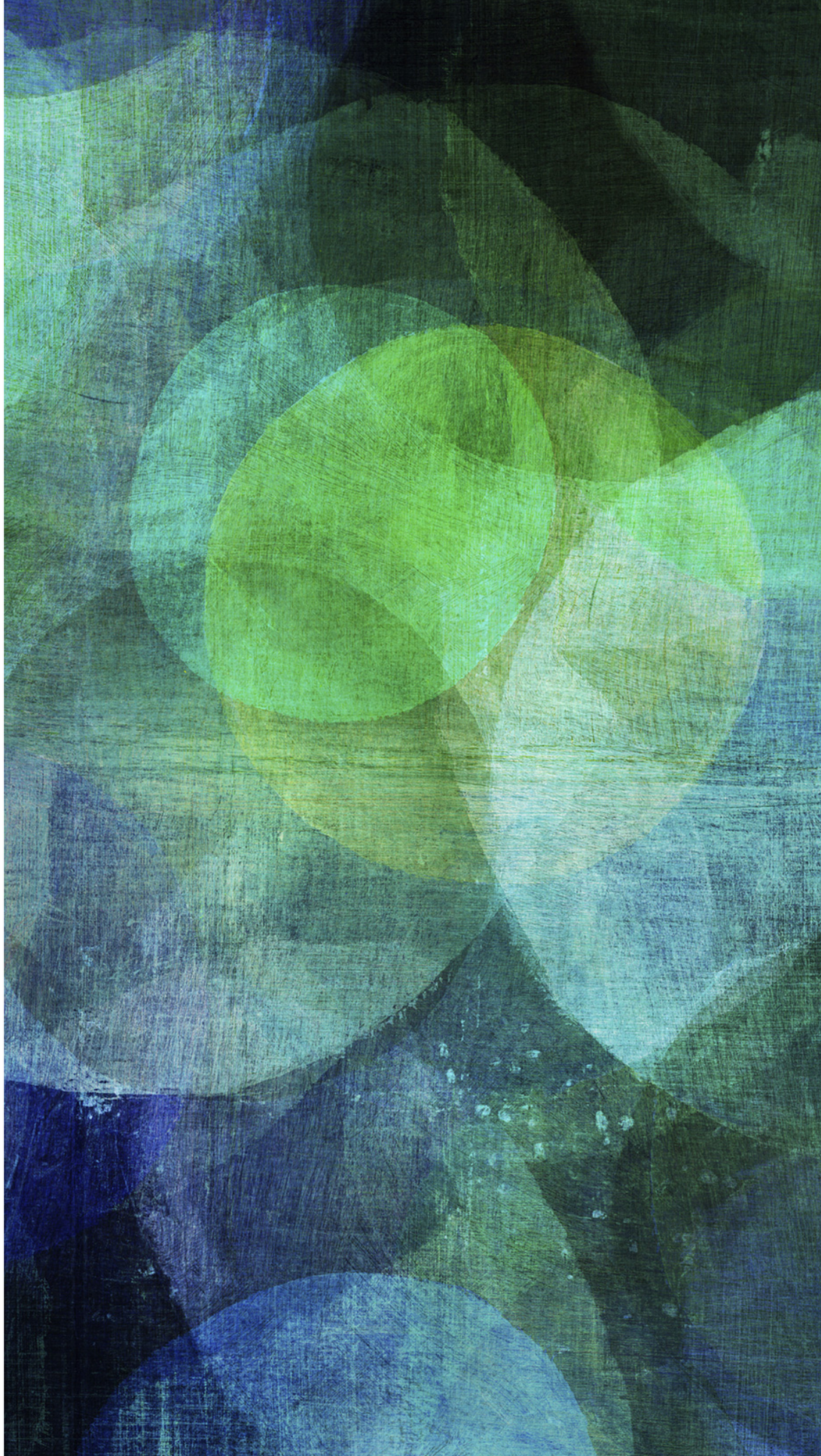
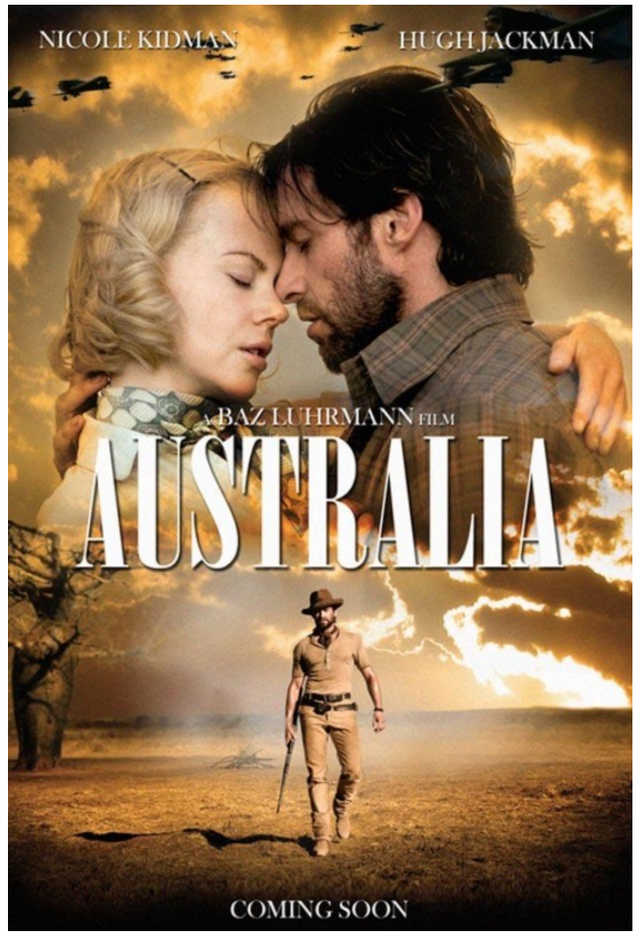
MOVIE REVIEWS SENTIMENT ANALYSIS

Samuel Bolivar




DON'T LET GO
SANDRA BULLOCK
GEORGE CLOONEY
GRAVITY
10.4.13


KEANU REEVES
LAURENCE FISHBURNE
THE MATRIX


NICOLE KIDMAN
HUGH JACKMAN
A BAZ LUHRMANN FILM
AUSTRALIA
COMING SOON


WHEN KEVIN'S FAMILY LEFT FOR VACATION THEY FORGOT ONE MINOR DETAIL: KEVIN.
BUT DON'T WORRY...
HE COOKS. HE CLEANS.
HE KICKS SOME BUTT.
FROM JOHN HUGHES
HOME ALONE
A FAMILY COMEDY WITHOUT THE FAMILY.


SCHINDLER'S LIST
whoever saves one life
saves the world entire
ezlightweb.de


STAR WARS
THE EMPIRE STRIKES BACK


Forrest Gump


JURASSIC PARK
WWW.WALLPAPER3RENDER1.COM

AGENDA

- Data Acquisition and Preprocessing
- Classification Models
- Logistic Regression
- Natural Language Processing (NLP)
- Clustering Analysis
- Predicting Movie Sentiment
    - Predicting on Out Of Sample (OOS) data
- Conclusions
- Questions

