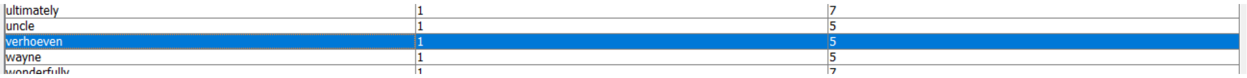
Rees Braam
INLS 613
2/24/2020

Assignment 2

Q1

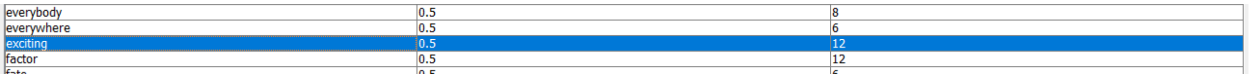


ultimately	1	7
uncle	1	5
verhoeven	1	5
wayne	1	5
wonderfully	1	7

The term "verhoeven" has a precision of 1 for the positive class and had 5 total hits in the training set.

Near the bottom of the list of terms with perfect precision lies the term "verhoeven." This isn't a word I've ever seen, so I initially wondered why it showed up on the list at all. A quick google search of the term shows that it's the last name of a Dutch director who has directed many famous films and has received nine academy award nominations. From this, I can easily reach the conclusion that he's a pretty good director. Combine that with the fact that there's only 5 hits on this term in the whole set (and spread among only 3 rows at that), and it's no longer a surprise that this term shows up as perfectly positive. In fact, if you look in-depth at the reviews that mention the name Verhoeven, it becomes clear that all of them are for his movie The Fourth Man, a movie that has generally good reviews. Due to the small sample size, this unsurprisingly results in a perfect positive correlation for Verhoeven.

Q2

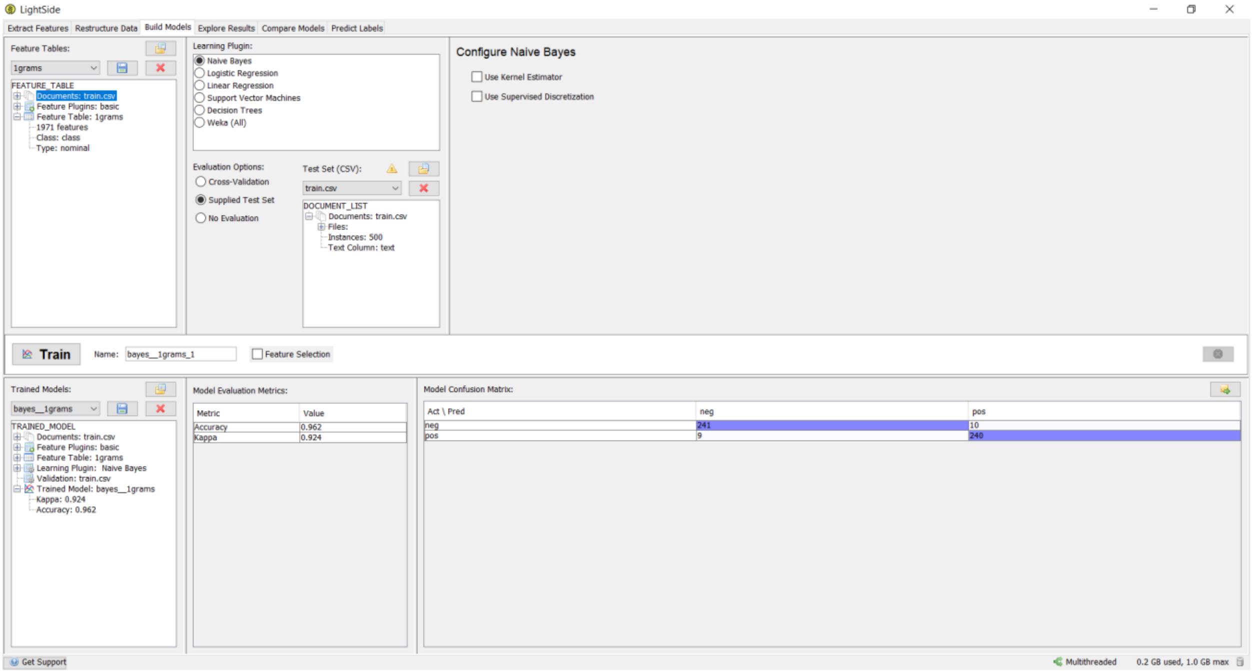


everybody	0.5	8
everywhere	0.5	6
exciting	0.5	12
factor	0.5	12
fate	0.5	6

The term "exciting" has a precision of 0.5 for the positive class and had 12 total hits in the training set.

