Project 3: Web APIs and NLP

r/improv vs r/StandUpComedy

Agenda

- Introduction/Problem Statement
- Model used and scoring metrics
- Web scraping and data cleaning
- Pre-processing
- EDA and Visualization
- Modeling
- Conclusions / Limitations
- Recommendations

Introduction

- Improv vs Stand Up Comedy, 2 forms of artistic expression that are commonly and erroneously used interchangeably.

"Three key differences between the two to underline:

1) Stand-ups are alone on stage, whereas improv folks are onstage with 2-3 other teammates.
2) Stand-ups craft repeatable messaging designed to sell our audiences on ideas & things we find funny. Improv teams are creating stories on the spot, which forever disappear the moment they're completed – all to never be recreated again.
3) Improv folks are positive."

~ Jon Selig

Source: https://www.linkedin.com/pulse/necessary-distinction-jon-selig/

Problem Statement

A performing arts center is thinking of adding courses to their roster for revenue and is looking to compile feedback on Improv and StandUpComedy.

1. The performing arts center would like to be able to separate comments from each subreddit for easy reference.

2. They would like to see the terms that are commonly associated with each performing art in order to better formulate their syllabus.

3. As a secondary concern, they would also like to see which art form may respond better to classes and structured instruction.

Models

- Multinomial Naive Bayes and Logistic Regression

- Logistic regression is easier to implement, interpret, and very efficient to train.

- Naive Bayes doesn't require as much training data. It handles both continuous and discrete data. It is highly scalable with the number of predictors and data points. It is fast and can be used to make real-time predictions.

Scoring metrics

- Accuracy

1. improv and StandUpComedy observations are pretty much even
2. objective is just to correctly and accurately classify the posts to garner more information

As opposed to

- Sensitivity
- Specificity
- Precision

Web Scraping and data cleaning

- Using Pushshift's API
- Subreddits: r/improv and r/StandUpComedy
- 2000 of the latest posts scraped from each subreddit for a total of 4000 posts
- After removing duplicates and dealing with null values, remaining 3920 observations.
- Used a combination of tools to remove unwanted text and symbols from data.

Preprocessing

- Tokenize text
- Remove stopwords
- Stemming
- Lemmatizing

Used ‘English’ stopwords and created custom list of stopwords.

Proceeded with the next steps using the lemmatized text.

Stemming caused words to become nonsensical.

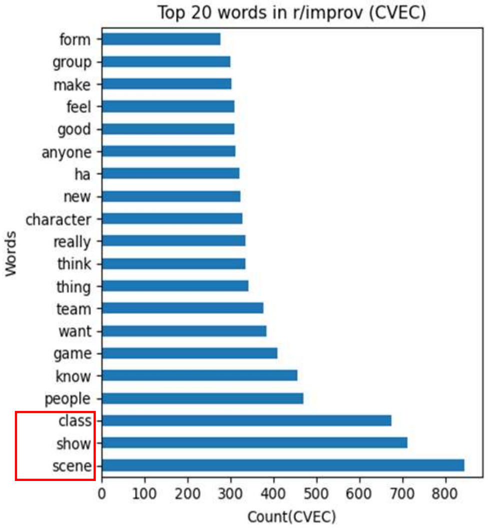
EDA and Visualization

- Plot bar charts using

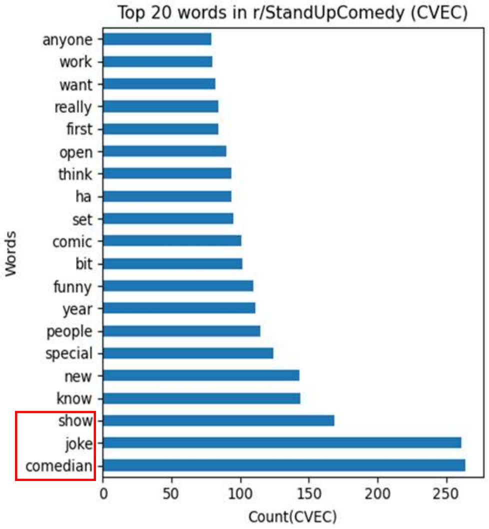
1. CVEC (Count Vectorization)
2. TF-IDF (Term Frequency-Inverse Document Frequency)

Plot for unigram, bigram and trigram.

