A Taste of Applied Machine Learning

Carolyn Penstein Rosé

Language Technologies Institute/ Human-Computer Interaction Institute

Carolyn Rosé

PhD in Language and Information Technologies, 1998

Enjoys Israeli folk dancing, playing piano, long walks in the woods, and knitting, crocheting, and spinning yarn

- Joint appointment between Language Technologies and Human-Computer Interaction
- President of the International Society of the Learning Sciences
- Co-Chair of the CSCL community committee
- Associate Editor of the International Journal of Computer Supported Collaborative Learning

Machine Learning: Conceptual Overview

How does machine learning work?

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunn			E	yes
rainy			E	yes
s				yes
c				yes
overcast			ALSE	yes
rainy			TRUE	no

A slightly ... rule learner will find t... the most informat... ass. What do you th... his case?

Clas...

What w... se?

Outlook:

Sunny -> No

Overcast -> Yes

Rainy-> Yes

<Feature Name>:

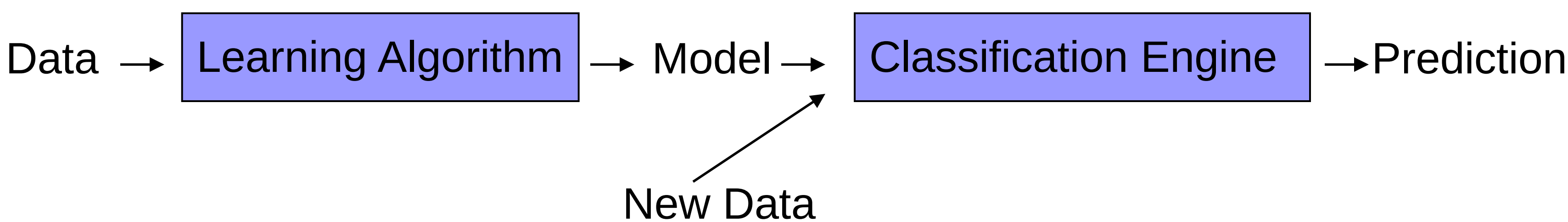
<value> -> <prediction>

<value> -> <prediction>

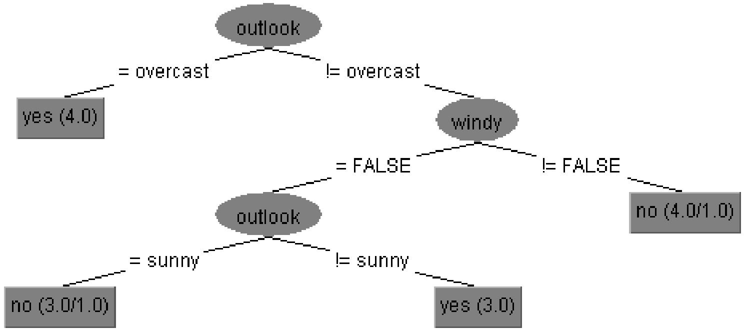
...

What is machine learning?

- Automatically or *semi-automatically*
  - Inducing concepts (i.e., rules) from data
  - Finding patterns in data
  - Explaining data
  - Making predictions



More Complex Algorithm…



* Only makes 2 mistakes!

Why is it better?

- Not because it is more complex
  - Sometimes more complexity makes performance worse
- What is different in what the three rule representations assume about your data?
  - 0R
  - 1R
  - Trees
- The best algorithm for your data will give you exactly the power you need

Why is it better?

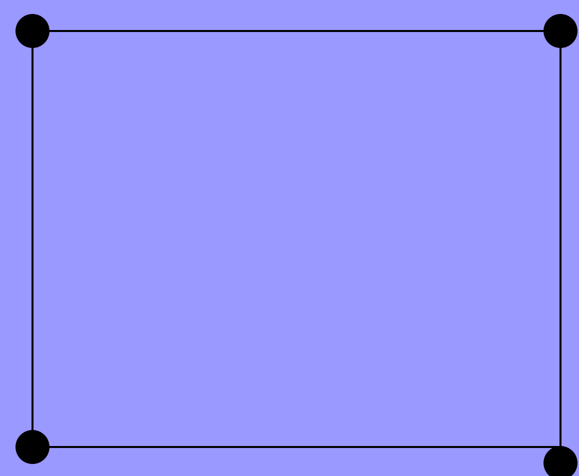
- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you ex

Lets say you don’t know the shape, what shape would you guess?

Why is it better?

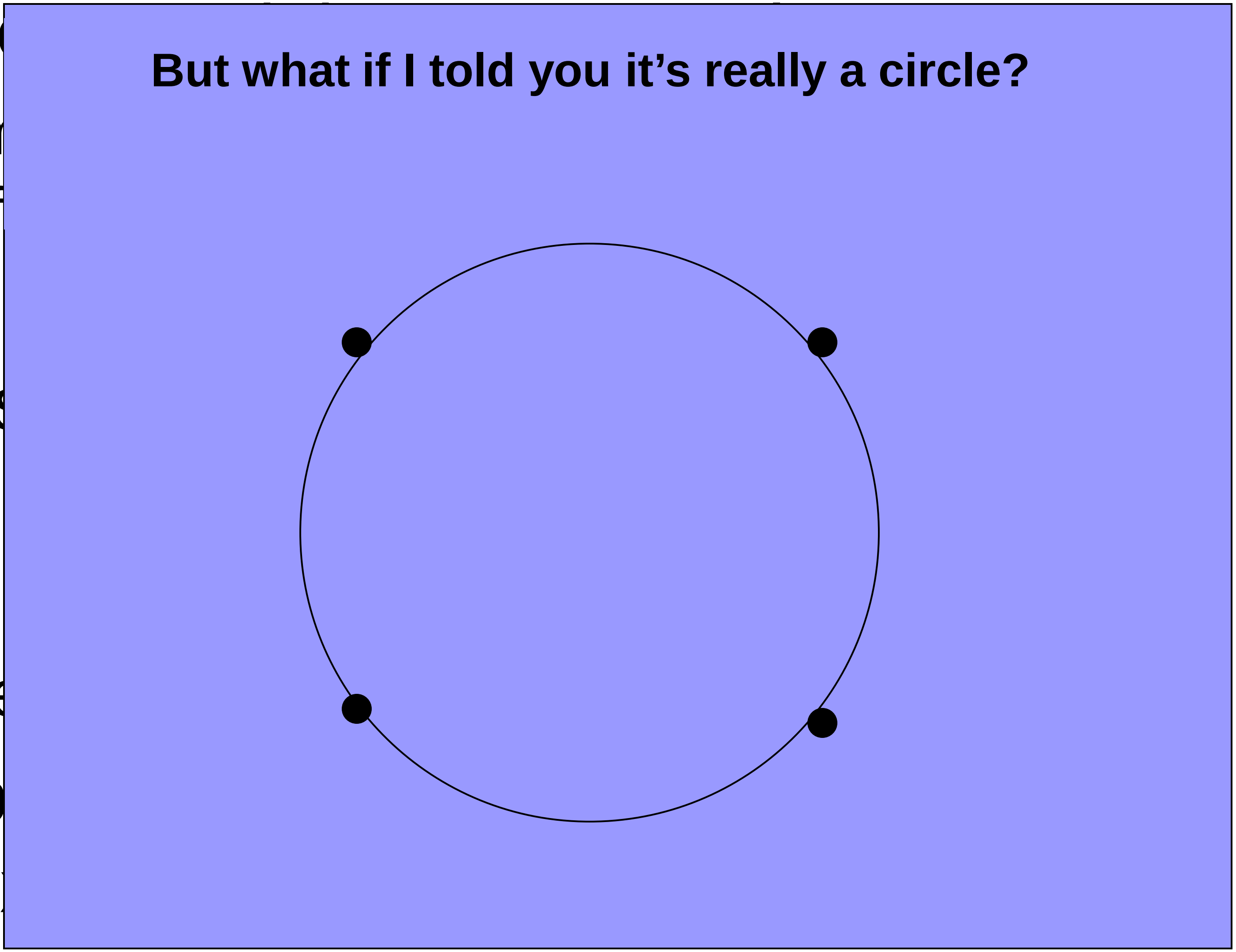
- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you e

Lets say you don’t know the shape, what shape would you guess?



Why is it better?

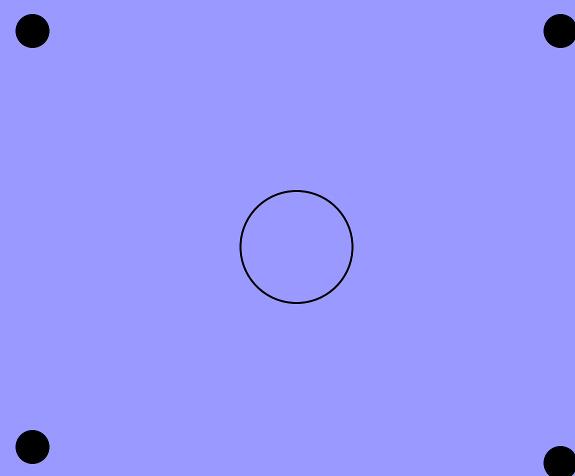
- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you e



Why is it better?

- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you e

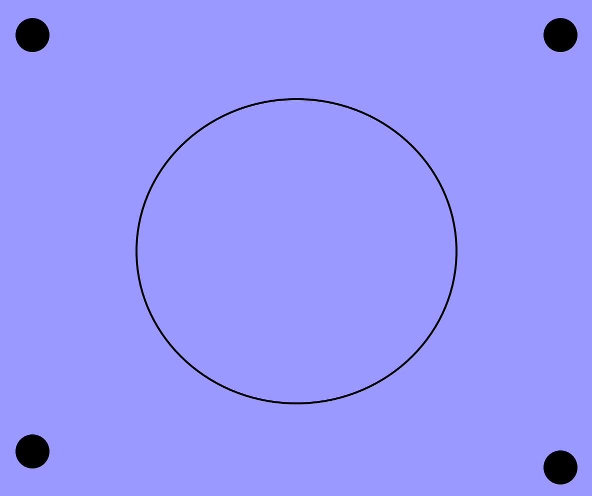
If you know the shape, you have fewer degrees of freedom – less room to make a mistake.



Why is it better?

- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you e

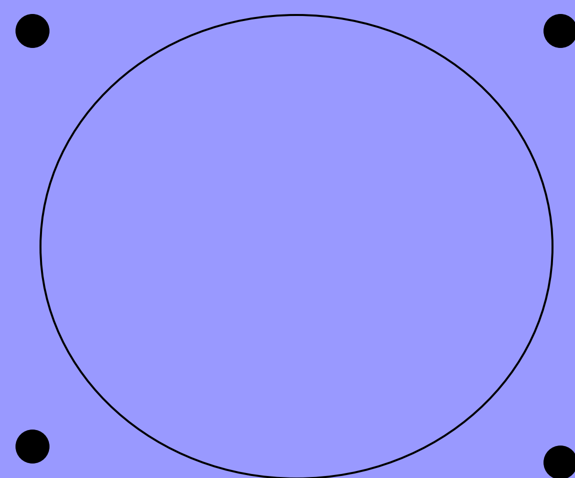
If you know the shape, you have fewer degrees of freedom – less room to make a mistake.



Why is it better?

- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you e

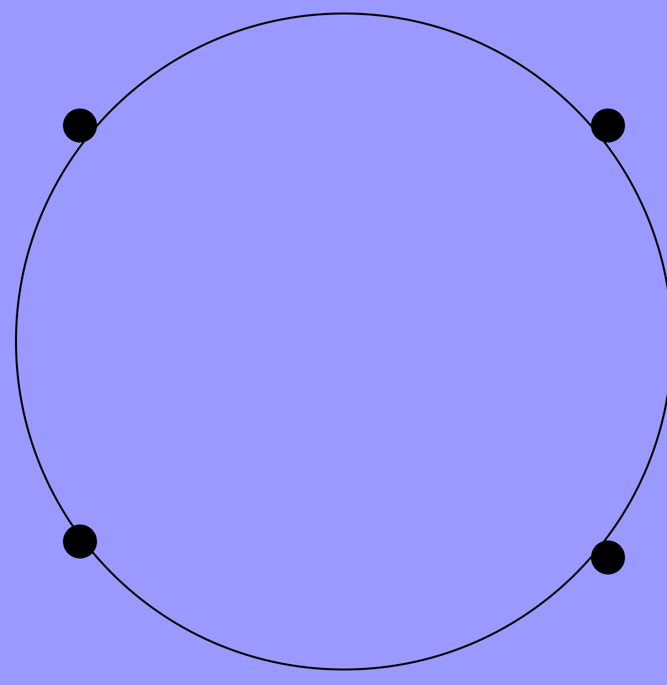
If you know the shape, you have fewer degrees of freedom – less room to make a mistake.



Why is it better?

- Not b
  - Som
    perf
- What
  repres
  - 0R
  - 1R
  - Tree
- The b
  you e

If you know the shape, you have fewer degrees of freedom – less room to make a mistake.



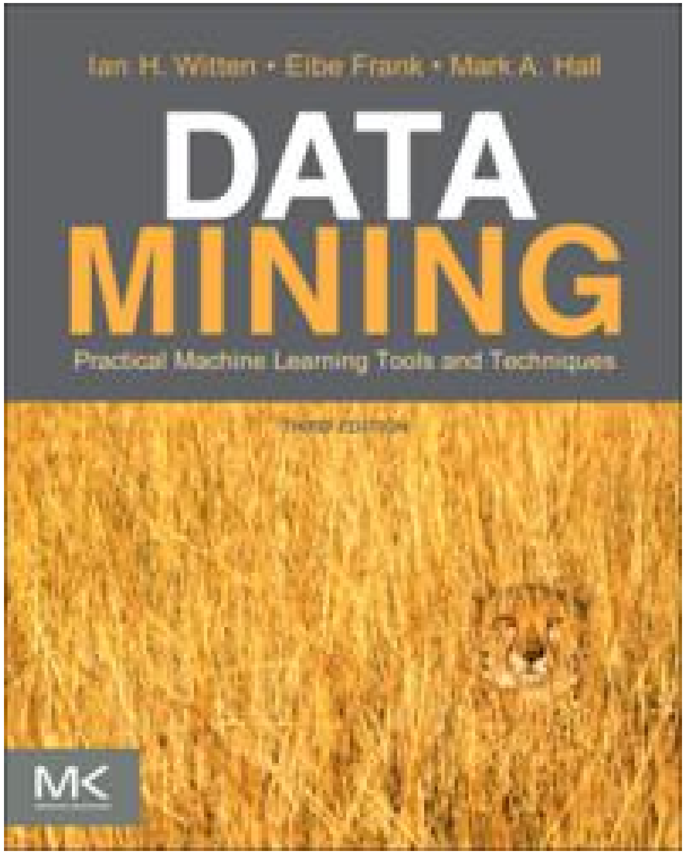
Why is it better?

