Tips for Hobbyist Detectives

Distinguishing posts from “web sleuth” reddit communities

Welcome Sleuths and Amateur Detectives!

Outline

- The Problem
- Data Capture
- EDA
- Models
- Conclusions and Recommendations

The Problem

Distinguishing posts from the reddit communities ‘UnsolvedMysteries’ and ‘UnresolvedMysteries’

You have a draft post for reddit! But where should you post it?

Data Capture

Let’s get technical!

- Pushshift API
- Retrieve posts from r/UnsolvedMysteries and r/UnresolvedMysteries
- Author, Awarders, Created UTC, Self Text, Subreddit (TARGET), Title
- API implementation function
- ~ 1000 valid posts per subreddit
- Save to file

EDA

What does our data look like?

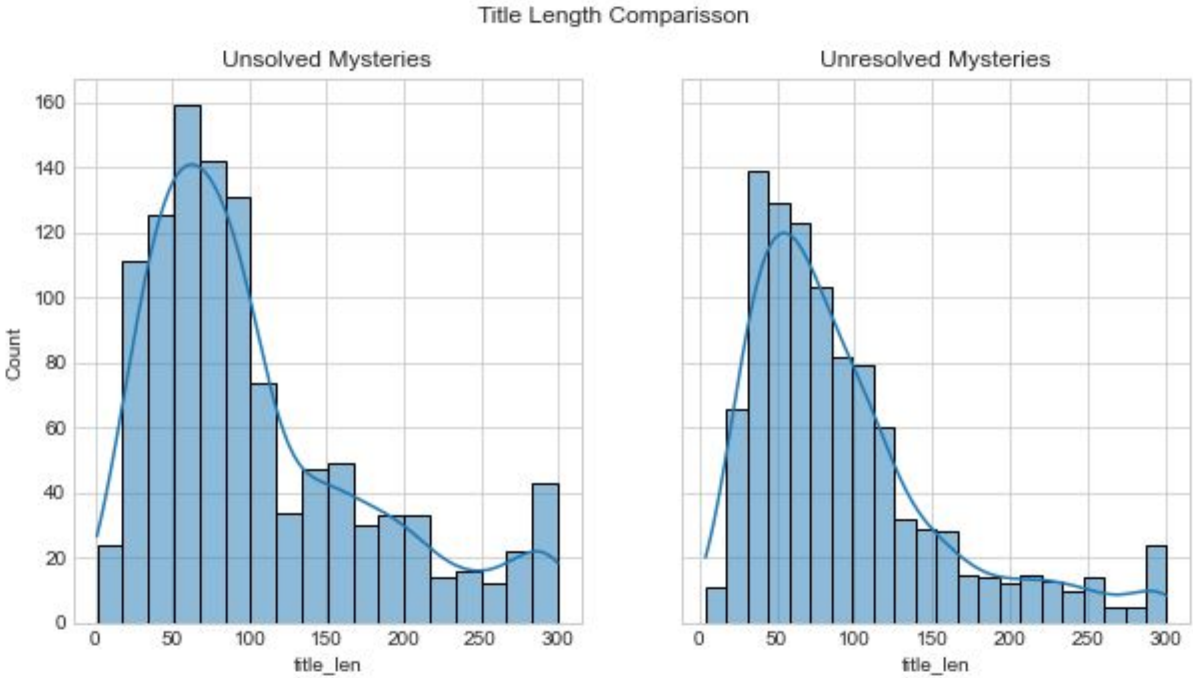
- Selftext and awarders are empty :(
- Concat DataFrames
- Binarize our TARGET: UnresolvedMysteries == 1, UnsolvedMysteries == 0