Scanning the list of terms that appear an equal number of times in both classes, I immediately noticed the term "exciting." This term, in my opinion, seems to be *extremely* out-of-place, and I would have expected it to have a high correlation with positive reviews. Looking through its occurrences in the training set, all the positive uses are exactly as expected. However, it seems that some movie reviewers like to use the term exciting to refer to either something that *should've* been exciting or something *else* that was exciting. Like the issue that I saw in A1, many generally positive or negative terms are not used to reference the movie itself, but instead to something else, or are just negated. In this case however, the former is the cause for the even occurrence of the term "exciting."

Q3



Feature Tables: 1grams

FEATURE_TABLE
- Documents: train.csv
- Feature Plugins: basic
- Feature Table: 1grams
  - 1971 features
  - Class: class
  - Type: nominal

Learning Plugin:
- ● Naive Bayes
- ○ Logistic Regression
- ○ Linear Regression
- ○ Support Vector Machines
- ○ Decision Trees
- ○ Weka (All)

Evaluation Options:
- ○ Cross-Validation
- ● Supplied Test Set
- ○ No Evaluation

Test Set (CSV): train.csv

DOCUMENT_LIST
- Documents: train.csv
  - Files:
  - Instances: 500
  - Text Column: text

Configure Naive Bayes
- □ Use Kernel Estimator
- □ Use Supervised Discretization

Name: bayes__1grams_1 □ Feature Selection

Trained Models: bayes__1grams

TRAINED_MODEL
- Documents: train.csv
- Feature Plugins: basic
- Feature Table: 1grams
- Learning Plugin: Naive Bayes
- Validation: train.csv
- Trained Model: bayes__1grams
  - Kappa: 0.924
  - Accuracy: 0.962

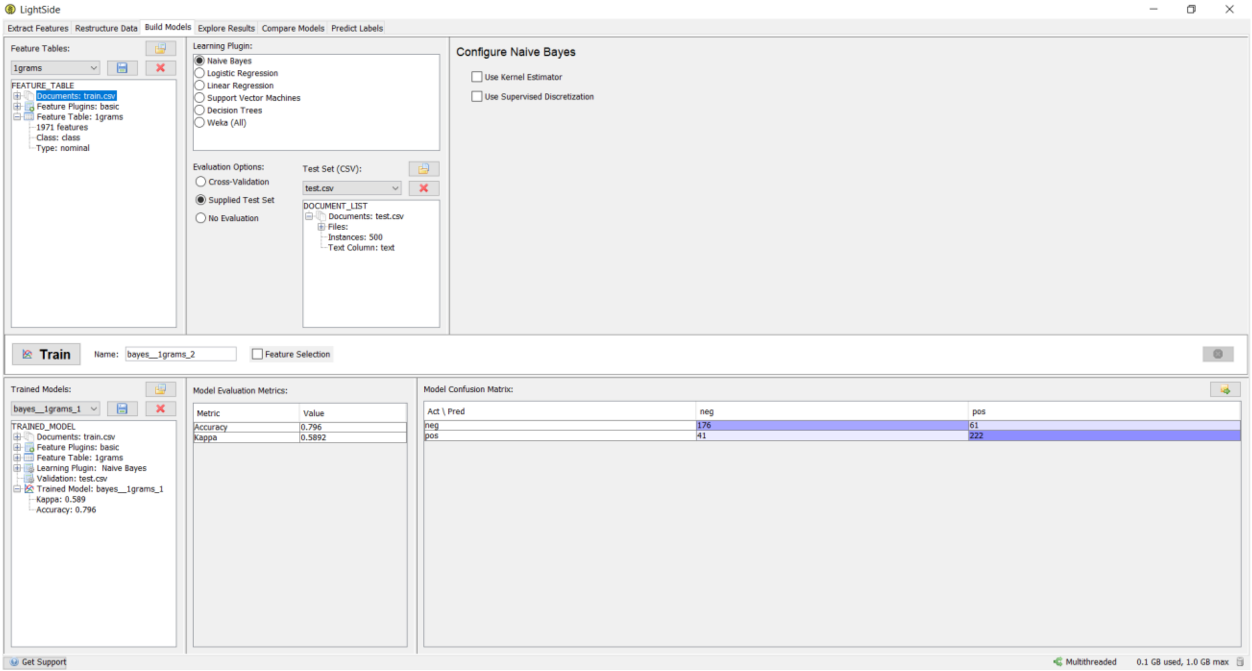
Model Evaluation Metrics:

Metric	Value
Accuracy	0.962
Kappa	0.924

Model Confusion Matrix:

Act \ Pred	neg	pos
neg	241	10
pos	9	240

Train model applied to train.csv



Feature Tables: 1grams

FEATURE_TABLE
- Documents: train.csv
- Feature Plugins: basic
- Feature Table: 1grams
  - 1971 features
  - Class: class
  - Type: nominal