- Not because it is more complex
  - Sometimes more complexity makes performance worse
- What is different in what the three rule representations assume about your data?
  - 0R
  - 1R
  - Trees
- The best algorithm for your data will give you exactly the power you need

Tools and Resources

Essential Reading

- Witten, I. H., Frank, E., Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, third edition, Elsevier: San Francisco



Other Suggested Readings

- Richard Cotton (2013). *Learning R*, O’Reilly and Associates
- Allen Downey (2013). *Think Bayes*, O’Reilly and Associates
- Mark Lutz (2013). *Learning Python*, O’Reilly and Associates
- Drew Conway & John White (2012). *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started*, O’Reilly Media

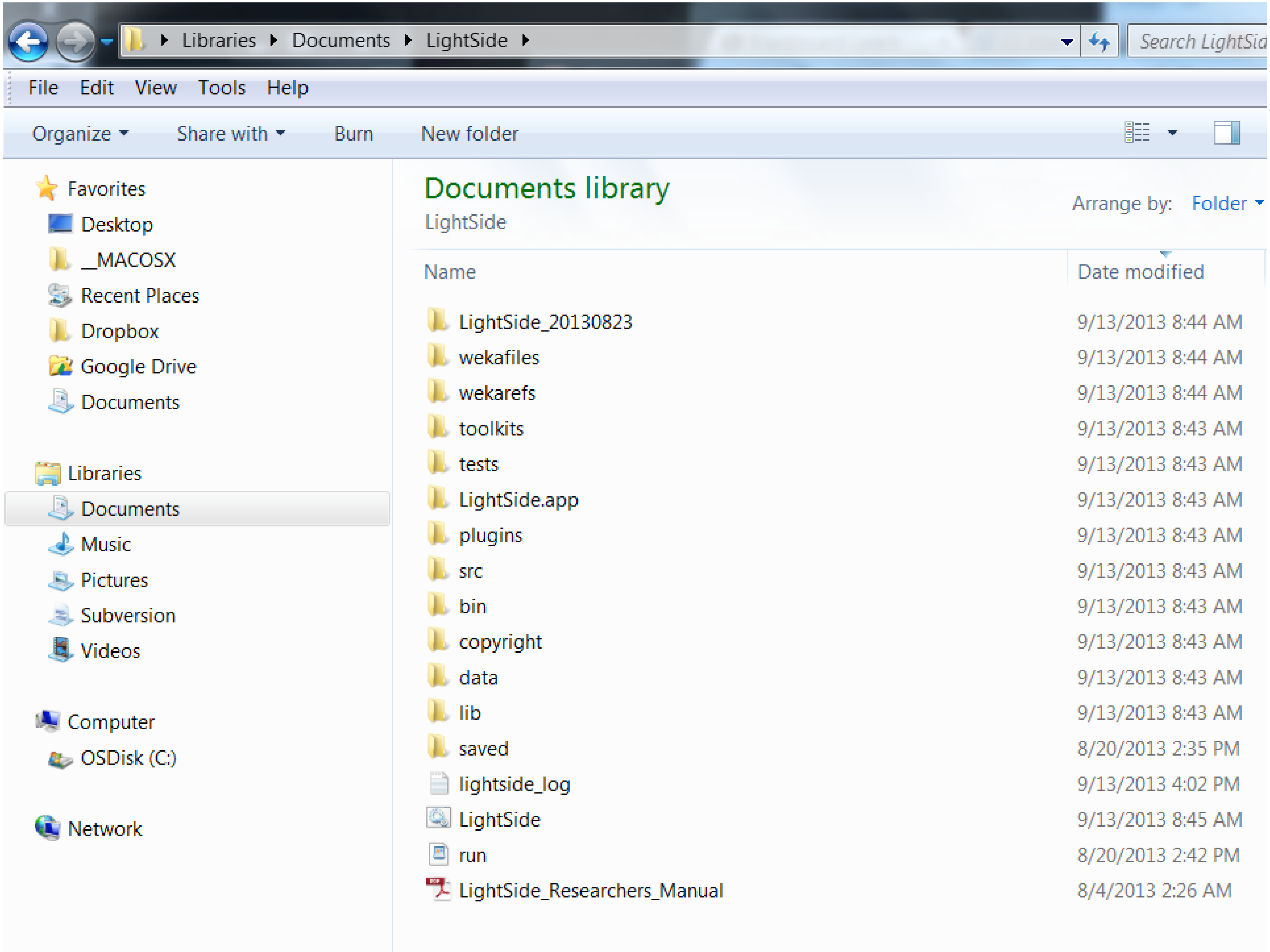
Software Tools

- Data manipulation tools
  - Whatever you are comfortable with
  - Scripting language like R, Python, Perl
  - Excel

- Weka (http://www.cs.waikato.ac.nz/ml/weka/)
  - Open source machine learning toolkit
  - Includes Java API

- LightSIDE (http://lightsidelabs.com/research)
  - Weka add-on for text processing
  - Developed at CMU!

lightsidelabs.com/research/


Libraries ▸ Documents ▸ LightSide ▸
Search LightSid
File Edit View Tools Help
Organize
Share with
Burn
New folder
Favorites
Desktop
__MACOSX
Recent Places
Dropbox
Google Drive
Documents
Libraries
Documents
Music
Pictures
Subversion
Videos
Computer
OSDisk (C:)
Network
Documents library
LightSide
Arrange by: Folder
Name
Date modified
LightSide_20130823 9/13/2013 8:44 AM
wekafiles 9/13/2013 8:44 AM
wekarefs 9/13/2013 8:44 AM
toolkits 9/13/2013 8:43 AM
tests 9/13/2013 8:43 AM
LightSide.app 9/13/2013 8:43 AM
plugins 9/13/2013 8:43 AM
src 9/13/2013 8:43 AM
bin 9/13/2013 8:43 AM
copyright 9/13/2013 8:43 AM
data 9/13/2013 8:43 AM
lib 9/13/2013 8:43 AM
saved 8/20/2013 2:35 PM
lightside_log 9/13/2013 4:02 PM
LightSide 9/13/2013 8:45 AM
run 8/20/2013 2:42 PM
LightSide_Researchers_Manual 8/4/2013 2:26 AM

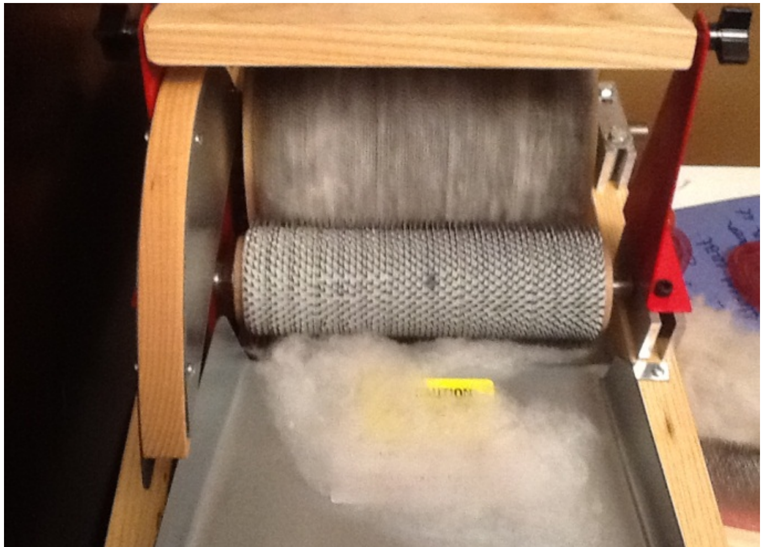
What is machine learning

- Algorithms?

Ch. 48 (multiple of 7 plus 6)

1. Right Side Row. Hdc in third ch from hook and in each remaining ch. Ch.1, turn.
2. And ALL Wrong Side Rows. Sc in each stitch across. Ch.2, turn.
3. *Work hdc in 4 sc, work post dc as follows: yo, insert hook right to left under post of hdc below next sc, draw up loop, (yo and draw through 2 loops) twice to complete dc. Skp sc behind post dc, work hdc in next sc, work post dc in hdc below next sc*, skip sc behind post dc; repeat from * to * across, ending with hdc in last 4 sc. Ch 1, turn.
4. Wrong side row, repeat row 2.
5. *Hdc in 4 sc, skip firt post dc, and work post dc on second post dc. Ch. 1, now work post dc around the skipped post dc (crossover made, skip the 3 sc behind crossover *; repeat from *to* across, ending with hdc in last 4 sc.
6. Wrong side row, repeat row 2.
7. * Hdc in 4 sc, work post dc in post dc below next sc (always skip sc behind each post dc), hdc in next sc, post dc in post dc below next sc*. Repeat from * to * ending with hdc in last 4 sc. Ch 1, turn. Repeat Rows 2 through 7 for pattern.

Sampling, Cleaning, Reformatting…

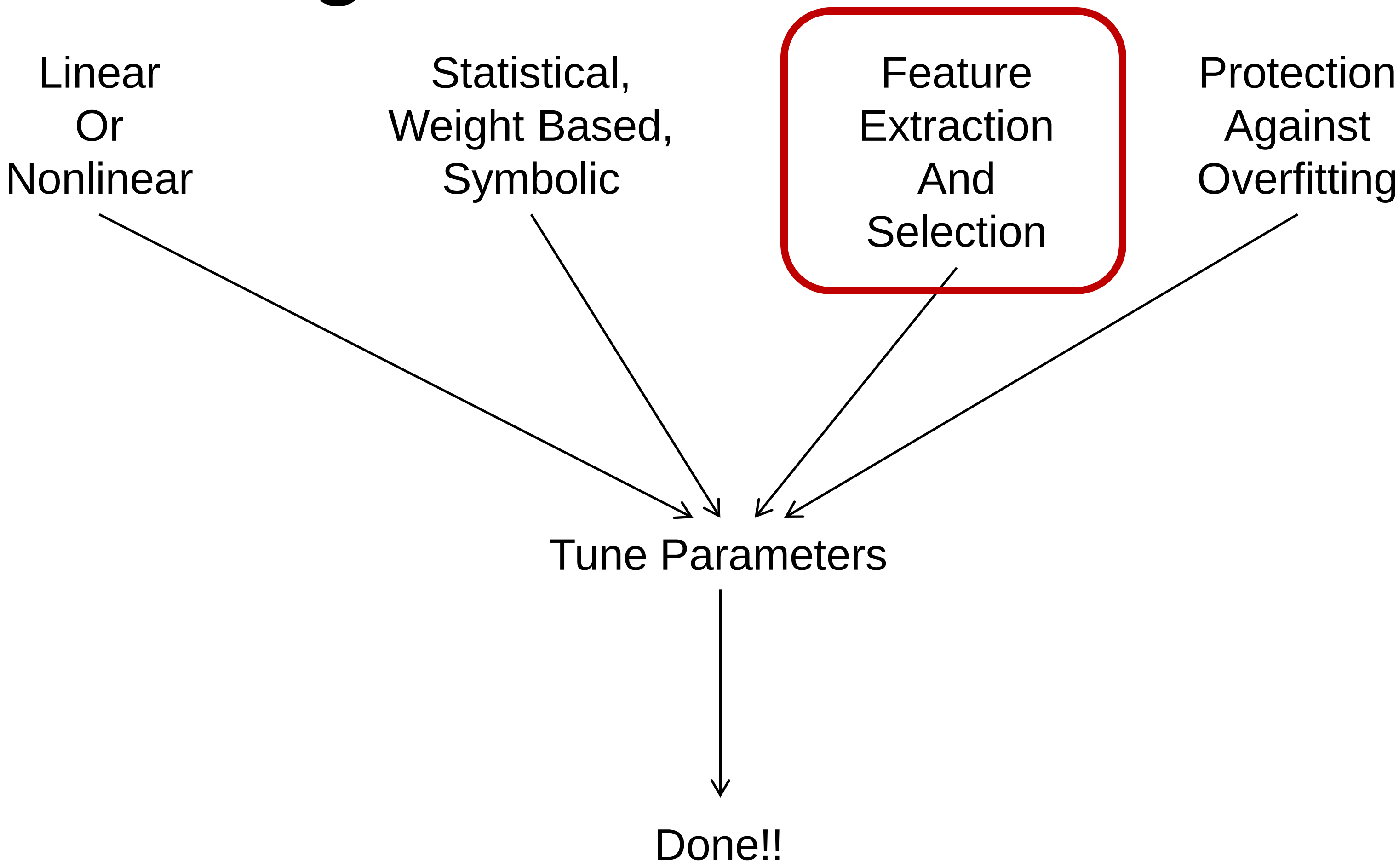


The rectangular batts
(like csv files with raw texts)
that come out of
the drum carder
are still not ready
to make something out of

Here's where you come in!

Text Teaser

Decisions about Machine Learning Methods



Consider this simple example…

SimpleExample.xls

	A	B
1	Code	text
2	Question	Tell me what your favorite color is.
3	Statement	I tell you my favorite color is blue.
4	Question	Where do you live?
5	Statement	I live where my family lives.
6	Question	Which kinds of baked goods do you prefer
7	Statement	I prefer to eat wheat bread.
8	Question	Which courses should I take?
9	Statement	You should take my applied machine learning course.
10	Question	Tell me when you get up in the morning.
11	Statement	I get up early.

Look for what distinguishes Questions and Statements in this dataset.

What clues do you see?

What are good features for text categorization?

SimpleExample.xls

	A	B
1	Code	text
2	Question	Tell me what your favorite color is.
3	Statement	I tell you my favorite color is blue.
4	Question	Where do you live?
5	Statement	I live where my family lives.
6	Question	Which kinds of baked goods do you prefer
7	Statement	I prefer to eat wheat bread.
8	Question	Which courses should I take?
9	Statement	You should take my applied machine learning course.
10	Question	Tell me when you get up in the morning.
11	Statement	I get up early.

What distinguishes Questions and Statements?

Not all questions end in a question mark.

What are good features for text categorization?

SimpleExample.xls

	A	B
1	Code	text
2	Question	Tell me what your favorite color is.
3	Statement	I tell you my favorite color is blue.
4	Question	Where do you live?
5	Statement	I live where my family lives.
6	Question	Which kinds of baked goods do you prefer
7	Statement	I prefer to eat wheat bread.
8	Question	Which courses should I take?
9	Statement	You should take my applied machine learning course.
10	Question	Tell me when you get up in the morning.
11	Statement	I get up early.

What distinguishes Questions and Statements?

I versus you is not a reliable predictor

What are good features for text categorization?

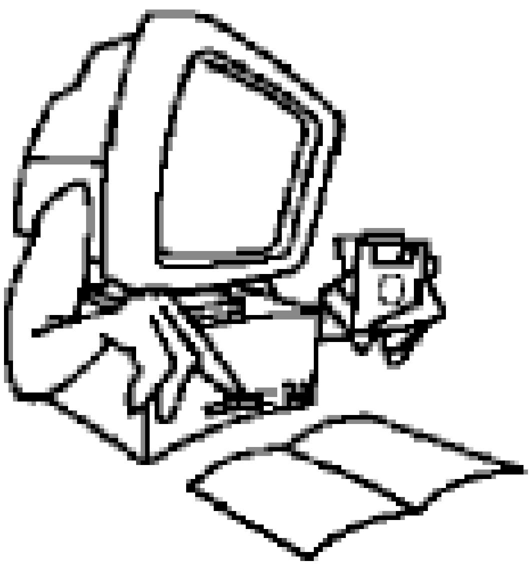
SimpleExample.xls

	A	B
1	Code	text
2	Question	Tell me what your favorite color is.
3	Statement	I tell you my favorite color is blue.
4	Question	Where do you live?
5	Statement	I live where my family lives.
6	Question	Which kinds of baked goods do you prefer
7	Statement	I prefer to eat wheat bread.
8	Question	Which courses should I take?
9	Statement	You should take my applied machine learning course.
10	Question	Tell me when you get up in the morning.
11	Statement	I get up early.

