The Reviews Are In

Text & Sentiment Analysis Of Movie Reviews

Motivation

r for someone to "win," but also hope
eone to kiss in the end?

Comment Share

Smith

ere more light saber battles. I loved the
ted that there were only two light saber
xt movie should be good.

Comment Share

Hong
utes ago ·

st of the Academy Award contenders
t would be The Wolf of Wall Street. 12
is a must-see.

Comment Share

Meagher

e best Disney animated movie, either
without it, since Toy Story 3 in 2010.

Dillon Meagher
4 hrs ·

I like my “alone in space” movies to be far weirder and more cerebral than The Martian. Scott or Damon can stay on Mars; fly me to Sam Rockwell and the Moon or directly into the Sun with Danny Boyle.

15 Likes

Like Comment Share

Rebecca Meagher
1 hr ·

Though not at all innovative or original, is at least a worthwhile action fantasy. Plus, JLaw is in it #MayTheOddsBeInYourFavor

2 Likes

Like Comment Share

Haley Huang
10 minutes ago ·

The movie is oddly blithe about its central premise. How can this movie set up this monstrous society and this brutal game of genocide and expect us to not only cheer for someone to "win," but also hope she finds someone to kiss in the end?

4 Likes

Like Comment Share

Taylor Swift
5 minutes ago ·

OMG, Twilight is like the BEST movie ev
acting is sooooo bad. It is difficult to wa

2.5k Likes

Like Comment

Tom Meagher
2 hrs ·

Zootopia is the best Disney animated m
with Pixar or without it, since Toy Story

2 Likes

Like Comment

Tyler Reese
2 hrs ·

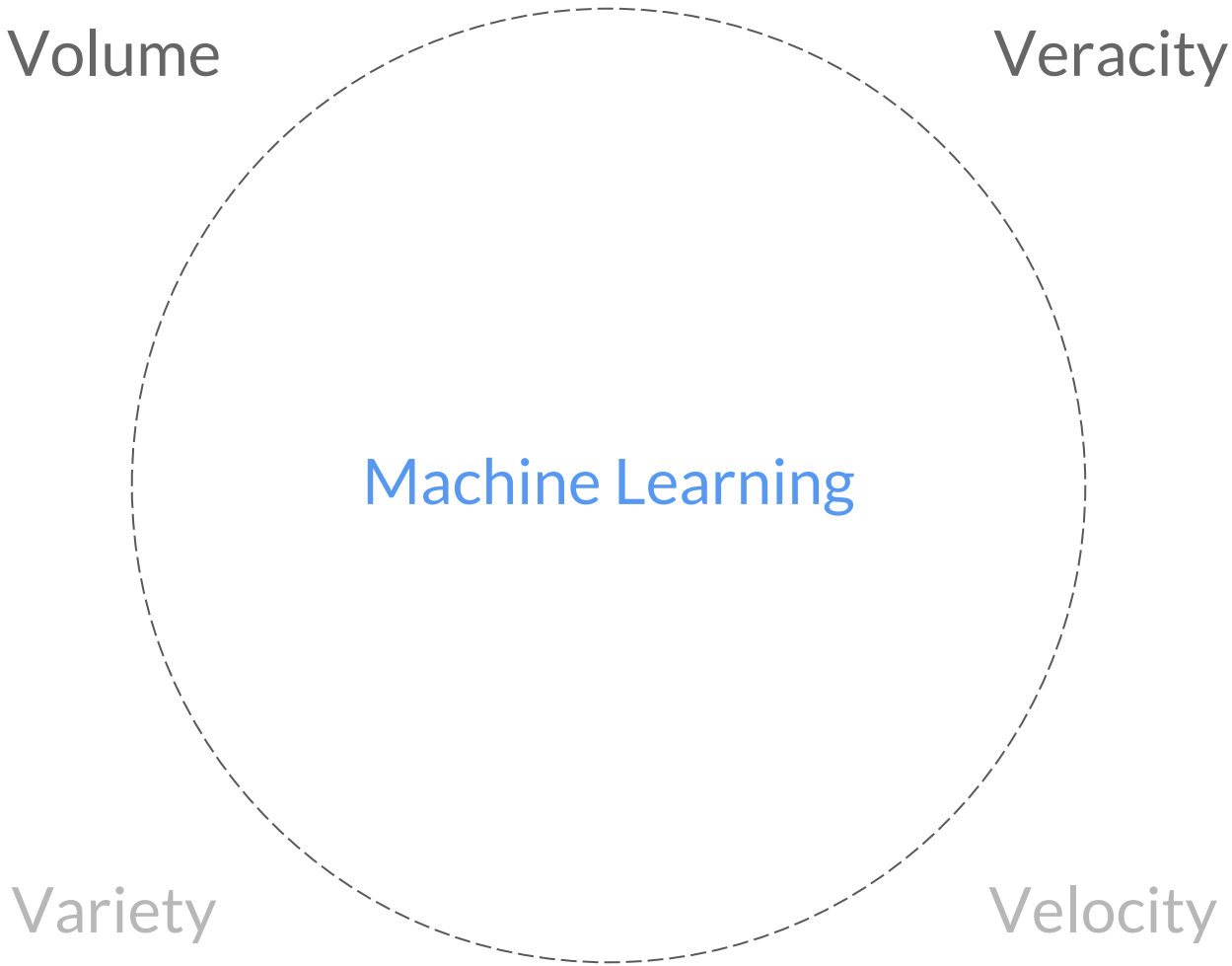
I hate when someone recounts a story i
was not always present, but even if the
like witnessing the sweet revenge of a b
or victim?), this intense movie of evokin
grows even more compelling when sho
lengths that people can go to survive a
ordeal.