DATA ACQUISITION

Completed • Knowledge • 578 teams

Bag of Words Meets Bags of Popcorn

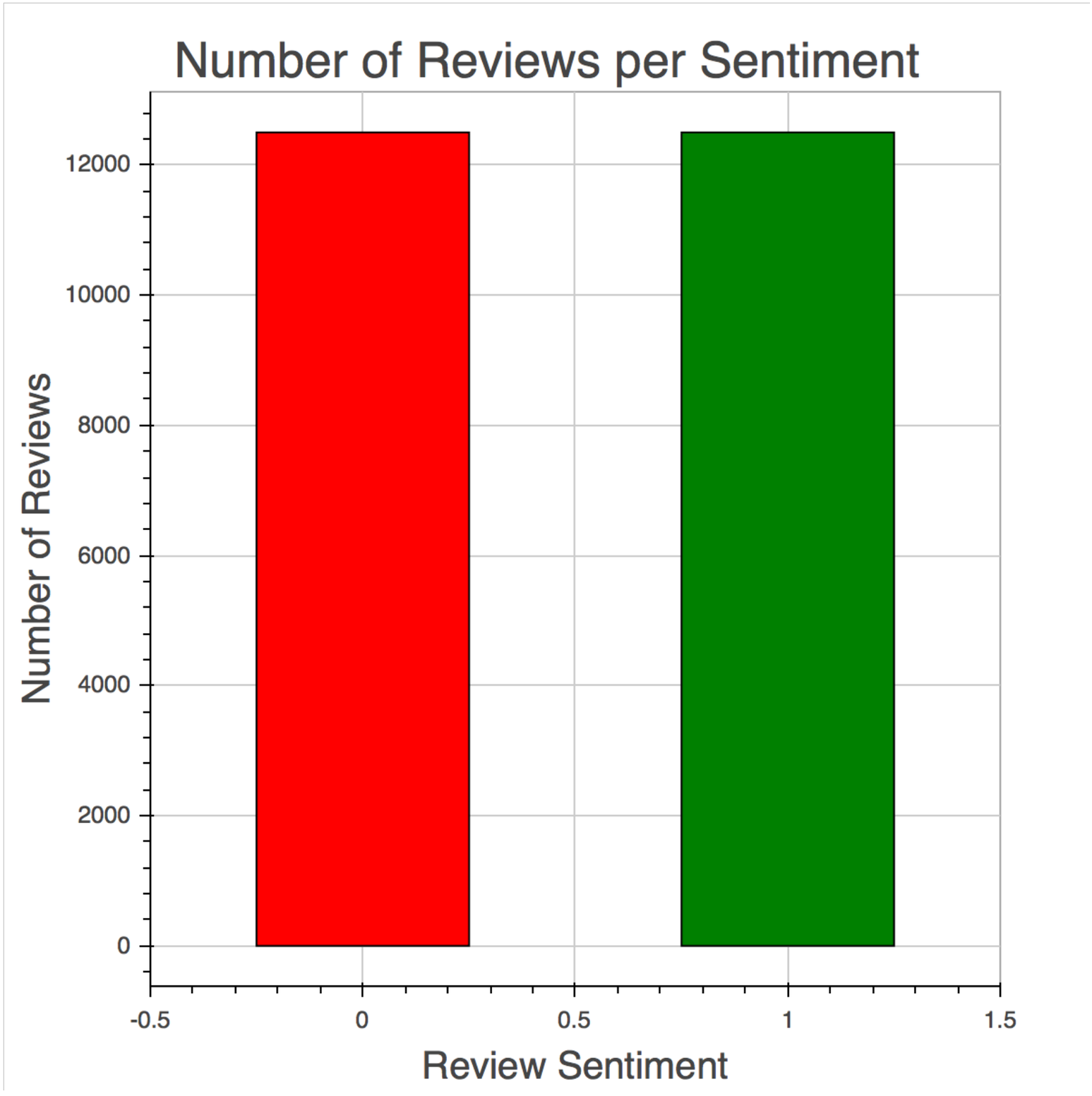
Tue 9 Dec 2014 – Tue 30 Jun 2015 (13 months ago)

- Train data comprises 25000 IMDB movie reviews
  - Binary labeled
    - 1 means Good Review (Rating >7)
    - 0 means Bad Review (Rating < 5)
- Test data comprises 25000 IMDB movie reviews
- Out Of Sample (OOS) data comprises 50000 IMDB movie reviews
- All the above extracted from labeled data from Kaggle
- No individual movie has more than 30 reviews

DATA PREPROCESSING

- Checked each Data Set for null values on any of the fields
- Inspected a random review sample from each Data Set to confirm structure and language
- Changed all data to lower case letters
- Checked Train Data for Imbalance of the categories we want to predict.
- Used NLTK and BeatifulSoup later for NLP processing

DATA PREPROCESSING



CLASSIFICATION MODELS

CLASSIFICATION MODELS

- Used Dummy with default values, but with strategy=‘most_frequent'
- Did not use the GaussianNB; after some research, found MultinomialNB which is ideal for my problem.

```
Init signature: MultinomialNB(self, alpha=1.0, fit_prior=True, class_prior=None)
Docstring:
Naive Bayes classifier for multinomial models

The multinomial Naive Bayes classifier is suitable for classification with
discrete features (e.g., word counts for text classification). The
multinomial distribution normally requires integer feature counts. However,
in practice, fractional counts such as tf-idf may also work.
```

- Used Decision Trees and Random Forests
- Used K Nearest Neighbours (KNN)
- Used Logistic Regression

CLASSIFICATION MODELS

	Accuracy Score	Cross Validation Score (CV=100)
Dummy	0.4962	0.5038
Naive Bayes	0.8584	0.8444
Decision Trees	0.7178	0.8444
Random Forests	0.8570	0.8474
KNN	0.5796	0.5692
Logistic Regression	0.8802	0.8596

MULTINOMIAL NAIVE BAYES (MULTINOMIALNB)

➤ Confusion Matrix

```
Results of Multinomial Naive Bayes:
                Predicted Class 0  Predicted Class 1
Actual Class 0               2175                306
Actual Class 1                402               2117

Precision: 0.873710276517
Recall: 0.840412862247
```

➤ Classification Report

	precision	recall	f1-score	support
Class 0	0.84	0.88	0.86	2481
Class 1	0.87	0.84	0.86	2519
avg / total	0.86	0.86	0.86	5000

RANDOM FORESTS

➤ Confusion Matrix

```
Results of Random Forests:
                 Predicted Class 0  Predicted Class 1
Actual Class 0                2122                359
Actual Class 1                 356               2163

Precision: 0.857652656622
Recall: 0.858674077015
```

➤ Classification Report

	precision	recall	f1-score	support
Class 0	0.86	0.86	0.86	2481
Class 1	0.86	0.86	0.86	2519
avg / total	0.86	0.86	0.86	5000

LOGISTIC REGRESSION

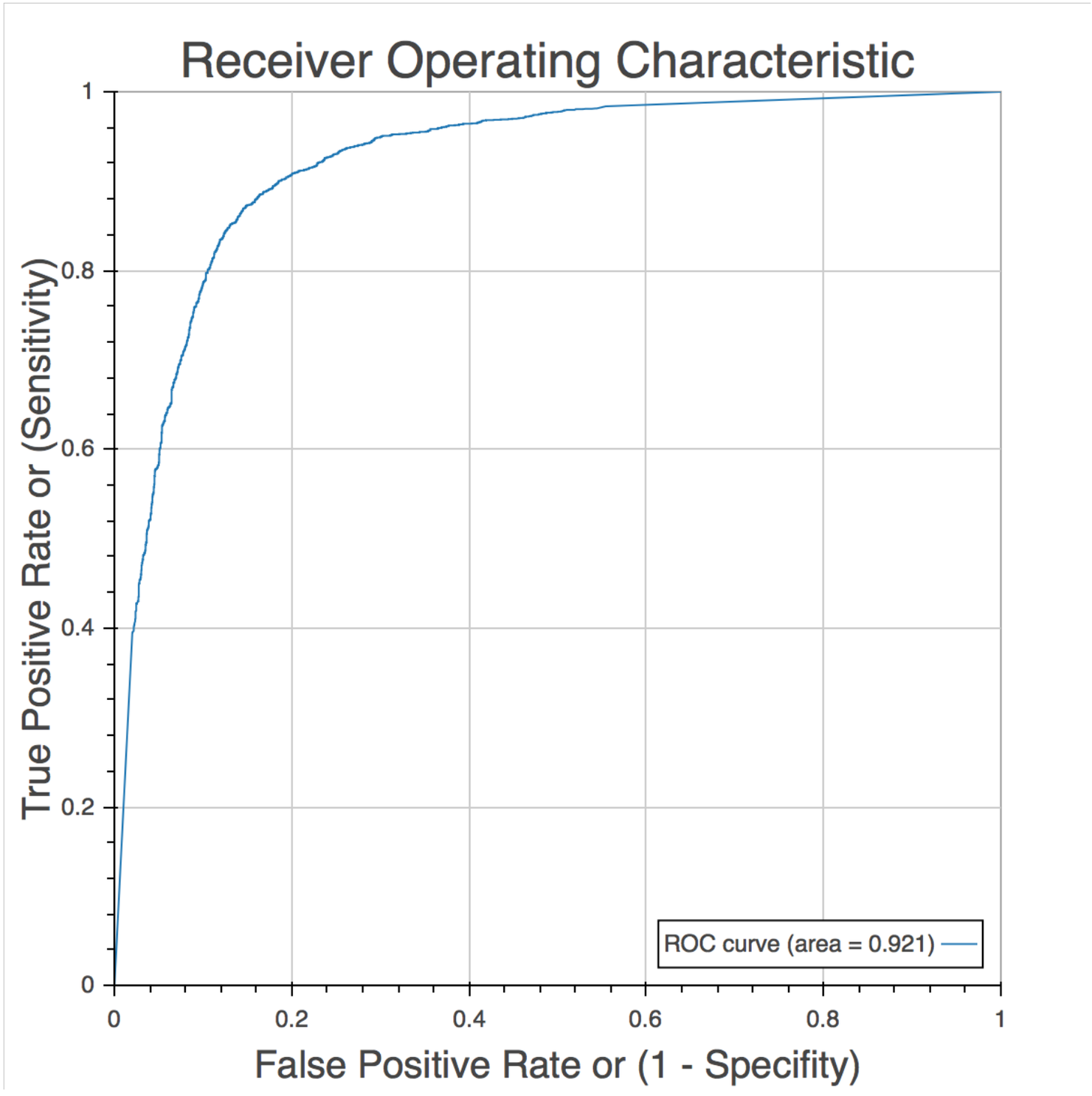
➤ Confusion Matrix

```
Results of Logistic Regression:
                 Predicted Class 0  Predicted Class 1
Actual Class 0                2157                324
Actual Class 1                 275               2244

Precision: 0.873831775701
Recall: 0.890829694323
```

