

# Using NLP to Sort between r/Games and r/IndieGaming Subreddits

Julia Tsaltas  
July 16th, 2021

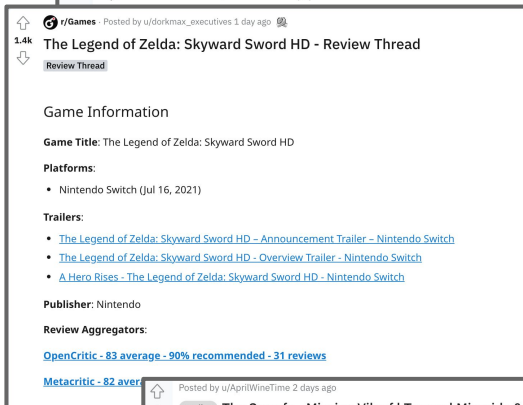
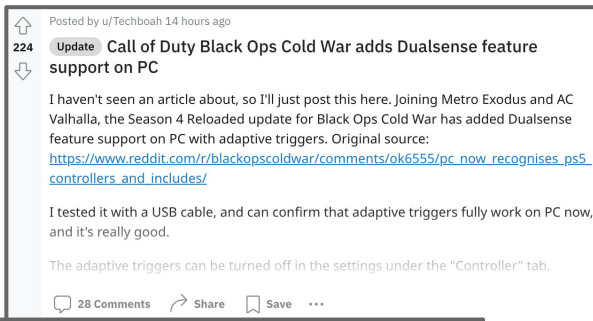
# Outline

- Gaming discussions
- Subreddit data
- Model setup
- CountVectorizer insights

# Gaming Discussion

r/Games

AAA = Games built by a large studio with a big budget



AAA  
Discussion

Reviews

Trailers

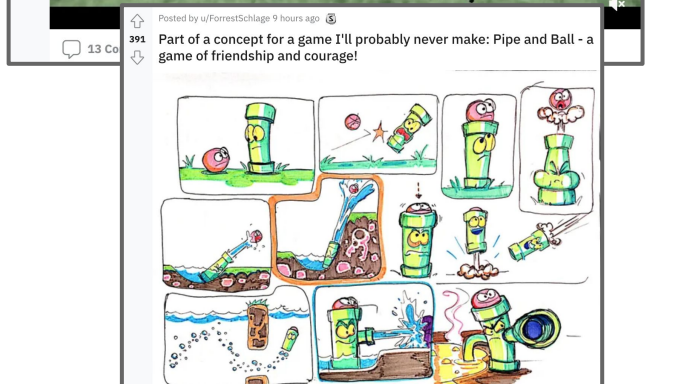
# Gaming Discussion

r/IndieGaming

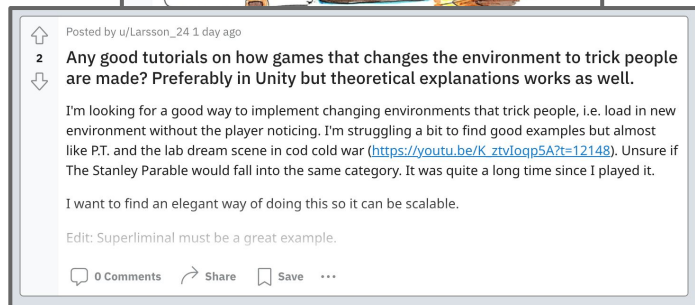
Indie = Game developed by a small group not supported by a large developer



**Progress**



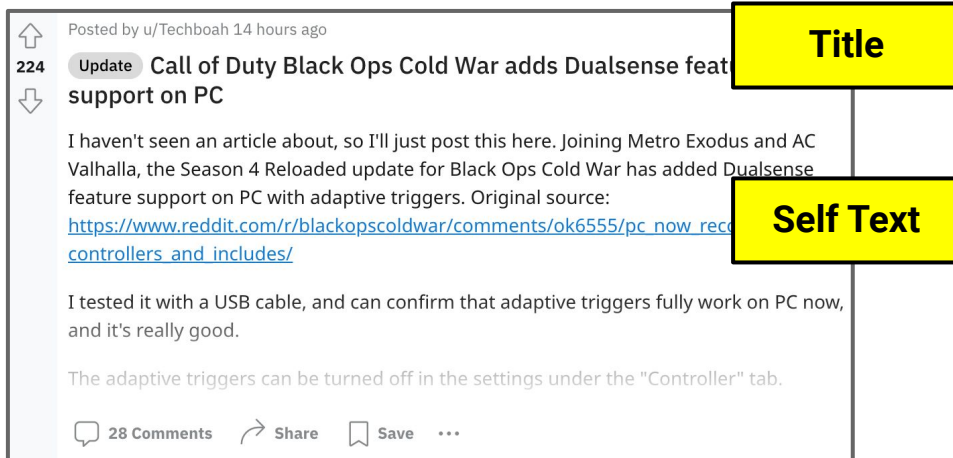
**Concept Art**



**Advice**

# Subreddit Data

- 1126 posts / subreddit
- Title + selftext
- 20,000+ unique terms
- Pipeline =  
CountVectorizer + Model



# Model Results

Best Accuracy	Best Variance/Bias	Worst Performer
<b>Random Forest</b>  <b>99.9%</b> - Train <b>86.0%</b> - Test  max_depth = None min_sample_split = 4 N_estimators = 150	<b>Logistic Regression</b>  <b>81.2%</b> - Train <b>80.8%</b> - Test  C = 0.001 penalty = Ridge	<b>KNN</b>  <b>71.5%</b> - Train <b>64.7%</b> - Test  k = 15

# CountVectorizer Insights

- ↓max\_features ↑variance ↑score
- ↑max\_df ↓variance ↓score
- ngram range → best 2 or 3

# Summary

- Models do capture differences between gaming forums
- All models performed better than baseline
- Overfitting present on all models