Study:
Reddit NLP Classification

Daniel Schlant, Data Scientist
NLP Consultants

Our Study

Data

- Post titles, scraped using Reddit Pushshift API
- Posts that included self-text
- Subreddit inception to present

Model Goals

- Classify Reddit submissions
- Learn about user's outlook via their created text

The Reddit Threads

Discussion regarding the potential collapse of global civilization, defined as a significant decrease in human population and/or political/economic/social complexity over a considerable area, for an extended time. We seek to deepen our understanding of collapse while providing mutual support, not to document every detail of our demise.

Welcome to r/Futurology, a subreddit devoted to the field of Future(s) Studies and speculation about the development of humanity, technology, and civilization.

Modeling Process

Baseline Accuracy: 50%

Removed holdout set (20% of overall)

Ran modeling analysis on 30% of remaining data (24% of overall)

Final Model: TF-IDF Text Vectorization/Multinomial Naive Bayes

- Expanded Custom Stopwords List
- Lemmatization
- Bigrams
- 64,000 Features

Performance with Some Custom Stopwords, for Comparison (Smaller Dataset)	
Model	**Validation Set Accuracy (%)**
Multinomial Naive Bayes	83.19
Gaussian Bayes	69.00
Logistic Regression	79.57
KNN	64.23
Decision Tree	58.40
Random Forest	77.55
AdaBoost	72.27
Ensemble (LR,NB,RFC)	80.39