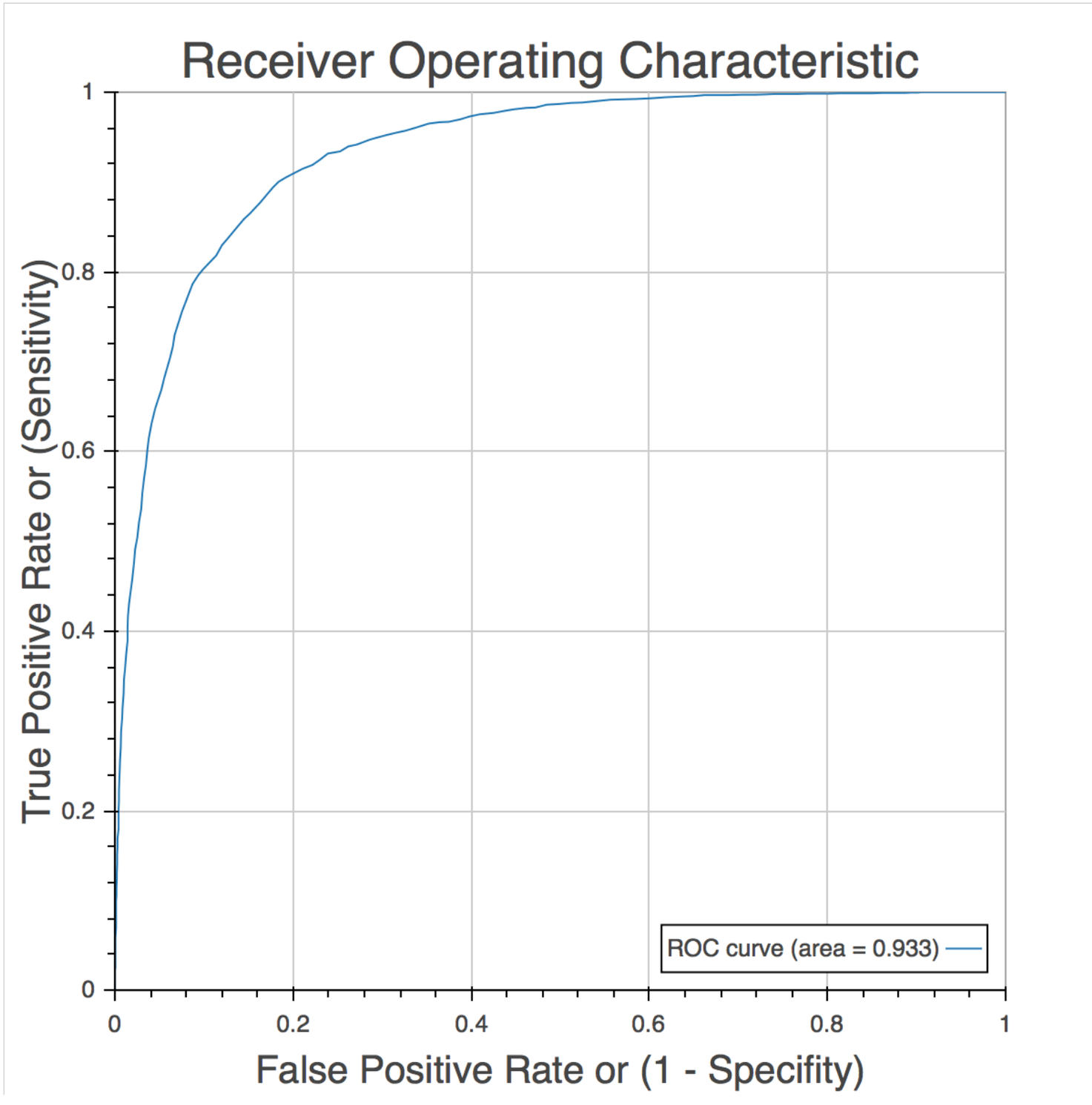
➤ Classification Report

```
              precision    recall  f1-score   support

    Class 0       0.89      0.87      0.88      2481
    Class 1       0.87      0.89      0.88      2519

avg / total       0.88      0.88      0.88      5000
```

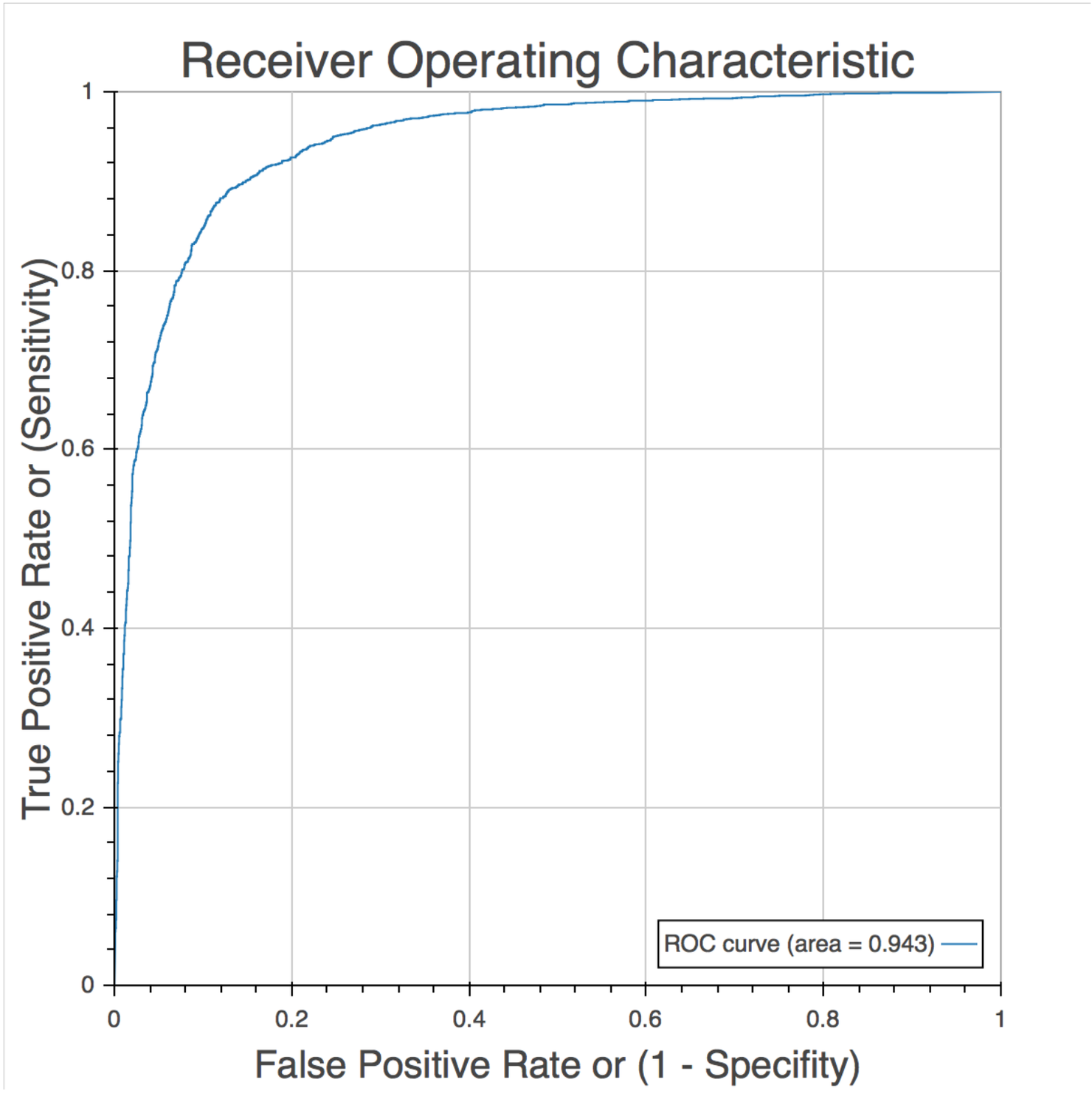
MULTINOMIALNB ROC AND AUC



RANDOM FORESTS ROC AND AUC



LOGISTIC REGRESSION ROC AND AUC



LOGISTIC REGRESSION

Further Exploration

GRID SEARCH

- Originally ran Logistic Regression with default parameters
- What about the parameter C?

```
C : float, optional (default=1.0)
    Inverse of regularization strength; must be a positive float.
    Like in support vector machines, smaller values specify stronger
    regularization.
```

- Ran Grid Search with the following
  - parameters = {'C':[1000, 100, 10, 1, 0.1, 0.01, 0.001, 0.0001]}
  - Kept CV=100 (as originally ran)

GRID SEARCH RESULTS

- Obtained the following Scores:

```
[mean: 0.85985, std: 0.02450, params: {'C': 1000},
 mean: 0.86425, std: 0.02459, params: {'C': 100},
 mean: 0.86995, std: 0.02264, params: {'C': 10},
 mean: 0.87580, std: 0.02253, params: {'C': 1},
 mean: 0.88360, std: 0.02209, params: {'C': 0.1},
 mean: 0.87605, std: 0.02049, params: {'C': 0.01},
 mean: 0.84460, std: 0.02212, params: {'C': 0.001},
 mean: 0.78405, std: 0.02739, params: {'C': 0.0001}]
```