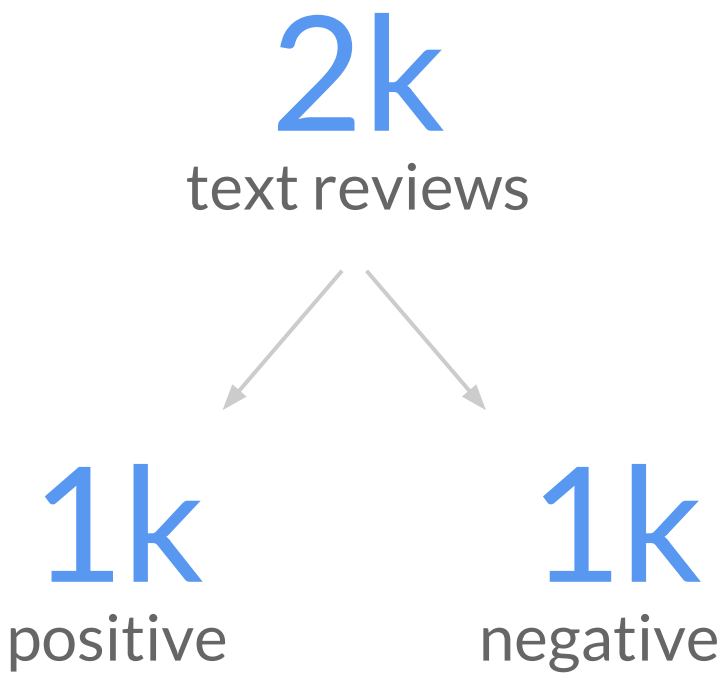
2 Likes

Like Comment

“

Comprehension of this unstructured text informs our decisions.


Volume
Veracity
Machine Learning
Variety
Velocity


2k
text reviews
1k
positive
1k
negative

Objectives

1. Preliminary sentiment analysis on movie reviews

Methodology

- Randomly split 2k movie reviews into two groups

- Built vectorizer-classifier pipeline (TfidVectorizer)
  - Filtered out rare or too frequent tokens
  - Fit Linear Support Vector Classifier with relatively high penalty
- Determined grid search token set for text files
  - words (1-grams) or words and word pairs (1- and 2-grams)
- Performed grid search cross-validation

Results

Grid Search CV scores

n-gram Range	Score
(1,1)	0.82533
(1,2)	0.84733

On **training data**, the linear **SVC** pipeline is **more accurate** when it **considers** both **words** and **pairs** of words.