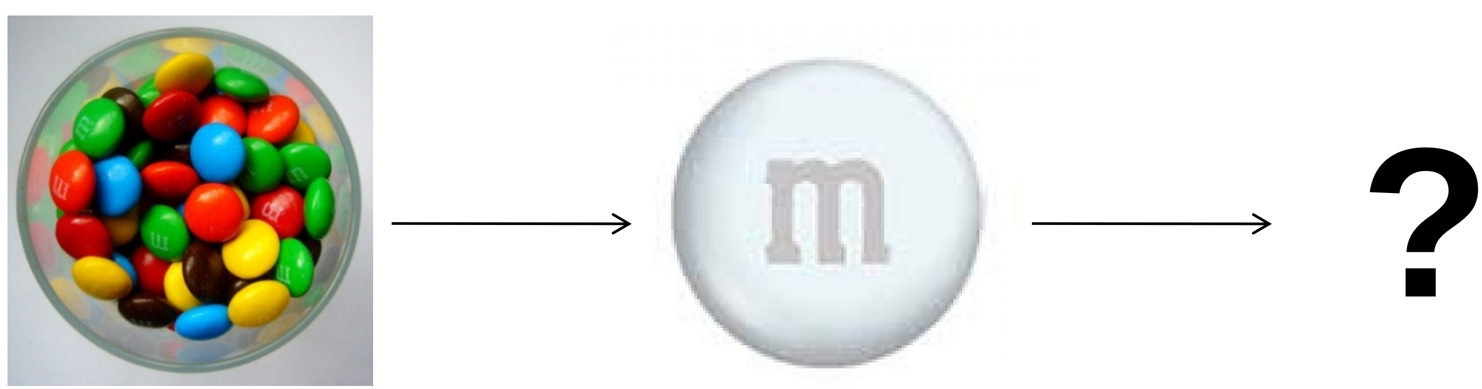
What distinguishes Questions and Statements?

Not all WH words occur in questions

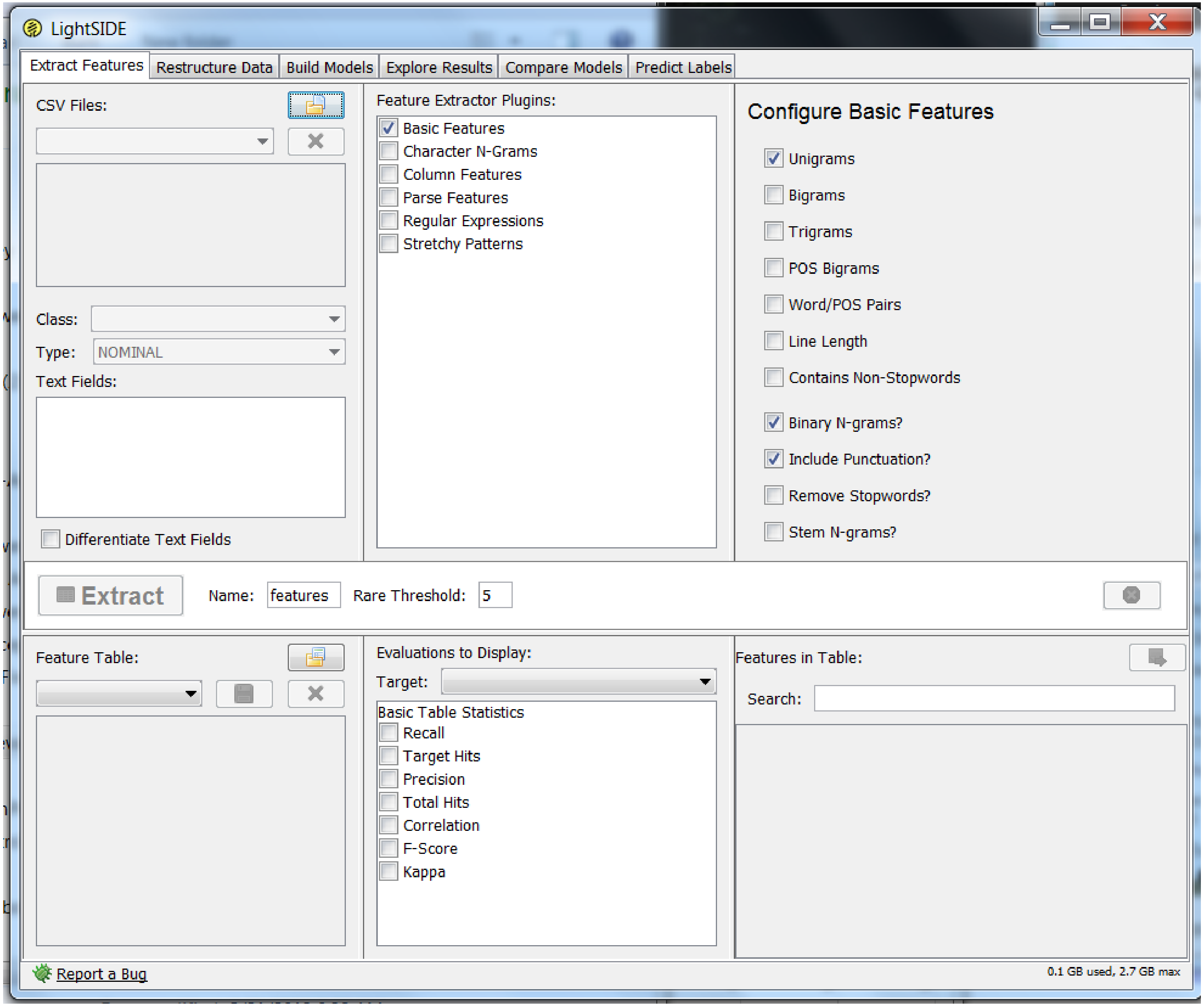
Effective data representations make problems learnable…

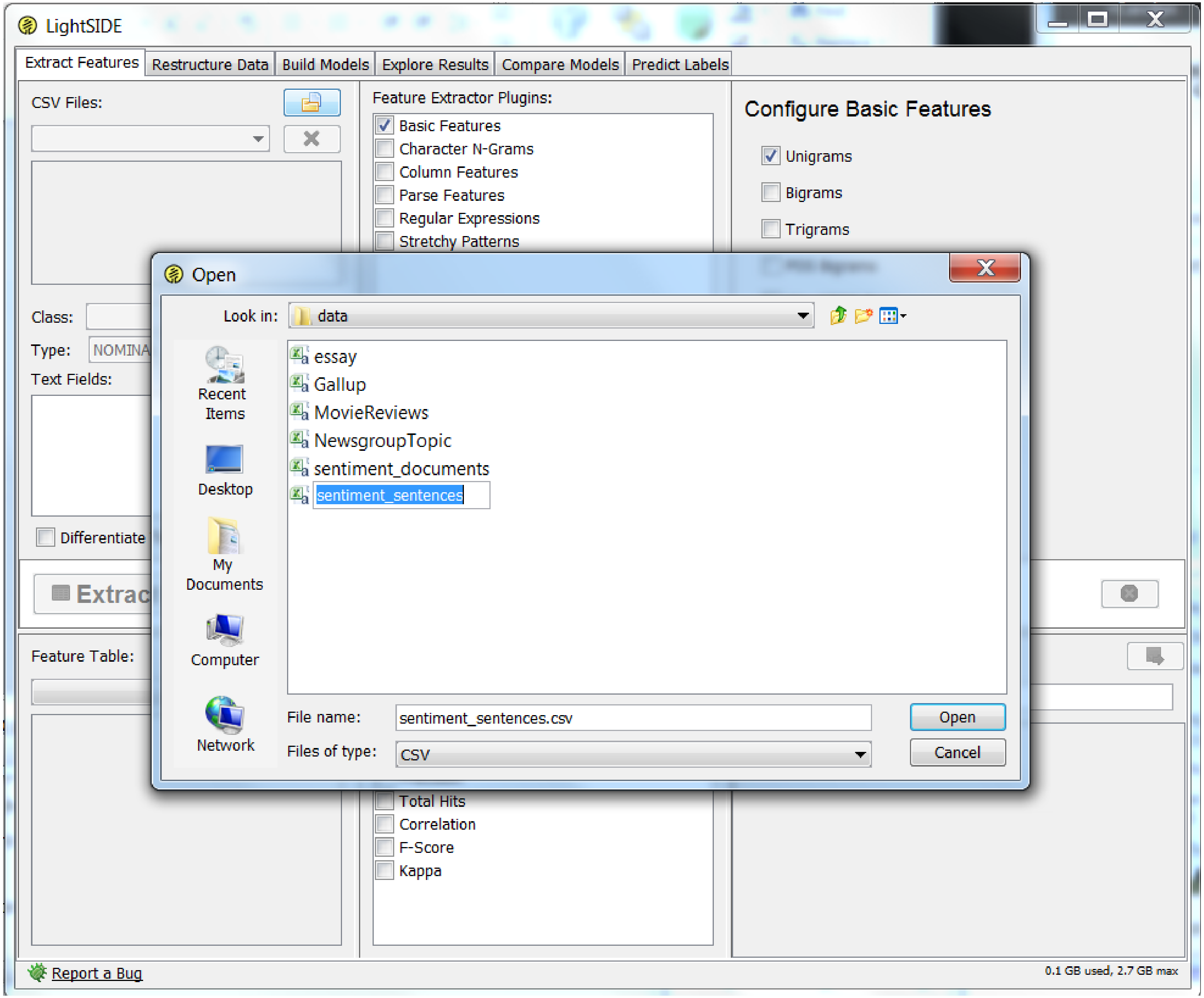


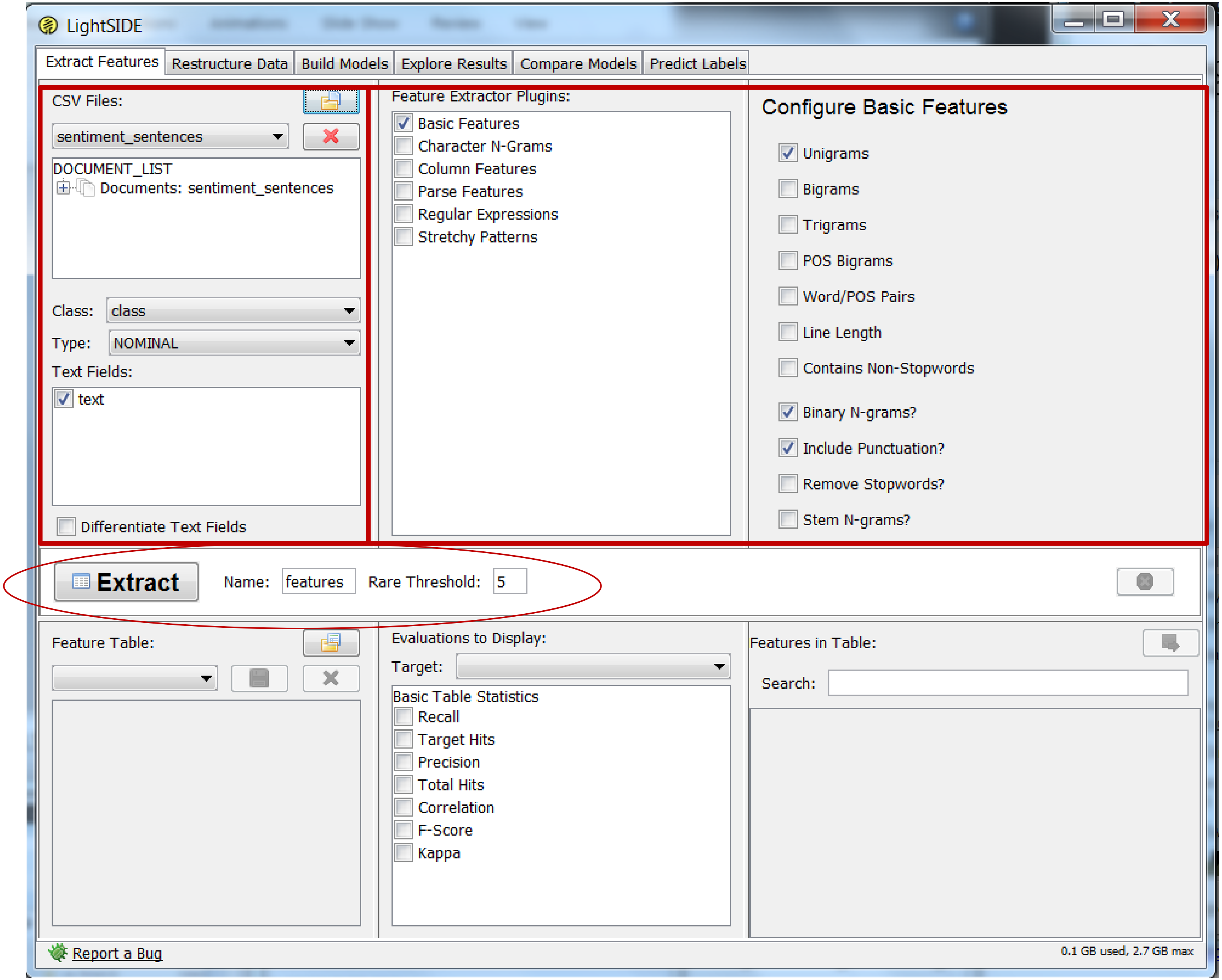
- Machine learning isn't magic
- But it ***can*** be useful for identifying meaningful patterns in your data when used properly
- Proper use requires insight into your data