Learning Plugin:
- ● Naive Bayes
- ○ Logistic Regression
- ○ Linear Regression
- ○ Support Vector Machines
- ○ Decision Trees
- ○ Weka (All)

Evaluation Options:
- ○ Cross-Validation
- ● Supplied Test Set
- ○ No Evaluation

Test Set (CSV): test.csv

DOCUMENT_LIST
- Documents: test.csv
  - Files:
  - Instances: 500
  - Text Column: text

Configure Naive Bayes
- □ Use Kernel Estimator
- □ Use Supervised Discretization

Name: bayes__1grams_2 □ Feature Selection

Trained Models: bayes__1grams_1

TRAINED_MODEL
- Documents: train.csv
- Feature Plugins: basic
- Feature Table: 1grams
- Learning Plugin: Naive Bayes
- Validation: test.csv
- Trained Model: bayes__1grams_1
  - Kappa: 0.589
  - Accuracy: 0.796

Model Evaluation Metrics:

Metric	Value
Accuracy	0.796
Kappa	0.5892

Model Confusion Matrix:

Act \ Pred	neg	pos
neg	176	61
pos	41	222

Train model applied to test.csv

Unsurprisingly, the model trained on train.csv performed unnaturally well when tested against train.csv, obtaining an unreasonably high 96% accuracy score, a score you told us we should realistically never see in a task as complex as this unless we messed something up. Compared to the performance the same model had when tested against test.csv, the pictures above show that the model had a more realistic accuracy score of ~80% here, which is substantially lower for obvious reasons.

“Reed Diamond plays a man suffering from amnesia who's been in a mental asylum for over a decade after he was found wondering the back roads with blood on his hands. The doctors want to test out an experimental new drug that'll return his lost memories if it works. But when the drugs give him hallucinations of a demon he chooses to escape instead. While outside he befriends a young boy whose stepfather (Greg Grunberg) mistreats his mother won't let her near the darkroom in his basement & acts suspicious in general. While the general 'mystery' of the film is a tad easy to identify way before it's revealed I found Mr. Diamond's acting to be enthralling enough to keep my attention throughout. (In the interest of full disclosure I've been a huge fan of his since Homicide and his brief but extremely pivotal role in The Shield up through Journeyman & Dollhouse) Not a great film nor a good one but serviceable enough. Although I did like it better than the previous films that I've seen from Director/writer Michael Hurst (Room 6 Pumkinhead 4 Mansquito) Eye Candy: one fleeting pair of boobs in a hallucination My Grade: C-” – Cell 78 from train.csv

The above review was falsely predicted positive by the model trained and test on train.csv, putting it among the less than 4% of reviews that were falsely classified in this test. However, reading the review, we can see that the first half of the review is just a summary of the movie, and so is not at all indicative of the true class for the review. In the second half, the reviewer does not really write anything bad about the move aside from the part where he gives it a C-. Even worse, the part where it’s mentioned that the film is not good, the reviewer does not use a word with a negative connotation and instead opts to negate a positive word (“not a a great film”). Overall, nothing in this review really gives away that it’s a negative review except for the grade at the end. However, most reviewers use a scale from 1-10 and not the letter grade scale, so it is unlikely that the model would be able to pick out the grade anyways. Overall, this review appears to just be an edge case, and it’s clear why it was misclassified in my opinion.

Q4

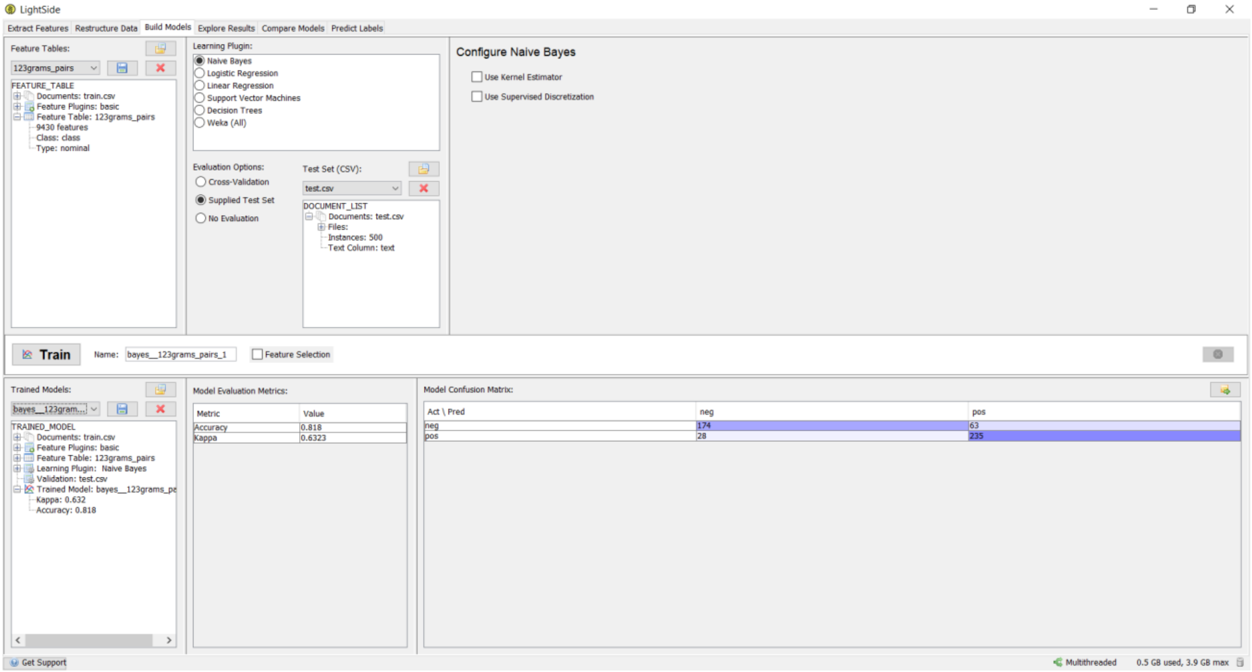
threshold	training set accuracy	test set accuracy
1	99.2%	78%
2	98.6%	80%
3	97.8%	79.4%
4	97.6%	80.6%
5	96.2%	79.6%

