Road to Revolution: Socialism vs Communism

Subreddit Classification via Natural Language Processing

Ashley White | December 21, 2018 | DSI #6

Objective & Methodology

Objective: Use NLP and classification algorithms to distinguish between two similar subreddit posts: Communism v. Socialism

Methodology:

1. Query PushShift API to retrieve submissions
2. Clean & pre-process text
3. Vectorize / tokenized text
4. Gridsearch to optimize hyperparameters across two classification algorithms

Baseline Statistics

Socialism

From each according to his ability, to each according to his contribution.

42%

posts

Communism

From each according to his ability, to each according to his needs.

58%

posts

Text Pre-Processing

1. Removed extraneous tags: ‘removed’ , moderator posts, hyperlinks, non-letter characters; dropped rows < 10 words

2. Lemmatized text to reduce duplicates and better compare similarities

3. Vectorized data using TF-IDF:
   - 75K vectors
   - Removed / edited stopwords
   - N-Gram Range: (1, 2)

Communism

Frequent	*Rare*
Propaganda machine	Communist history
Provisional government	Accurate
Proletarian revolution	Good

Socialism

Frequent	*Rare*
Proxy war	Modern politics
Provide resource	Propaganda
Profit state	Like CNN

Combined

Frequent	*Rare*
Provision covering	Revisionism
Provisional government	Eastern bloc
Provocateur FBI	Nationalism

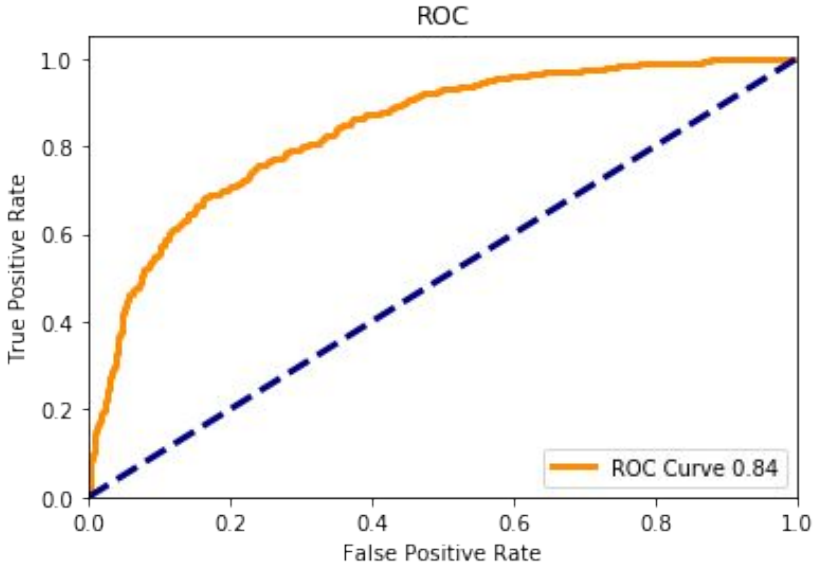
Model Selection & Scoring

Logistic Regression

Ridge Regularization

75% Training Accuracy

72% Test accuracy



Random Forest

200+ N-Estimators
Entropy - Information Gain

73% Training Accuracy

68% Test accuracy