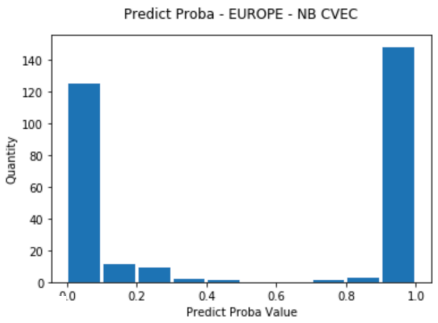
Models can more easily categorize certain posts.

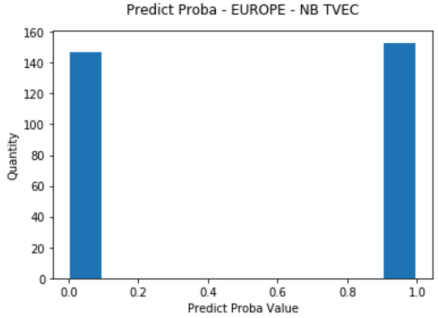
Logistic Regression Models





Naïve Bayes Models





- A very high or very low predict probability indicates the model has a high certainty of correctly sorting the posts.
- As can be seen form the above charts, each post has a different probability of being accurately sorted. The majority of posts have a very high likelihood of being accurately sorted.

Source: *https://www.reddit.com/*

Model Interpretation – Accuracy Scores

As can be seen below, the various models we worked with provided different degrees of accuracy.

Baseline Accuracy:

- The baseline accuracy is the simplest possible prediction.
- Recommends that we always select “USA” as the predicted subreddit.

	Baseline Accuracy
USA	0.53
EUROPE	0.47

Models:

- Several models were used to find the one that would most accurately sort the Reddit posts.
- It was determined that the TVEC models were in general more successful than the CVECs. However, the increased accuracy is marginal.

	Training Data	Testing Data	Delta
TVEC Logistic Regression	1.0	0.85	0.15
TVEC Logistic Regression Pipeline	1.0	0.85	0.15
TVEC Naive Bayes	1.0	0.84	0.16
CVEC Logistic Regression	1.0	0.83	0.17
CVEC Logistic Regression Pipeline	1.0	0.83	0.17
CVEC Naive Bayes	1.0	0.82	0.18
CVEC Naive Bayes Pipeline	1.0	0.82	0.18

Source: *https://www.reddit.com/*