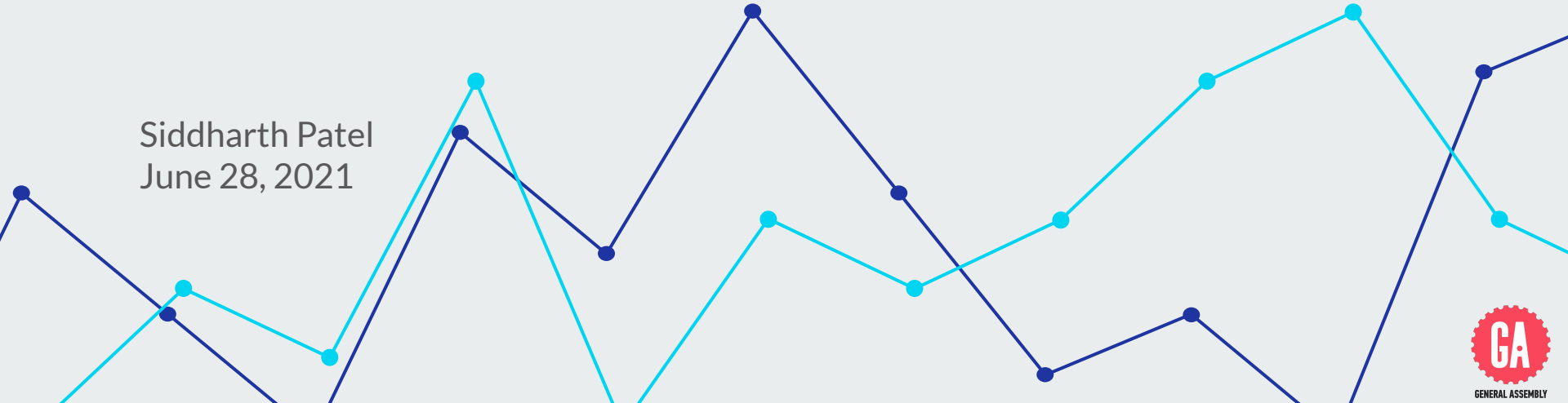
Reddit API & Subreddit Classification

Siddharth Patel
June 28, 2021



GENERAL ASSEMBLY

Overview

Using the text classification capabilities of some machine learning models, we try to analyse data collected from a public website on two similar but different topics.

Problem Statement

To resolve a bet among a group of friends, who is more creative – Filmmakers or Screenwriting?