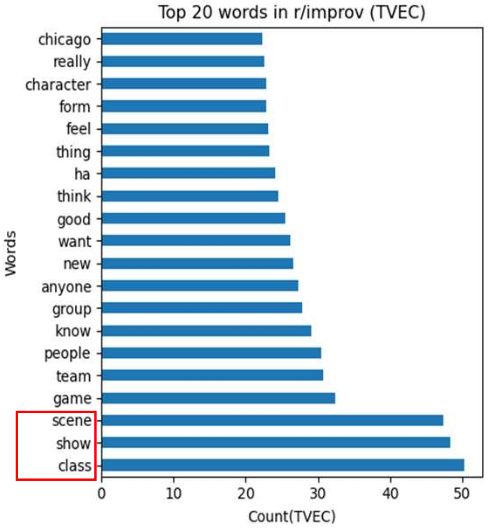
Unigram



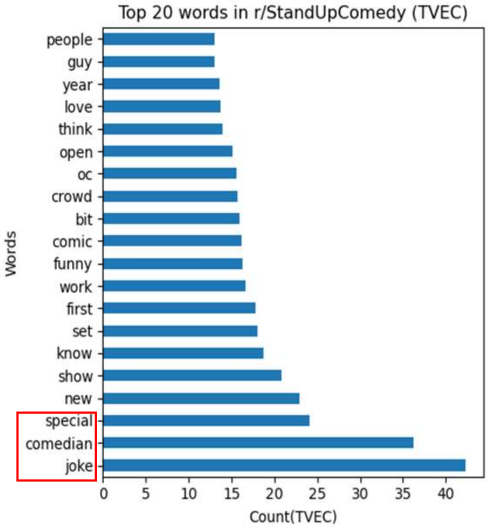
Top 20 words in r/improv (CVEC)
form
group
make
feel
good
anyone
ha
new
character
really
think
thing
team
want
game
know
people
class
show
scene
Words
0
100
200
300
400
500
600
700
800
Count(CVEC)



Top 20 words in r/StandUpComedy (CVEC)
anyone
work
want
really
first
open
think
ha
set
comic
bit
funny
year
people
special
new
know
show
joke
comedian
Words
0
50
100
150
200
250
Count(CVEC)

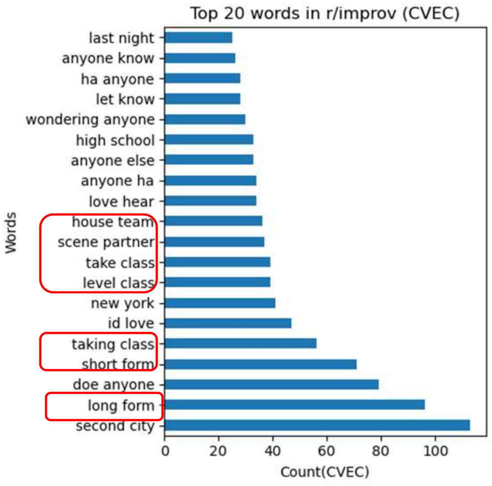


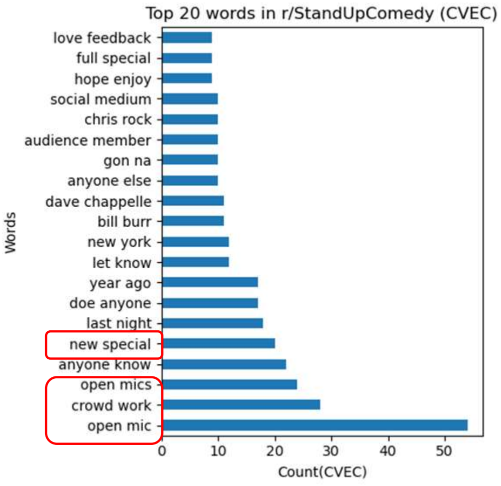
Top 20 words in r/improv (TVEC)
chicago
really
character
form
feel
thing
ha
think
good
want
new
anyone
group
know
people
team
game
scene
show
class
Words
0
10
20
30
40
50
Count(TVEC)