- Best Estimator is

```
LogisticRegression(C=0.1, class_weight=None, dual=False,
fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear',
tol=0.0001,
          verbose=0, warm_start=False)
```

- With Best Score

```
0.8836000000000005
```

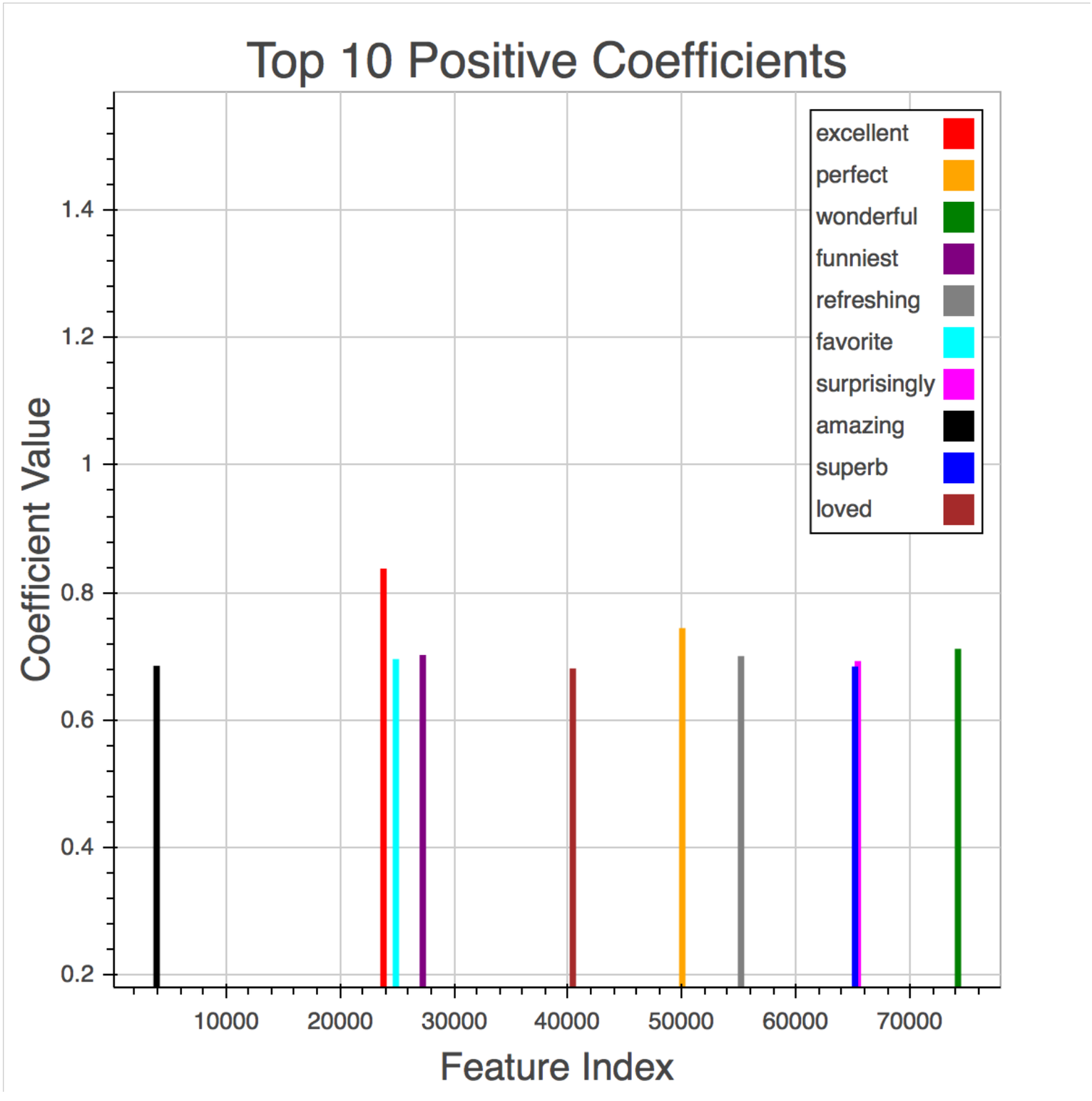
RUN AGAIN MODEL WITH NEW PARAMETER C=0.1

- New Cross-Validation Score of 0.8667
    - Previous was 0.8596
- New Accuracy Score of 0.8844
    - Previous was 0.8802
- New Precision is 0.877186164011
    - Previous was 0.873831775701
- New AUC is 0.949
    - Previous was 0.943

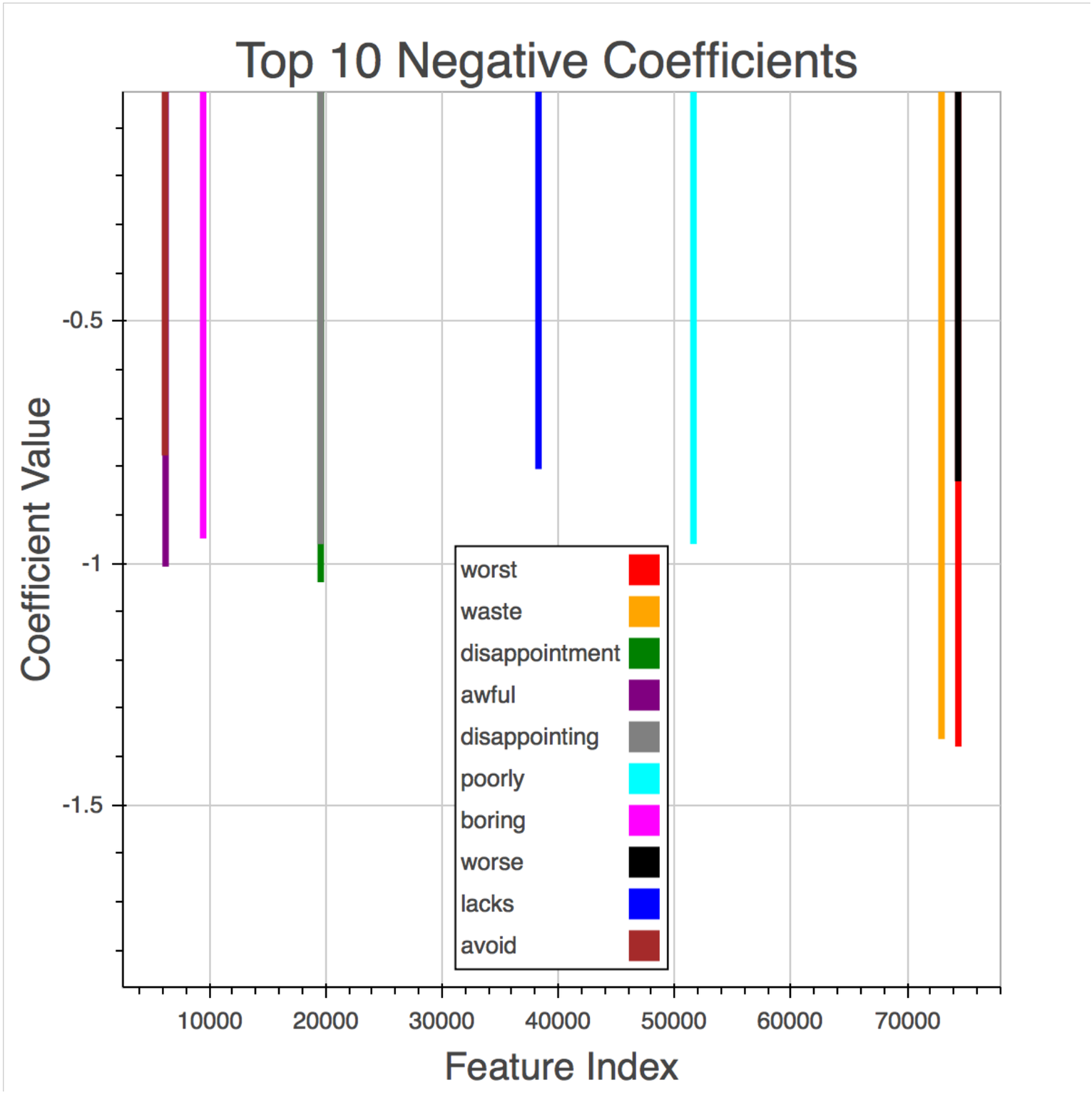
MODEL COEFFICIENTS

- Positive coefficients seem to be related with words contained on positive reviews!!
- Opposite of the above for Negative coefficients
- After vectorising the reviews on train data, I obtained a matrix of (25000, 74539) vectors; being the columns the different words used on all train reviews.
    - How important is each word on generalising review sentiment?