Gathering Data

Reddit API

/r/Filmmakers

1.8M followers

740 posts

/r/Screenwriting

1.1M followers

982 posts

Exploring Data

/r/Filmmakers

Target = 0

382 NaN

/r/Screenwriting

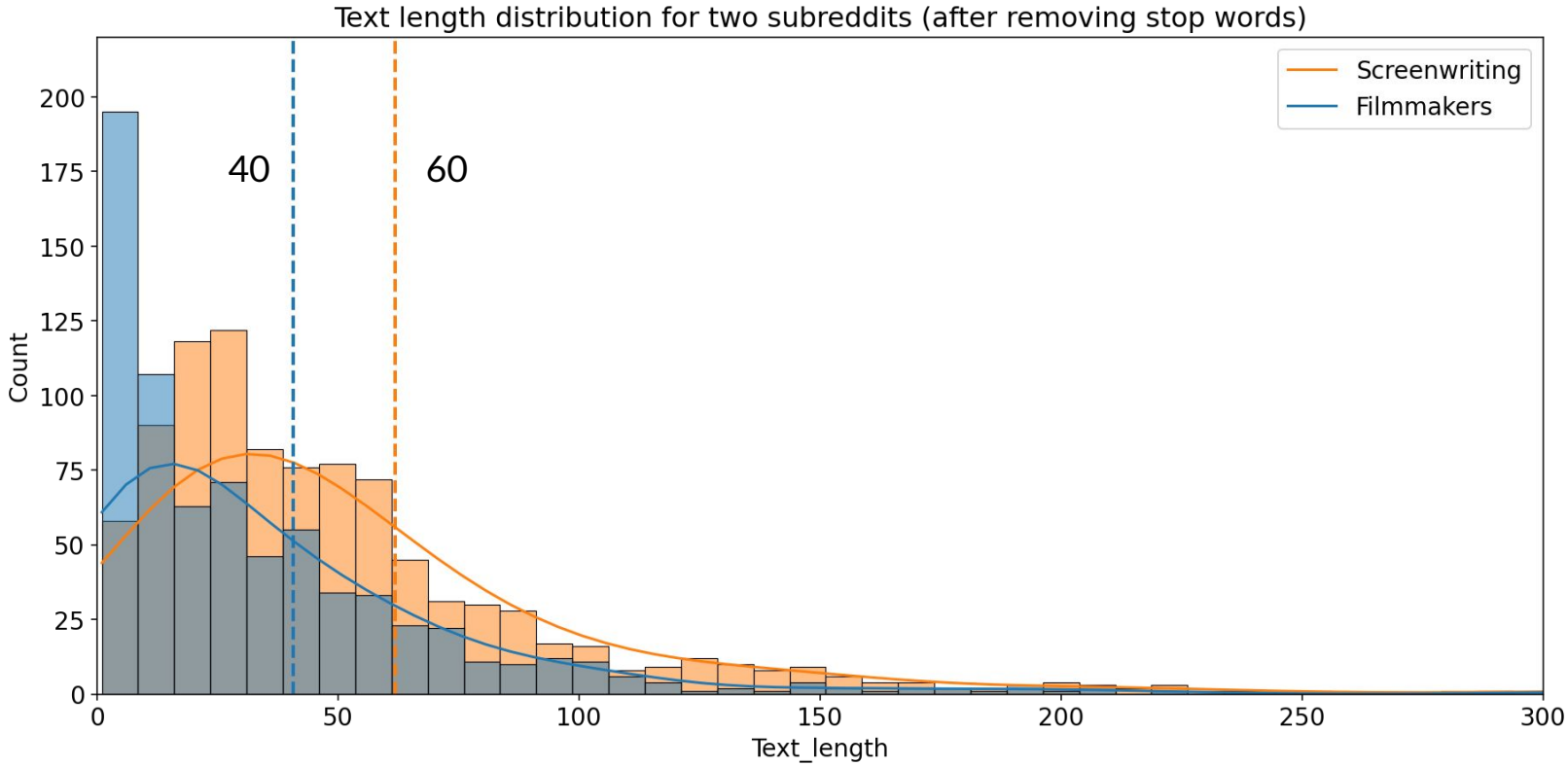
Target = 1

Cleaning steps:

1. Replacing missing data with ‘ ’
2. Combine post title & text to single column
3. Keep only english text characters, remove everything else
4. Remove subreddit topics from the text
5. Remove stop words

GENERAL ASSEMBLY

Text length distribution



Model Building

Logistic Regression

Multinomial Naive-Bayes

Support Vector Machines

CountVectorizer

TfidfVectorizer

Model Comparison

CountVectorizer	Train accuracy	Cross val accuracy	Test accuracy
Logistic Regression	86.43	85.43	85.12
Multinomial-NB	86.12	86.43	82.09
SVM	81.47	81.78	83.49

TfidfVectorizer	Train accuracy	Cross val accuracy	Test accuracy
Logistic Regression	87.83	87.91	86.05
Multinomial-NB	83.95	83.72	83.72
SVM	87.6	87.67	84.65

