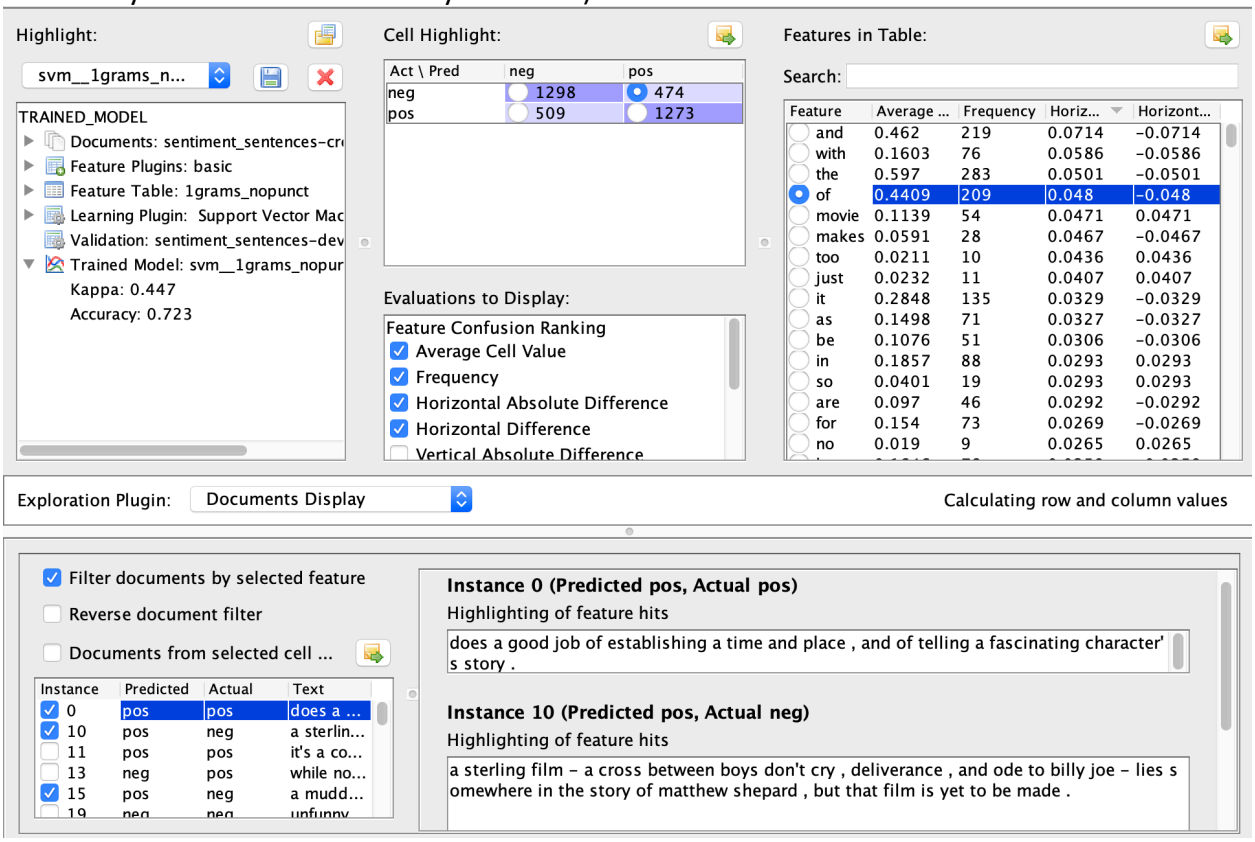
1. Done.
2. Done.
3. Baseline Model Evaluation Metrics:

