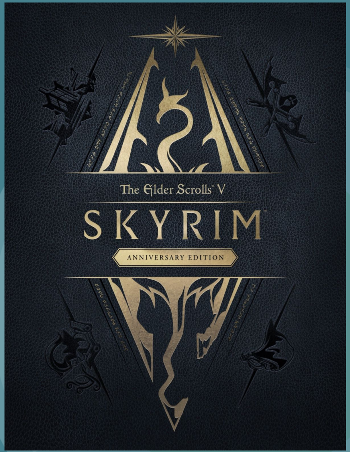
https://en.wikipedia.org/wiki/Metal_Gear_Solid_V:_The_Phantom_Pain#/media/File:Metal_Gear_Solid_V_The_Phantom_Pain_cover.png



https://www.google.com/url?sa=i&url=https%3A%2F%2Felderscrolls.bethesda.net%2Fen%2Farticle%2FREybYU3Gy2InVlHmus4UC%2Fskyrim-10th-anniversary-fan-celebration&psig=AOvVaw1enj4vO4SenDCP8kp6YJ95&ust=1634732745765000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCPjbh5q81vMCFQAAAAAdAAAAABAD

Metal Gear Solid and Skyrim

David Adoni

Problem Statement

- Konami wants to know about Metal Gear Solid on Reddit
- Find model that can classify the subreddits well

Process

1. Collect subreddit data using pushshift.io API
2. Clean data and EDA
3. Preprocessing and modeling
4. Evaluation
5. Conclusions

I.Data Collection

Data Collection

- Pushift API on r/metalgearsolid and r/skyrim
- 2017-2021 100 title posts collected for each year, then merged.
- Let’s take a look at some of the findings

