Natural Language Processing (NLP)

Colby Tse | Data Scientist

Now with 100% more Reddit

Let me get something off my chest...

Utilizing NLP classification models can I predict(or suggest) whether or not a post belongs in one subreddit over another?

- Subreddits: r/AmItheAsshole and r/TrueOffMyChest

Approach

A lot of iterations of TF-IDF and LogReg

```
Pipeline
  CountVectorizer
  LogisticRegression
```

```
Pipeline
  TfidfVectorizer
  LogisticRegression
```

```
GridSearchCV
GridSearchCV(cv=5, estimator=RandomForestClassifier(random_state=42), n_jobs=-1,
             param_grid={'max_depth': [3, 5, 10],
                         'min_samples_leaf': [3, 5, 10],
                         'min_samples_split': [5, 10, 15, 20],
                         'n_estimators': [100, 200, 300]})
  estimator: RandomForestClassifier
    RandomForestClassifier
    RandomForestClassifier(random_state=42)
```

```
GridSearchCV
GridSearchCV(cv=5, estimator=ExtraTreesClassifier(random_state=42), n_jobs=-1,
             param_grid={'max_depth': [3, 5, 10],
                         'min_samples_leaf': [3, 5, 10],
                         'min_samples_split': [5, 10, 15, 20],
                         'n_estimators': [100, 200, 300]})
  estimator: ExtraTreesClassifier
    ExtraTreesClassifier
    ExtraTreesClassifier(random_state=42)
```

How’d it go?

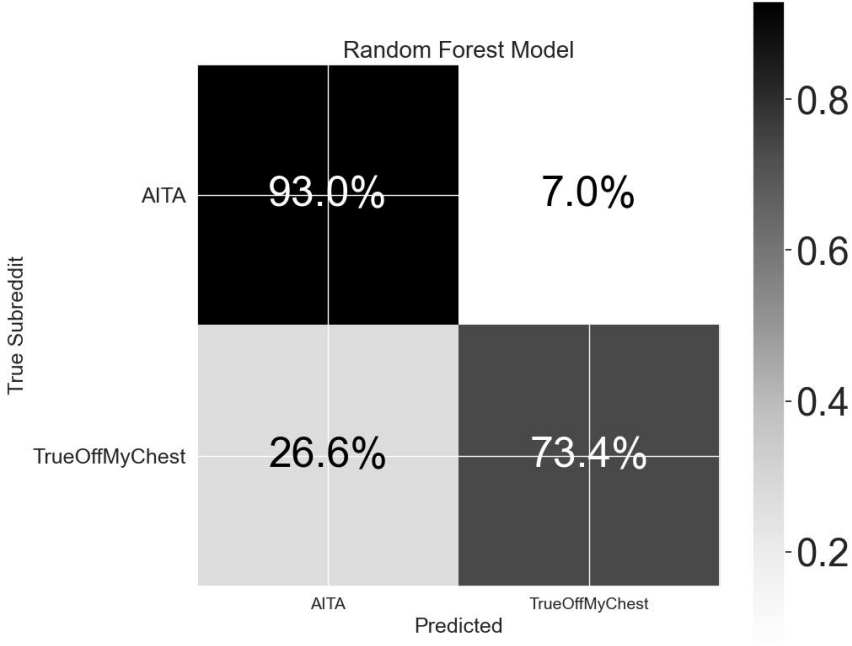
TF-IDF Vectorization and Logistic Regression performed the best but is slightly more overfit than Random Forest.

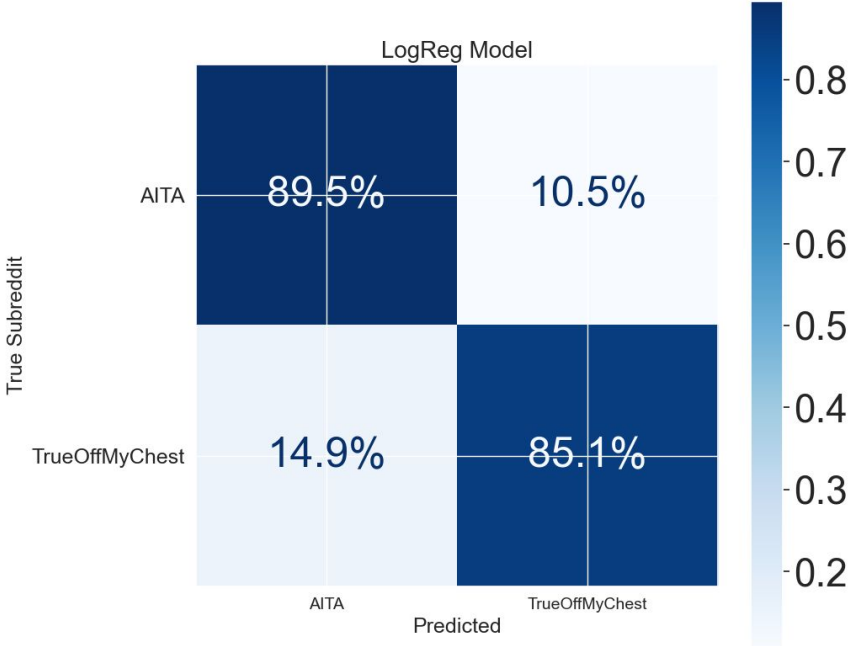
Model	Train	Test
CV/LogReg	99.8%	87.9%
TFIDF/LogReg	**91.7%**	**87.4%**
Random Forest	86.4%	83.6%
Extra Trees	70.4%	68.6%
Lemmatization	97.6	86.7

One might argue Random Forest is better but...

How'd it go?

Random Forest has high bias to predicting on AITA.





Am I the A-hole?

"It was I who languised for years thinking of nothing but you, but for this moment, and now the perfect tool for my vengeance is in front of us. I never planned on killing you but I will make you share my pain."

"Justice is merely the construct of the current power base. A base which, according to my calculations, is about to change!"

Am I the A-hole?

Quote 1: True Off My Chest

Quote 2: Am I the Asshole

Both quotes from Darth Maul

Next Steps

Optimize some code to perform a more thorough parameter optimization.

Bring in additional data from other subreddits and other forums. Possibly non-story driven posts.

Sentiment analysis

Any Spaceballs fans?

Utilize additional Reddit functionality like tags (NTA vs TA)