Which Science?

Classifying internet messages using natural language processing

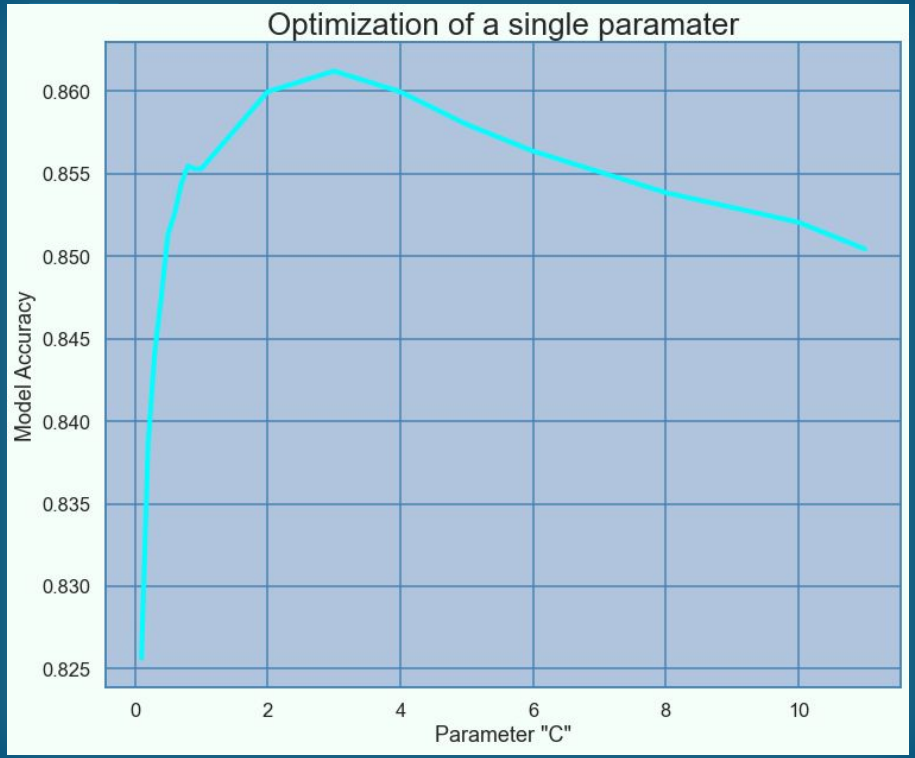
Goals

- Gather data from reddit.com ‘subreddits’: Biology, Chemistry, and Physics
- Train classification models to predict which subreddit a post belongs to.
- Optimize models to find the best parameters.
- Improve predictions by using an ensemble of models to ‘vote’ and choose the best prediction for each post.
- Consider the unique challenges of this classification problem.

Data Collection

- Webscraping with Pushshift API
- 10,000 posts from each subreddit
- Drop posts that were removed
  - Spam
  - Irrelevant
  - Rule-breaking
  - Duplicates
- Combine data and split into training/testing sets