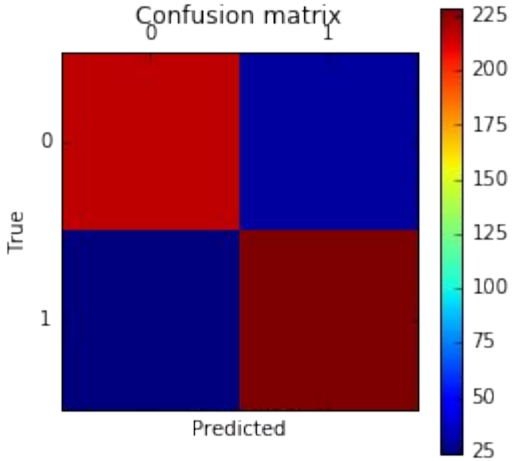
Classification Report

Class	Precision	Recall	f1-Score	Support
Negative	0.84	0.87	0.58	247
Positive	0.87	0.84	0.85	253

Results (continued)

- Number of false negatives and false positives are both small compared to the number of true positives and negatives.
- Model performed quite well on our test data set.
- Test accuracy ~89%

Confusion Matrix, Grid Search SVC



[[217 30]
[24 229]]

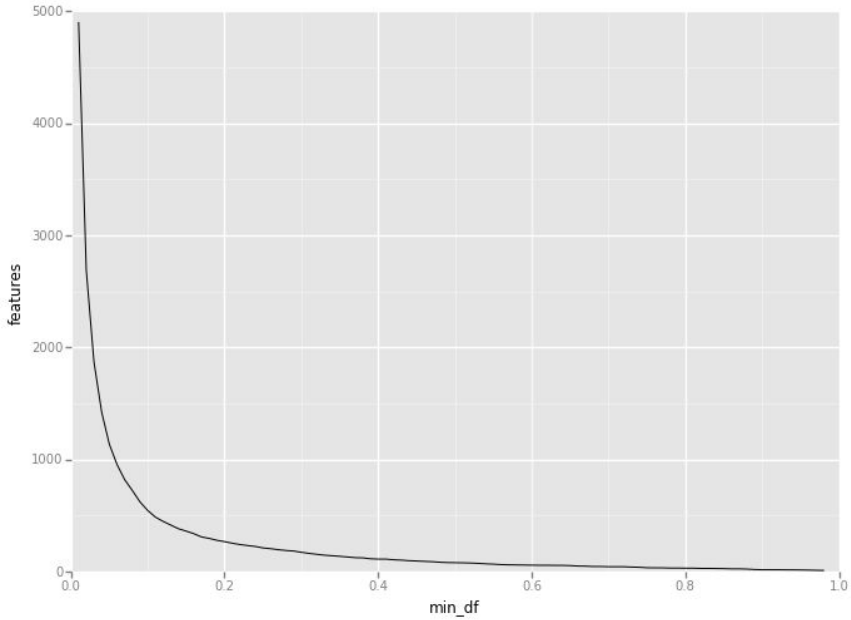
2. Explore the scikit-learn TfidVectorizer class