II. Cleaning & EDA

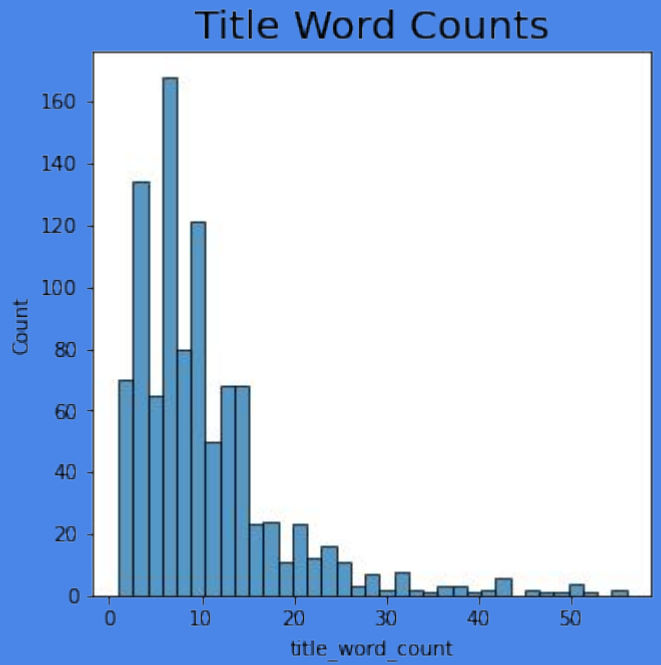
Cleaning

- Reg Exp Tokenizer, and Word Lemmatizer

- Dropped any NaNs

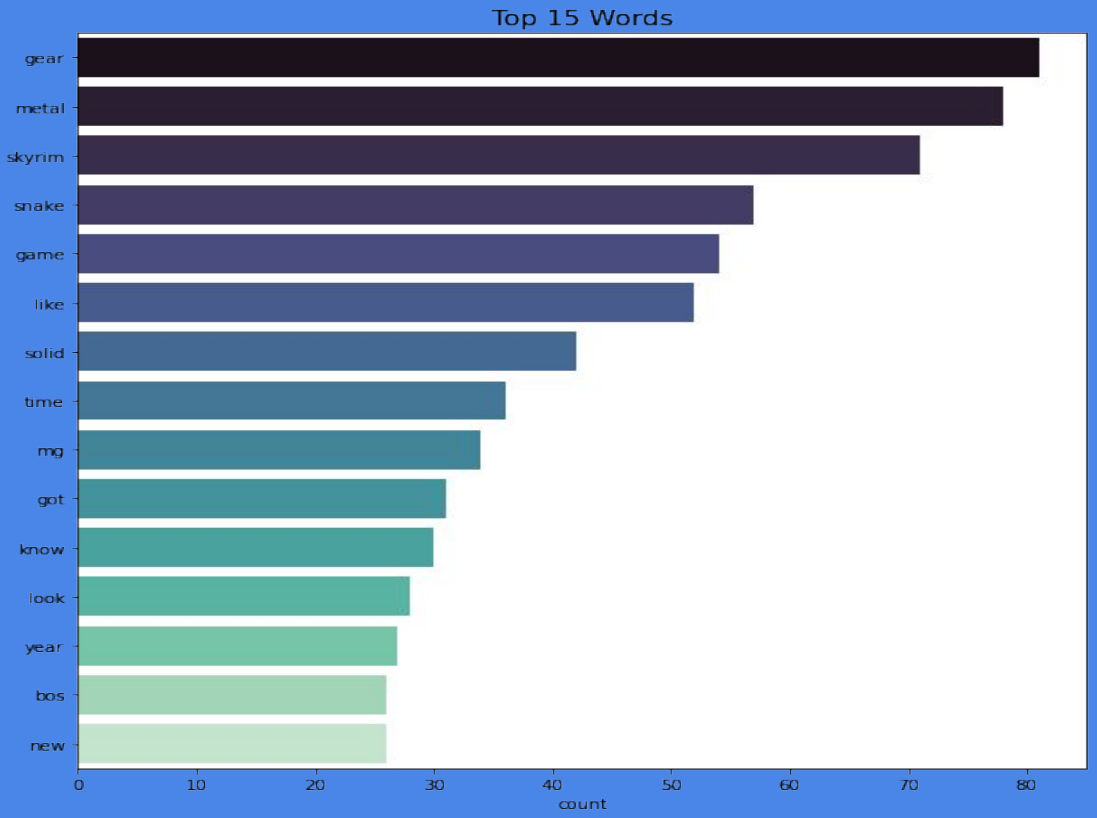
EDA

The amount of words in each title is shown



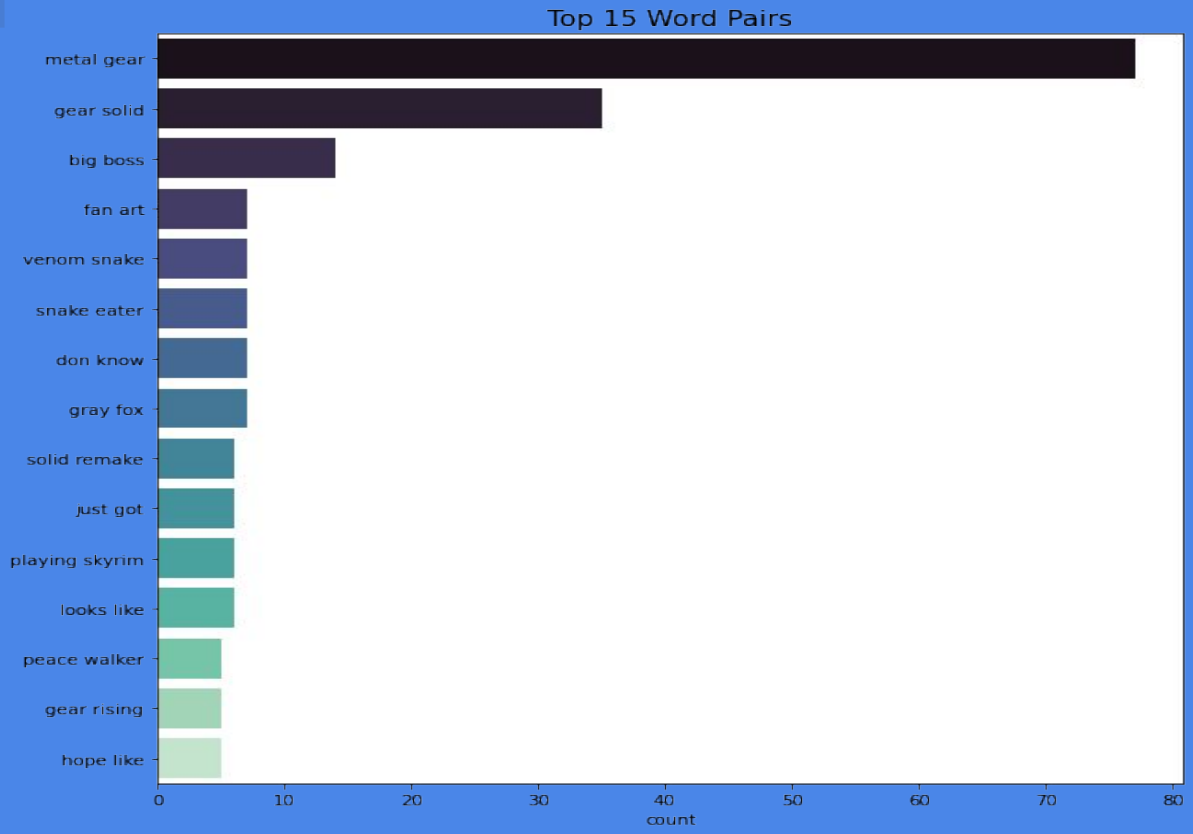
What are the most frequent words?

