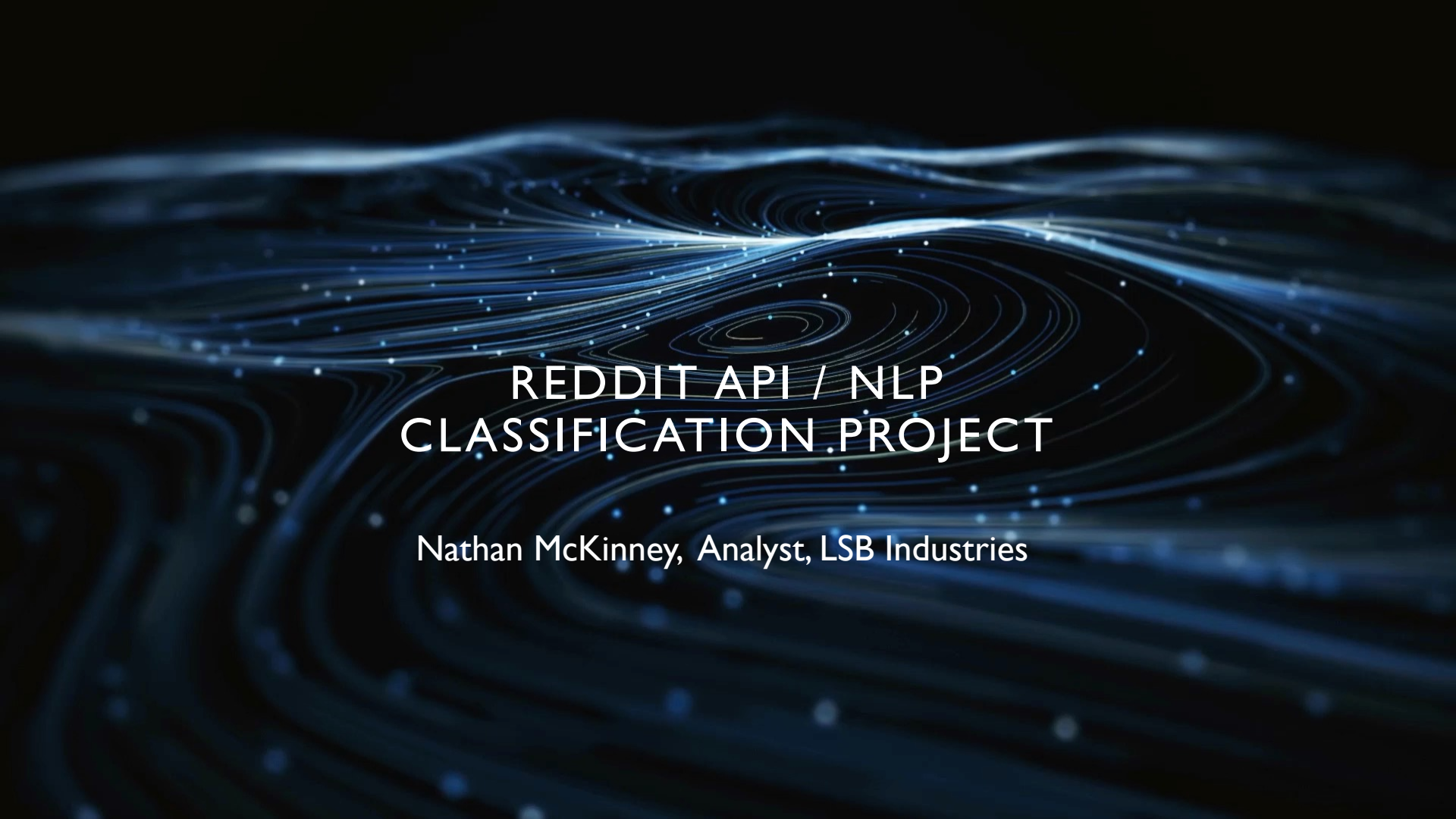


REDDIT API / NLP CLASSIFICATION PROJECT

Nathan McKinney, Analyst, LSB Industries

INTRODUCTION

The importance of digital media analytics and its presence here at LSB?

Are we able to use predictive modeling on Reddit data to better predict which post belongs to what subreddit?