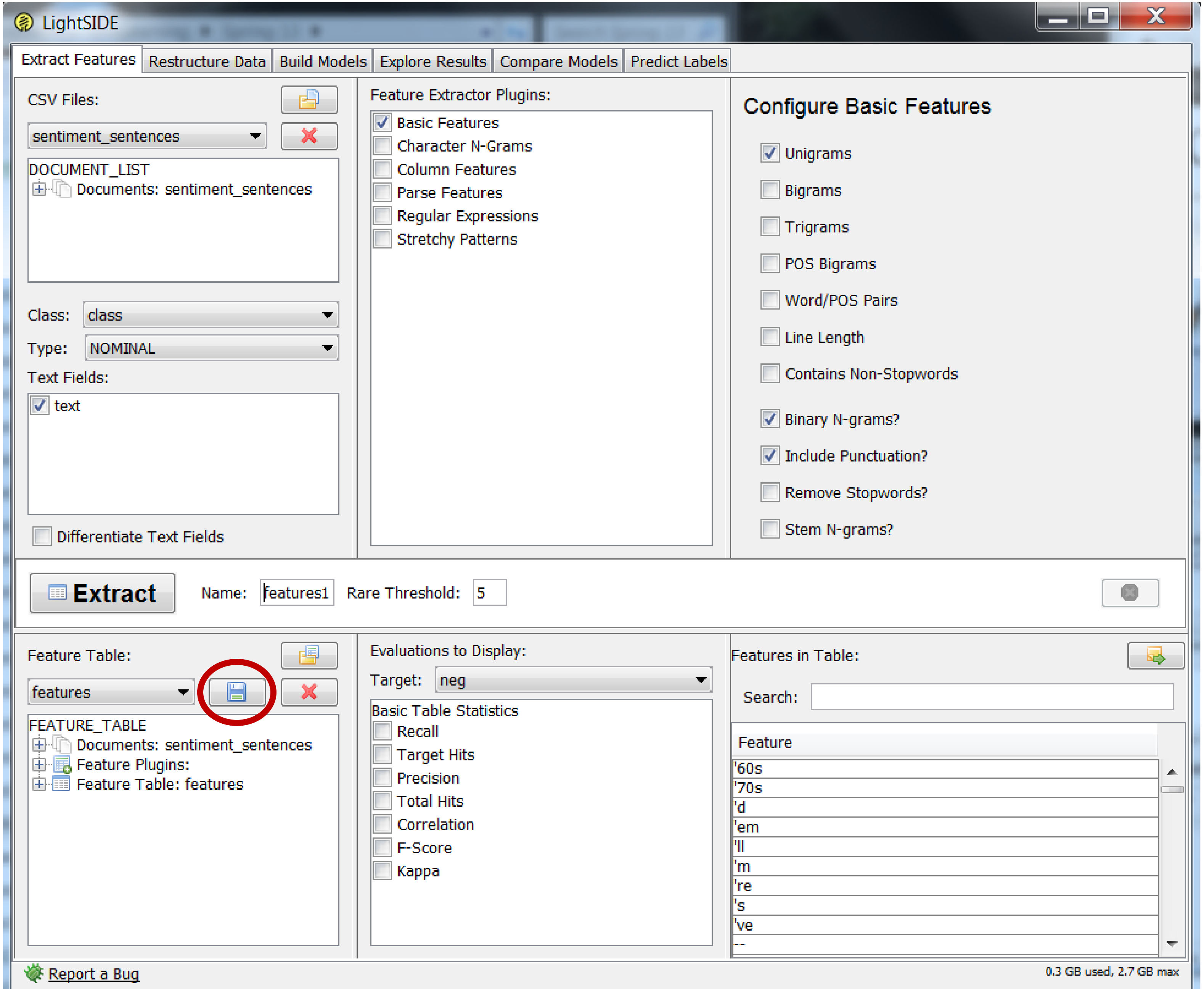


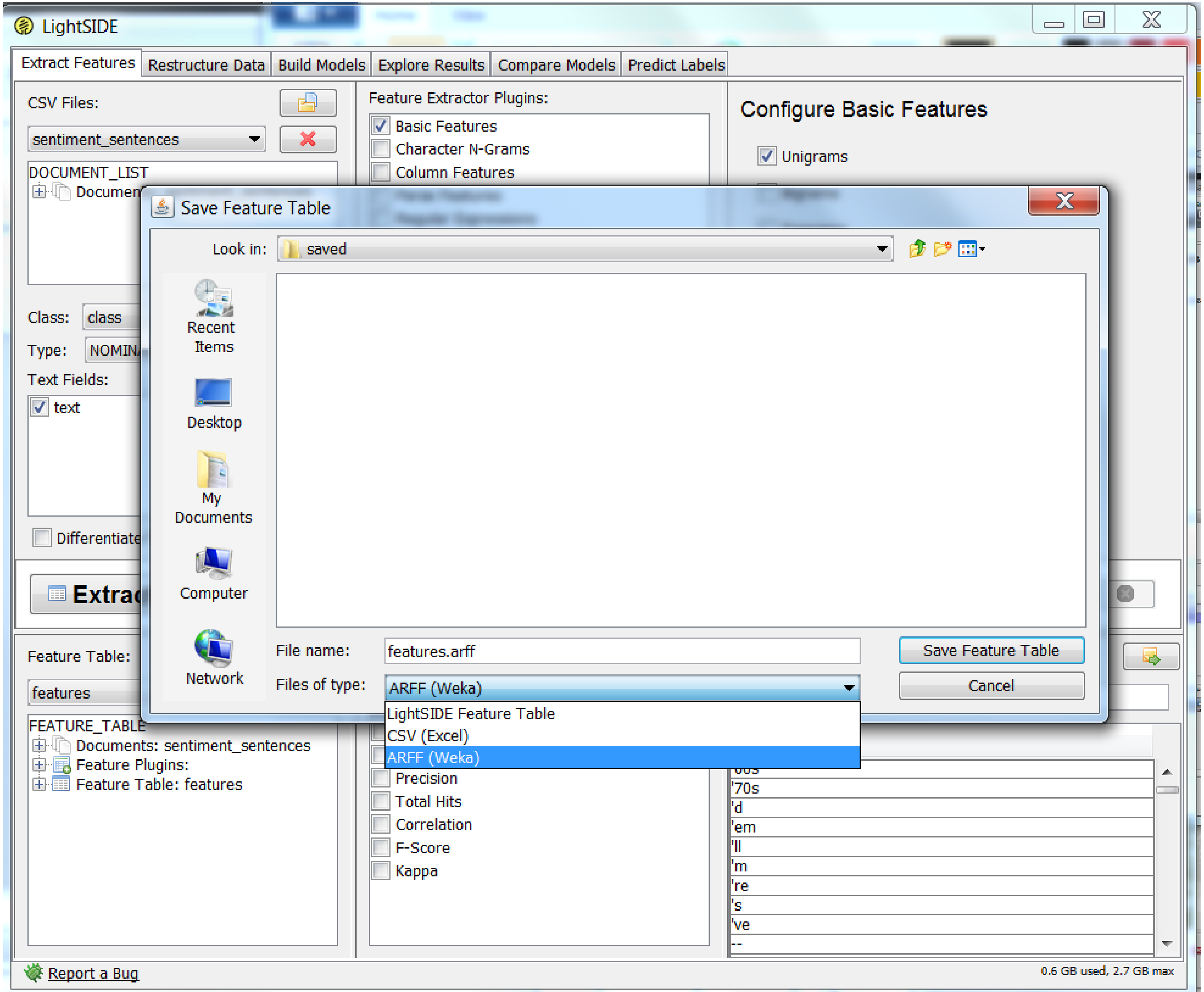
LightSIDE:
A quick tour


LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
Class:
Type: NOMINAL
Text Fields:
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Binary N-grams?
Include Punctuation?
Remove Stopwords?
Stem N-grams?
Extract
Name: features
Rare Threshold: 5
Feature Table:
Evaluations to Display:
Target:
Basic Table Statistics
Recall
Target Hits
Precision
Total Hits
Correlation
F-Score
Kappa
Features in Table:
Search:
Report a Bug
0.1 GB used, 2.7 GB max

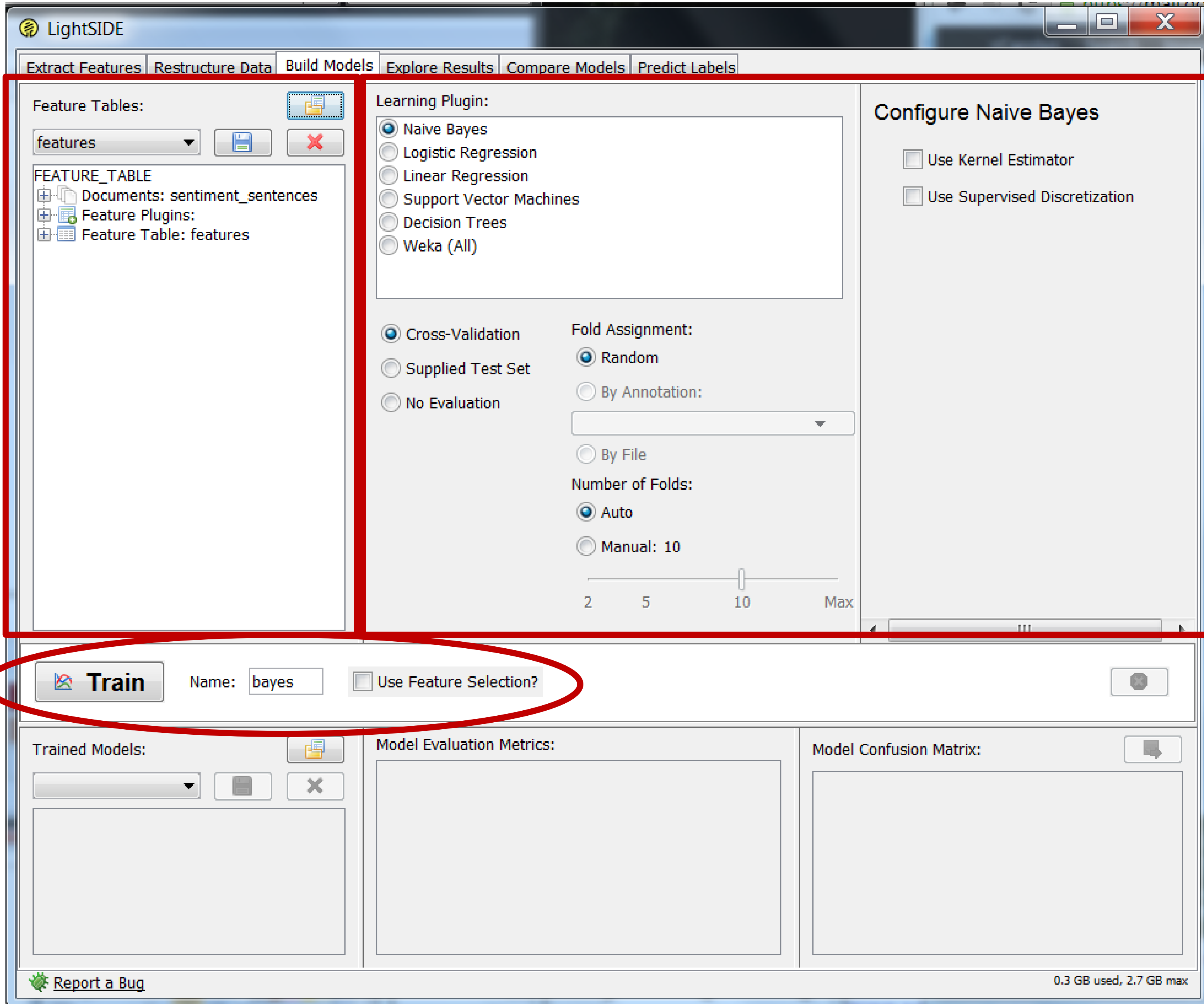

LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
Class:
Type:
NOMINA
Text Fields:
Differentiate
Extrac
Feature Table:
Open
Look in:
data
Recent Items
Desktop
My Documents
Computer
Network
essay
Gallup
MovieReviews
NewsgroupTopic
sentiment_documents
sentiment_sentences
File name:
sentiment_sentences.csv
Files of type:
CSV
Open
Cancel
Total Hits
Correlation
F-Score
Kappa
Report a Bug
0.1 GB used, 2.7 GB max

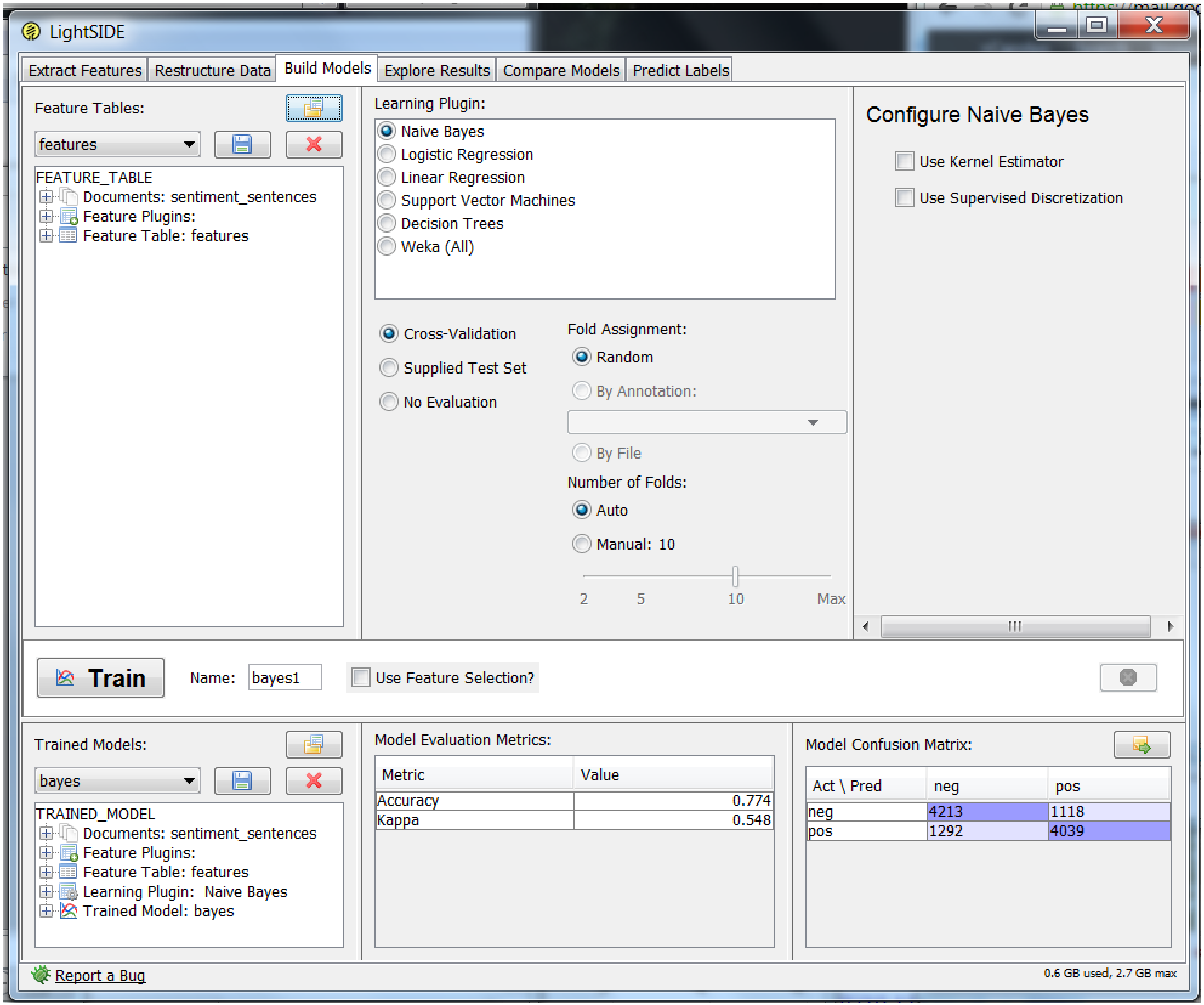

LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences
DOCUMENT_LIST
Documents: sentiment_sentences
Class: class
Type: NOMINAL
Text Fields:
text
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Binary N-grams?
Include Punctuation?
Remove Stopwords?
Stem N-grams?
Extract
Name: features
Rare Threshold: 5
Feature Table:
Evaluations to Display:
Target:
Basic Table Statistics
Recall
Target Hits
Precision
Total Hits
Correlation
F-Score
Kappa
Features in Table:
Search:
Report a Bug
0.1 GB used, 2.7 GB max


LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences
DOCUMENT_LIST
Documents: sentiment_sentences
Class: class
Type: NOMINAL
Text Fields:
text
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Binary N-grams?
Include Punctuation?
Remove Stopwords?
Stem N-grams?
Extract
Name: features1
Rare Threshold: 5
Feature Table:
features
FEATURE_TABLE
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Evaluations to Display:
Target: neg
Basic Table Statistics
Recall
Target Hits
Precision
Total Hits
Correlation
F-Score
Kappa
Features in Table:
Search:
Feature
'60s
'70s
'd
'em
'll
'm
're
's
've
--
Report a Bug
0.3 GB used, 2.7 GB max



LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences
DOCUMENT_LIST
Class:
class
Type:
Text Fields:
text
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Configure Basic Features
Unigrams
Save Feature Table
Look in:
saved
Recent Items
Desktop
My Documents
Computer
Network
File name:
features.arff
Files of type:
ARFF (Weka)
LightSIDE Feature Table
CSV (Excel)
ARFF (Weka)
Save Feature Table
Cancel
Feature Table:
features
FEATURE_TABLE
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Precision
Total Hits
Correlation
F-Score
Kappa
'70s
'd
'em
'll
'm
're
's
've
--
Report a Bug
0.6 GB used, 2.7 GB max