| | |
|---|---|
| Accuracy | 0.7138 |
| Kappa | 0.4277 |

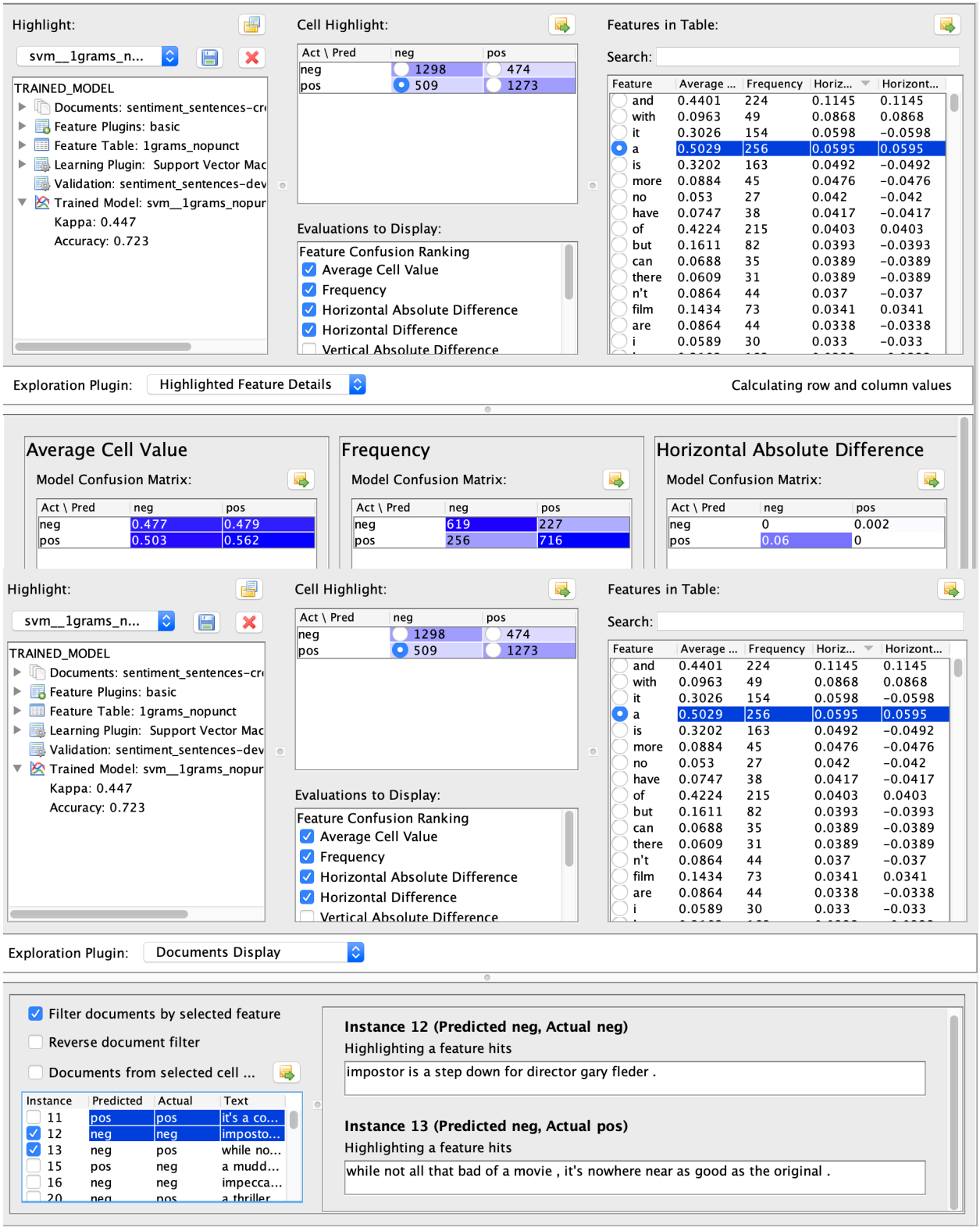
4. Done.
5. 

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).