TOP 10 POSITIVE WORDS



TOP 10 NEGATIVE WORDS



MOST FREQUENT TOKENS (FEATURES)

Token ID	Count	Token
8672	101871	br
43990	44047	movie
24445	40159	film
38612	20281	like
35649	17774	just
27960	15147	good
66668	12727	time
63195	11988	story
53637	11738	really
5469	9308	bad

NATURAL LANGUAGE PROCESSING (NLP)

NATURAL LANGUAGE PROCESSING (NLP)

- Used NLTK and BeautifulSoup to further parse the content of all positive and negative reviews.
- Used the most frequent non-meaningful words and add them to the Stopwords

WORD CLOUD FROM GOOD REVIEWS

WORD CLOUD FROM BAD REVIEWS

NLP ANALYSIS

- Concordance of ‘excellent’

```
                                Displaying 25 of 1635 matches:
s.The cinematography speaks of the excellent skills of Josef Werching that acce
ave to say that The Mother does an excellent job of explaining the sexual desir
e . This movie brought to mind the excellent movies that Branagh made with Emma
lm started and ended so well , had excellent acting and writing , it 's hard no
ow , and she and Celeste have been excellent support networks to each other for
 Jim and his bride-to-be all do an excellent job of fitting into stereotypes of
id-fifties.All in all , this is an excellent anime series to watch if you are a
esel . Micheals and Perfect had an excellent match here , but it was Diesel who
 simply because some segments were excellent and covered issues that usually ge
ost movies since . Martin Sheen is excellent , and though Nick Nolte has a smal
has a small part , he too provides excellent support . Vic Morrow as the villai
 elf . First of all , the acting is excellent , especially the leads . Although
o responsible for Ran ) , and some excellent performances from the actors , and
acting , superb cinematography and excellent writing.8 out of 10 , kids . '' ''
enty years ! ) Bill Nighy gives an excellent performance as the off kilter lead
 ed production I have to give it an excellent rating as with the exception with
otherhood . Jennifer Beals does an excellent job as a straight Cinderella , esp
th his scary performance and he 's excellent also but I really thought Trebor s
de of Star Trek 's third season is excellent and a highlight of the much malign
ievable . Spock 's acting here was excellent as Freiberger candidly admitted to
uture ' trilogy , 'Bill and Ted 's Excellent Adventure ' , 'Groundhog Day ' and
 me.In the case of Terry Gilliam 's excellent film , '12 Monkeys ' , it 's hard
tte Hartley and G. D. Spradlin are excellent in their supporting roles . And Pe
 for a topping . The production is excellent and the pacing is fast so it 's ea
er ride . The special effects were excellent and the costumes Uma Thurman wore
```

NLP ANALYSIS

- Concordance of ‘worst’

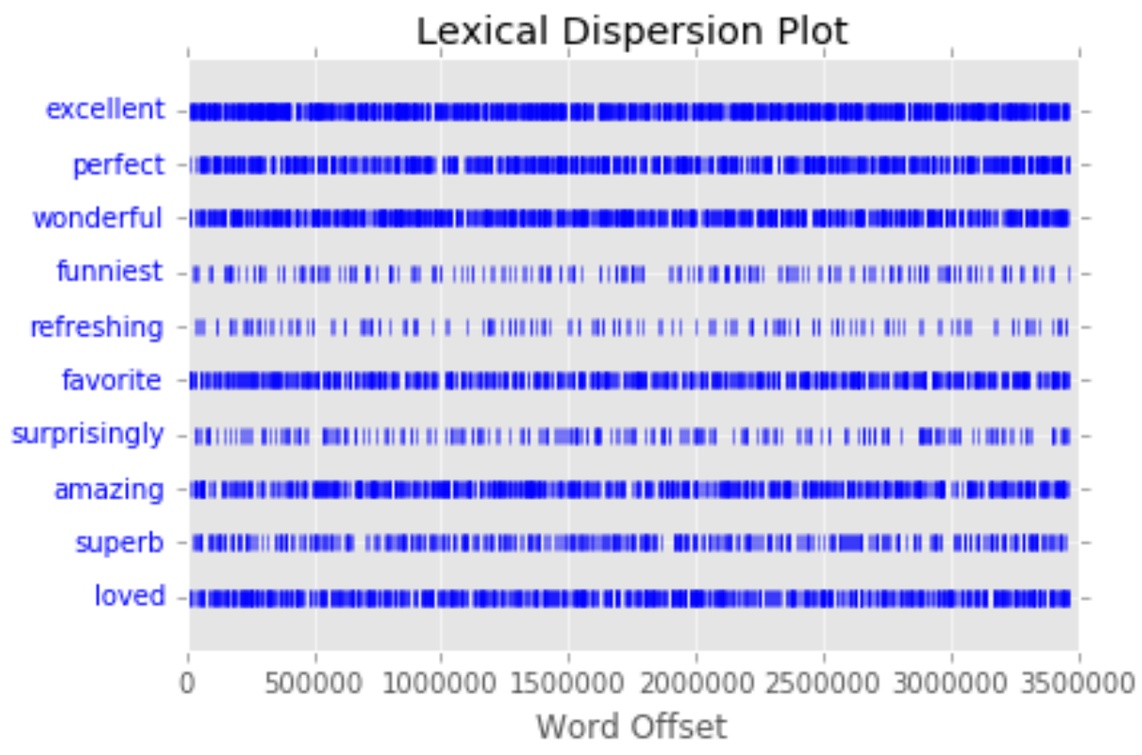
```
                                Displaying 25 of 2430 matches:
s from his mouth he just vomited the worst film of all time . '' '' '' `` `` ''
 way for a while , with the absolute worst element being Midkiff 's worthless p
 big groan-worthy twist at the end . Worst of all is the overlong `Zion Rave '
shreds ! ! ! ! ! \ '' '' That 's the worst part about this movie is , this shar
t BEN AND ARTHUR is quite simply the worst film I have ever sat through in my l
 nt of view it is probably one of the worst films I have encountered absolutely
cast Jeroen Krabbé because he 's the worst actor and every character he 's play
 'm faced with here right ? It is the worst film ever because he 's supposed to
not without competition for title of worst film so it has to sink pretty low to
 . '' '' '' `` `` '' This is n't the worst movie I 've ever seen , but I really
 s just as well they have some of the worst - and not just the human characters
big science fiction.It 's one of the worst movies I have ever seen ... Simply .
'' '' '' This is without a doubt the worst movie I have ever seen . It is not f
ry plain women leads in the film.The worst film I have seen in years , & hopefu
 and killing.This is Ridley Scott 's worst movie in my opinion and there are no
`` `` '' This is probably one of the worst French movies I have seen so far , a
Oscar caliber . But to me the single worst flaw was the total misrepresentation
he head in what is easily one of the worst effects ever . The Vietnam scenes ar
ose to the truth of scriptures . The worst part of it was I really wanted it to
 `` '' I can not say that Aag is the worst Bollywood film ever made , because I
 all compatible with that song . 6 ) Worst of all : Not only does the movie sho
 seen this dire comedy is by far the worst . 3/10 '' '' '' '' '' '' A patient e
 vorite type ) This film is among the worst . For me , an idea drives a movie .
tally lame.This got to be the second worst movie Lindsey is ever in since Confe
 , but this one is up there with the worst of them . Plot troubling to follow .
```

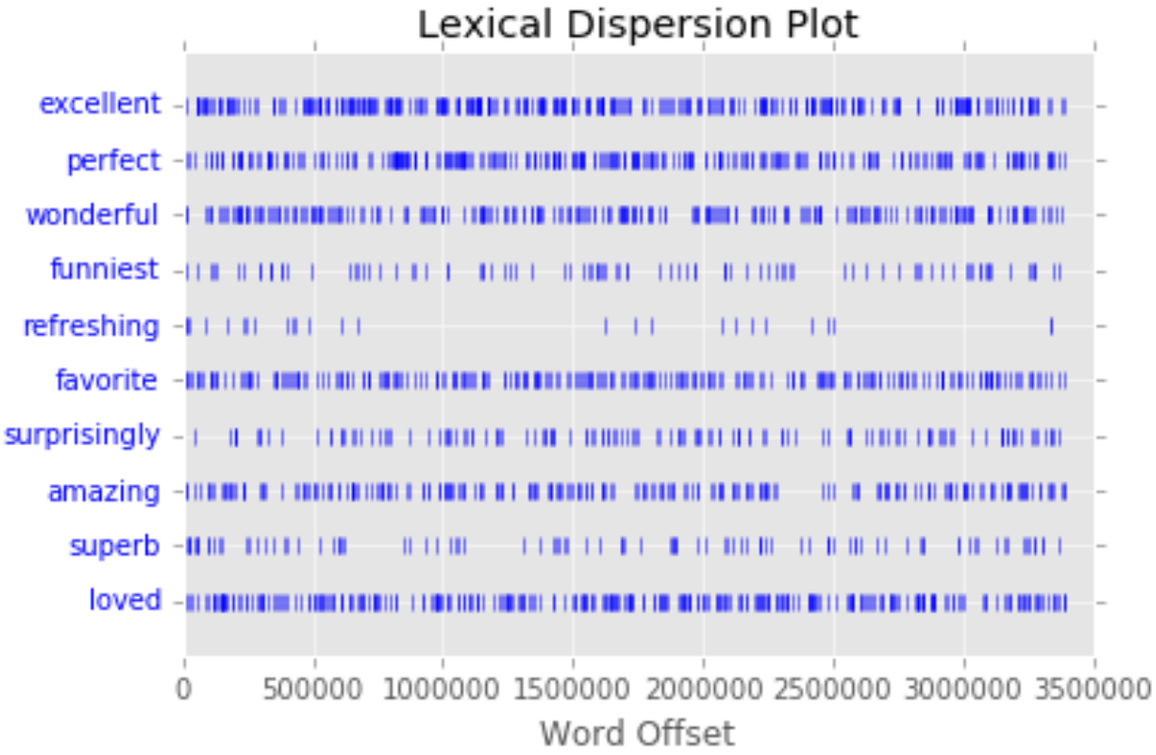
NLP ANALYSIS

- Are longer reviews more positive or negative?
    - There are 134715 sentences on all positive reviews
        - 10.7772 sentences per positive review (average)
    - There are 128369 sentences on all negative reviews
        - 10.2695 sentences per negative review (average)
- Is sophistication of the language related?
    - Positive Reviews have a lexical dispersion of 0.0313
    - Negative Reviews have a lexical dispersion of 0.0318