NLP USE CASES

- Risk-Sensing
- Sentiment Analysis
- Competitor Movements

SUBREDDITS USED

r/Economics

- Global Economic News and Discussions
- 1.6 million members

r/Finance

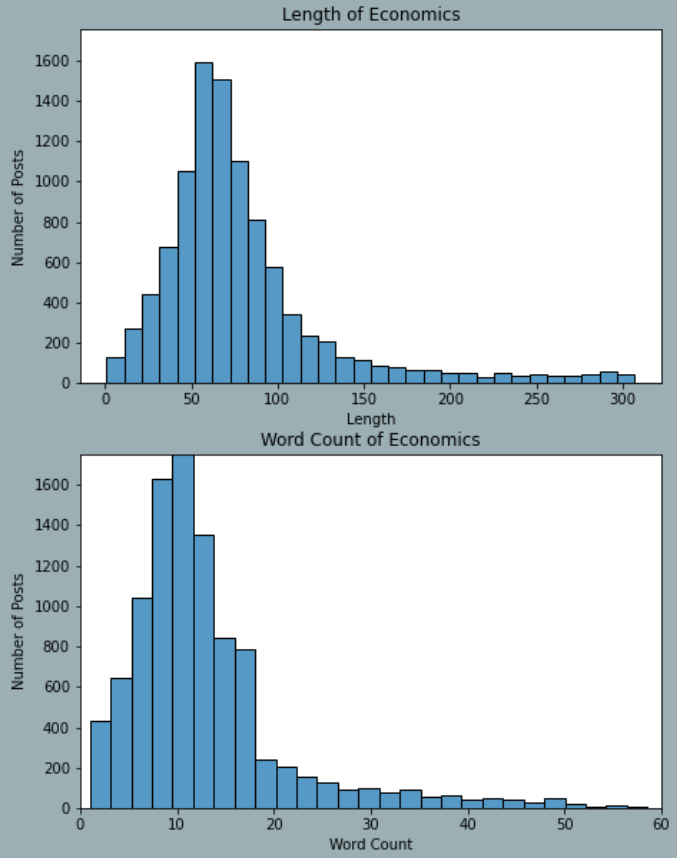
- Global Financial News and Views
- 509,000 members

DATA EXTRACTION PROCESS

- Pushshift API via Python
  - 100 posts at a time, consecutive iterations
- Process is reproducible for classification of the posts of two subreddits

EXPLORATORY ANALYSIS

- Significant need to standardize text data
  - ❖ Remove hyperlinks and file names, white space, common words (and, but, etc.)
  - ❖ Remove video submissions
- Post count is a 1:1 ratio, with different start dates
  - ❖ Null Hypothesis = 50% Accuracy



MODELING

- Ran various classification models
  - Text transformed to numerical matrix via Vectorizers
  - Performed GridSearch on Multinominal, Random Forest, and K Neighbors
  - Generally overfit, while accuracy was consistent between most models

Model	Accuracy	Train Score	Test Score
LogR (Vect)		97%	76%
LogR (Tfid)		92%	77%
MNB (Tfid)	76.6%		
MNB (Tfid) GRD	73.6%	74%	73%
RF (Tfid)	75.0%		
RF (Tfid) GRD	75.1%	98%	75%
K Neighbors (Vect)	74.0%	82%	72%
K Neighbors GRD	74.0%	74%	74%

CHOSEN MODEL

- K-neighbors with Grid Search

  - Params:
    Nearest neighbor = 71
    Metric = 'Euclidean'
    Cross validation folds = 10

  - Best Fit

  - Reasonable degree of accuracy in predicting true values

	MNB (Tfid) GRD	RF (Tfid) GRD	K Neighbors GRD
True Positive	1971	2155	2020
True Negative	2448	2360	2435
False Positive	553	641	566
False Negative	1029	845	980
Accuracy	73.6%	75.2%	74.2%
Recall	65.7%	71.8%	67.3%
Precision	78.1%	77.1%	78.1%

	Training Score	Test Score
K Neighbors	74%	74%
RF	98%	75%

CONCLUSIONS

- We are able to make accurate predictions above our null hypothesis (50%)

- Our chosen model, K-Neighbors with a grid search, is 74% accurate in predicting which subreddit a post is from.

- We accept this iteration of the model, but recognize that there are possibly improvements to overfitting models (Random Forest) that could change our chosen model

FUTURE IMPROVEMENTS

- Add reddit comments from the respective subreddits to the model flow, so that we can evaluate both at once

- Code that allows one to search for all mentions of a specified word and pull out the most relevant posts and comments

- Reproducible sentiment analysis given a beginning date. (Capturing and evaluating most recent posts)