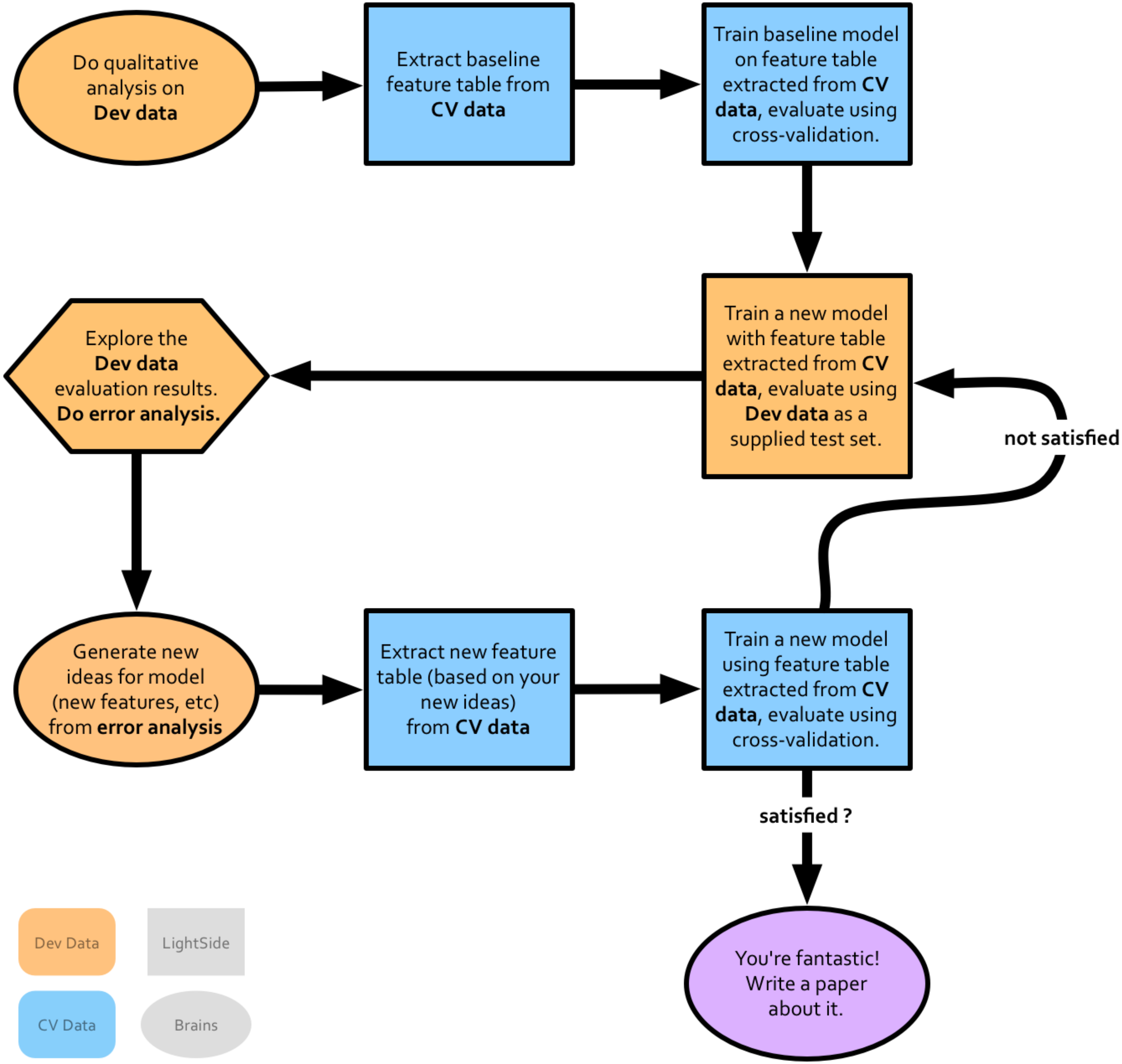

LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Configure Naive Bayes
Use Kernel Estimator
Use Supervised Discretization
Train
Name:
bayes
Use Feature Selection?
Trained Models:
Model Evaluation Metrics:
Model Confusion Matrix:
Report a Bug
0.3 GB used, 2.7 GB max


LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Configure Naive Bayes
Use Kernel Estimator
Use Supervised Discretization
Train
Name:
bayes1
Use Feature Selection?
Trained Models:
bayes
TRAINED_MODEL
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Learning Plugin: Naive Bayes
Trained Model: bayes
Model Evaluation Metrics:
Metric
Value
Accuracy
0.774
Kappa
0.548
Model Confusion Matrix:
Act \ Pred
neg
pos
neg
4213
1118
pos
1292
4039
Report a Bug
0.6 GB used, 2.7 GB max

Effective Development and Evaluation Process in LightSIDE

Avoiding Overfitting!

- Separate data for evaluation from data for exploration
- We will refer to the exploration set as the Dev Set
- We will refer to the evaluation set as the cross-validation set
- You should also have a final test set you never look at until you think you are done!


Do qualitative analysis on Dev data
Extract baseline feature table from CV data
Train baseline model on feature table extracted from CV data, evaluate using cross-validation.
Train a new model with feature table extracted from CV data, evaluate using Dev data as a supplied test set.
Explore the Dev data evaluation results. Do error analysis.
not satisfied
Generate new ideas for model (new features, etc) from error analysis
Extract new feature table (based on your new ideas) from CV data
Train a new model using feature table extracted from CV data, evaluate using cross-validation.
satisfied ?
You're fantastic! Write a paper about it.
Dev Data
LightSide
CV Data
Brains

Remember!!!!

- Use your development data for:
  - Qualitative analysis before ML
  - Error analysis
  - Ideas for design of new features
- Use your cross validation data for:
  - Evaluating your performance
- ***Never*** include the data you are testing on in the data you do feature selection with!!!

Basic Text Feature Extraction

Represent text as a vector where each position corresponds to a term

This is called the “bag of words” approach

Cheese
Cows
Eat
Hamsters
Make
Seeds

- Cows make cheese.
- 110010

- Hamsters eat seeds.
- 001101

Represent text as a vector where each position corresponds to a term

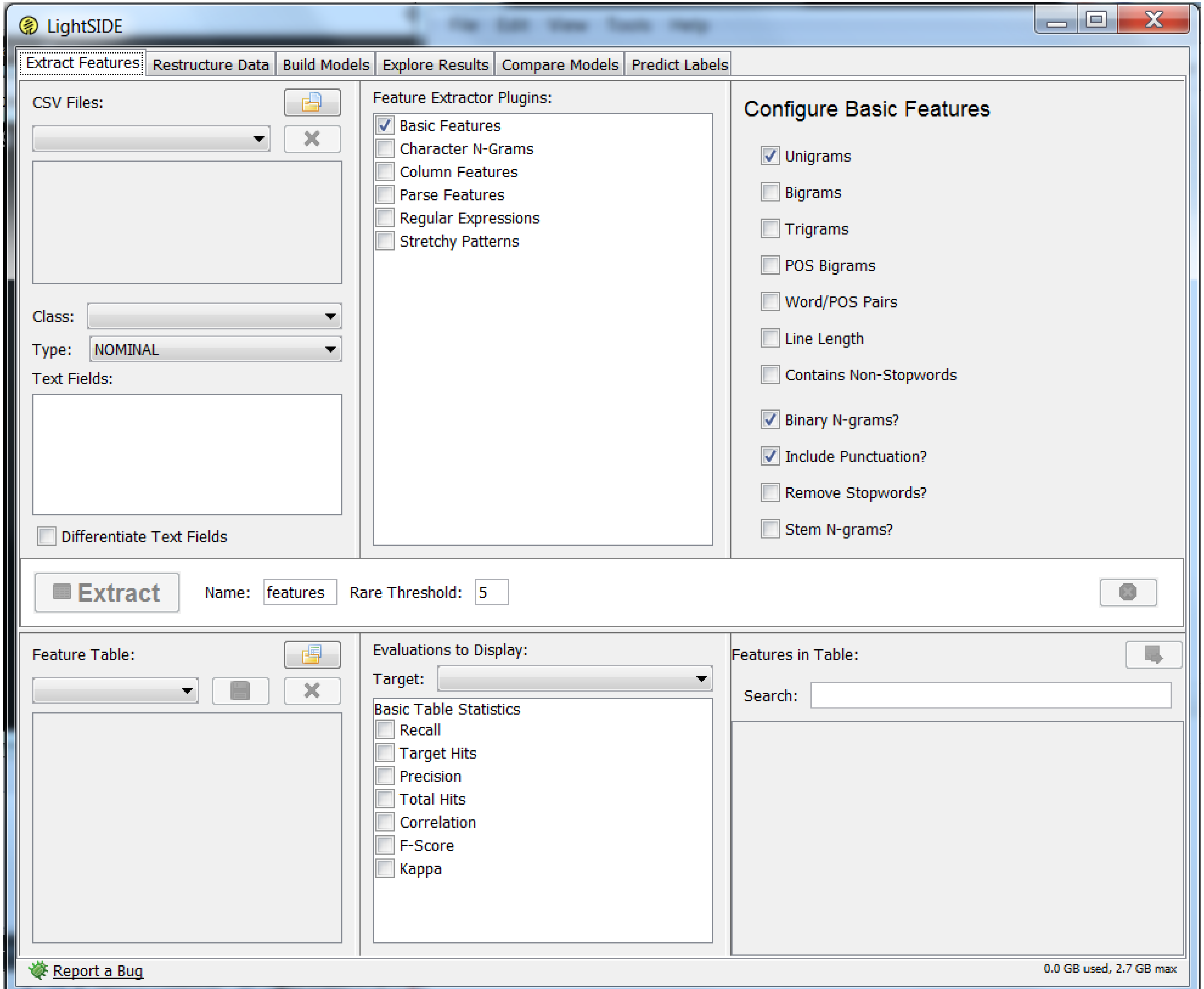
This is called the "bag of words" approach

But same representation for "Cheese makes cows."!

Cheese
Cows
Eat
Hamsters
Make
Seeds

- Cows make cheese.
- 110010

- Hamsters eat seeds.
- 001101



LightSIDE
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
Class:
Type: NOMINAL
Text Fields:
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Binary N-grams?
Include Punctuation?
Remove Stopwords?
Stem N-grams?
Extract
Name: features
Rare Threshold: 5
Feature Table:
Evaluations to Display:
Target:
Basic Table Statistics
Recall
Target Hits
Precision
Total Hits
Correlation
F-Score
Kappa
Features in Table:
Search:
Report a Bug
0.0 GB used, 2.7 GB max

U.S. government shuts down as Congress can't agree on spending bill

By **Tom Cohen**, CNN

updated 12:43 AM EDT, Tue October 1, 2013

Shutdown could hurt nation's veterans

STORY HIGHLIGHTS

- **NEW:** The House is expected to vote again overnight, including on appointing House negotiators
- "We will not go to conference with a gun to our head," says Sen. Harry Reid
- Obama says troops will get paid on time, but civilians may get more furloughs
- Conservatives wanted to undermine Obamacare before its private exchanges take effect Tuesday

Washington (CNN) -- The U.S. government shut down at 12:01 a.m. ET Tuesday after lawmakers in the House and the Senate could not agree on a spending bill to fund the government.

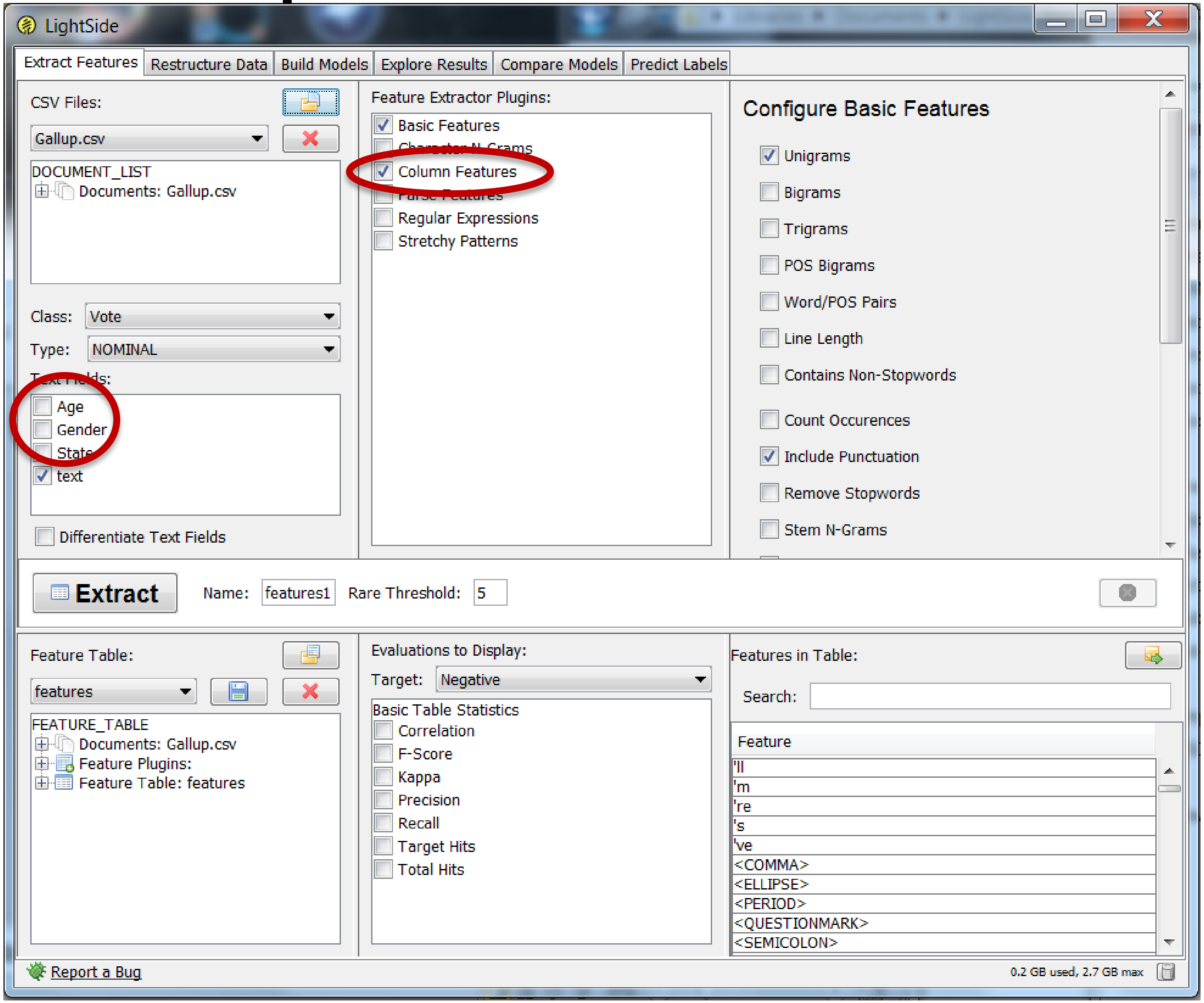
The two sides bickered and blamed each other for more than a week over Obamacare, the president's signature health care law. House Republicans insisted the spending bill include anti-Obamacare amendments. Senate Democrats were just as insistent that it didn't.

Federal employees who are considered essential will continue working. But employees deemed non-essential -- close to 800,000 -- will be furloughed.