Top 15 words (single words)



CountVectorizer was used to find these with n_grams set to (1,1)

Top 15 words (word pairs)



CountVectorizer was used to find these with n_grams set to (2,2)

III.Models

Models included

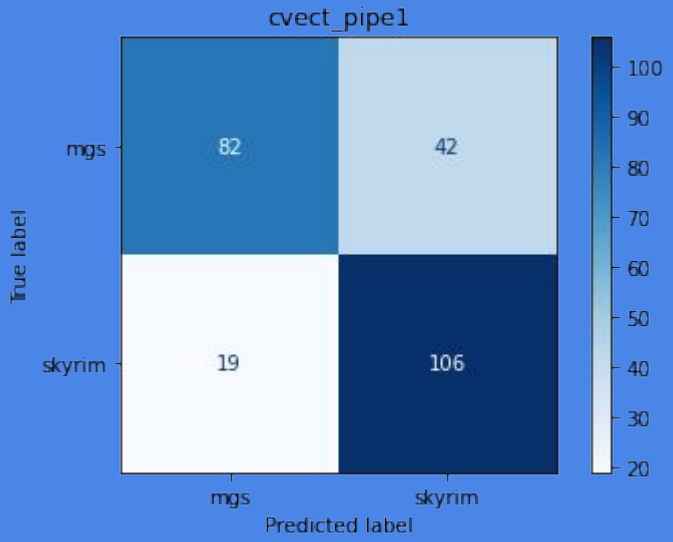
Only models that scored better than the baseline will be included

Baseline model

“is_mgs” =1 0.502518

“is_mgs” =0 0.497482

Model 1

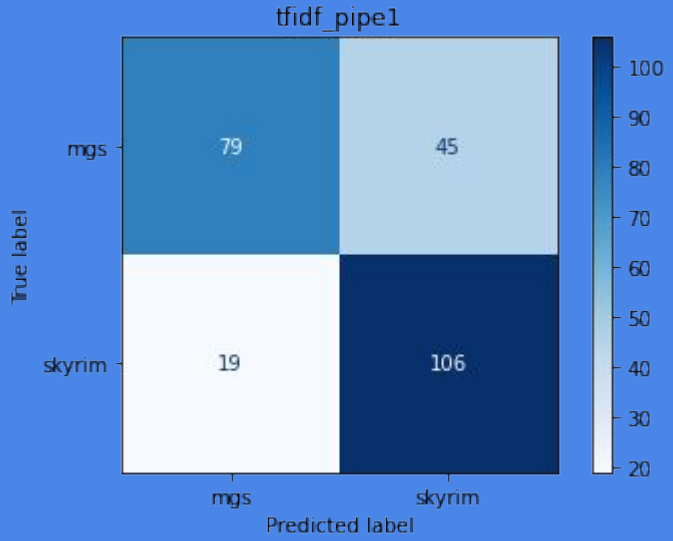


This model used CountVectorizer and MultinomialNB.