Modeling and Optimization



Natural Language Processing

- Vectorizer: *TfidfVectorizer*
- Max Features, ngram range

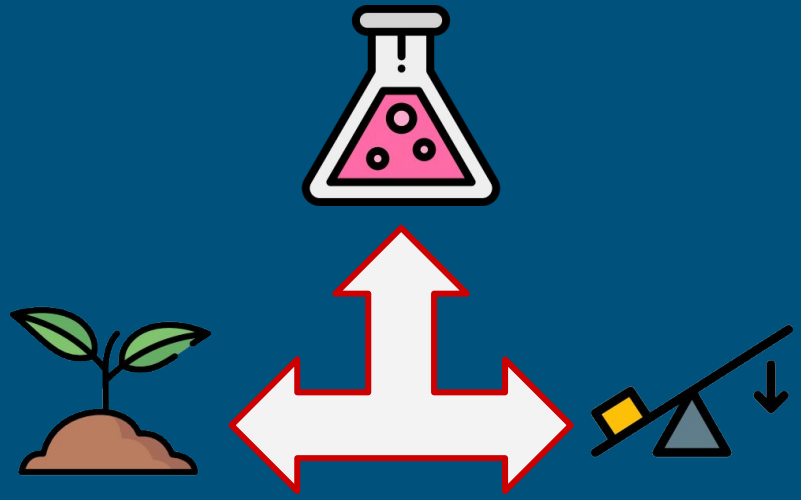
Logistic Regression

- Regularization Alpha (1/C)
- Regularization Type

Initial Results: 86% Accuracy

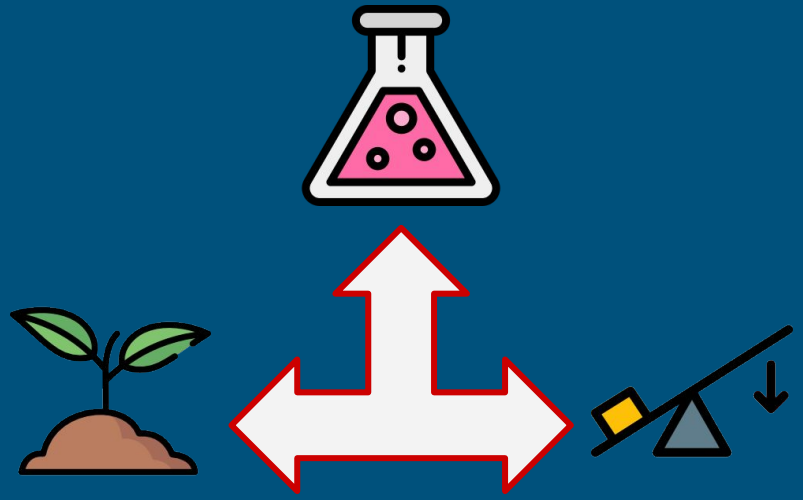
Models Considered

- Logistic Regression 86.3%
- Bagging Classifier(LR) 85.9%
- Naive Bayes 86.5%
- Linear SVC 86.9%
- Decision Tree 78.5%
- AdaBoost Classifier 80.8%