Examples from Gallup Poll Data

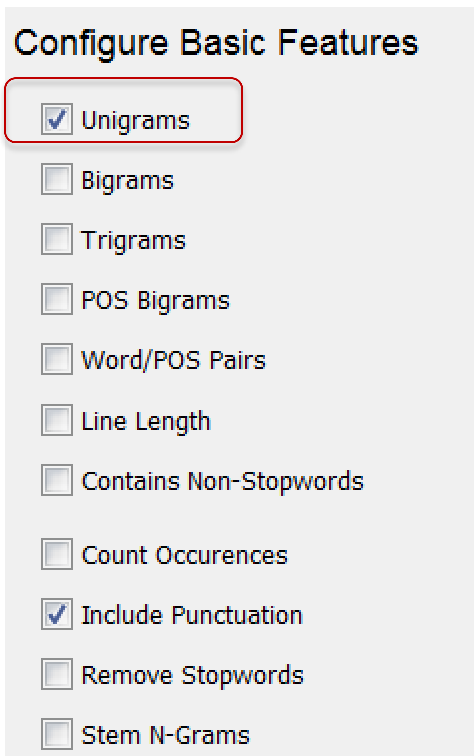
- Male from Virginia, age 30, negative: “I think it’ll increase costs for everyone.”
- Female from Illinois, unknown age, positive: “Because the cost of healthcare is just outta sight crazy”
- Male from Michigan, age 70, positive: “the cost”

The Gallup Poll Dataset



Basic Types of Features

"Because the cost of healthcare is just outta sight crazy"



Basic Types of Features

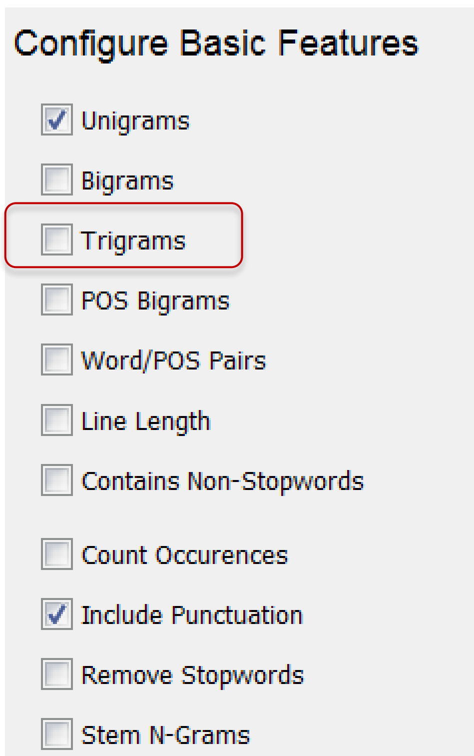
"Because the cost of healthcare is just outta sight crazy"

Configure Basic Features

- [x] Unigrams
- [ ] Bigrams
- [ ] Trigrams
- [ ] POS Bigrams
- [ ] Word/POS Pairs
- [ ] Line Length
- [ ] Contains Non-Stopwords
- [ ] Count Occurences
- [x] Include Punctuation
- [ ] Remove Stopwords
- [ ] Stem N-Grams

Basic Types of Features

“Because the cost of healthcare is just outta sight crazy”



Basic Types of Features

“the cost of healthcare”

DT NN PRP NN

Configure Basic Features

- [x] Unigrams
- [ ] Bigrams
- [ ] Trigrams
- [ ] POS Bigrams
- [ ] Word/POS Pairs
- [ ] Line Length
- [ ] Contains Non-Stopwords
- [ ] Count Occurences
- [x] Include Punctuation
- [ ] Remove Stopwords
- [ ] Stem N-Grams

Part of Speech Tagging

http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition/subord
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PP Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative

Part of Speech Tagging

http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund/present participle
30. VBN Verb, past participle
31. VBP Verb, non-3rd ps. sing. present
32. VBZ Verb, 3rd ps. sing. present
33. WDT wh-determiner
34. WP wh-pronoun
35. WP Possessive wh-pronoun
36. WRB wh-adverb

Basic Types of Features

"the cost of healthcare"

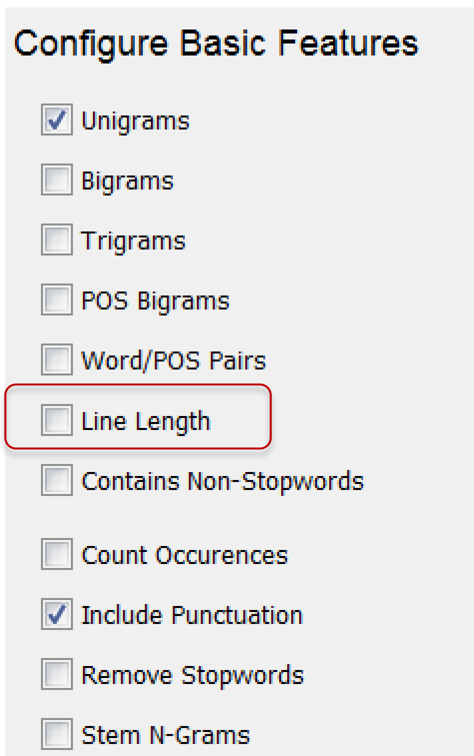
the	cost	of	healthcare
DT	NN	PRP	NN

Configure Basic Features

- [x] Unigrams
- [ ] Bigrams
- [ ] Trigrams
- [ ] POS Bigrams
- [ ] Word/POS Pairs
- [ ] Line Length
- [ ] Contains Non-Stopwords
- [ ] Count Occurences
- [x] Include Punctuation
- [ ] Remove Stopwords
- [ ] Stem N-Grams

Basic Types of Features

“the cost of healthcare” **4**



Basic Types of Features

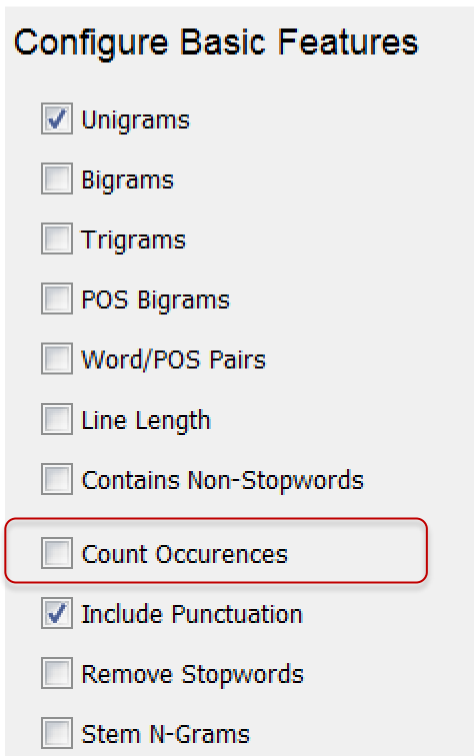
"the cost of healthcare" **YES**

Configure Basic Features

- [x] Unigrams
- [ ] Bigrams
- [ ] Trigrams
- [ ] POS Bigrams
- [ ] Word/POS Pairs
- [ ] Line Length
- [ ] Contains Non-Stopwords
- [ ] Count Occurences
- [x] Include Punctuation
- [ ] Remove Stopwords
- [ ] Stem N-Grams

Basic Types of Features

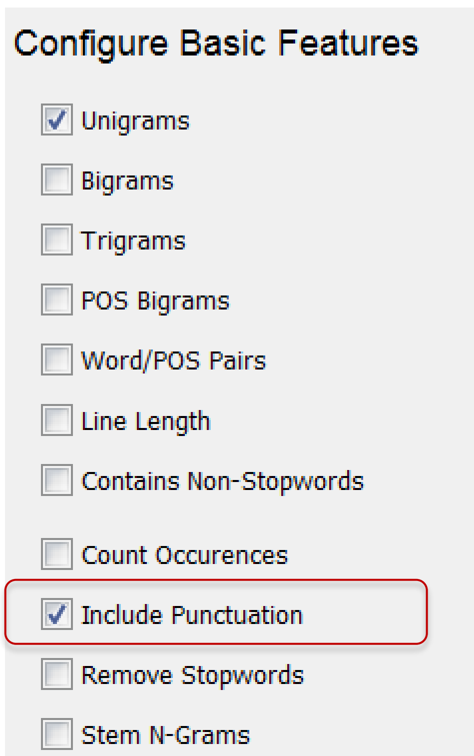
"the cost is too great. The cost is immense!"



The value of the feature is the number of times it occurs, rather than 1 if it occurs or 0 otherwise, which is the default.

Basic Types of Features

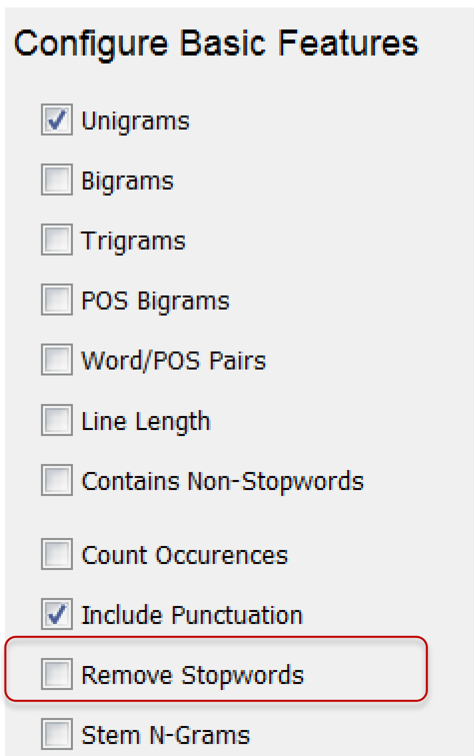
“the cost is too great. The cost is immense!”



If you uncheck this, punctuation will be ignored and stripped out of the representation.

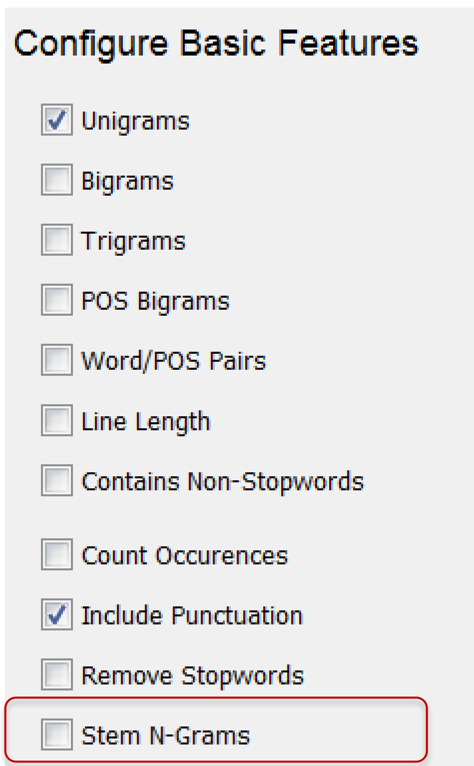
Basic Types of Features

“~~the~~ cost ~~of~~ healthcare”



Basic Types of Features

“healthcare cost**s**” → “healthcare cost”



Clarification on Basic text feature extractor

- POS tagging happens before stemming or stopword removal
- POS bigrams are not affected by stopword removal – POS tags for stopwords will still be included
- On word n-grams, the only n-grams that will be dropped in the case of stopword removal are ones that consist only of stopwords

Feature Space Customizations

- Feature Space Design
  - Think like a computer!
  - Machine learning algorithms look for features that are good predictors, not features that are necessarily meaningful
  - Look for approximations
    - If you want to find questions, you don’t need to do a complete syntactic analysis
    - Look for question marks
    - Look for wh-terms that occur immediately before an auxilliary verb