Score Train: 0.9717741935483871

Score Test: 0.7550200803212851

Model 2



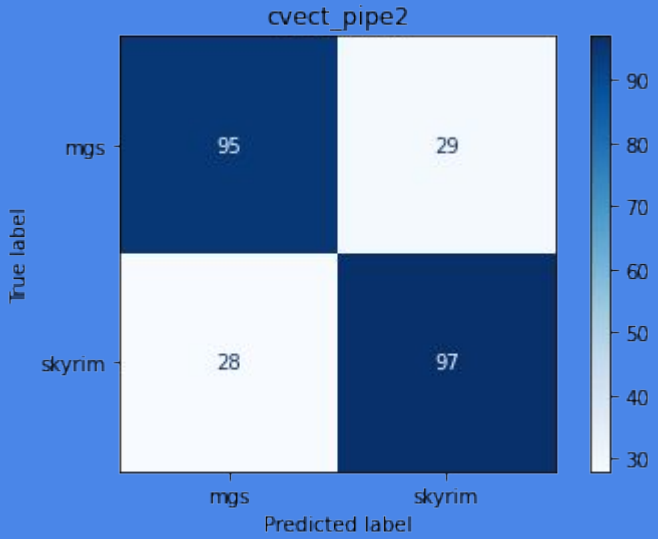
This model used GridsearchCV TfidfVectorizer and MultinomialNB.

Best Params: {'tfidfvectorizer__max_features': 2000, 'tfidfvectorizer__ngram_range': (1, 1), 'tfidfvectorizer__stop_words': 'english'}

Best Estimator Score Train: 0.9838709677419355

Best Estimator Score Test: 0.7429718875502008

Model 3



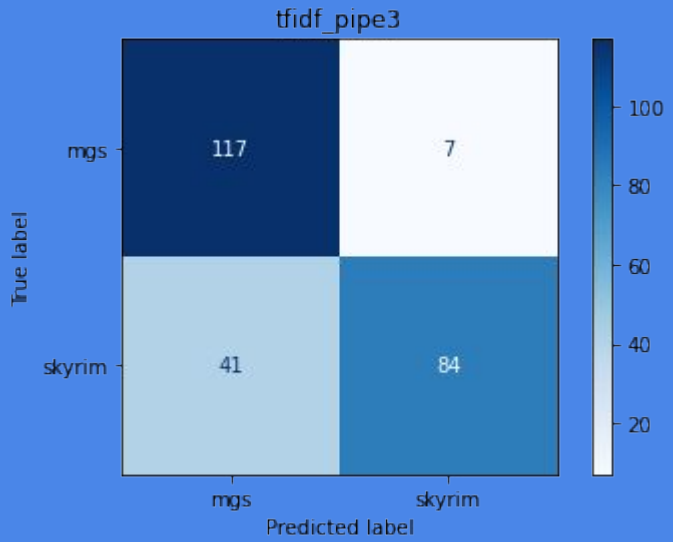
This model used GridsearchCV CountVectorizer, StandardScaler, and Logistic Regression.

Best Params: {'countvectorizer__max_features': 2000, 'countvectorizer__ngram_range': (1, 2), 'countvectorizer__stop_words': None}

Best Estimator Score Train: 0.9865591397849462

Best Estimator Score Test: 0.7710843373493976

Model 6



This model used GridsearchCV TfidfVectorizer, and RandomForestClassifier.

Best Params: {'randomforestclassifier__criterion': 'gini', 'randomforestclassifier__max_depth': 30, 'randomforestclassifier__n_estimators': 60, 'tfidfvectorizer__max_features': None, 'tfidfvectorizer__ngram_range': (1, 1), 'tfidfvectorizer__stop_words': 'english'}

Best Estimator Score Train: 0.8629032258064516

Best Estimator Score Test: 0.8072289156626506

IV.Conclusion

Conclusion

Overall Model 3 performed best on the train, but was only decent on test data.

Model 6 performed with a good balance but was lower on the train set and highest on test data.

Model 6 would be the model used for the classification.

Questions?