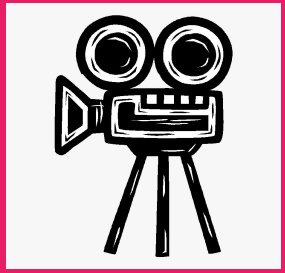
Review Sentiment Analysis



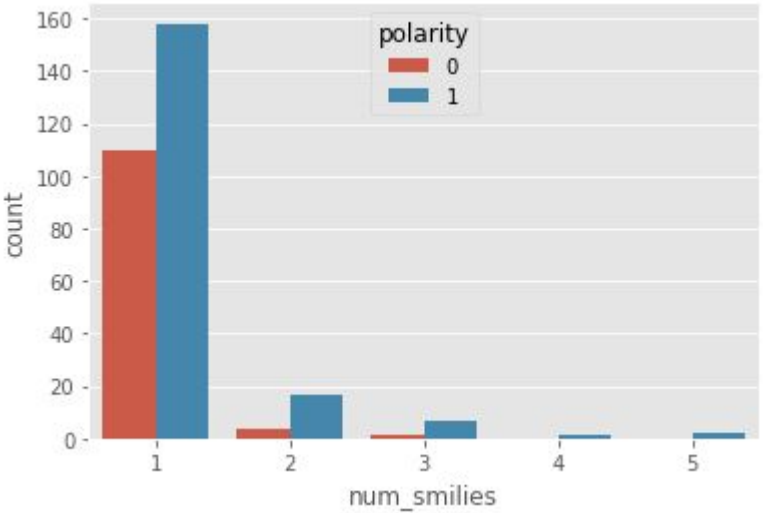
Group 5: Alex Diaz-Clark, Henry Wong, Morgan Sell, Sherry Duong

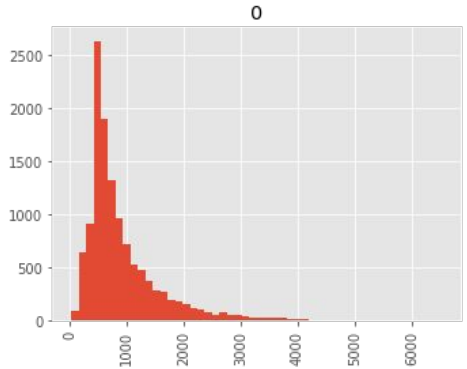
The Data

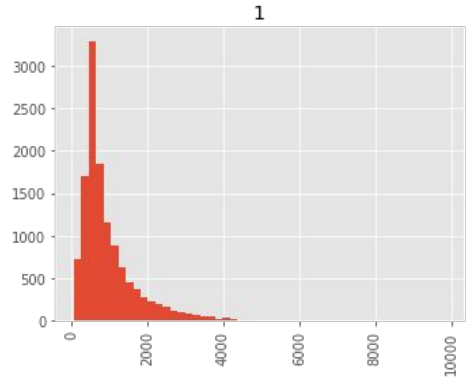
---

1. 50K Movie Reviews:
    i. 25K in train data
    ii. 25K in test data
2. All data was text reviews
3. Even split between positive & negative reviews in train data:
    i. No need for class balancing

EDA: Created some features to see if any key differences/correlations between polarity