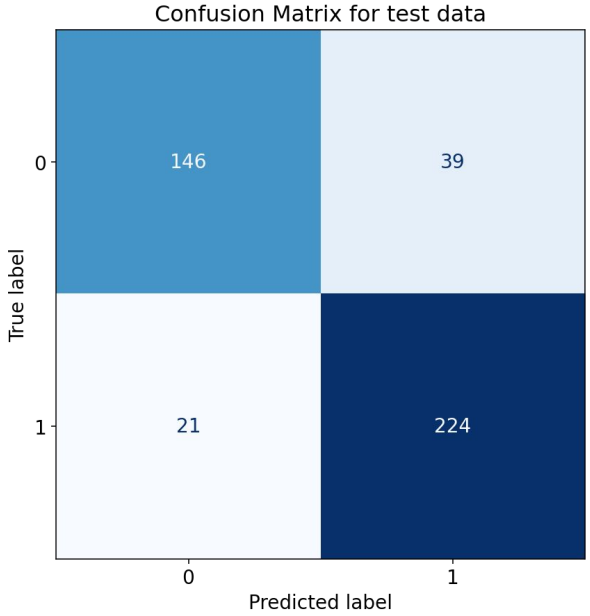
Best Model

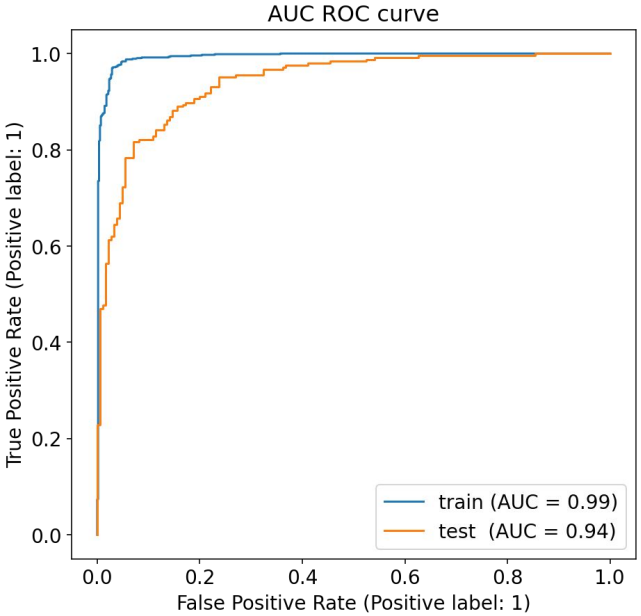
```
# start by defining the best model
best_model = gs_dict['gs_lr_tvec']
best_model.best_estimator_
```

```
Pipeline(steps=[('tvec',
                 TfidfVectorizer(max_df=0.3, max_features=5000, min_df=2,
                                 stop_words='english')),
                ('lr', LogisticRegression(max_iter=500))])
```

Filmmakers	Screenwriting
camera, video, director, editing, music, production, lighting, shots, documentary, sound, commercial	write, screenplay, read, character, draft, pilot, feedback, story, idea, scene

Evaluation





Misclassification

Hi all, I'm fairly new to this community and I apologize profusely in advance if this is the wrong place for my question. If that is the case, I'd appreciate any advice on where else to post it: I have a feature script that's intended for a relatively low budget indie production set in Florida. I am looking for resources or communities through which I can identify and contact indie producers/directors/production companies to pitch my script. Does anybody have any tips on where to find such resources? Thanks in advance for any advice/tips you can provide!!

Actual class: Screenwriting
Predicted class: Filmmakers

Misclassification

Motorcycles in anime: The unlikely history behind Akira, Sailor Moon, and more!

Actual class: Filmmakers
Predicted class: Screenwriting

Conclusions

1. Accuracy of about 87% for train and 86% for test data
2. 60 misclassifications out of the 430 test data observations
3. Other cleaning techniques like lemmatizing can be employed
4. Can collect more “good” data and use other models, like Random Forest, Boosting with Decision Trees
5. Both classes seemingly comprise of disparate "creative" words

Filmmakers and *Screenwriting* subreddits are creative in their own rights, both have some unique creative talking points which differentiates the two roles

Thank You!