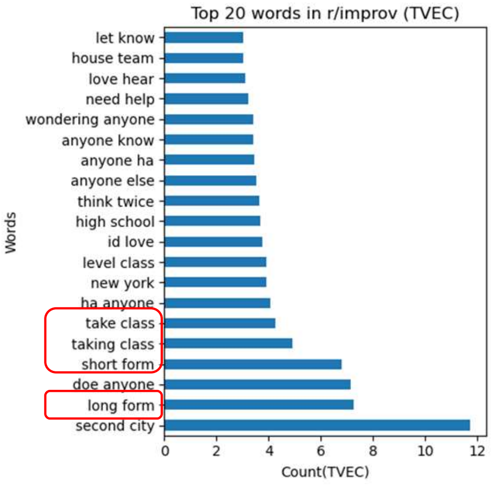


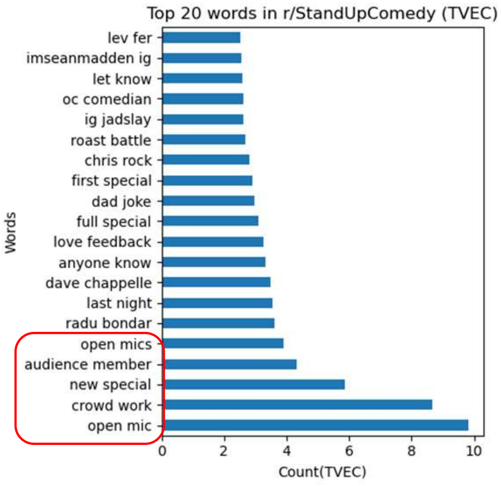
Top 20 words in r/StandUpComedy (TVEC)
people
guy
year
love
think
open
oc
crowd
bit
comic
funny
work
first
set
know
show
new
special
comedian
joke
Words
0
5
10
15
20
25
30
35
40
Count(TVEC)

Bigram

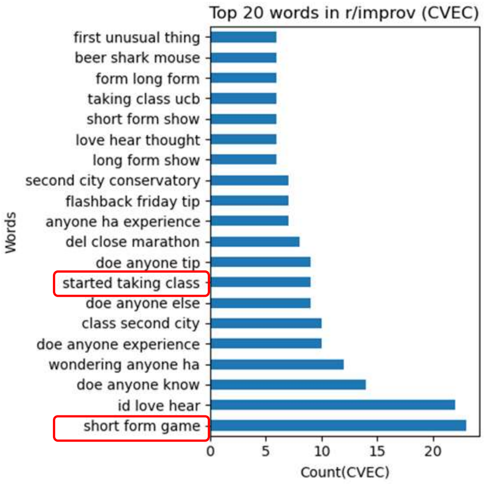

Top 20 words in r/improv (CVEC)
Words
last night
anyone know
ha anyone
let know
wondering anyone
high school
anyone else
anyone ha
love hear
house team
scene partner
take class
level class
new york
id love
taking class
short form
doe anyone
long form
second city
0
20
40
60
80
100
Count(CVEC)

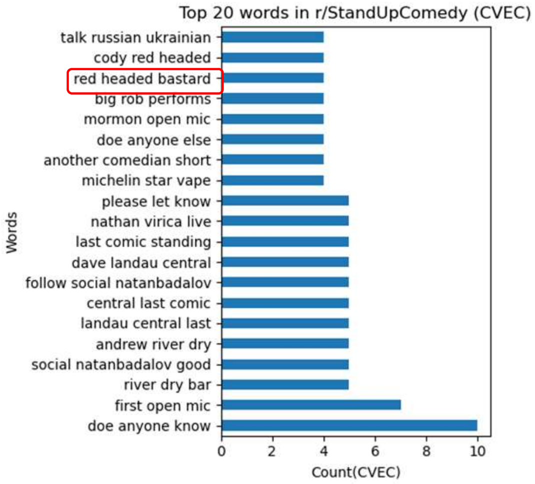

Top 20 words in r/StandUpComedy (CVEC)
Words
love feedback
full special
hope enjoy
social medium
chris rock
audience member
gon na
anyone else
dave chappelle
bill burr
new york
let know
year ago
doe anyone
last night
new special
anyone know
open mics
crowd work
open mic
0
10
20
30
40
50
Count(CVEC)