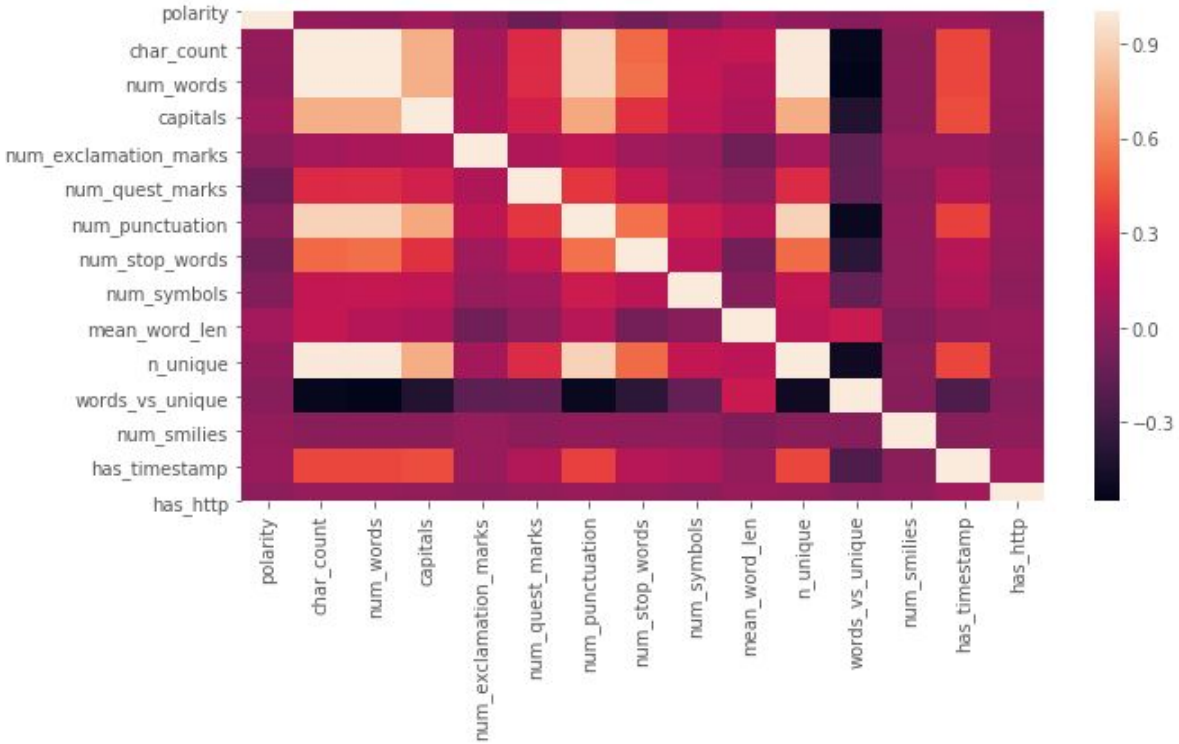






Length of Review between positive & negative

EDA: Created some features to see if any key differences/correlations between polarity



num_quest_marks	-0.123904
num_stop_words	-0.100506
num_symbols	-0.041137
num_punctuation	-0.020218
words_vs_unique	-0.018700
num_exclamation_marks	-0.009938
has_http	0.006827
n_unique	0.015342
num_words	0.017243
char_count	0.025597
num_smilies	0.030437
has_timestamp	0.046259
capitals	0.056332
mean_word_len	0.084105
polarity	1.000000

Clusters of Topics

Negative Review Clusters(with stopwords):

1. 'story', 'characters', 'good', 'character', 'much', 'one', 'original', 'great', 'actors', 'plot'
2. 'br', 'the', 'this', 'if', '10', 'in', 'and', 'as', 'what', 'but'
3. 'film', 'films', 'would', 'made', 'making', 'see', 'make', 'director', 'people', 'time'

Positive Review Clusters(with stopwords):

1. 'fly', 'focusing', 'marie', 'serving', 'wrapped', 'went', 'dog','images', 'seeking', 'me'
2. 'box', 'thing', 'thrilling', '000', 'incoherent', 'information', 'buy', 'neat', 'angle', 'also'
3. 'naked', 'name', 'watch', 'serving', 'marie', 'seat', 'wrapped','lloyd', 'remaining', 'three'

Sentiments Shared Keywords

Baseline Models

NaiveBayes
Normalized, removed stopwords.
-Validation Accuracy: 0.87
Normalized, did not remove stopwords.
-Validation Accuracy: 0.87

Logistic Regression
Normalized, removed stopwords.
-Validation Accuracy: 0.89
Normalized, did not remove stopwords.
-Validation Accuracy: 0.90
-Test Accuracy: 0.88

RandomForestClassifier
Normalized, removed stopwords.
-Validation Accuracy: 0.76
Normalized, did not remove stopwords.
-Validation Accuracy: 0.76

GradientBoostingClassifier
Normalized, removed stopwords.
-Validation Accuracy: 0.81
Normalized, did not remove stopwords.
-Validation Accuracy: 0.81

Improving model

Change N-grams = 2grams and 3grams

include/ not include stopwords

Keep/not keep punctuation

Created custom tokenizer inclusive of “br”

Used CountVectorizer vs TfidfVectorizer

Conclusion

Best-performing model: Logistic Regression

Further exploration:

- Create a more robust tokenizer inclusive of lamentizer and stemming.
- Take more time to optimize parameters of the classifier model (grid search).