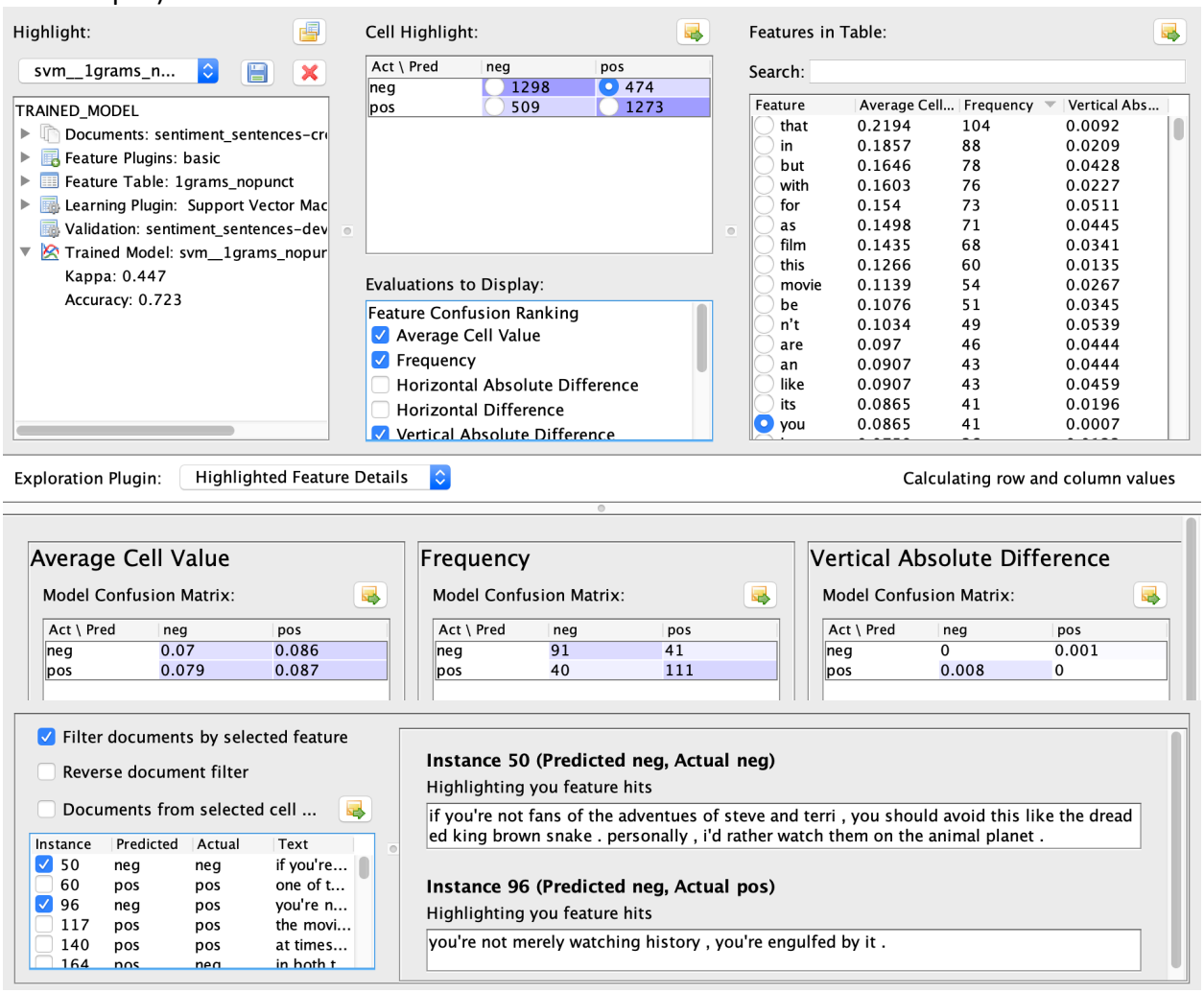

- **a**: Has a high absolute value in the error cell for documents classified as negative that are actually positive (lots of these incorrectly classified documents have this feature) and has a high horizontal difference (more documents with the positive label containing this feature are incorrectly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



