|  |
| --- |
| Classifying Reddit Subreddits |
|  |

|  |  |
| --- | --- |
| 05/04/2018 | r/politics r/Liberal. r/Conservative |

|  |
| --- |
| Classify posts on the r/politics subreddit according to political idealogy |

Classifying Reddit Subreddits

r/politics r/Liberal. r/Conservative

# Introduction

Reddit is the fourth most visited site in the United States in 2018, as reported by Alexa, a consultancy. It is the seventh most visited site in the entire world. Reddit is a community of like-minded individuals with a wide variety of interests who congregate via sub thread of their interest.

# Problem Definition

The primary goal was predicted on the hypothesis that the submissions on r/Politics subreddit can be bifurcated to liberal and conservative adjacent posts with a nominal error term to capture the submission which fall outside this binary paradigm. An algorithm trained to accurately classify the subreddit origin of a post from a combined dataset from r/Liberal and r/Conservative can be reasonably expected to classify the submission on the r/politics subreddit as effectively.

The secondary goal was to understand the factors which contribute to the success of posts in each of these three subreddits.

# DATASET

The dataset was scrapped from the three subreddits over a one week period from both the “Hot” and “Popular” sub categories as defined by Reddit. The r/Liberal contained 1658 unique posts, r/Conservative contained 2184 and r/politics had 3382.

The feature engineered dataset contained 11 numerical and 9 categorical features.

# MethoDology

First, the title of each post was vectorized in preparation for machine learning use and sentiment analysis was conducted to create another variable. Then the best performing algorithm per dataset was chosen among many variants using the test validation score.

r/politics: RandomForestClassifier &LogisticRegression (ensemble) 0.80

r/Conservative GradientBoostingClassifier: 0.72

r/Liberal BernoulliNBinomial 0.63

# Findings

r/politics submissions cannot be effectively classified using the common attributes on r/Liberal and r/Conservative threads. This may be due to the heterodoxy of the prediction data caused by the presence of a wider spectrum of political opinion than the project definition of Liberal and Conservative thought.

Each sub-thread contains enough specific and uniquely intrinsic attributes to successfully predict the success of a particular post, where the success is defined as attracting more than the median amount of comments or scores.