Showerthoughts

Vs.

StonerPhilosophy

Overview

Data and Custom Features

Grid Searching

- Parameters
- Best Model

Modeling with Context

- Confusion Matrices
- ROC Curves
- Examples

Next Steps

- Using custom features
- More models
- More data
- Include Selftext

Custom Features:

- Word_count
  - Showerthought mean ~15
  - Stonerphilosophy mean ~12
- Sentiment
  - not_lofty_words = ['real', 'prove', 'fact', 'reality', 'conscious', 'mortal']
  - lofty_words = ['universe', 'world', 'exist', 'theory', 'hypothetical', 'existence', 'unseen', 'immortal', 'god']
  - Showerthought mean ~0.03
  - Stonerphilosophy mean ~0.06

Lowest sentiment examples:

Are thoughts real?

Fog is just real life render distance

Highest sentiment examples:

Do thoughts exist?

A theory on existence, no drugs required

*caveat-it's hard to perform sentiment analysis with such short documents

Brief GridSearching

```
{'cvec__lowercase': False,
 'cvec__max_df': 0.9,
 'cvec__max_features': 3000,
 'cvec__min_df': 2,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': None,
 'cvec__tokenizer': <__main__.LemmaTokenizer at 0x1a1e33aa10>}
```

SVM accuracy score: ~ 0.698

MNB accuracy score: ~ 0.688

LR accuracy score: ~ 0.685

Modeling with Context

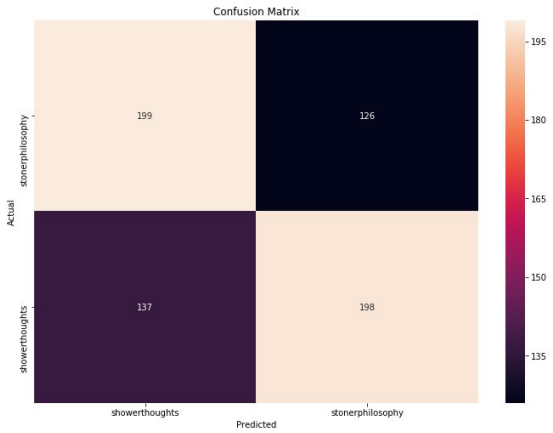
Stop Words: modified(‘english’) + ['stoner', 'weed', 'marijuana', 'high', 'baked']