Model Results

82.8% accuracy rate on validation set

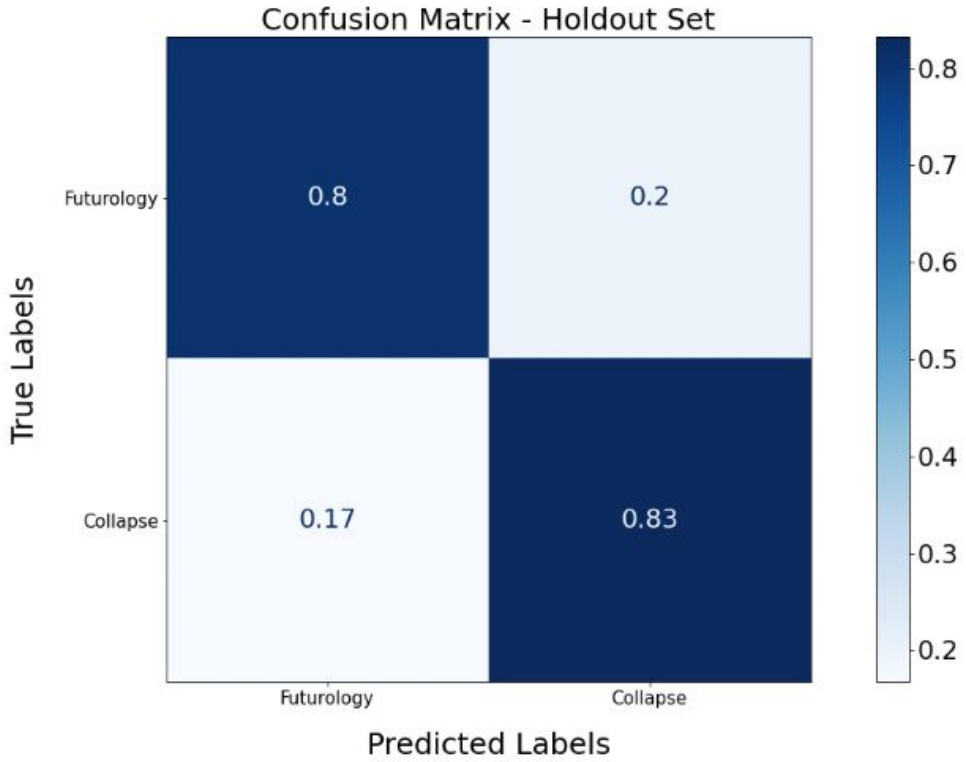
- 80% of overall data

82% accurate on holdout set

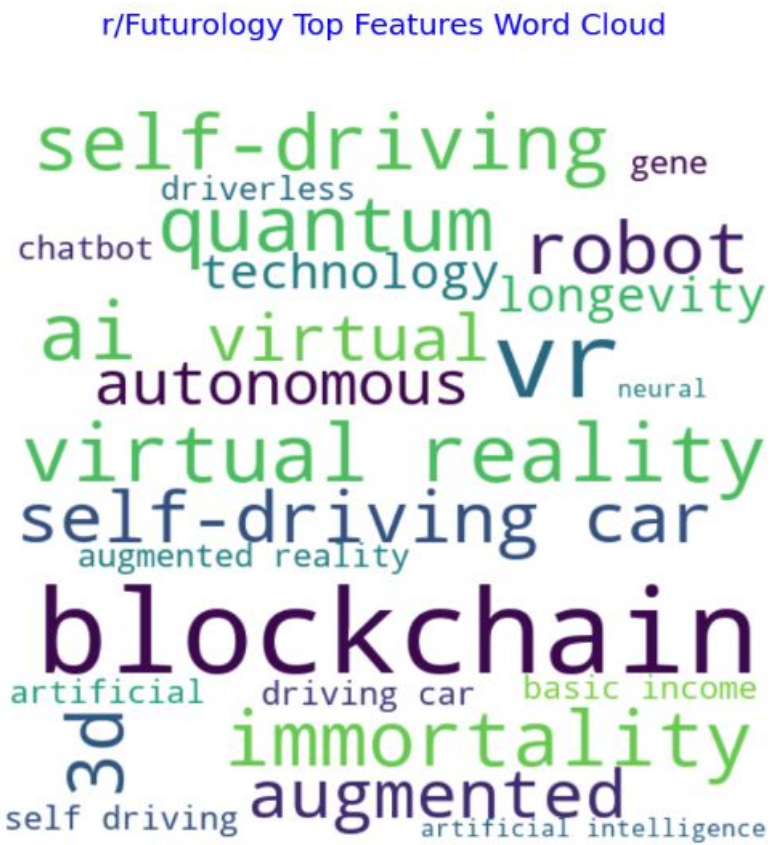
- 20% of overall data
- 50.3% baseline accuracy

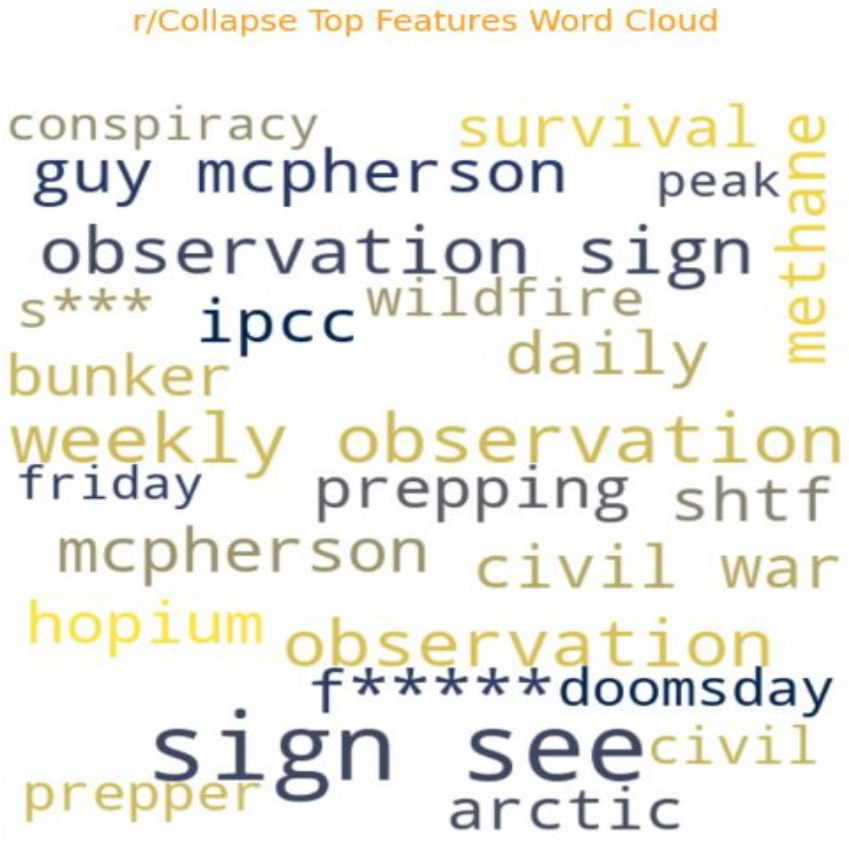
Holdout: 80% accurate on Futurology posts, 83% on Collapse posts

- Climate change posts



Key Features, by Model Importance





Misclassifications

Post from r/collapse

Autonomous vehicles, how can they not be classified as possibly-autonomous weaponry by the UN

Model estimate: 91% probability Futurology

Post from r/Futurology

Apocalyptic times due to climate change - haven't we had worse? Where's the hope?

Model estimate: 97% probability collapse

Conclusion & Further Research and Study

Model successfully classified posts at 82% rate. Further research needs to be done to assess whether this model can be applied to other forums, platforms to predict a user's worldview/future outlook.

Include Self-Text Data

- Self-text had been removed for over 30% of total submissions
- Using similar model, with slight tuning, improved accuracy rate to 85% on holdout set when available self-text added

Sentiment Analysis

Expand Research to Other Subreddits

- Further develop NLP model on other self-identifying communities

Thank you for your time,
any questions?