Methodology

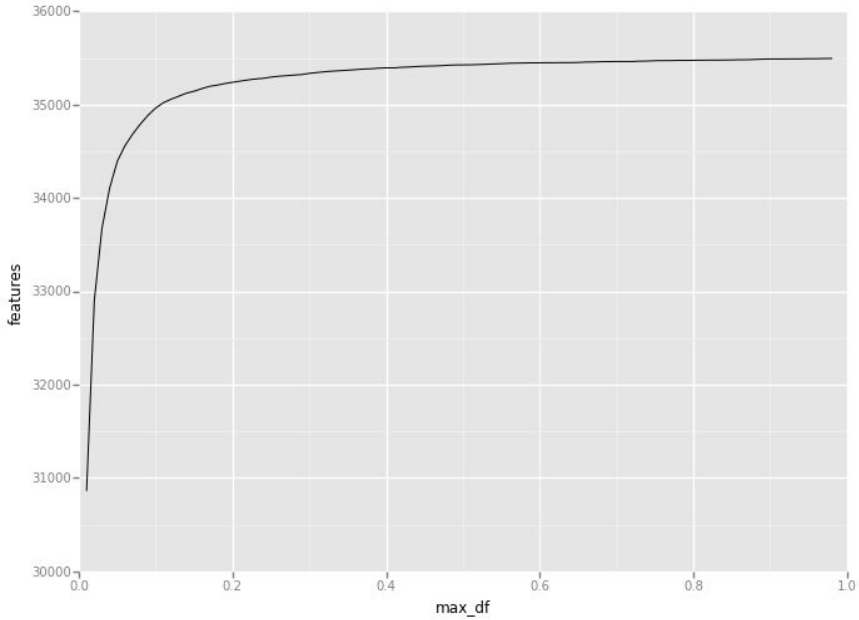
- Define the term frequency-inverse document frequency (TF-ID) statistic.
  - measures how important a word is to a document in a particular collection of documents
- Run the TfidVectorizer class on the training data.
  - *min_df*: filter terms with lower frequency
  - *max-df*: filter terms with greater frequency; stop words.
  - *n-gram range*: how many n-gram words are to considered

Results

Min_df vs. TfidVectorizer Features



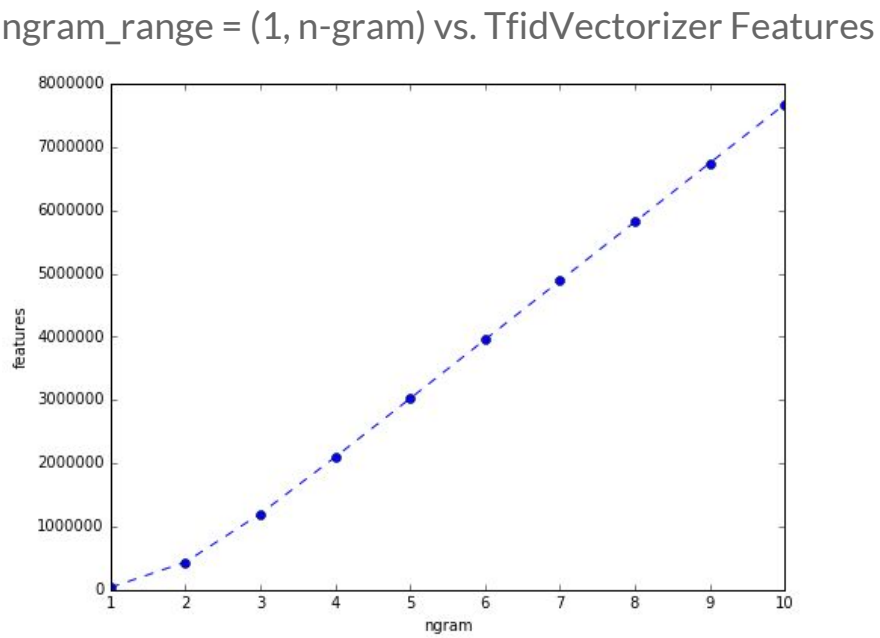
Max_df vs. TfidVectorizer Features



- Relationship between number of features in our vocabulary and values of *min_df* and *max_df*
- Vocabulary size inversely related to *min_df*, directly related to *max_df*