EDA

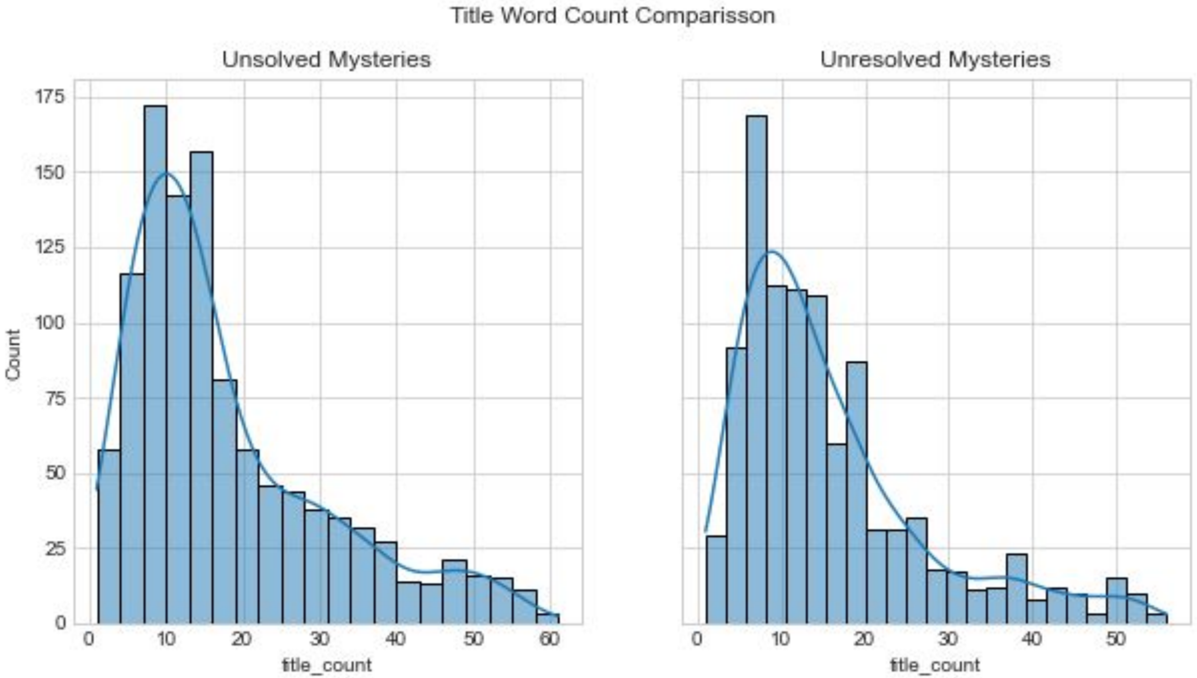
What does our data look like?

- Focus on Title
- Add title length
- Add title word count
- Add sentiment
- Compare our subreddits!

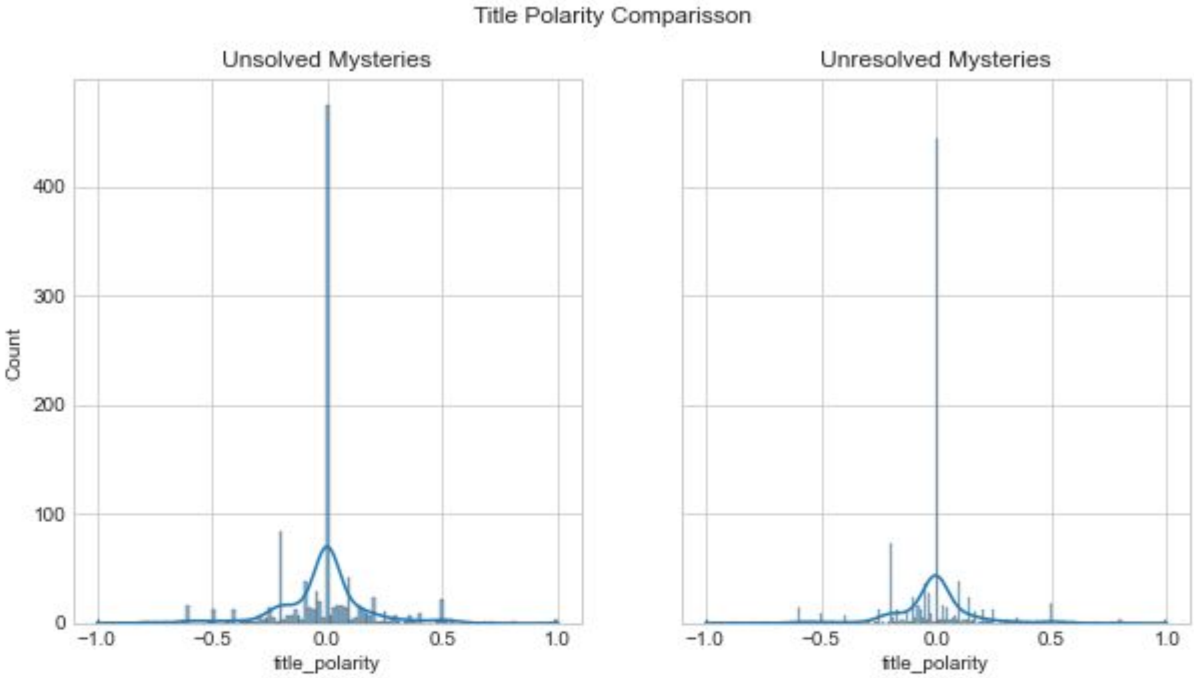
EDA - Comparing our Subreddits on Title Length