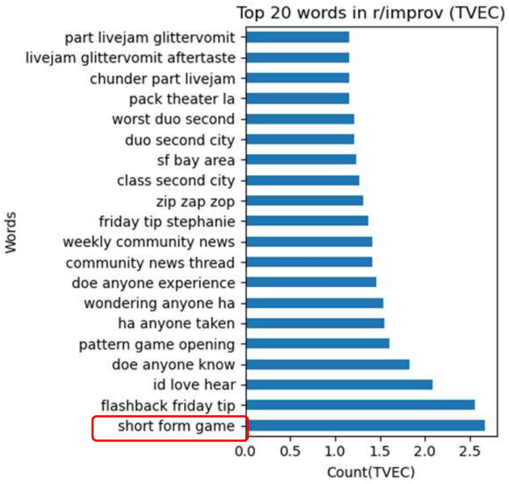

Top 20 words in r/improv (TVEC)
Words
let know
house team
love hear
need help
wondering anyone
anyone know
anyone ha
anyone else
think twice
high school
id love
level class
new york
ha anyone
take class
taking class
short form
doe anyone
long form
second city
0
2
4
6
8
10
12
Count(TVEC)

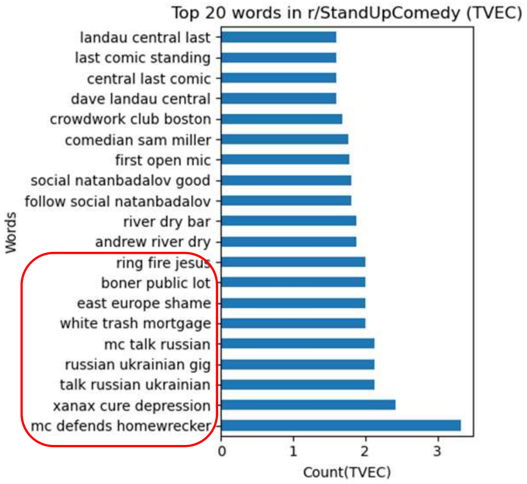

Top 20 words in r/StandUpComedy (TVEC)
Words
lev fer
imseanmadden ig
let know
oc comedian
ig jadslay
roast battle
chris rock
first special
dad joke
full special
love feedback
anyone know
dave chappelle
last night
radu bondar
open mics
audience member
new special
crowd work
open mic
0
2
4
6
8
10
Count(TVEC)

Trigram


Top 20 words in r/improv (CVEC)
first unusual thing
beer shark mouse
form long form
taking class ucb
short form show
love hear thought
long form show
second city conservatory
flashback friday tip
anyone ha experience
del close marathon
doe anyone tip
started taking class
doe anyone else
class second city
doe anyone experience
wondering anyone ha
doe anyone know
id love hear
short form game
Words
0
5
10
15
20
Count(CVEC)

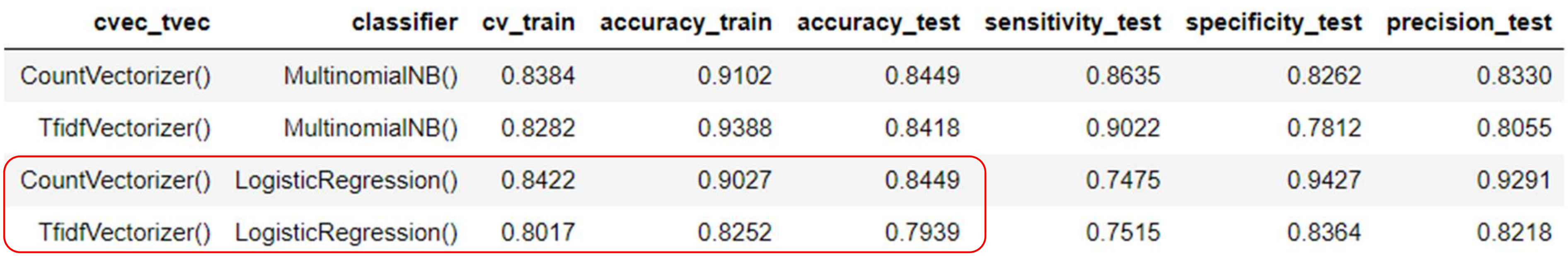

Top 20 words in r/StandUpComedy (CVEC)
talk russian ukrainian
cody red headed
red headed bastard
big rob performs
mormon open mic
doe anyone else
another comedian short
michelin star vape
please let know
nathan virica live
last comic standing
dave landau central
follow social natanbadalov
central last comic
landau central last
andrew river dry
social natanbadalov good
river dry bar
first open mic
doe anyone know
Words
0
2
4
6
8
10
Count(CVEC)