Results

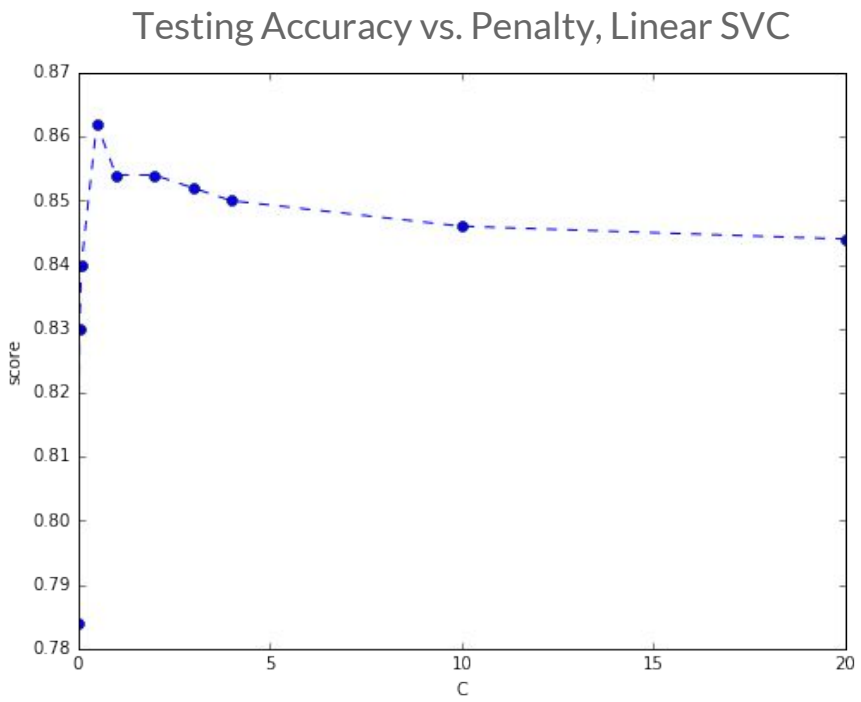
- Number of features in TdifVectorizer vocabulary increases as n-gram is increased in the form (1 , n-gram)
- Growth appears to be roughly linear.



3. Machine Learning Algorithms

Methodology

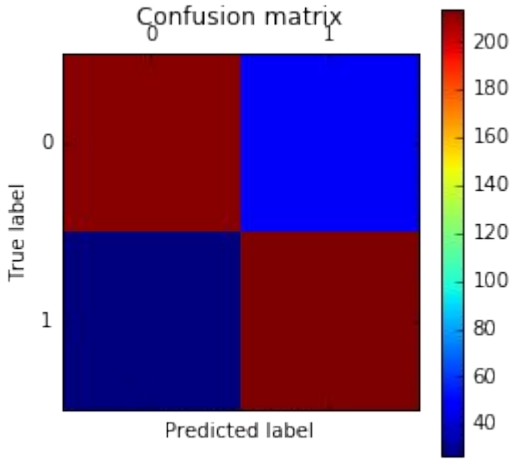
- Linear Support Vector Classifier (SVC)
  - *penalty* parameter ({0.01, 0.05, 0.1, 0.5, 1 , 2, 3, 4, 10, 20})



Results

- Comparatively small number of false positives and negatives against a much larger amount of true positives and negatives
- SVC does well predicting review polarity (test accuracy ~85%)

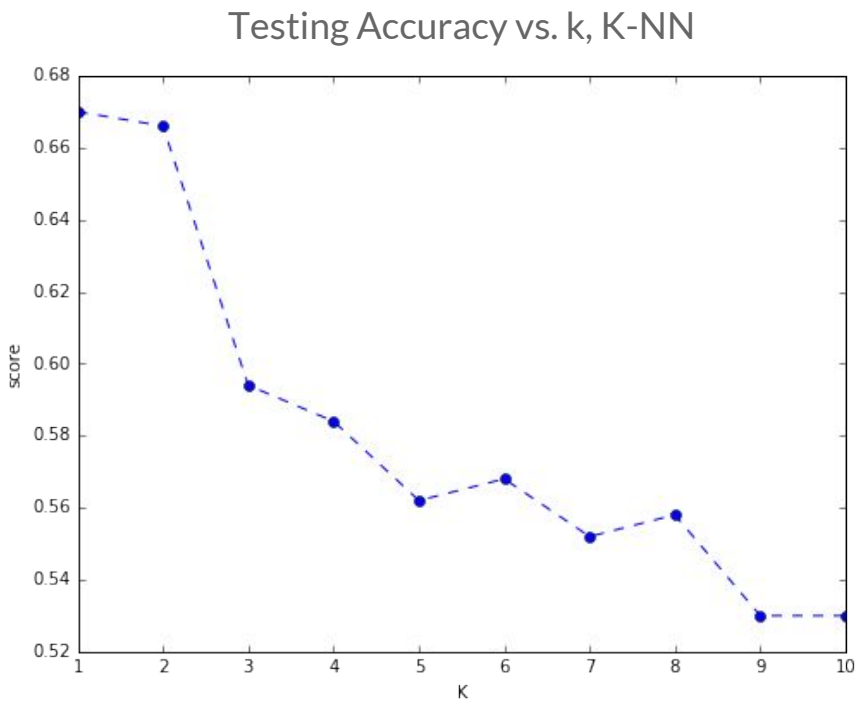
Confusion Matrix, Linear SVC model with C = 10