From the table above, we can see that as the threshold increases, training set accuracy decreases linearly, while test set accuracy seems to remain about constant. The reason that this occurs is that the test set does not care about words that occur few times, as the output model does not map 1:1 to the set it is being tested on. Therefore, the inclusion of these words does not affect the accuracy of the model on the test set. However, the training set does gain from having these rare occurrences, because it does contain all the rare words. Therefore, their inclusion allows the training set to account for rare words that the test set essentially ignores either way. This is why the training set accuracy is hurt by excluding the rare words but the test set accuracy remains largely unchanged.

Q5

With the options laid out so nicely, it was quite easy to use many different combinations with no direction just to try to brute force a lucky combination with a high accuracy. However, as it turns out, this is much more difficult than one would think, and it turns out that it’s really better to have an idea of the direction you want to go if you want to get a good accuracy score, let alone one that is actually *better* than the base ~80% that using just unigrams gives. I will say, however, that I was able to marginally increase the base unigram accuracy to ~81% by increasing the threshold with unigrams to 6. Another interesting experiment I tried was enabling everything, which decreased the accuracy a significant amount, probably the most in fact.

After playing around a little bit to almost no success, I decided to go in the direction of adding both bigrams and trigrams. At this point, I was quite convinced that bigrams and trigrams were important to increasing the accuracy score but could not get it to actually improve. In this case, the key turned out to be the threshold. However, unlike the base unigram case where increasing the threshold helped, it turned out that decreasing was what helped here. This makes sense as well, since bigrams and trigrams would occur less frequently. However, the magic number was 3 – any lower or higher and the accuracy decreased. From here, it seemed that the only other change I made that could increase the accuracy was to include Word/POS Pairs as a basic feature. Interestingly, this option is not mentioned in the manual, but had the only positive impact of all the POS options, at least in my experience. Thus, my highest accuracy was 81.8%, which was better than the default case, but not by much.



Feature Tables: 123grams_pairs

FEATURE_TABLE
Documents: train.csv
Feature Plugins: basic
Feature Table: 123grams_pairs
9430 features
Class: class
Type: nominal

Learning Plugin:
● Naive Bayes
○ Logistic Regression
○ Linear Regression
○ Support Vector Machines
○ Decision Trees
○ Weka (All)

Evaluation Options:
○ Cross-Validation
● Supplied Test Set
○ No Evaluation

Test Set (CSV): test.csv

DOCUMENT_LIST
Documents: test.csv
Files:
Instances: 500
Text Column: text

Configure Naive Bayes
□ Use Kernel Estimator
□ Use Supervised Discretization

Name: bayes__123grams_pairs_1 □ Feature Selection

Trained Models: bayes__123gram...

TRAINED_MODEL
Documents: train.csv
Feature Plugins: basic
Feature Table: 123grams_pairs
Learning Plugin: Naive Bayes
Validation: test.csv
Trained Model: bayes__123grams_pa
Kappa: 0.632
Accuracy: 0.818

Model Evaluation Metrics:

Metric	Value
Accuracy	0.818
Kappa	0.6323

Model Confusion Matrix:

Act \ Pred	neg	pos
neg	174	63
pos	28	235

In addition, even though it was not part of the assignment, I tried the other learning plugins in the “Build Models” section. However, I was not able to beat the Naïve Bayes classifier. I find this a little interesting, since the assumption that movie reviews are independent of each other seems like it would be false here.