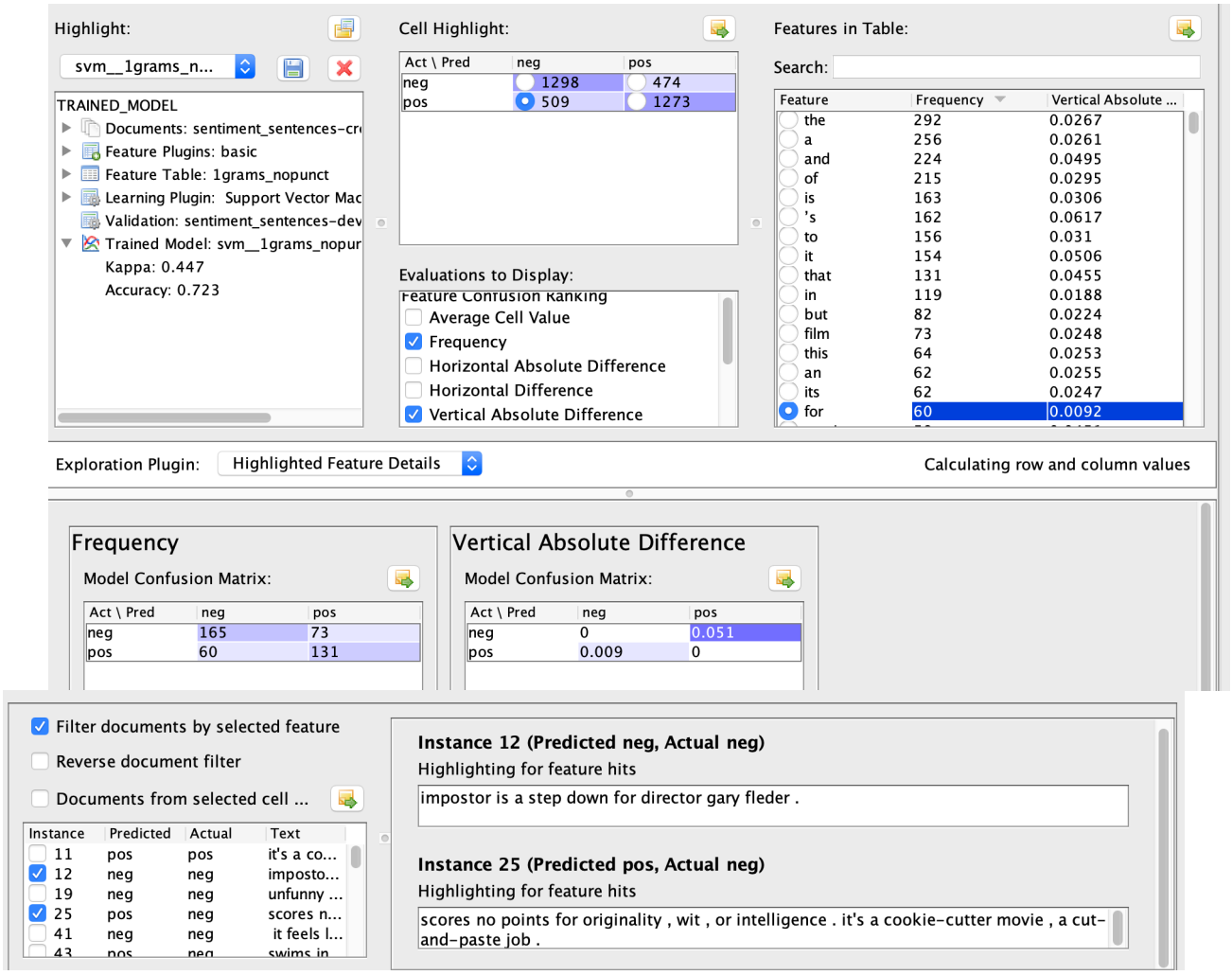

Problematic Features by Vertical Comparison:

- **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).




- **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).