NLP ANALYSIS

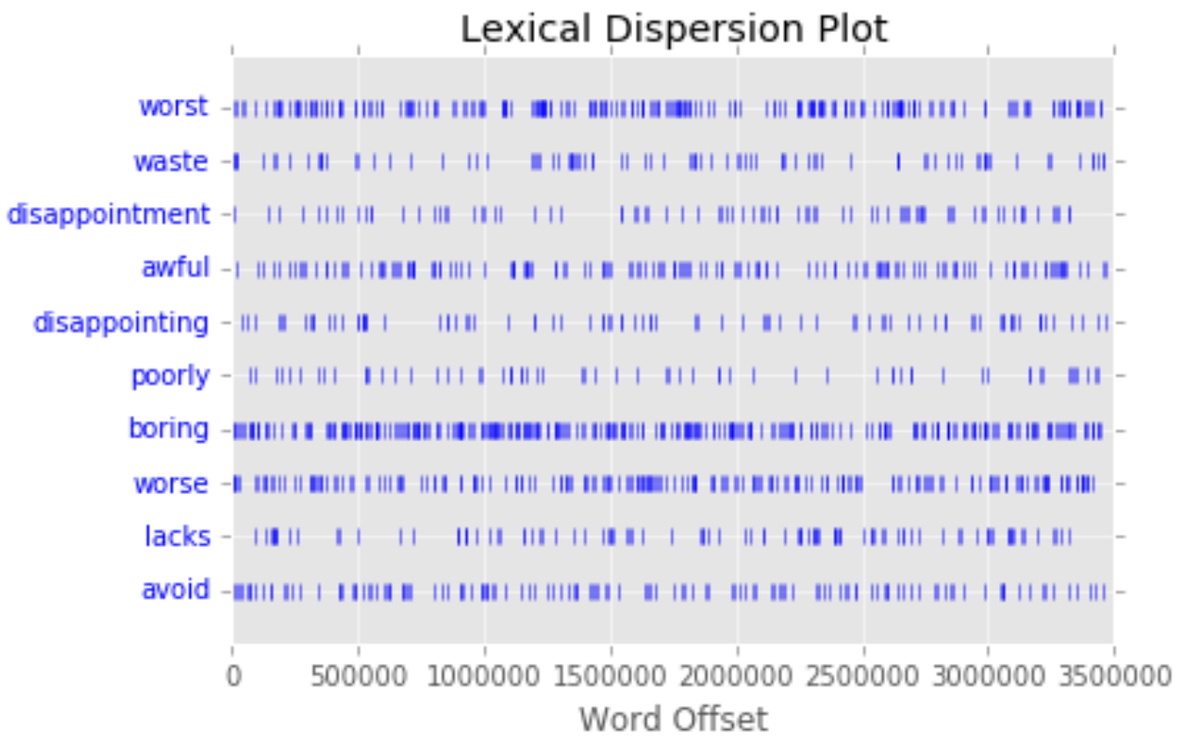
- How disperse are our Top 10 Positive Words on training reviews