Top 20 words in r/improv (TVEC)
part livejam glittervomit
livejam glittervomit aftertaste
chunder part livejam
pack theater la
worst duo second
duo second city
sf bay area
class second city
zip zap zop
friday tip stephanie
weekly community news
community news thread
doe anyone experience
wondering anyone ha
ha anyone taken
pattern game opening
doe anyone know
id love hear
flashback friday tip
short form game
Words
0.0
0.5
1.0
1.5
2.0
2.5
Count(TVEC)


Top 20 words in r/StandUpComedy (TVEC)
landau central last
last comic standing
central last comic
dave landau central
crowdwork club boston
comedian sam miller
first open mic
social natanbadalov good
follow social natanbadalov
river dry bar
andrew river dry
ring fire jesus
boner public lot
east europe shame
white trash mortgage
mc talk russian
russian ukrainian gig
talk russian ukrainian
xanax cure depression
mc defends homewrecker
Words
0
1
2
3
Count(TVEC)

Modeling

- Baseline taken using value counts of each subreddit
- Improv 50.1%
- StandUpComedy 49.9%



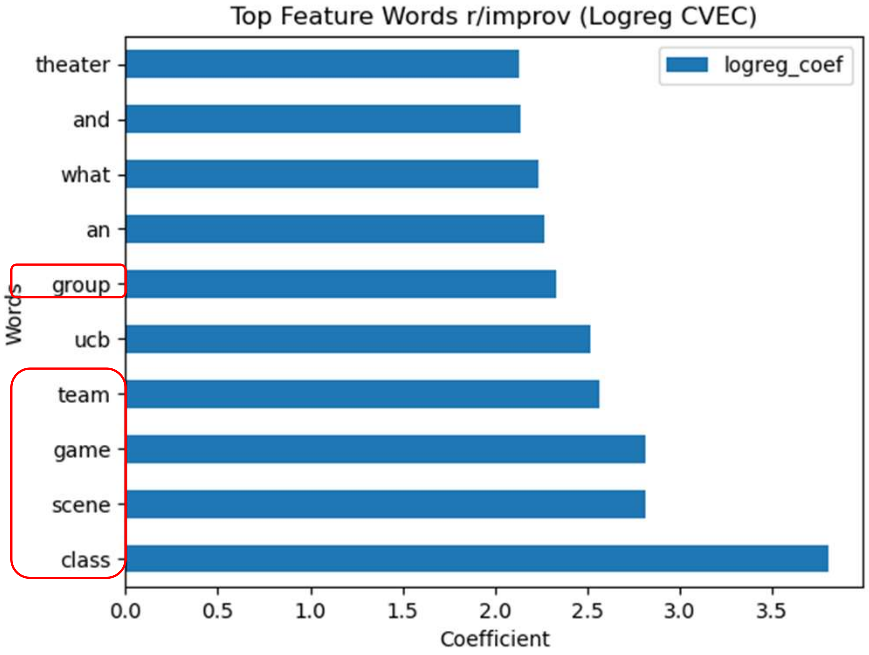
cvec_tvec	classifier	cv_train	accuracy_train	accuracy_test	sensitivity_test	specificity_test	precision_test
CountVectorizer()	MultinomialNB()	0.8384	0.9102	0.8449	0.8635	0.8262	0.8330
TfidfVectorizer()	MultinomialNB()	0.8282	0.9388	0.8418	0.9022	0.7812	0.8055
CountVectorizer()	LogisticRegression()	0.8422	0.9027	0.8449	0.7475	0.9427	0.9291
TfidfVectorizer()	LogisticRegression()	0.8017	0.8252	0.7939	0.7515	0.8364	0.8218