Error Analysis

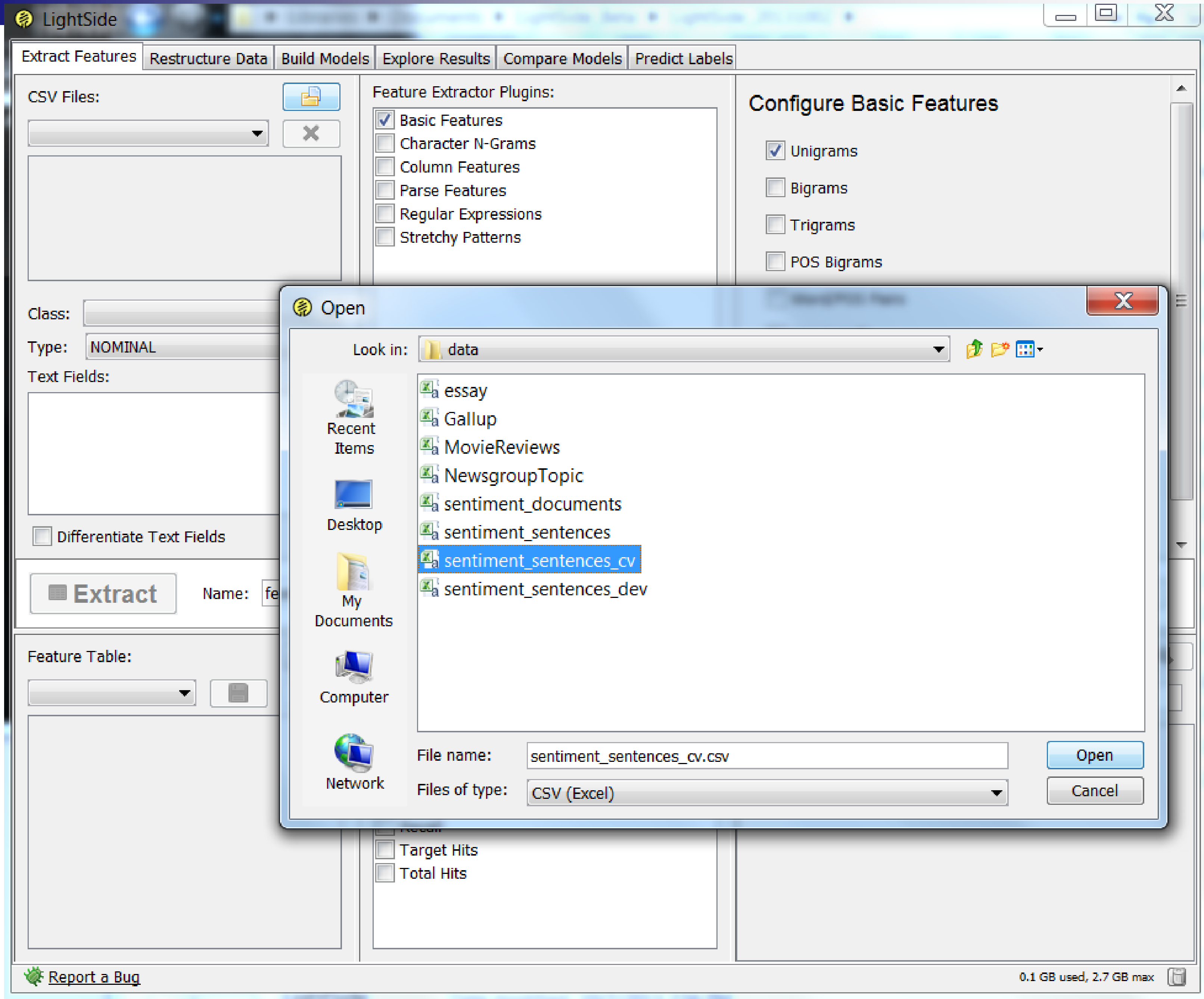
Error Analysis Process

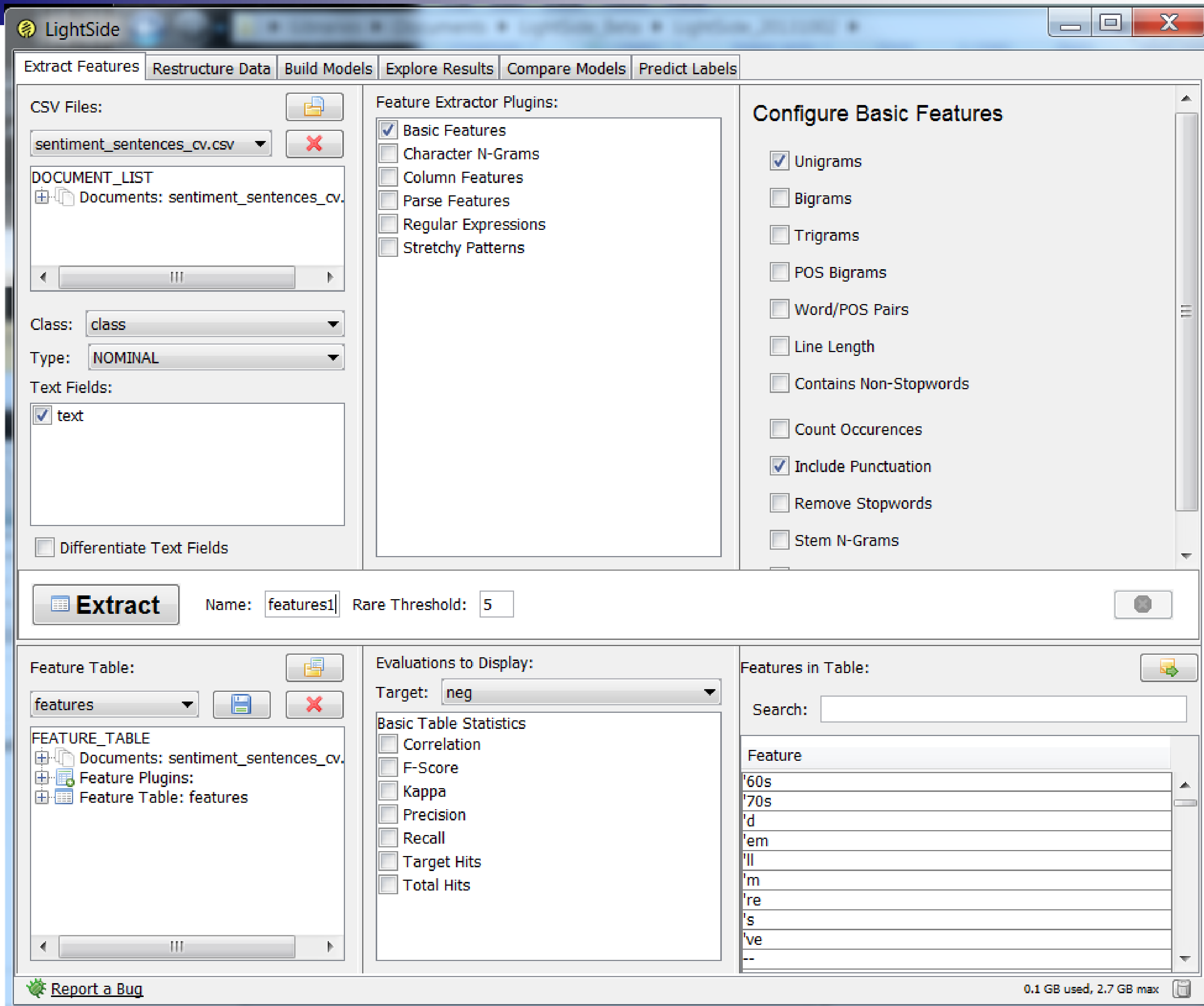
High Level Overview

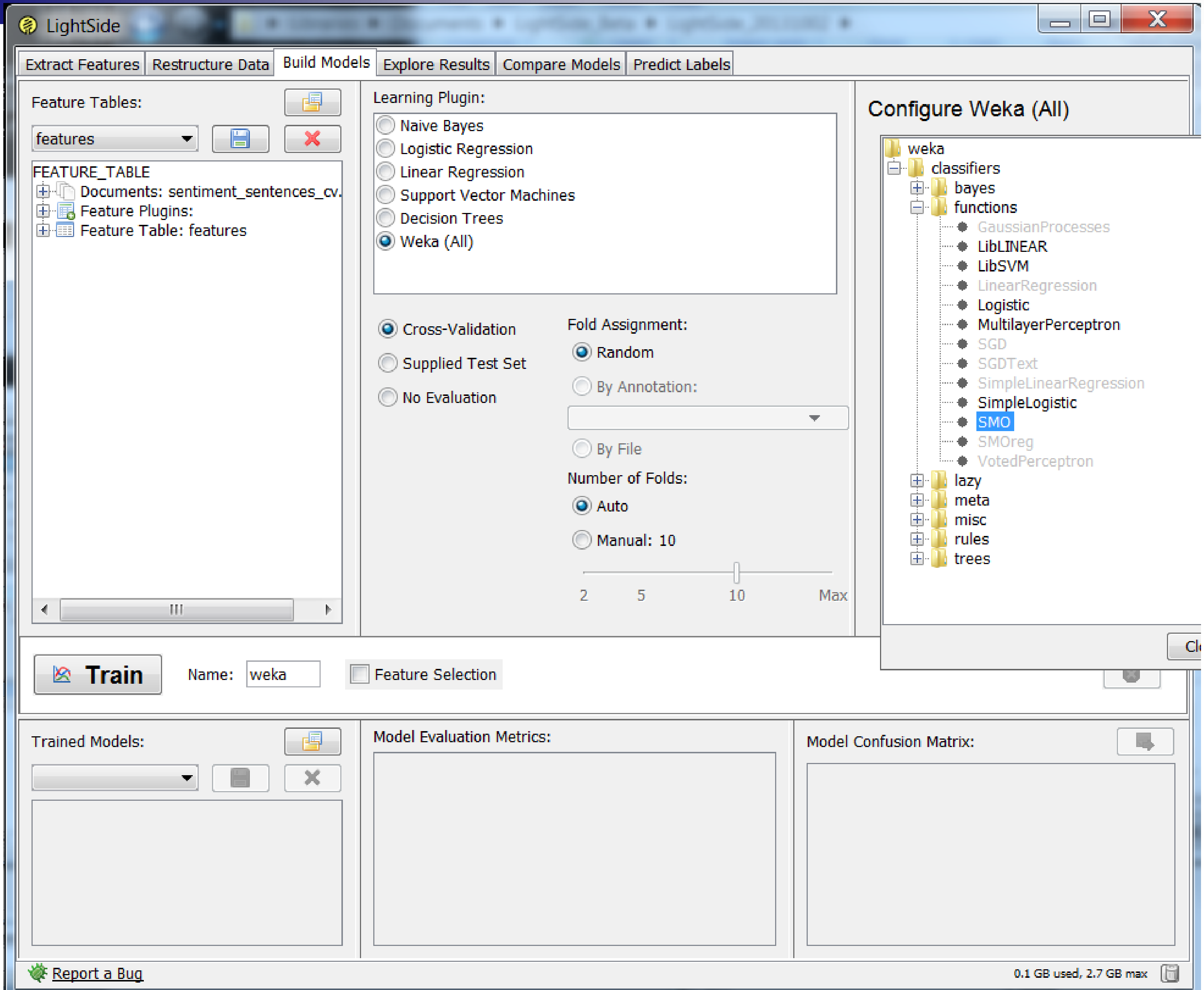
```
=== Confusion Matrix ===

   a   b   <-- classified as
  44  25 |
  10 149 |
```

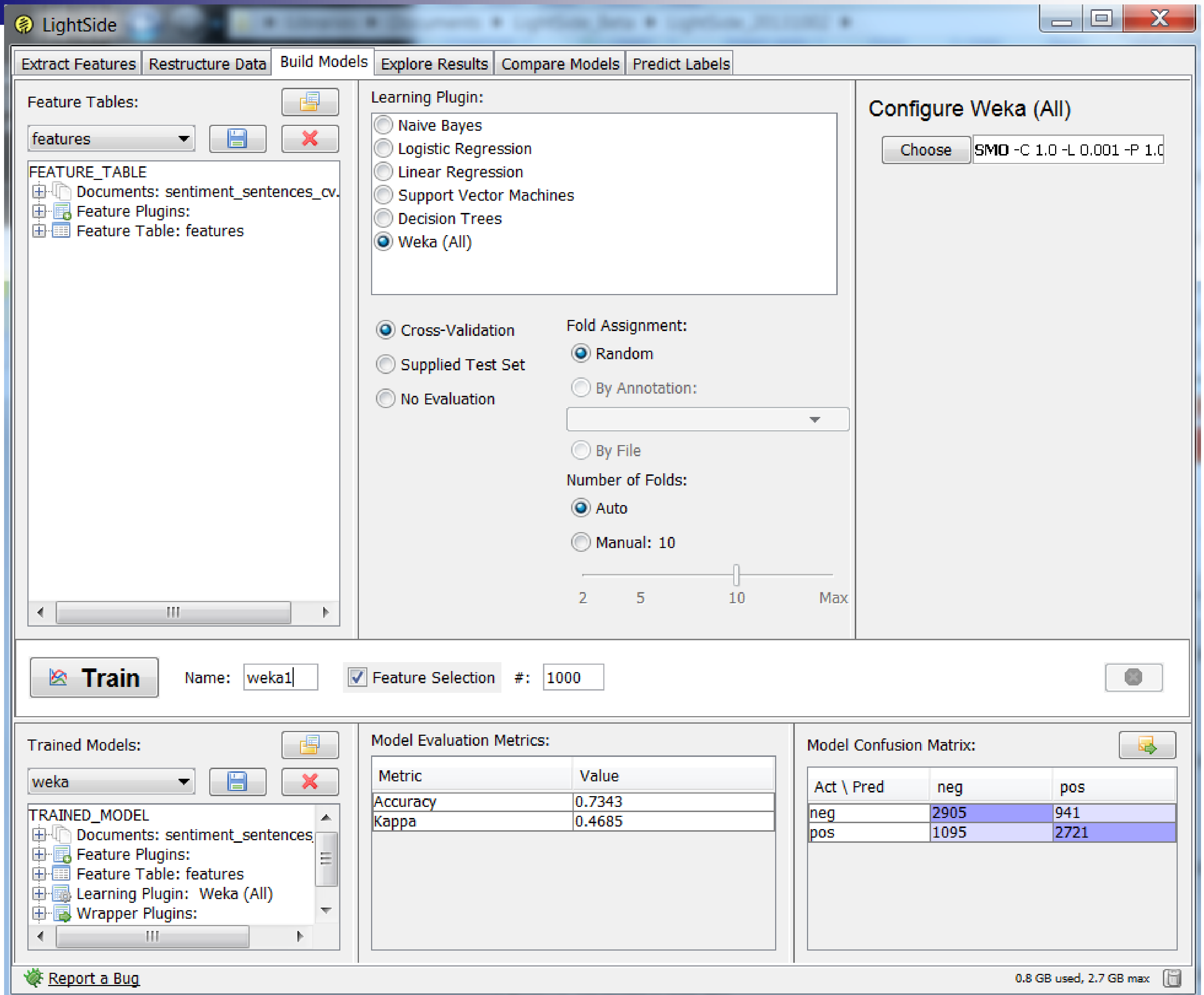
- Identify large error cells
- Make comparisons
  - Ask yourself how it is similar to the instances that were correctly classified with the same class **(vertical comparison)**
  - How it is different from those it was incorrectly not classified as (**horizontal comparison**)

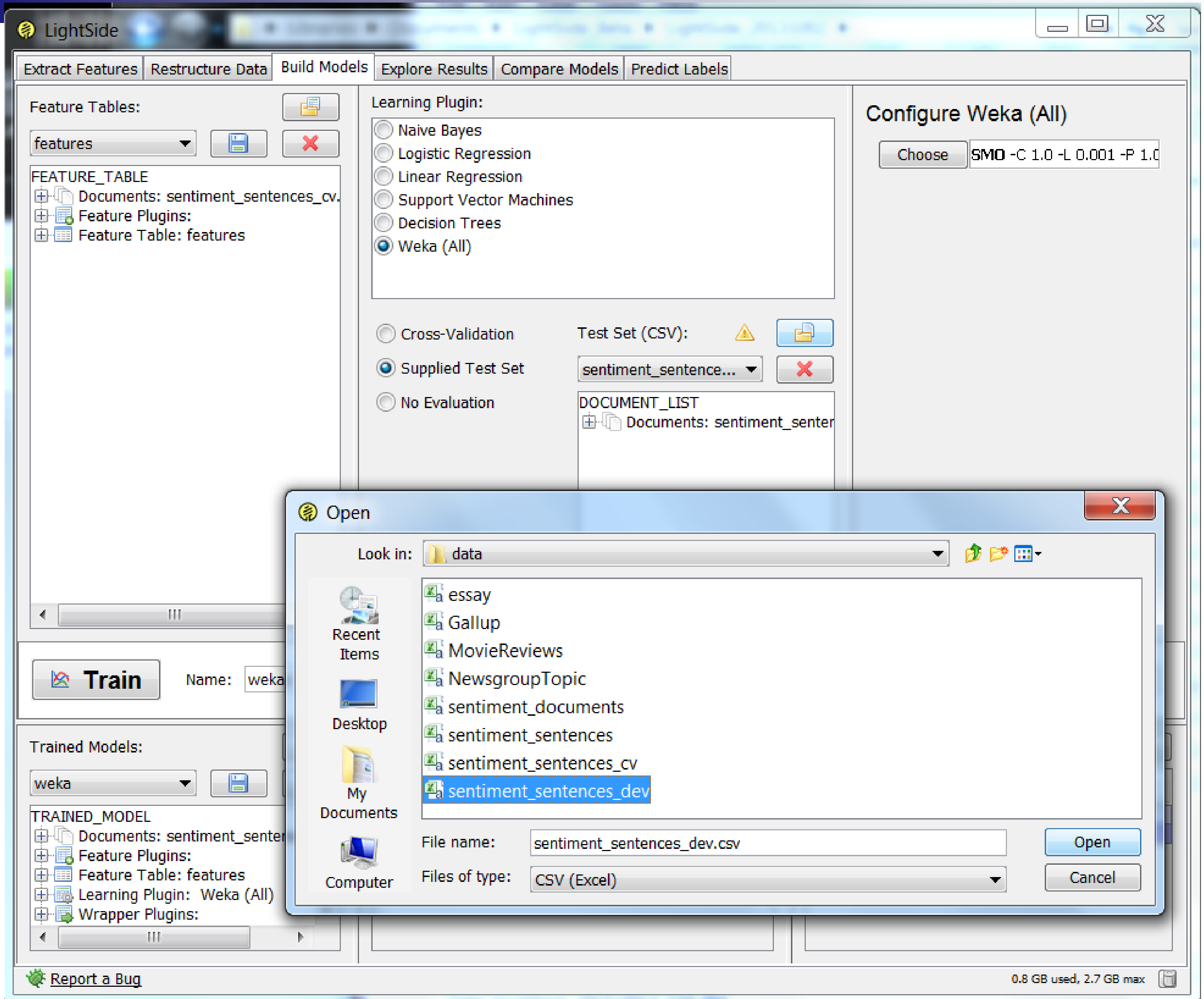
Goal: We want to discover how to re-represent the data so that instances with the same class value look more similar to one another and instances with different class values look more different