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 |

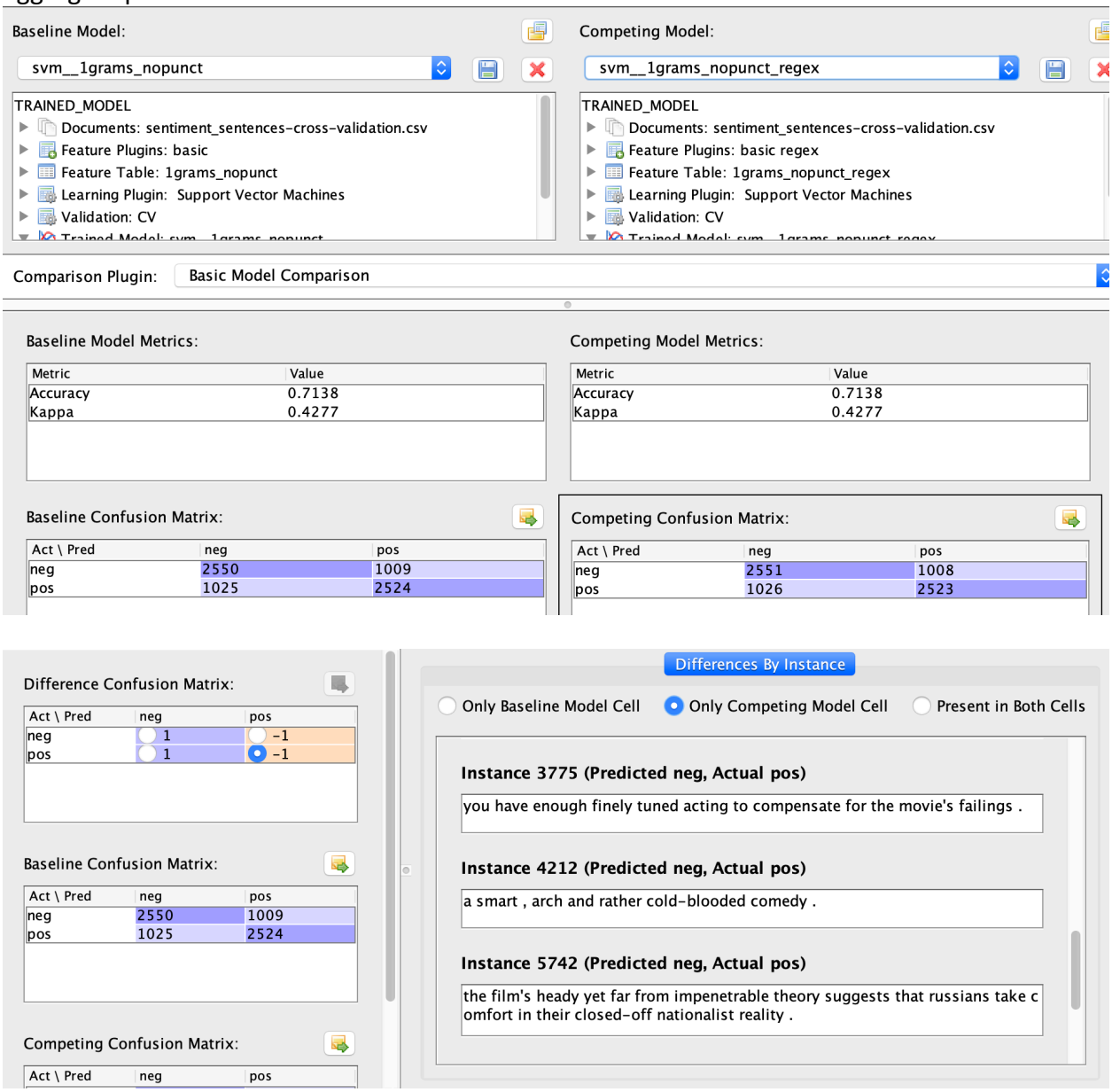
Including unigrams and regex with counting (outlined above) using 10-fold cross-validation:

| | |
|---|---|
| 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.




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.