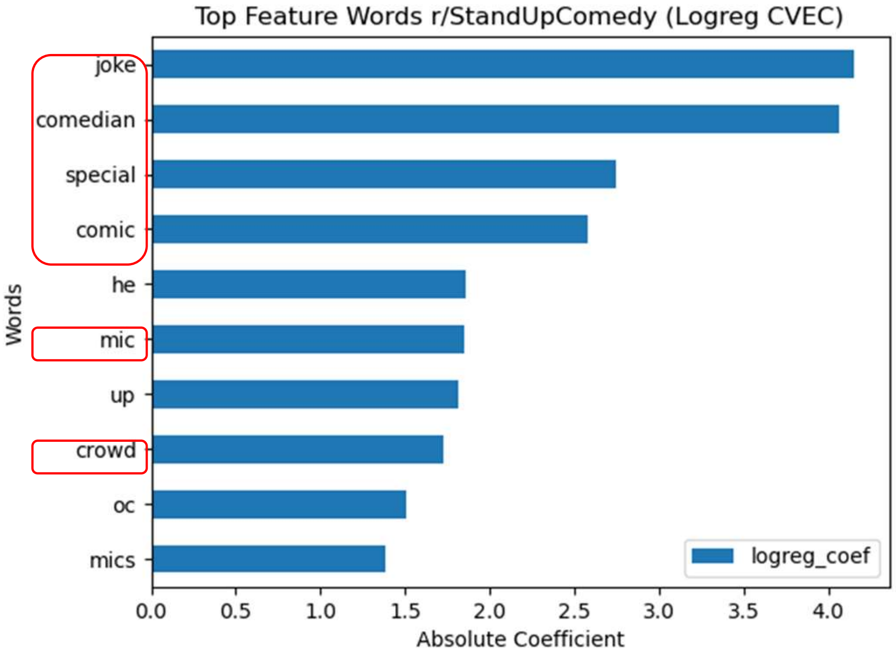
Modeling cont’d

- Best models and best params:

1. CVEC LogisticRegression() Best Params: {'cvec__max_features': 10000, 'cvec__ngram_range': (1, 3), 'logreg__C': 0.1, 'logreg__penalty': 'l2', 'logreg__solver': 'liblinear'}