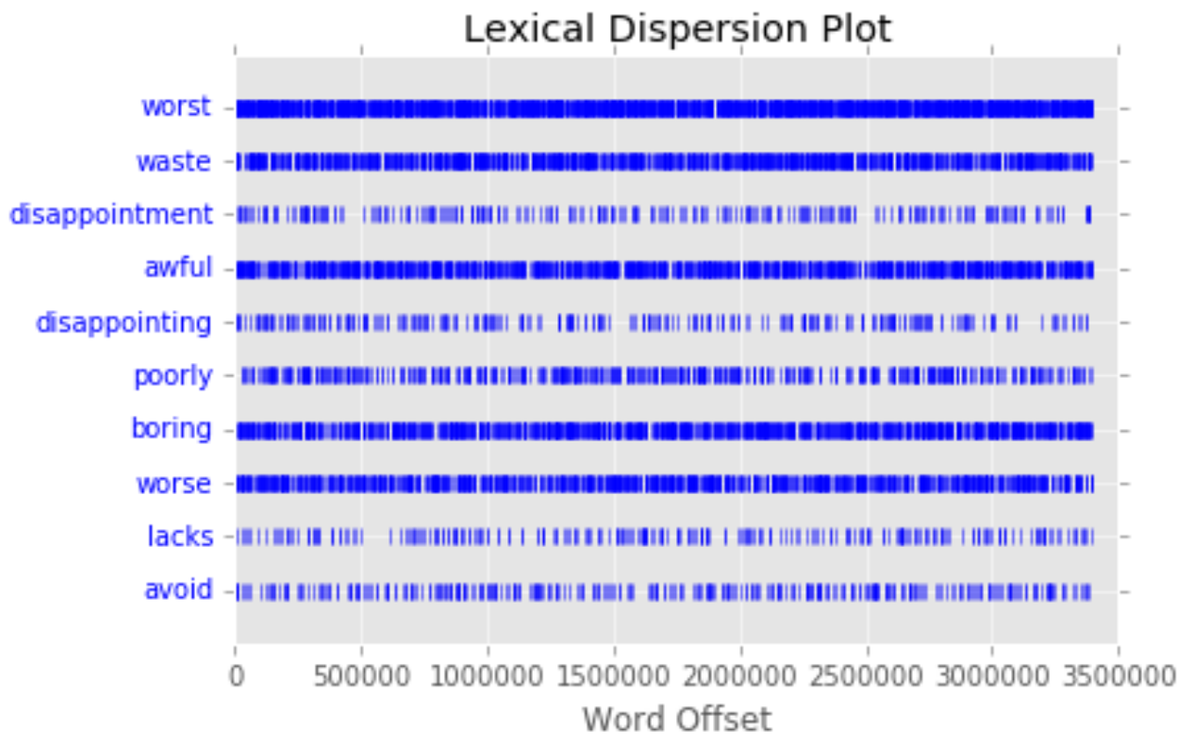




NLP ANALYSIS

- How disperse are our Top 10 Negative Words on training reviews





CLUSTERING ANALYSIS

Better called Attempt instead

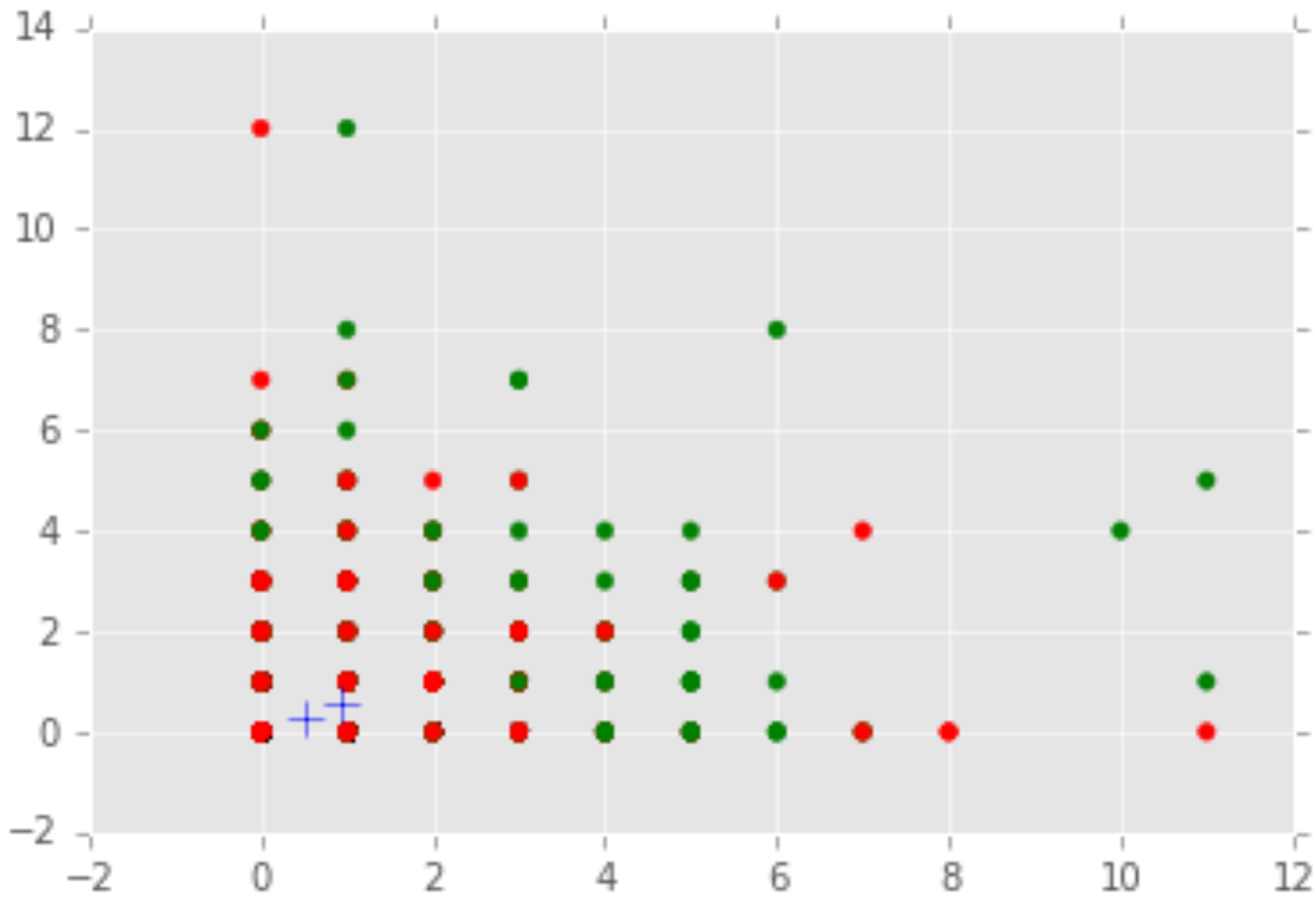
K-MEANS CLUSTER

- On the data set there are 25000 movie reviews (test data) that are not labeled
    - Up until now the test data used was produced by train_test_split of the original train data (25000 labeled reviews)

- Attempted K-Means analysis using same Count Vectorizer
    - **init='k-means++'** and **n_clusters=2**
        - It took very long to do something!

K-MEANS CLUSTER

- Reduced the Count Vectoriser to process 5000 reviews
- Obtained some results, but could not make sense of them



PREDICTING SENTIMENT

On Unlabelled OOS data

LOGISTIC REGRESSION MODEL BUILT

- The model built is based on a Count Vectoriser obtained from the train_test_split based on the original Train Data
    - It contains up to 74539 features
- Original OOS data is unlabelled (same as Test Data)
- Count Vectoriser on Test Data produces 102751 features
- When attempting a **predict** on Test Data on the **built and improved Logistic Regression model**, obtained error:

```
ValueError: X has 102751 features per sample; expecting 74539
```

- So what do do now? (Panic in Sunday afternoon!)

PIPELINE

- Investigated and researched about Pipeline usage
- Prepared a Pipeline Classifier using:
    - Same Count Vectoriser parameters as before
    - Logistic Regression optimised
- Trained the Pipeline with the train_test_split
- Obtained accuracy score of 0.8912
    - Logistic Regression optimised accuracy is 0.8844
- Obtained Cross Validation Score of 0.8588
    - Optimised Logistic Regression Cross Val Score is 0.8667

OPTIMISED LOGISTIC REGRESSION

➤ Confusion Matrix

```
Results of Logistic Regression:
                 Predicted Class 0  Predicted Class 1
Actual Class 0                2165                316
Actual Class 1                 262               2257

Precision: 0.877186164011
Recall: 0.89599047241
```

➤ Classification Report

```
              precision    recall  f1-score   support

    Class 0       0.89      0.87      0.88      2481
    Class 1       0.88      0.90      0.89      2519

avg / total       0.88      0.88      0.88      5000
```

PIPELINE

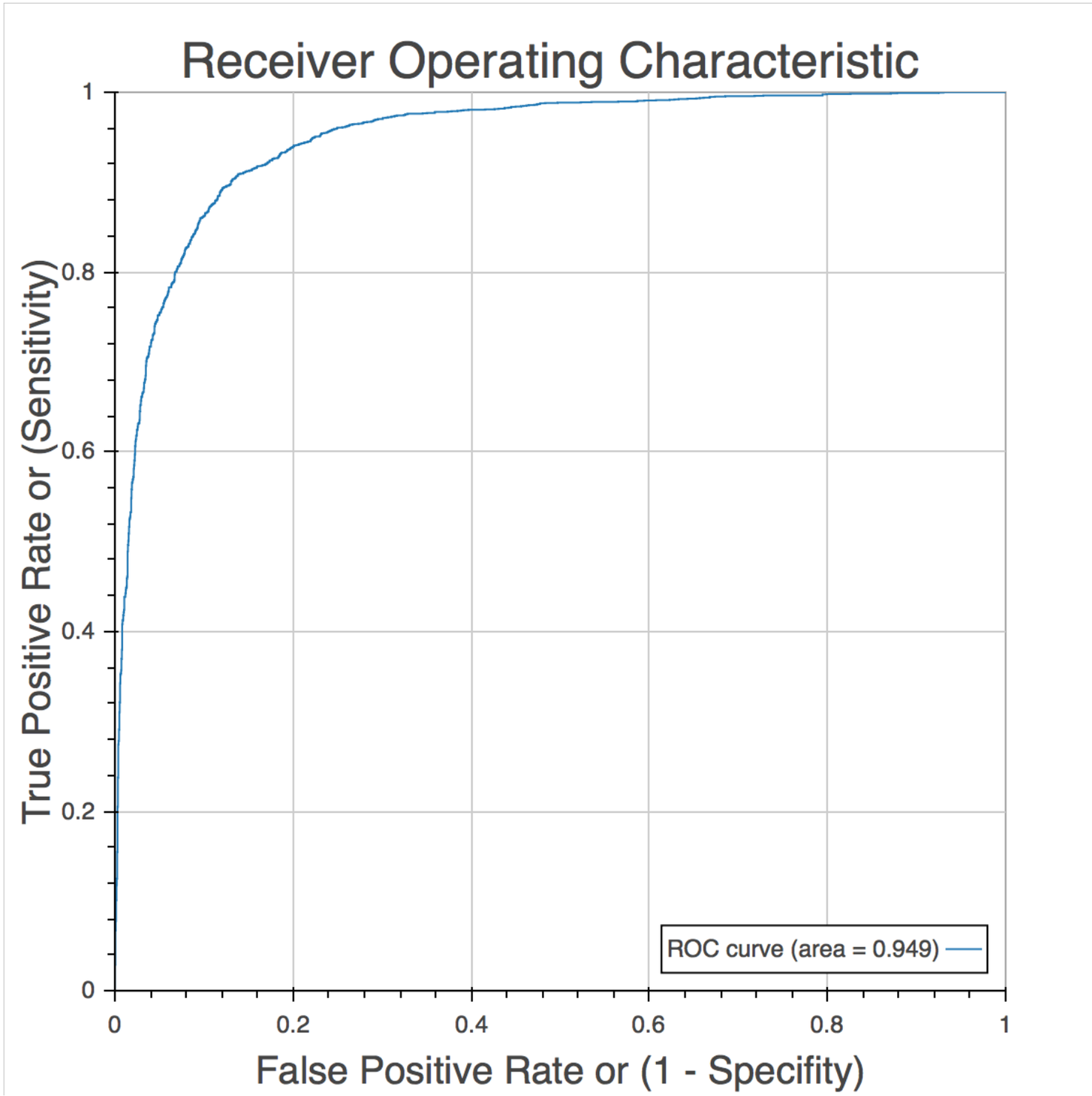
➤ Confusion Matrix

```
Results of Pipeline:
                 Predicted Class 0  Predicted Class 1
Actual Class 0                2190                291
Actual Class 1                 253               2266

Precision: 0.886194759484
Recall: 0.899563318777
```