Tokenizer: w/Regular Expression--kept punctuation

Stemmed instead of Lemmatized

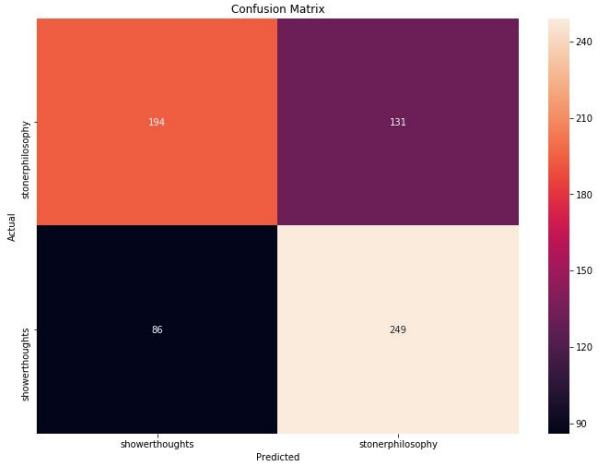
Confusion Matrices

Multinomial Naive Bayes



Sensitivity Rate: 0.38769230769230767
Specificity Rate: 0.408955223880597

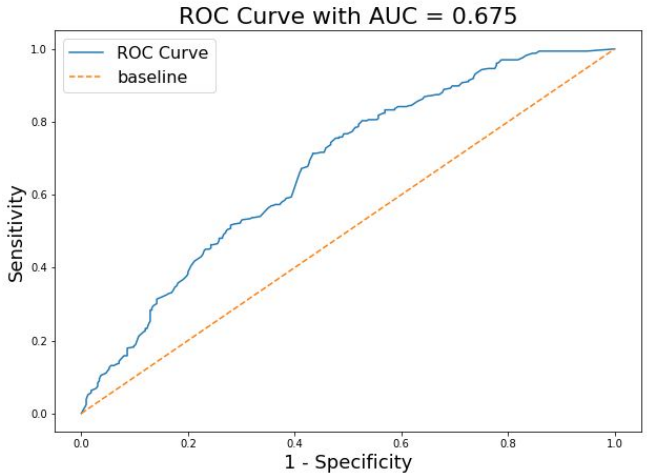
Logistic Regression



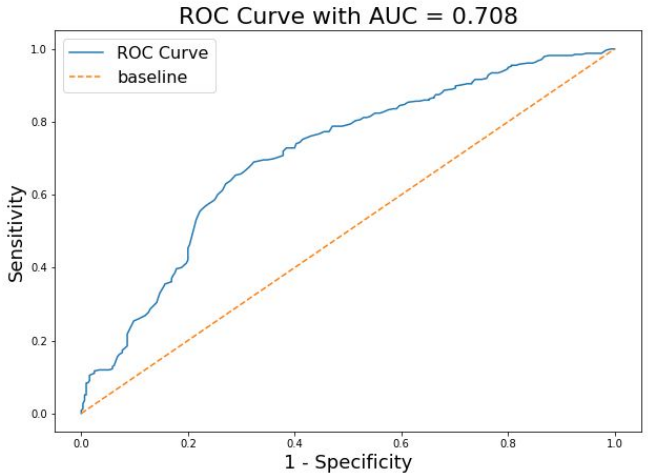
Sensitivity Rate: 0.40307692307692305
Specificity Rate: 0.25671641791044775

Receiver Operating Curve

Multinomial Naive Bayes



Logistic Regression



Examples

Both Models Correct

- **Showerthoughts**
  - When you say to somebody "Don\'t tell me what to do", you are telling them what to do.
  - Homer Simpson has had around 200 jobs and his kids have so many problems at age 10 and 8
- **StonerPhilosophy**
  - if string theory one day proves that the universe is made out of tiny loops of stringy stuff, do you think that means our entire universe might just be a big tapestry in some gigantic alien's spacemansion?
  - A yodel is just a well-placed voice crack

Both Models Incorrect

- **Showerthoughts**
  - If tomatoes are a fruit... Isn't ketchup a smoothie?
  - Peanut butter doesn't get enough credit for how creamy it really is for being made out of legumes.
- **StonerPhilosophy**
  - We’re all up in arms about global warming right now, when we should have caught on a lot sooner since Smash Mouth has been warning us since 1999.
  - Airlines could save a lot of money if they could find passengers who are attracted to each other.

Examples

MNB Correct

- **How many pictures am I in the background of a stranger's picture?**
- **Some people are scared of automatic subway, but they are just an horizontal elevator.**

***both titles came from showerthoughts**

LR Correct

- **Alice in Wonderland was just a story about Alice’s crazy edibles trip**
- **The following sentence is false. The preceding sentence is true.**

***both titles came from stonerphilosophy**

Next steps

Feature Engineering

Use wordcount and sentiment in the model and see if it performs better

More Models

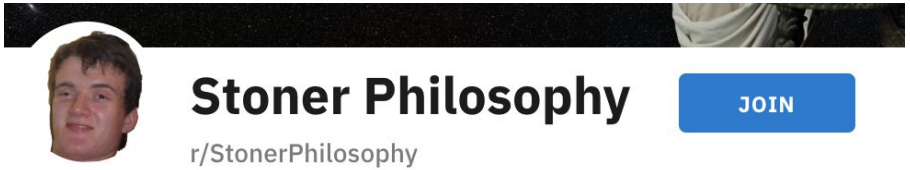
- SVM's
- Random Forests
- Bootstrapping

More Data

Include Selftext

Thank you!

?



https://stackoverflow.com/questions/47423854/sklearn-adding-lemmatizer-to-countvectorizer