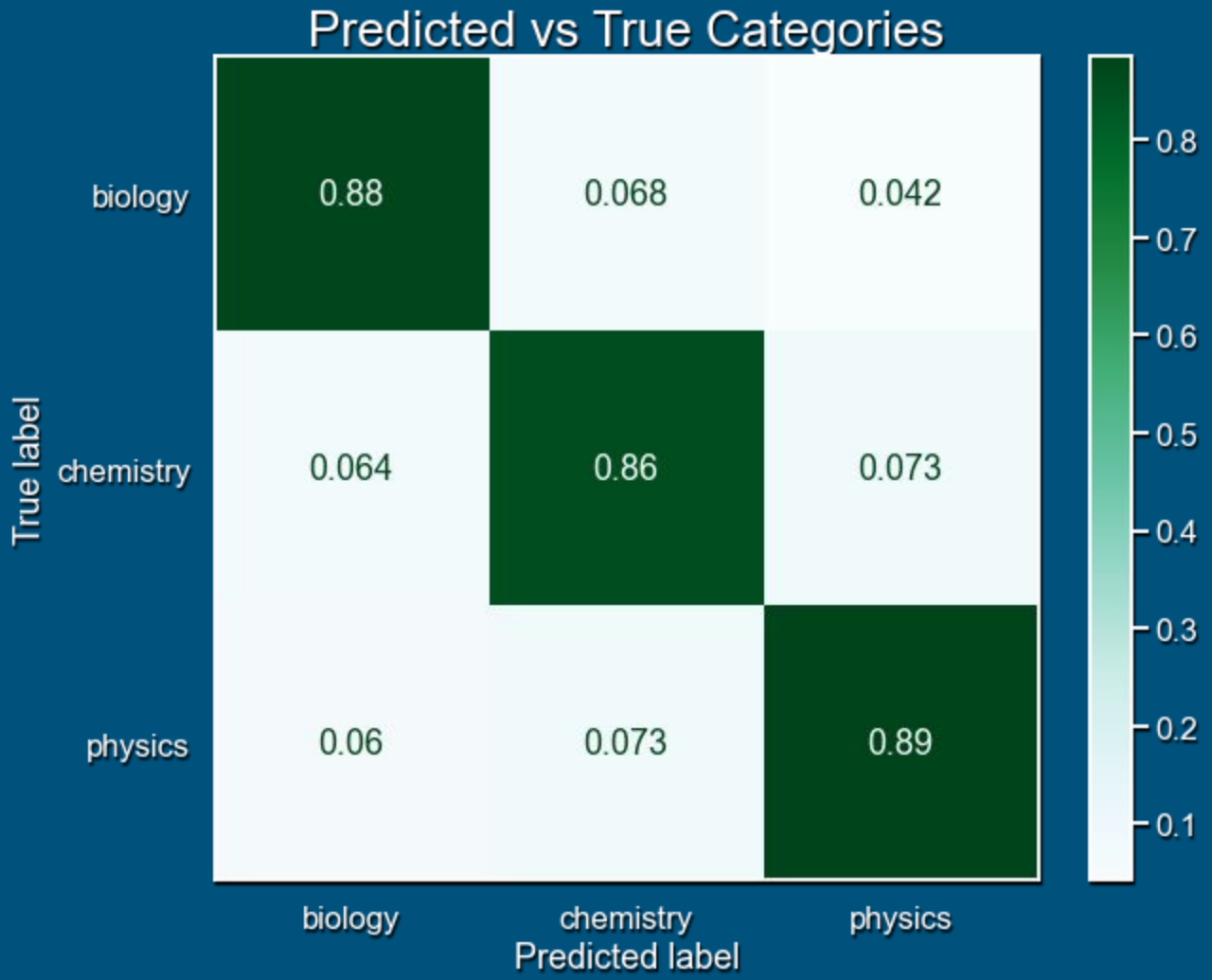
Models Considered

~~Logistic Regression~~	~~86.3%~~
~~Bagging Classifier(LR)~~	~~85.9%~~
Naive Bayes	86.5%
Linear SVC	86.9%
~~Decision Tree~~	~~78.5%~~
AdaBoost Classifier	80.8%



87.3%

Accuracy of ensemble vote predictions


Predicted vs True Categories
biology
chemistry
physics
True label
0.88
0.068
0.042
0.064
0.86
0.073
0.06
0.073
0.89
biology
chemistry
physics
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Classification Difficulties:

Spam, and community discussion posts

PREDICTION:

ACTUAL:

“Software Apps with Professionally Designed Sales Pages and Start Making Sales!”

PREDICTION:

ACTUAL:

“Scientific Research Survey”

“Hi there, i’m creating a business plan for a Science research based cloud service. I’m surveying those in the science community . . .”

Classification Difficulties:

Overlap in fields of science

PREDICTION:

ACTUAL:

“What’s the difference between the nitroglycerin in dynamite and medicine?”

PREDICTION:

ACTUAL:

“Can magnetic waves sanitize water from parasites?”

Classification Difficulties:

Room for Improvement

PREDICTION:

ACTUAL:

"Bioactive peptides and carbohydrates from seaweed for food applications: Natural occurrence, isolation, purification, and identification"

PREDICTION:

ACTUAL:

"Just wondering"

"How is potential difference across a fully charged capacitor related to its capacitance?"

Classification Difficulties:

Context!

PREDICTION: 

ACTUAL: 

“Astrophysicist gets magnets stuck up nose while inventing coronavirus device”

Conclusions

- Most models performed best with only 1-word features and no “stop words”.
  - The exception is Decision-Tree models, which work well with 2- and 3-word features.
  - Normal English stop words did not work well on scientific language.
  - It may be useful to create a stop-words list specifically for this kind of data.
- Spam and off-topic discussions are a problem with this data source
  - When the text has no relevance to any field of science, classification is difficult, and not particularly meaningful.

Thank You!

Resources:

- *Pushshift API:* https://pushshift.io/
- *Reddit:* https://www.reddit.com/
- *Icons:*  http://www.freepik.com/
  https://www.flaticon.com/authors/freepik
  https://www.flaticon.com/authors/pixel-perfect