[[212 47]
[27 214]]

Methodology

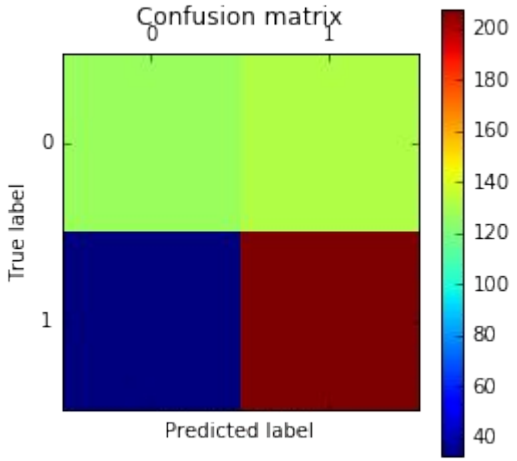
- K-Nearest Neighbors
  - *neighbor* parameter, k({1, 2, 3, 4, 5, 6, 7, 8, 9, 10})



Results

- Model has more false positives than it does true negatives
  - Predicted that more than half of (actually) negative reviews were positive
- Test accuracy ~67%

Confusion Matrix, K-NN with k = 1



[[127 132]
[33 208]]

Conclusion

- SVC: Little or no penalty models are more lenient to choosing linear separators, which misclassify training points.
- K-NN: As *k*, number of neighbors considered increases, testing accuracy decreases
- Linear SVC classifier performed much better than K-NN
  - worst SVC testing accuracy: ~83%
  - best K-NN testing accuracy: ~67%

4. Finding the *right* plot

Do negative reviews always end with a short final sentence?

Do positive reviews always begin with long, glowing sentences?

Do confused viewers (writing negative reviews) ask more questions?

Do positive reviews use more words in general?

Methodology

- **Conjecture** The polarity (positive/negative tendencies) of a given text can be determined based on the *structure* of the text itself.
- Structure-based predictors can be computed for all reviews (regardless of their word content)
  - Restricted to a much smaller, relevant number
  - Easier to interpret

Methodology (continued)