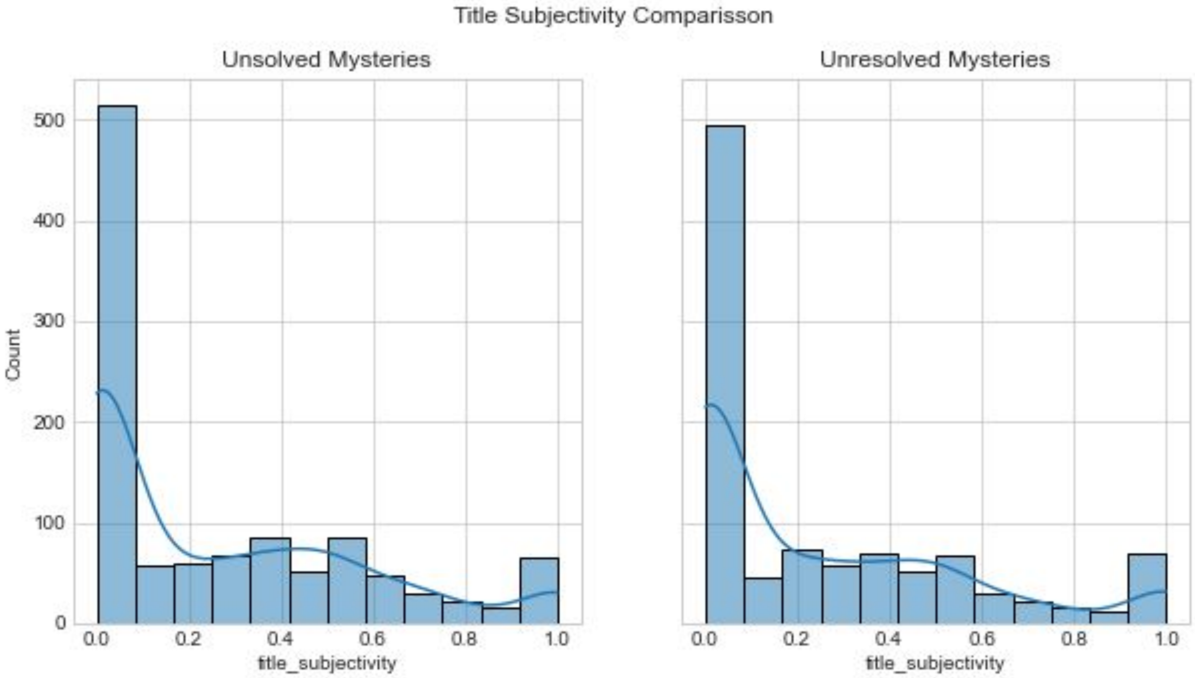
EDA - Comparing our Subreddits on Word Count



