

1. Done.
2. Done.
3. Baseline Model Evaluation Metrics:

Accuracy 0.7138

Kappa 0.4277

4. Done.
- 5.

Problematic Features by Horizontal Comparison:

- **of**: Has a high absolute value in the error cell for documents classified as positive that are actually negative (lots of these incorrectly classified documents have this feature) and has a high horizontal difference (more documents with the negative label containing this feature are incorrectly classified than correctly classified).

The screenshot displays the Orange3 software interface with the following components:

- Highlight:** svm_1grams_n...
- Cell Highlight:**

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 1298 | 474 |
| pos | 509 | 1273 |
- Features in Table:**

| Feature | Average ... | Frequency | Horiz... | Horizont... |
|-----------|---------------|------------|--------------|---------------|
| and | 0.462 | 219 | 0.0714 | -0.0714 |
| with | 0.1603 | 76 | 0.0586 | -0.0586 |
| the | 0.597 | 283 | 0.0501 | -0.0501 |
| of | 0.4409 | 209 | 0.048 | -0.048 |
| movie | 0.1139 | 54 | 0.0471 | 0.0471 |
| makes | 0.0591 | 28 | 0.0467 | -0.0467 |
| too | 0.0211 | 10 | 0.0436 | 0.0436 |
| just | 0.0232 | 11 | 0.0407 | 0.0407 |
| it | 0.2848 | 135 | 0.0329 | -0.0329 |
| as | 0.1498 | 71 | 0.0327 | -0.0327 |
| be | 0.1076 | 51 | 0.0306 | -0.0306 |
| in | 0.1857 | 88 | 0.0293 | 0.0293 |
| so | 0.0401 | 19 | 0.0293 | 0.0293 |
| are | 0.097 | 46 | 0.0292 | -0.0292 |
| for | 0.154 | 73 | 0.0269 | -0.0269 |
| no | 0.019 | 9 | 0.0265 | 0.0265 |
- Evaluations to Display:**
 - ☒ Feature Confusion Ranking
 - ☒ Average Cell Value
 - ☒ Frequency
 - ☒ Horizontal Absolute Difference
 - ☒ Horizontal Difference
 - ☐ Vertical Absolute Difference
- Exploration Plugin:** Documents Display
- Calculating row and column values**
- Filter documents by selected feature**
 - ☒ Filter documents by selected feature
 - ☐ Reverse document filter
 - ☐ Documents from selected cell ...
- Instance 0 (Predicted pos, Actual pos)**

Highlighting of feature hits

does a good job of establishing a time and place , and of telling a fascinating character's story .
- Instance 10 (Predicted pos, Actual neg)**

Highlighting of feature hits

a sterling film – a cross between boys don't cry , deliverance , and ode to billy joe – lies somewhere in the story of matthew shepard , but that film is yet to be made .

- **a**: Has a high absolute value in the error cell for documents classified as negative that are actually positive (lots of these incorrectly classified documents have this feature) and has a high horizontal difference (more documents with the positive label containing this feature are incorrectly classified than correctly classified).

Highlight: svm_1grams_n...

Cell Highlight:

Features in Table:

Search:

Feature Table:

| Feature | Average ... | Frequency | Horiz... | Horizont... |
|----------|---------------|------------|---------------|---------------|
| and | 0.4401 | 224 | 0.1145 | 0.1145 |
| with | 0.0963 | 49 | 0.0868 | 0.0868 |
| it | 0.3026 | 154 | 0.0598 | -0.0598 |
| a | 0.5029 | 256 | 0.0595 | 0.0595 |
| is | 0.3202 | 163 | 0.0492 | -0.0492 |
| more | 0.0884 | 45 | 0.0476 | -0.0476 |
| no | 0.053 | 27 | 0.042 | -0.042 |
| have | 0.0747 | 38 | 0.0417 | -0.0417 |
| of | 0.4224 | 215 | 0.0403 | 0.0403 |
| but | 0.1611 | 82 | 0.0393 | -0.0393 |
| can | 0.0688 | 35 | 0.0389 | -0.0389 |
| there | 0.0609 | 31 | 0.0389 | -0.0389 |
| n't | 0.0864 | 44 | 0.037 | -0.037 |
| film | 0.1434 | 73 | 0.0341 | 0.0341 |
| are | 0.0864 | 44 | 0.0338 | -0.0338 |
| i | 0.0589 | 30 | 0.033 | -0.033 |

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-------|-------|
| neg | 0.477 | 0.479 |
| pos | 0.503 | 0.562 |

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-----|-----|
| neg | 619 | 227 |
| pos | 256 | 716 |

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|------|-------|
| neg | 0 | 0.002 |
| pos | 0.06 | 0 |

Evaluations to Display:

- ☒ Feature Confusion Ranking
- ☒ Average Cell Value
- ☒ Frequency
- ☒ Horizontal Absolute Difference
- ☒ Horizontal Difference
- ☐ Vertical Absolute Difference

Exploration Plugin: Highlighted Feature Details

Calculating row and column values

Highlight: svm_1grams_n...