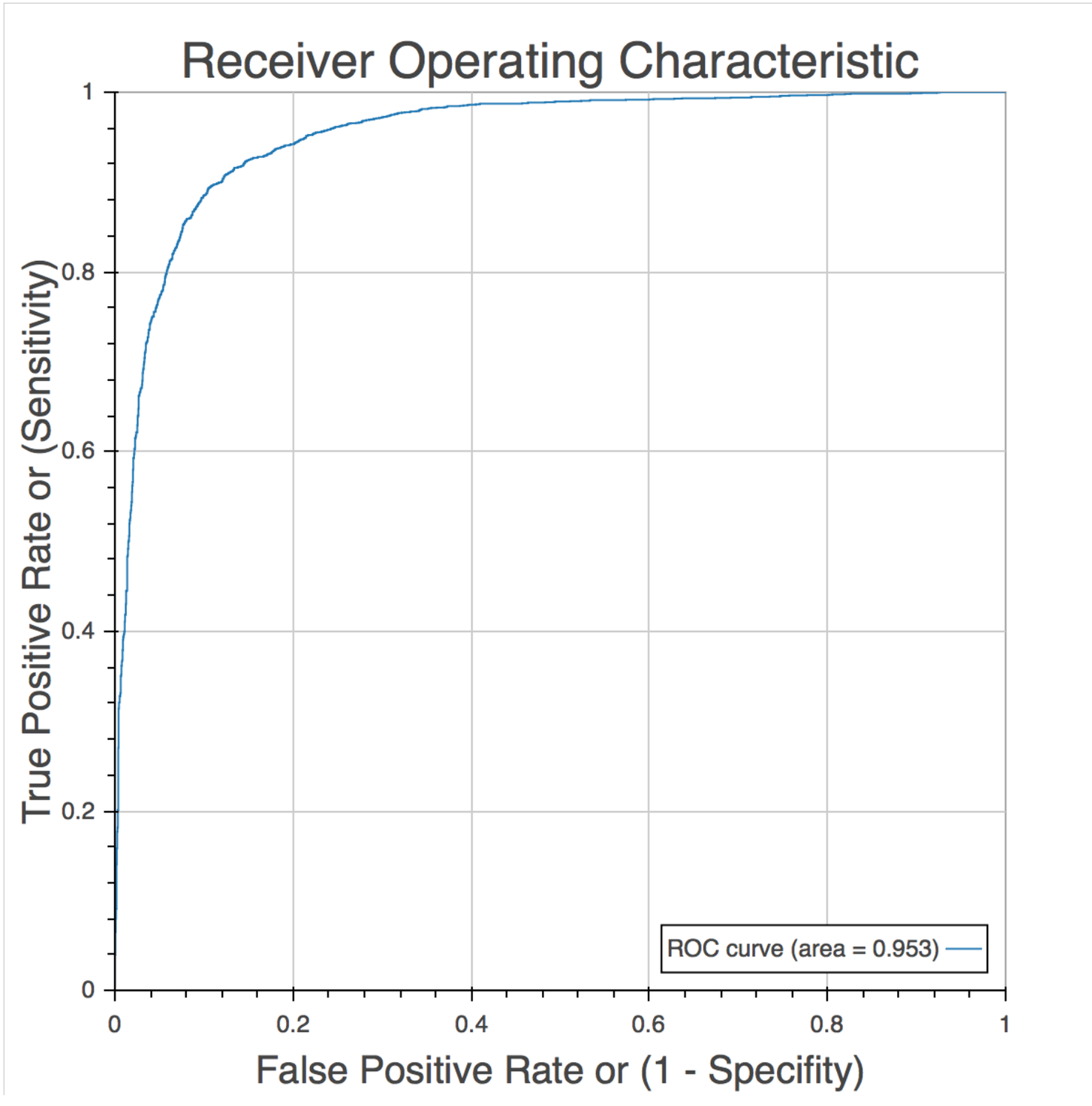
➤ Classification Report

```
              precision    recall  f1-score   support

    Class 0       0.90      0.88      0.89      2481
    Class 1       0.89      0.90      0.89      2519

avg / total       0.89      0.89      0.89      5000
```

OPTIMISED LOGISTIC REGRESSION ROC AND AUC



PIPELINE ROC AND AUC

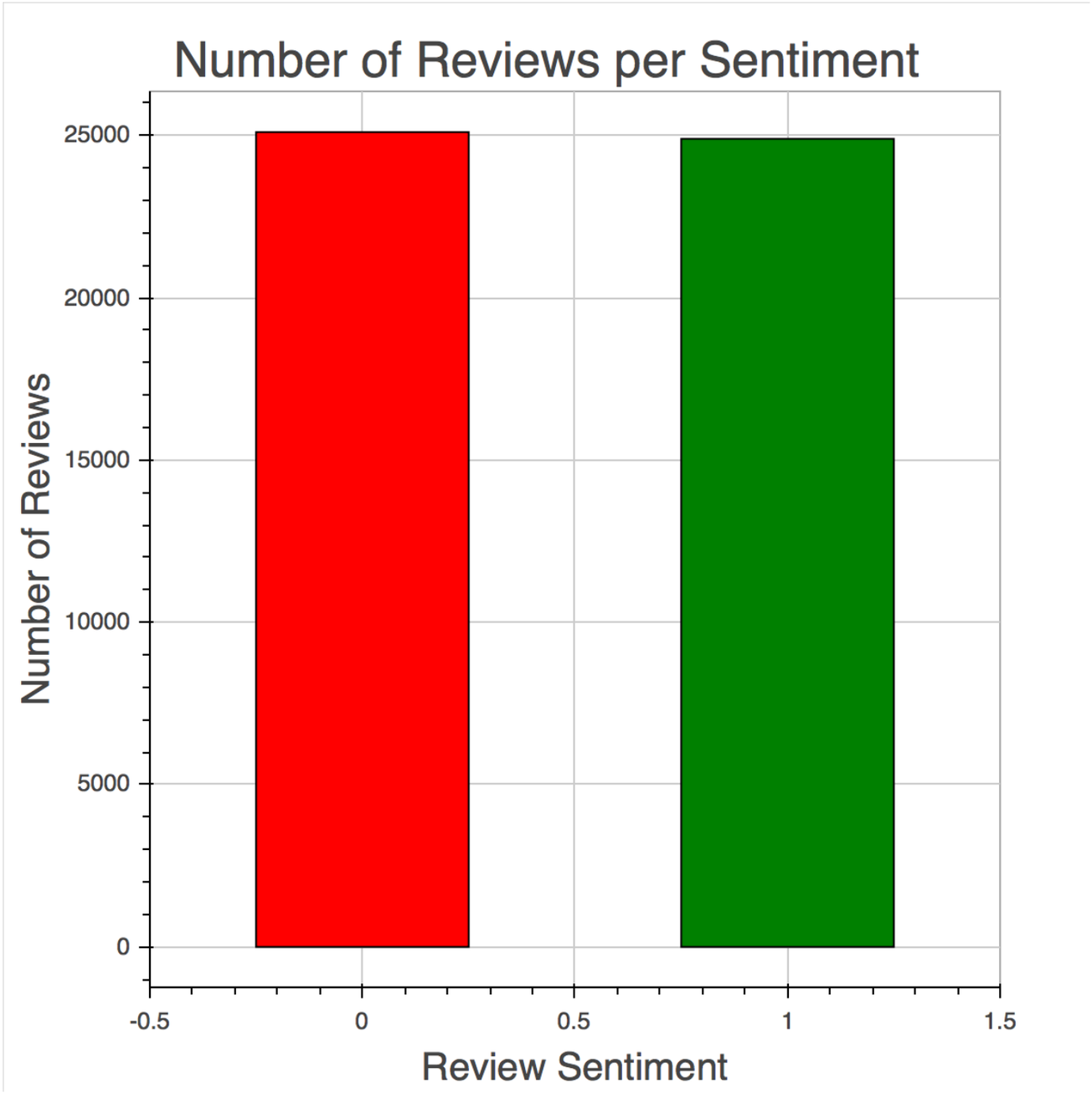


LET'S PREDICT ON OOS DATA!

Using Pipeline

CLASSIFICATION RESULTS

- 24896 Positive Reviews 25104 and Negative Reviews



LET'S HAVE A LOOK AT SOME OF THE PREDICTIONS

➤ Classified as Negative Review

```
'"Incredibly dumb and utterly predictable story of a rich teen girl
who, not given love by her parents, starts a girl gang. They rob gas
stations, rape guys (!!!) and kill a policeman.<br /><br />All the \
\"teenagers\\" in this film are easily in their late 20s/early 30s,
the acting is all horrible and the script has every cliche imaginable
with hilarious dialogue--it comes as no surprise that it was written
by the immortal Ed Wood Jr.!<br /><br />Worth seeing for laughs. Best
lines--\\"They\'re shooting back!\\" and \\"It ain\'t supposed to be
like this.\\""'
```

LET'S HAVE A LOOK AT SOME OF THE PREDICTIONS

➤ Classified as Positive Review

'"A well made, gritty science fiction movie, it could be lost among hundreds of other similar movies, but it has several strong points to keep it near the top. For one, the writing and directing is very solid, and it manages for the most part to avoid many sci-fi cliches, though not all of them. It does a good job of keeping you in suspense, and the landscape and look of the movie will appeal to sci-fi fans.

If you\'re looking for a masterpiece, this isn\'t it. But if you\'re looking for good old fashioned post-apoc, gritty future in space sci-fi, with good suspense and special effects, then this is the movie for you. Thoroughly enjoyable, and a good ending.

"'

CONCLUSIONS

- Developed a predictive classification model able to categorise movie reviews into positives or negatives based on their text content.
- Studied NLP futures confirming the reasons why the model selected the top important features.
- Learned about Count Vectoriser and Pipeline usage.
- Model can be improved with more work on parsing the reviews and other exploration ideas :)

QUESTIONS?

Please be gentle :)