EDA - Comparing our Subreddits on Polarity

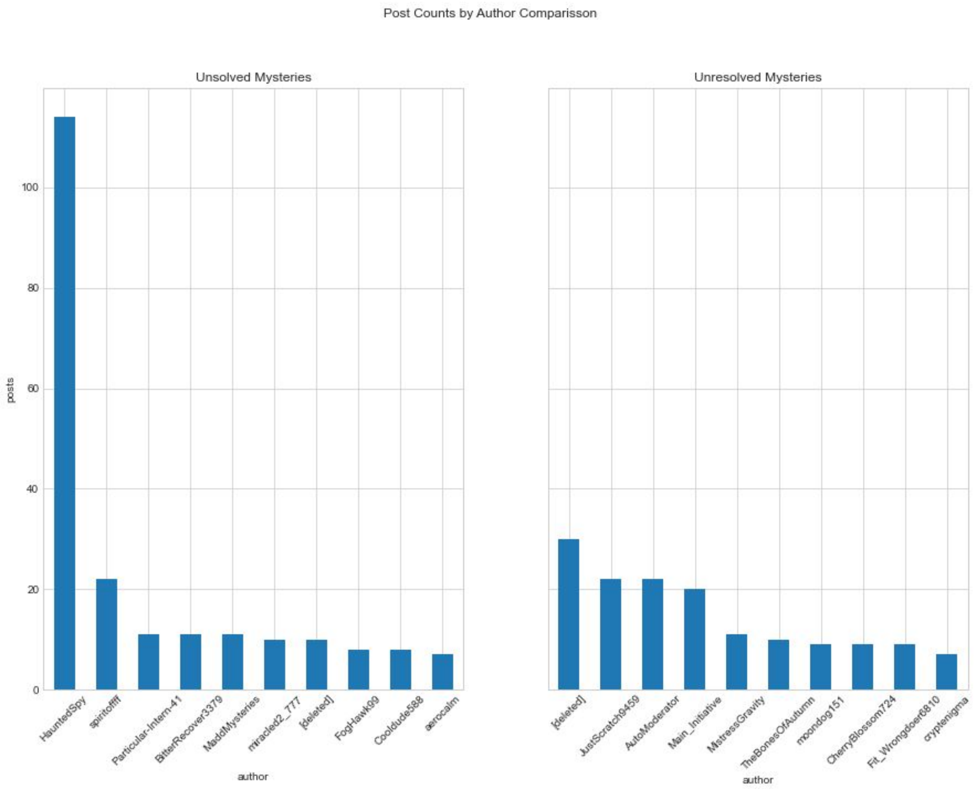


EDA - Comparing our Subreddits on Subjectivity



Titles appear to be very similar for both subreddits!

EDA - Comparing our Subreddits by Author

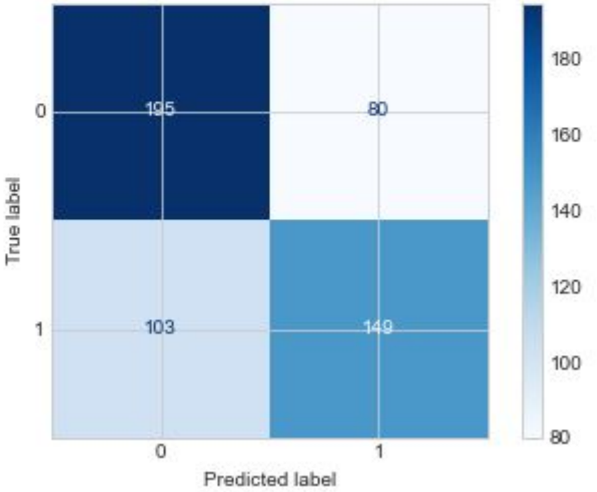


Models

More technical stuff!