Cell Highlight:

Features in Table:

Search:

Feature Table:

| Feature | Average ... | Frequency | Horiz... | Horizont... |
|----------|---------------|------------|---------------|---------------|
| and | 0.4401 | 224 | 0.1145 | 0.1145 |
| with | 0.0963 | 49 | 0.0868 | 0.0868 |
| it | 0.3026 | 154 | 0.0598 | -0.0598 |
| a | 0.5029 | 256 | 0.0595 | 0.0595 |
| is | 0.3202 | 163 | 0.0492 | -0.0492 |
| more | 0.0884 | 45 | 0.0476 | -0.0476 |
| no | 0.053 | 27 | 0.042 | -0.042 |
| have | 0.0747 | 38 | 0.0417 | -0.0417 |
| of | 0.4224 | 215 | 0.0403 | 0.0403 |
| but | 0.1611 | 82 | 0.0393 | -0.0393 |
| can | 0.0688 | 35 | 0.0389 | -0.0389 |
| there | 0.0609 | 31 | 0.0389 | -0.0389 |
| n't | 0.0864 | 44 | 0.037 | -0.037 |
| film | 0.1434 | 73 | 0.0341 | 0.0341 |
| are | 0.0864 | 44 | 0.0338 | -0.0338 |
| i | 0.0589 | 30 | 0.033 | -0.033 |

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-------|-------|
| neg | 0.477 | 0.479 |
| pos | 0.503 | 0.562 |

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-----|-----|
| neg | 619 | 227 |
| pos | 256 | 716 |

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|------|-------|
| neg | 0 | 0.002 |
| pos | 0.06 | 0 |

Evaluations to Display:

- ☒ Feature Confusion Ranking
- ☒ Average Cell Value
- ☒ Frequency
- ☒ Horizontal Absolute Difference
- ☒ Horizontal Difference
- ☐ Vertical Absolute Difference

Exploration Plugin: Documents Display

Filter documents by selected feature ☒

Reverse document filter ☐

Documents from selected cell ...

| Instance | Predicted | Actual | Text |
|--|-----------|--------|--------------|
| <input type="checkbox"/> 11 | pos | pos | it's a co... |
| <input checked="" type="checkbox"/> 12 | neg | neg | imposto... |
| <input checked="" type="checkbox"/> 13 | neg | pos | while no... |
| <input type="checkbox"/> 15 | pos | neg | a mudd... |
| <input type="checkbox"/> 16 | neg | neg | impecca... |
| <input type="checkbox"/> 20 | neg | pos | a thriller |

Instance 12 (Predicted neg, Actual neg)
Highlighting a feature hits
impostor is a step down for director gary fleder .

Instance 13 (Predicted neg, Actual pos)
Highlighting a feature hits
while not all that bad of a movie , it's nowhere near as good as the original .

Problematic Features by Vertical Comparison:

- **you**: has high frequency in both error cells and a vertical absolute difference close to 0 in the error cell for documents classified as positive that are actually negative (ie. It shows up in

lots of documents that are correctly classified as pos and documents that are incorrectly classified as pos).

Highlight: svm__1grams_n...

Cell Highlight:

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 1298 | 474 |
| pos | 509 | 1273 |

Features in Table:

| Feature | Average Cell... | Frequency | Vertical Abs... |
|---------|-----------------|-----------|-----------------|
| that | 0.2194 | 104 | 0.0092 |
| in | 0.1857 | 88 | 0.0209 |
| but | 0.1646 | 78 | 0.0428 |
| with | 0.1603 | 76 | 0.0227 |
| for | 0.154 | 73 | 0.0511 |
| as | 0.1498 | 71 | 0.0445 |
| film | 0.1435 | 68 | 0.0341 |
| this | 0.1266 | 60 | 0.0135 |
| movie | 0.1139 | 54 | 0.0267 |
| be | 0.1076 | 51 | 0.0345 |
| n't | 0.1034 | 49 | 0.0539 |
| are | 0.097 | 46 | 0.0444 |
| an | 0.0907 | 43 | 0.0444 |
| like | 0.0907 | 43 | 0.0459 |
| its | 0.0865 | 41 | 0.0196 |
| you | 0.0865 | 41 | 0.0007 |