Interpreting coefficients



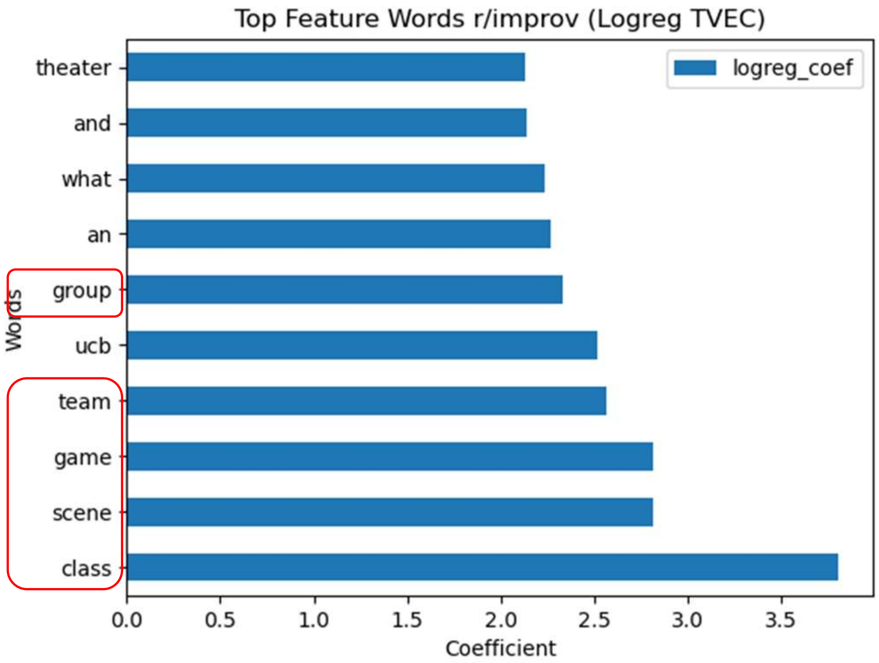


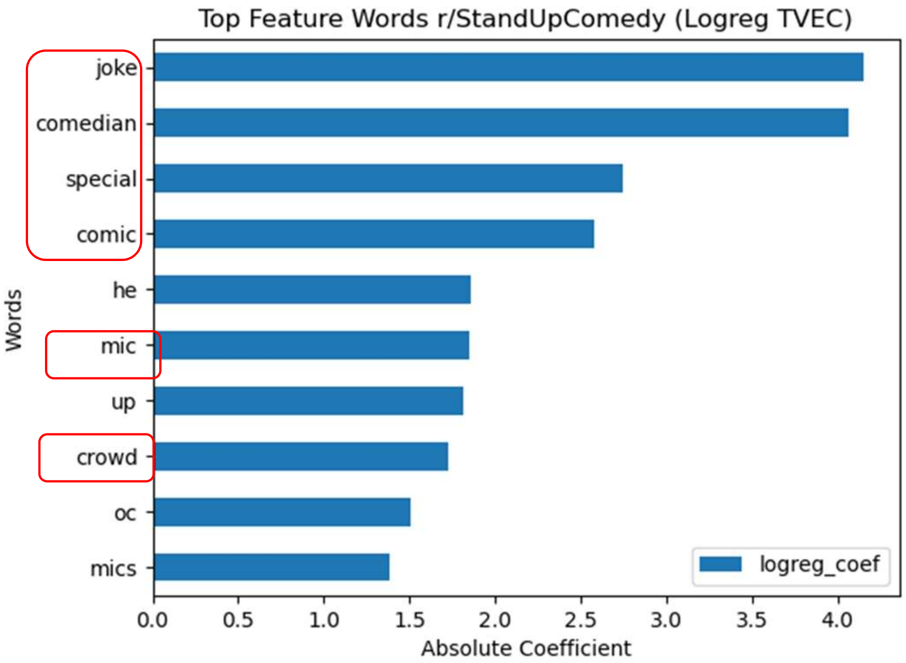
Modeling cont’d

- Best models and best params:

2. TVEC LogisticRegression() Best Params: {'logreg__C': 0.1, 'logreg__penalty': 'l2', 'logreg__solver': 'liblinear', 'tvec__max_features': 15000, 'tvec__ngram_range': (1, 3)}

Interpreting coefficients





Conclusion

- Although not in the top predictors for either subreddit, improv has a generally more positive tone to their posts, while StandUpComedy was a bit more cynical.

- StandUpComedy: is mainly a solo performance which involves using humour and comedy as it's main entertainment medium. Some of the words associated with it indicate that it is a more popular in mainstream media, with words like mic, special(as in Netflix special or some other network special or comedy special) would indicate.

- improv: is mainly a group activity and is commonly associated with scenes, long and short forms, and classes, and can be associated with not just comedy, but also any other form of artistic expression. It is also generally very positively portrayed as a team and group effort.

- The words with the strongest coefficients make more sense for the LogisticRegression model with CountVectorizer.

- All models had very similar run times that were relatively fast.

Limitations/Next steps

- Time constraints, I was not able to explore Decision Tree and Random Forest models.

- Introducing sentiment analysis to the reddit posts to further gauge the prevailing attitudes towards both forms performing arts.

Recommendations

- Use the Logistic Regression model with CountVectorizer
- Improv: Is more likely to be a class that people will sign up for. It is not generally associated with mainstream media. Can be sold both to individuals or as a group package. As a class, it would be easier to have a high student to teacher ratio as it can be taught in groups.
- Stand Up Comedy: Is more likely to be for solo performers. There is likely to be a people taking courses out of interest, but it is also likely that people who want to make it a profession. Because it is a solo activity, the student to teacher ratio would most likely have to be smaller. Words like "special", "comedian" and "crowd" also references stand up comedy in popular culture. It could be worth cashing in on the popularity.

Recommendations

- Pricing can also be higher for the Stand Up Comedy class because of the resources required in a lower student to teacher ratio, and also because it is more likely that people would be willing to pay a bit more for classes in something that could eventually be a career.
- Class syllabus for improv could be structured more towards building synergy and responding to other people in a group/team.
- Class syllabus for Stand Up Comedy could focus more on cultivating the individual's sense of humour and developing the ability of individuals to properly tell their jokes from the setup, to the body and finally the punchline.
- If your performing arts centre is concerned about maintaining a clean image, you might want to steer clear of Stand Up Comedy and focus instead on just doing improv. However, if you are of the opinion that all attention is good attention no matter how controversial, then definitely add in Stand Up Comedy classes.

Questions?