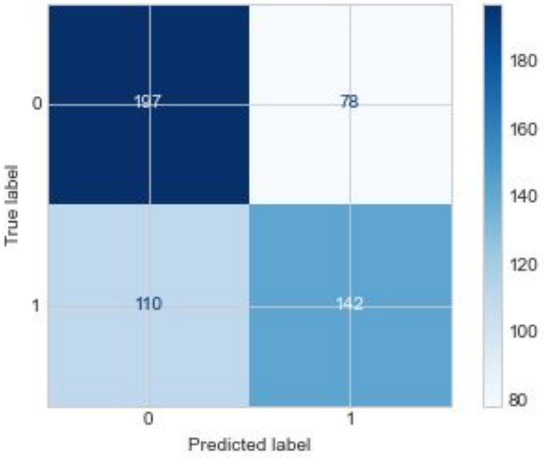
- Baseline (52.18%)
- Grid Search function
- Title Models
- Author Models

Title Models - Count Vectorizer + Naive Bayes



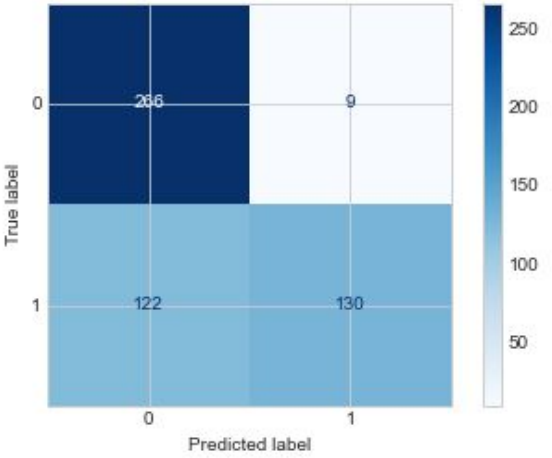
- Grid Search Best Training Set Score: 0.6494
- Best Grid Search Parameters:
- countvectorizer__max_df: 0.45
- countvectorizer__min_df: 1
- countvectorizer__ngram_range: (1, 2)
- countvectorizer__stop_words: english
- multinomialnb__alpha: 1
- Grid Search Best Test Set Score: 0.6528

Title Models - Count Vectorizer + Logistic Reg.



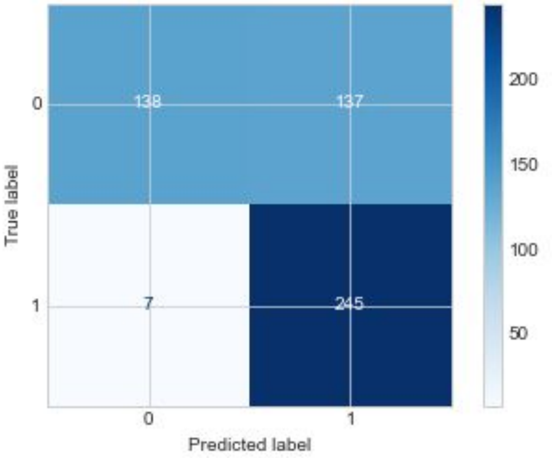
- Grid Search Best Training Set Score: 0.6323
- Best Grid Search Parameters:
- countvectorizer__max_df: 0.45
- countvectorizer__max_features: 2000
- countvectorizer__ngram_range: (1, 1)
- countvectorizer__stop_words: english
- logisticregression__C: 0.1
- Grid Search Best Test Set Score: 0.6433

Author Models - Count Vectorizer + Logistic Reg.



- Grid Search Best Training Set Score: 0.6924
- Best Grid Search Parameters:
- countvectorizer__max_df: 0.45
- countvectorizer__stop_words: english
- logisticregression__C: 0.1
- Grid Search Best Test Set Score: 0.7514

Author Models - Count Vect. + Decision Tree



- Grid Search Best Training Set Score: 0.6854
- Best Grid Search Parameters:
- countvectorizer__max_df: 0.45
- countvectorizer__stop_words: english
- decisiontreeclassifier__max_depth: 1000
- Grid Search Best Test Set Score: 0.7268

Conclusions and Recommendations

- Who you are is the most important factor!
- Choose a subreddit and post consistently
- Titles are quite similar, but my models can help if you are undecided
- Be objective and neutral
- ~16 words per title, <100 characters.

Q&A