Trained Model: svm__1grams_nopur
Kappa: 0.447
Accuracy: 0.723

Evaluations to Display:

- ☒ Feature Confusion Ranking
- ☒ Average Cell Value
- ☒ Frequency
- ☐ Horizontal Absolute Difference
- ☐ Horizontal Difference
- ☒ Vertical Absolute Difference

Exploration Plugin: Highlighted Feature Details

Calculating row and column values

Average Cell Value

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-------|-------|
| neg | 0.07 | 0.086 |
| pos | 0.079 | 0.087 |

Frequency

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-----|-----|
| neg | 91 | 41 |
| pos | 40 | 111 |

Vertical Absolute Difference

Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-------|-------|
| neg | 0 | 0.001 |
| pos | 0.008 | 0 |

☒ Filter documents by selected feature

☐ Reverse document filter

☐ Documents from selected cell ...

| Instance | Predicted | Actual | Text |
|--|-----------|--------|--------------|
| <input checked="" type="checkbox"/> 50 | neg | neg | if you're... |
| <input type="checkbox"/> 60 | pos | pos | one of t... |
| <input checked="" type="checkbox"/> 96 | neg | pos | you're n... |
| <input type="checkbox"/> 117 | pos | pos | the movi... |
| <input type="checkbox"/> 140 | pos | pos | at times... |
| <input type="checkbox"/> 164 | pos | neg | in both t |

Instance 50 (Predicted neg, Actual neg)

Highlighting you feature hits

if you're not fans of the adventures of steve and terri , you should avoid this like the dread ed king brown snake . personally , i'd rather watch them on the animal planet .

Instance 96 (Predicted neg, Actual pos)

Highlighting you feature hits

you're not merely watching history , you're engulfed by it .

- **for**: has high frequency in both error cells and a vertical absolute difference close to 0 in the error cell for documents classified as negative that are actually positive (ie. It shows up in lots of documents that are correctly classified as neg and documents that are incorrectly classified as neg).

The screenshot displays a machine learning interface with several panels:

- Highlight:** Shows the selected model `svm_1grams_n...`.
- Cell Highlight:** A confusion matrix for the model:

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 1298 | 474 |
| pos | 509 | 1273 |
- Evaluations to Display:**
 - ☐ Feature Confusion Ranking
 - ☐ Average Cell Value
 - ☒ Frequency
 - ☐ Horizontal Absolute Difference
 - ☐ Horizontal Difference
 - ☒ Vertical Absolute Difference
- Features in Table:** A table showing feature frequencies and vertical absolute differences. The feature `for` is highlighted with a frequency of 60 and a vertical absolute difference of 0.0092.

| Feature | Frequency | Vertical Absolute ... |
|---------|-----------|-----------------------|
| the | 292 | 0.0267 |
| a | 256 | 0.0261 |
| and | 224 | 0.0495 |
| of | 215 | 0.0295 |
| is | 163 | 0.0306 |
| 's | 162 | 0.0617 |
| to | 156 | 0.031 |
| it | 154 | 0.0506 |
| that | 131 | 0.0455 |
| in | 119 | 0.0188 |
| but | 82 | 0.0224 |
| film | 73 | 0.0248 |
| this | 64 | 0.0253 |
| an | 62 | 0.0255 |
| its | 62 | 0.0247 |
| for | 60 | 0.0092 |
- Exploration Plugin:** Set to `Highlighted Feature Details`.
- Frequency:** Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-----|-----|
| neg | 165 | 73 |
| pos | 60 | 131 |
- Vertical Absolute Difference:** Model Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-------|-------|
| neg | 0 | 0.051 |
| pos | 0.009 | 0 |
- Filter documents by selected feature:**
 - ☒ Filter documents by selected feature
 - ☐ Reverse document filter
 - ☐ Documents from selected cell ...
- Instance List:**

| Instance | Predicted | Actual | Text |
|----------|-----------|--------|---------------|
| 11 | pos | pos | it's a co... |
| 12 | neg | neg | imposto... |
| 19 | neg | neg | unfunny ... |
| 25 | pos | neg | scores n... |
| 41 | neg | neg | it feels l... |
| 43 | pos | neg | swims in |
- Instance Details:**
 - Instance 12 (Predicted neg, Actual neg):** Highlighting for feature hits: `impostor is a step down for director gary fleder .`
 - Instance 25 (Predicted pos, Actual neg):** Highlighting for feature hits: `scores no points for originality , wit , or intelligence . it's a cookie-cutter movie , a cut-and-paste job .`

6.