- New predictor set

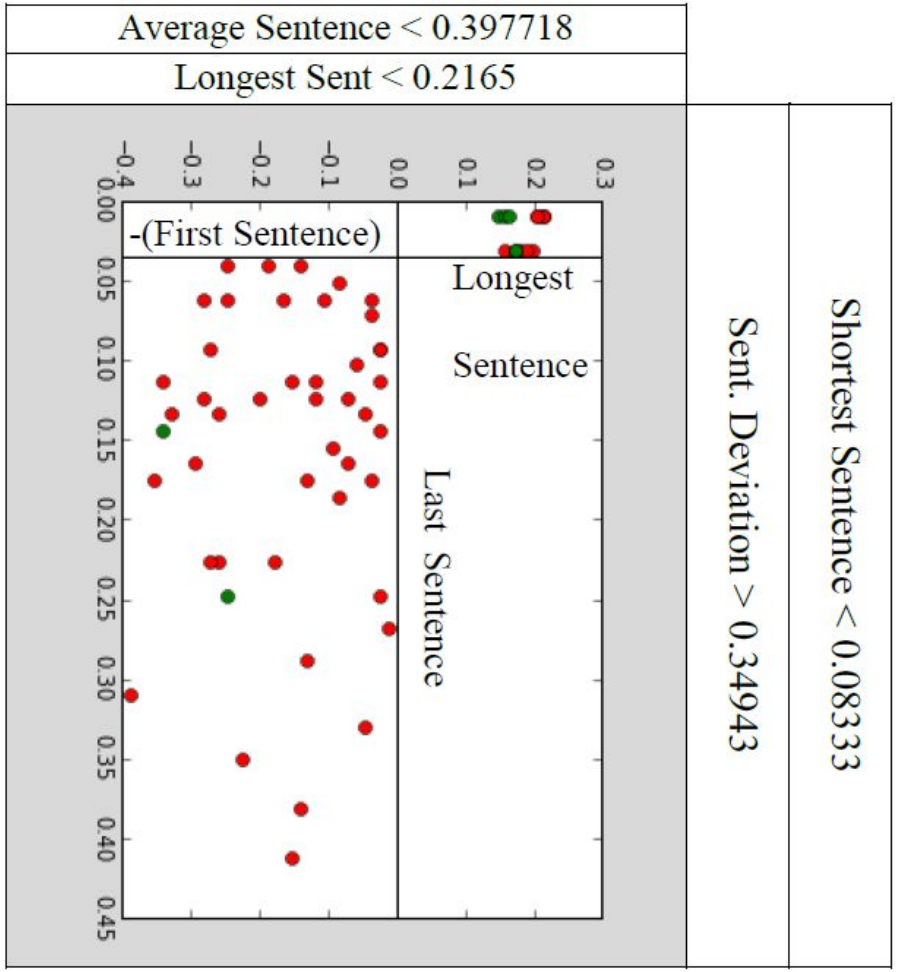
Total Words	First Sentence Length	Number of Contractions
Total Sentences	Longest Sentence	Number of Negative Prefixes
Number of “not” Contractions	Shortest Sentence	Total “You”
Total Number of “Not”	Sentence Deviation	Closest “Not”
Last Sentence Length	Number of Punctuation	

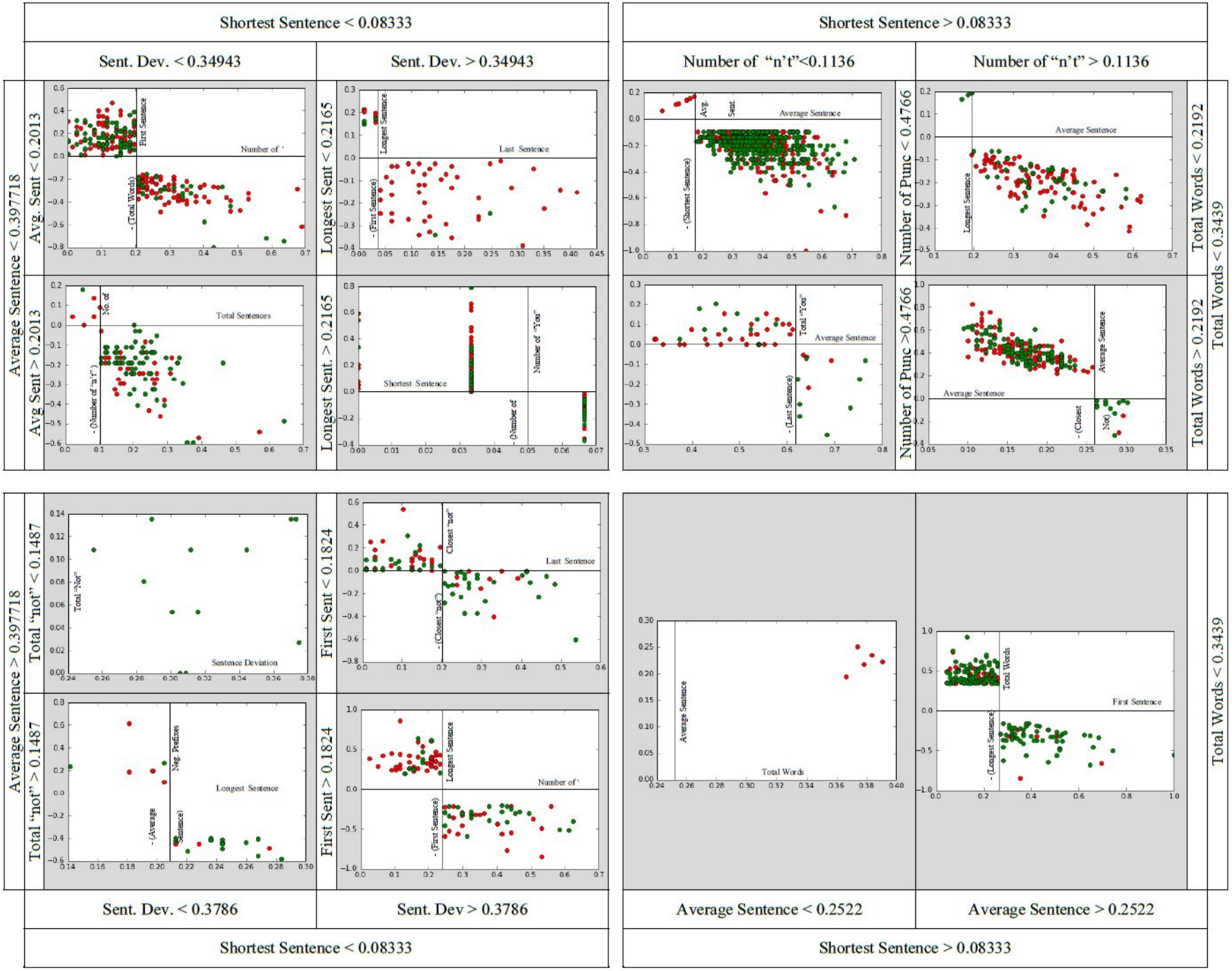
- Performed *Kernel PCA*, using scikit-learn kernel functions
  - linear, radial basis, and cosine