This support-vector machine seems to have issues with connective words such as “a”, “for”, and “as” as well as, more broadly, words which do not carry intrinsic context by themselves (such as “you”); thus, they appear in all documents, regardless of class. One simple improvement to the model then would be to include bigram or trigram features which would pair these connective words up with other words that carry meaning in many of the documents.

For example, as shown in the screenshots above, while the unigram “for” may not help different between a positive and negative review in document 12, the bigram “for director” carries more context. That said, a bigram could also confuse the situation further as is the case where the bigrams “points for” and “for originality” in document 25 may be weighed positively but, in fact, it proceeded by “no points” which actually implies a negative review.

In much the same way, a trigram would clarify the fairly ambiguous “you” in document 50 to “you should avoid”.

Thus, a more advanced solution would involve finding features that link these confusing modifiers to negative expressions based on the documents explored above such as the following regex features:

"no\s?.\s?.\s?for"

"not\s?.\s?.\s?as"

"you\s?.\s?.\s?avoid"

"you\s?.\s?.\s?not"

"not\s?.\s?.\s?a"

The \s?.\s?.\s? is to ensure that the two keywords are no more than two words apart (kind of like a specific use-case trigram).

7.

Including unigrams and bigrams using 10-fold cross-validation:

Accuracy 0.7112

Kappa 0.4223

Including unigrams and trigrams using 10-fold cross-validation:

Accuracy 0.7115

Kappa 0.4229

Including unigrams and regex with counting (outlined above) using 10-fold cross-validation:

Accuracy 0.7138

Kappa 0.4277

While the first two primitive approaches of increasing the n-gram degree actually perform worse than the baseline, the regex approach proposed offered exactly zero improvement.

8.

The addition of these regex features appears to have made no significant impact on the aggregate performance of the model or even the confusion matrix.

Baseline Model: svm__1grams_nopunct

Competing Model: svm__1grams_nopunct_regex

TRAINED_MODEL

- Documents: sentiment_sentences-cross-validation.csv
- Feature Plugins: basic
- Feature Table: 1grams_nopunct
- Learning Plugin: Support Vector Machines
- Validation: CV
- Trained Model: svm__1grams_nopunct

TRAINED_MODEL

- Documents: sentiment_sentences-cross-validation.csv
- Feature Plugins: basic regex
- Feature Table: 1grams_nopunct_regex
- Learning Plugin: Support Vector Machines
- Validation: CV
- Trained Model: svm__1grams_nopunct_regex

Comparison Plugin: Basic Model Comparison

Baseline Model Metrics:

| Metric | Value |
|----------|--------|
| Accuracy | 0.7138 |
| Kappa | 0.4277 |

Competing Model Metrics:

| Metric | Value |
|----------|--------|
| Accuracy | 0.7138 |
| Kappa | 0.4277 |

Baseline Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 2550 | 1009 |
| pos | 1025 | 2524 |

Competing Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 2551 | 1008 |
| pos | 1026 | 2523 |

Differences By Instance

☐ Only Baseline Model Cell ☒ Only Competing Model Cell ☐ Present in Both Cells

Instance 3775 (Predicted neg, Actual pos)

you have enough finely tuned acting to compensate for the movie's failings .

Instance 4212 (Predicted neg, Actual pos)

a smart , arch and rather cold-blooded comedy .

Instance 5742 (Predicted neg, Actual pos)

the film's heady yet far from impenetrable theory suggests that russians take comfort in their closed-off nationalist reality .

Difference Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|-----|-----|
| neg | 0 | -1 |
| pos | 1 | -1 |

Baseline Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 2550 | 1009 |
| pos | 1025 | 2524 |

Competing Confusion Matrix:

| Act \ Pred | neg | pos |
|------------|------|------|
| neg | 2551 | 1008 |
| pos | 1026 | 2523 |

From evaluating the differences by instance, it appears as though many of instances where the two models perform differently are those which do *not* include any of the new regex features.