Methodology (continued)

- Decision trees are a very flexible, and widely used, classification method
- There may be subsets of the predictor space where the data is separable
- **Conjecture** Can we use a high-level decision tree to build a sequence of scatter plots which separate the data?

Results


Average Sentence < 0.397718
Longest Sent < 0.2165
Shortest Sentence < 0.08333
Sent. Deviation > 0.34943
-(First Sentence)
Longest Sentence
Last Sentence
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45


Shortest Sentence < 0.08333
Shortest Sentence > 0.08333
Sent. Dev. < 0.34943
Sent. Dev. > 0.34943
Number of "n't"<0.1136
Number of "n't" > 0.1136
Average Sentence < 0.397718
Avg. Sent < 0.2013
Avg Sent > 0.2013
Longest Sent < 0.2165
Longest Sent. > 0.2165
Number of Punc < 0.4766
Number of Punc >0.4766
Total Words < 0.2192
Total Words > 0.2192
Total Words < 0.3439
Average Sentence > 0.397718
Total "not" < 0.1487
Total "not" > 0.1487
First Sent < 0.1824
First Sent > 0.1824
Sent. Dev. < 0.3786
Sent. Dev > 0.3786
Average Sentence < 0.2522
Average Sentence > 0.2522
Shortest Sentence < 0.08333
Shortest Sentence > 0.08333

Conclusion

- Somewhat successful, yet convoluted
- Some regions do an excellent job of separating the data, while others fail
- Whole data set scatter plots were always sufficiently “mixed up”

Business Intelligence & Decision Making

Business Intelligence & Decision Making

- Sentiment can be extremely valuable to movie studios.
  - Determine demographic performance
  - Make advertising decisions
- Future re-release of a film
- Movie studios can collect (and analyze) similar data for the movies of *other* studios.

Overall Conclusions

Overall Conclusions

- Conjectured review polarity could be determined based on the test structure
  - Further analysis required to affirm
- Did not find anything surprising in the data
  - Other than how difficult textual analysis can be
- Further, computationally-expensive analysis could reveal surprising trends

Questions

Helen, Haley, Tom, Tyler