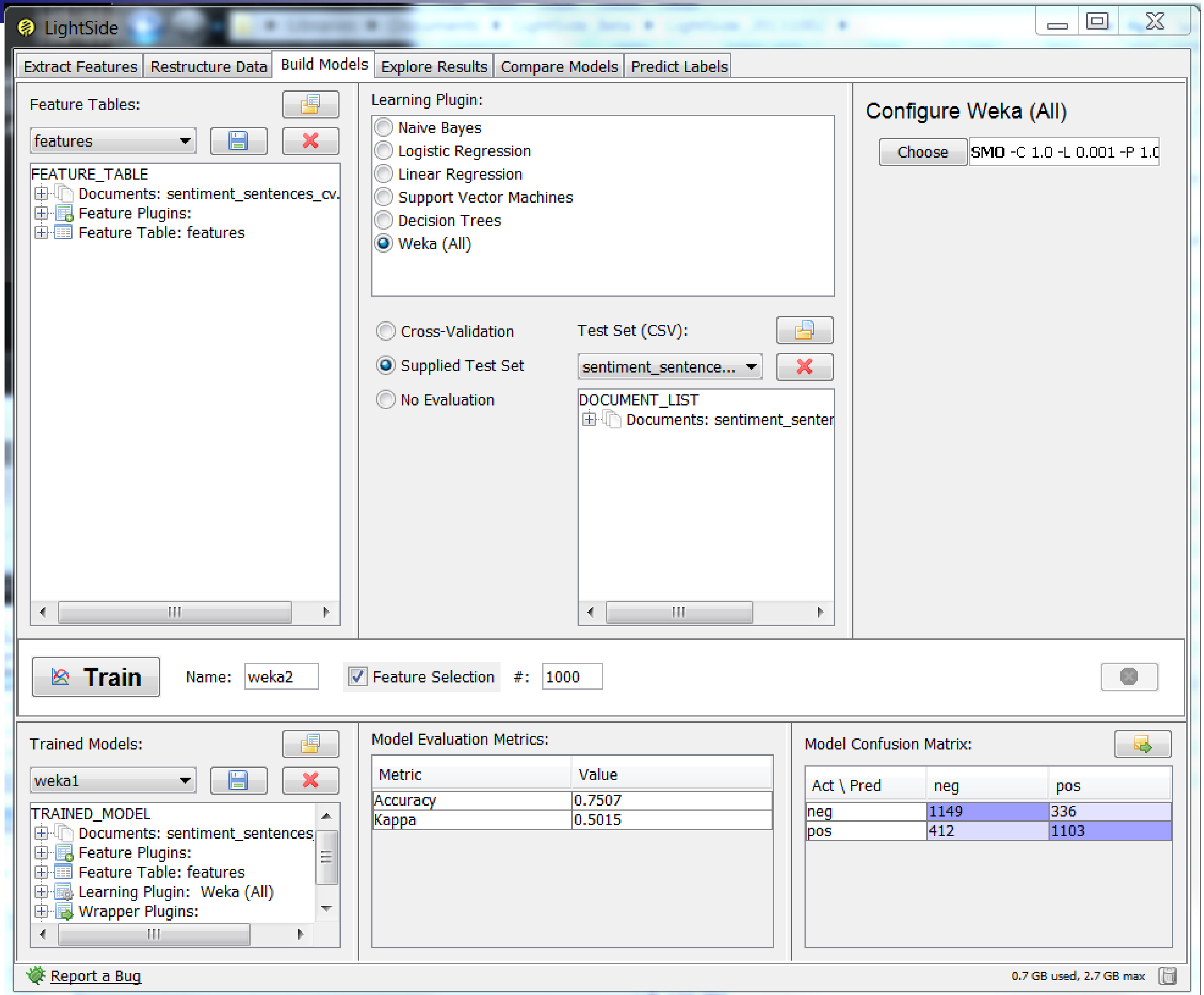

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
POS Bigrams
Class:
Type:
NOMINAL
Text Fields:
Differentiate Text Fields
Extract
Name:
Feature Table:
Open
Look in:
data
Recent Items
Desktop
My Documents
Computer
Network
essay
Gallup
MovieReviews
NewsgroupTopic
sentiment_documents
sentiment_sentences
sentiment_sentences_cv
sentiment_sentences_dev
File name:
sentiment_sentences_cv.csv
Files of type:
CSV (Excel)
Open
Cancel
Target Hits
Total Hits
Report a Bug
0.1 GB used, 2.7 GB max

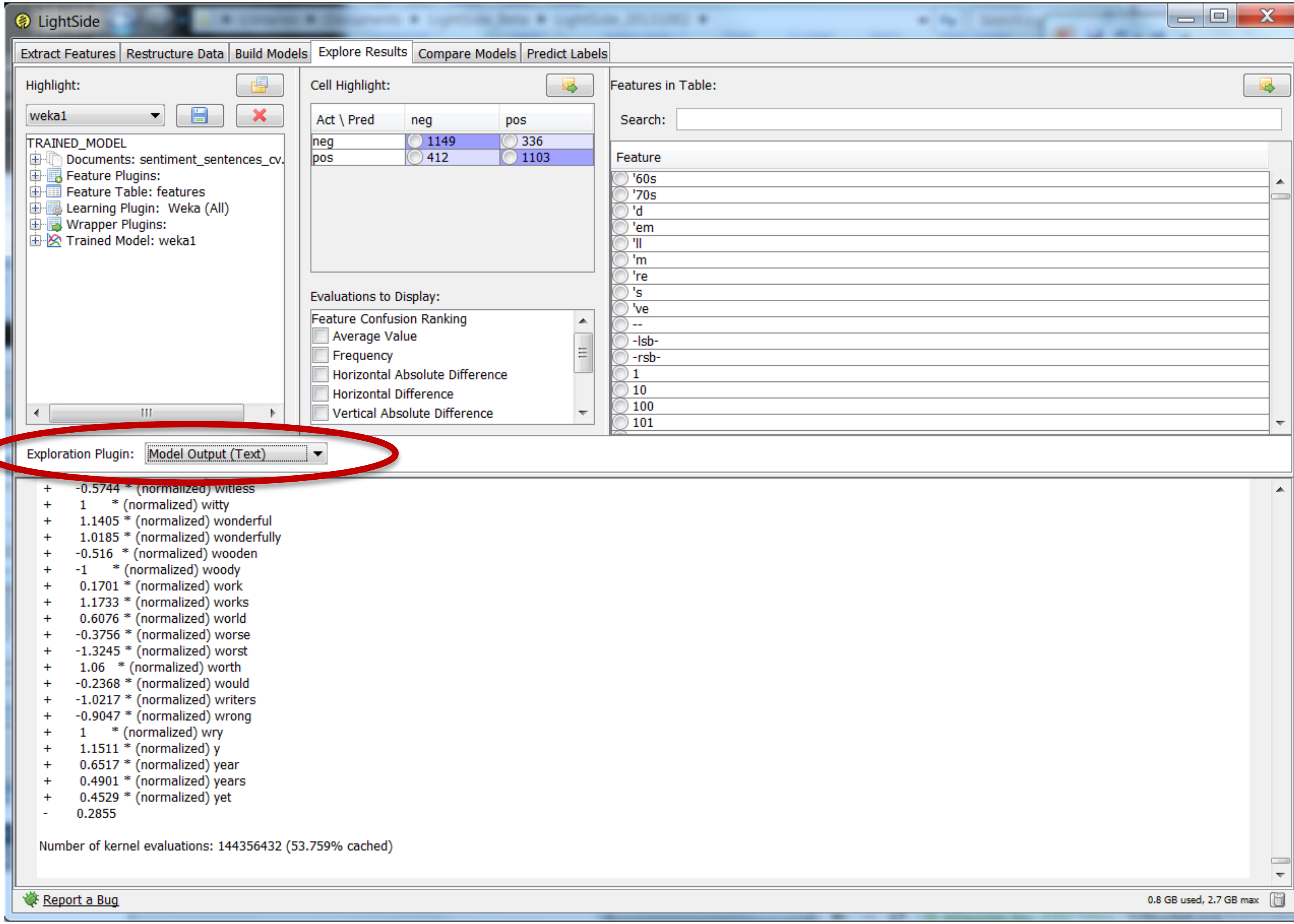

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences_cv.csv
DOCUMENT_LIST
Documents: sentiment_sentences_cv.
Class: class
Type: NOMINAL
Text Fields:
text
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Basic Features
Unigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Count Occurences
Include Punctuation
Remove Stopwords
Stem N-Grams
Extract
Name: features1
Rare Threshold: 5
Feature Table:
features
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Evaluations to Display:
Target: neg
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Feature
'60s
'70s
'd
'em
'll
'm
're
's
've
--
Report a Bug
0.1 GB used, 2.7 GB max

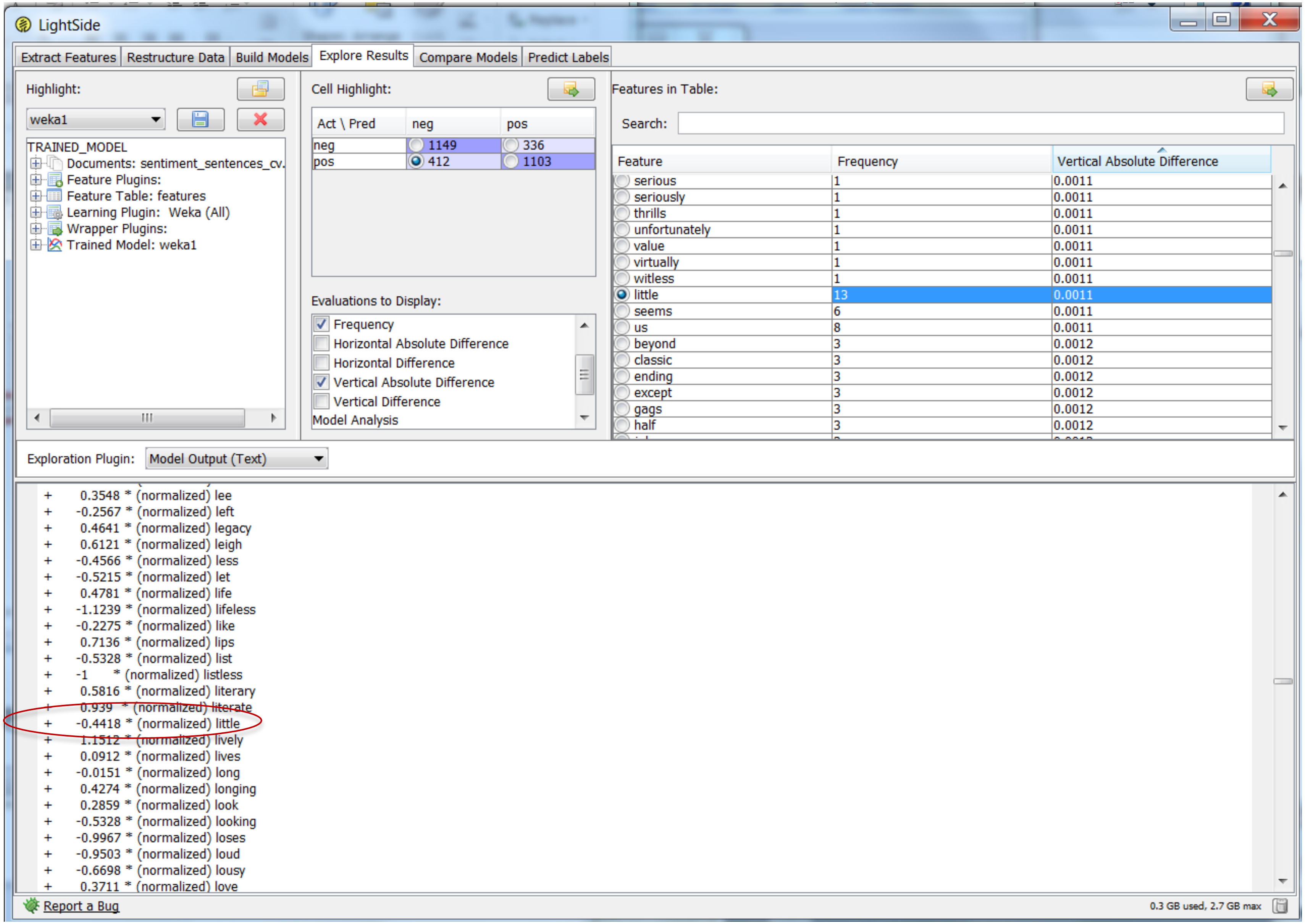

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Configure Weka (All)
weka
classifiers
bayes
functions
GaussianProcesses
LibLINEAR
LibSVM
LinearRegression
Logistic
MultilayerPerceptron
SGD
SGDText
SimpleLinearRegression
SimpleLogistic
SMO
SMOreg
VotedPerceptron
lazy
meta
misc
rules
trees
Train
Name:
weka
Feature Selection
Trained Models:
Model Evaluation Metrics:
Model Confusion Matrix:
Report a Bug
0.1 GB used, 2.7 GB max

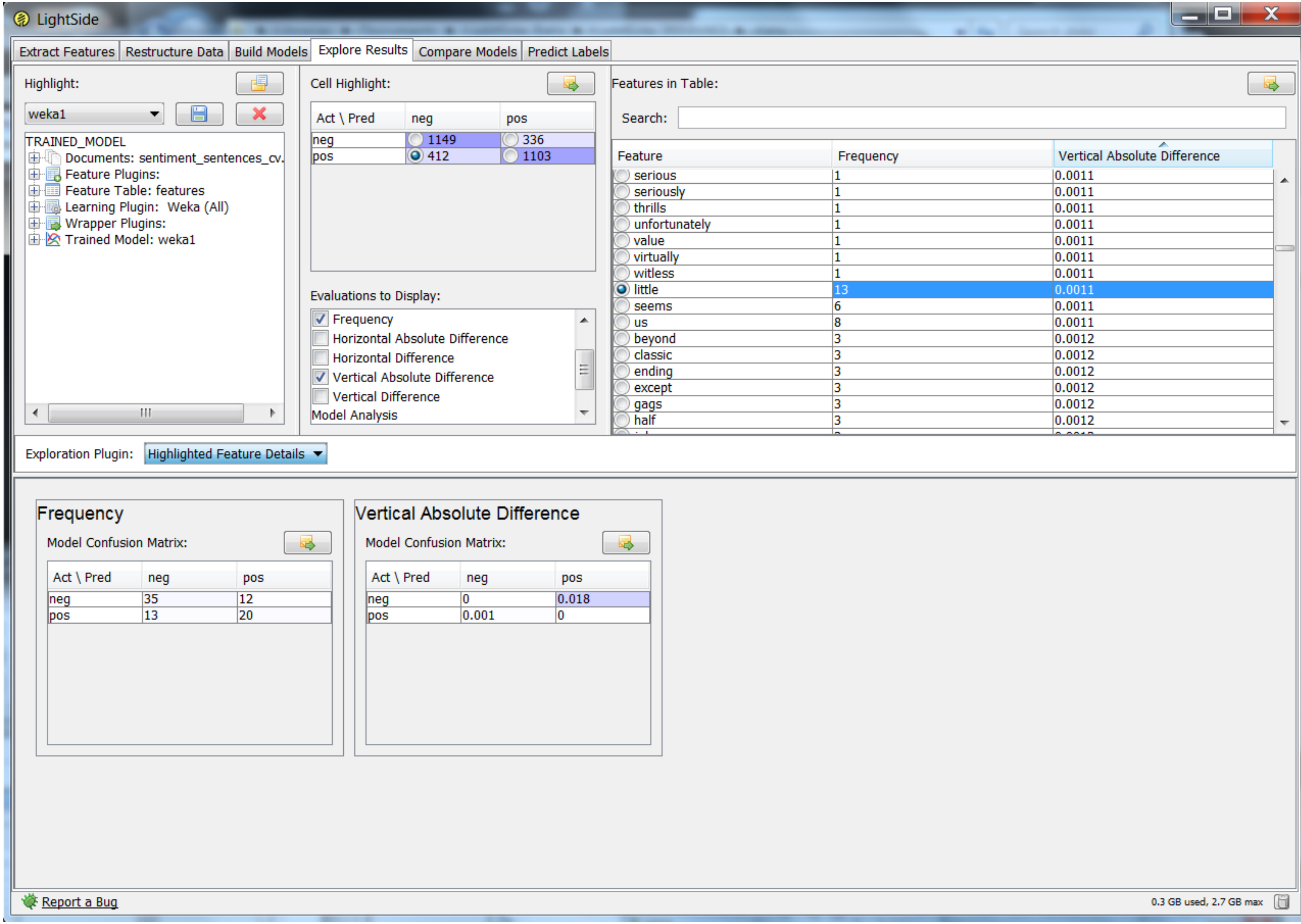

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Configure Weka (All)
Choose
SMO -C 1.0 -L 0.001 -P 1.0
Train
Name:
weka1
Feature Selection
#:
1000
Trained Models:
weka
TRAINED_MODEL
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Model Evaluation Metrics:
Metric
Value
Accuracy
0.7343
Kappa
0.4685
Model Confusion Matrix:
Act \ Pred
neg
pos
neg
2905
941
pos
1095
2721
Report a Bug
0.8 GB used, 2.7 GB max

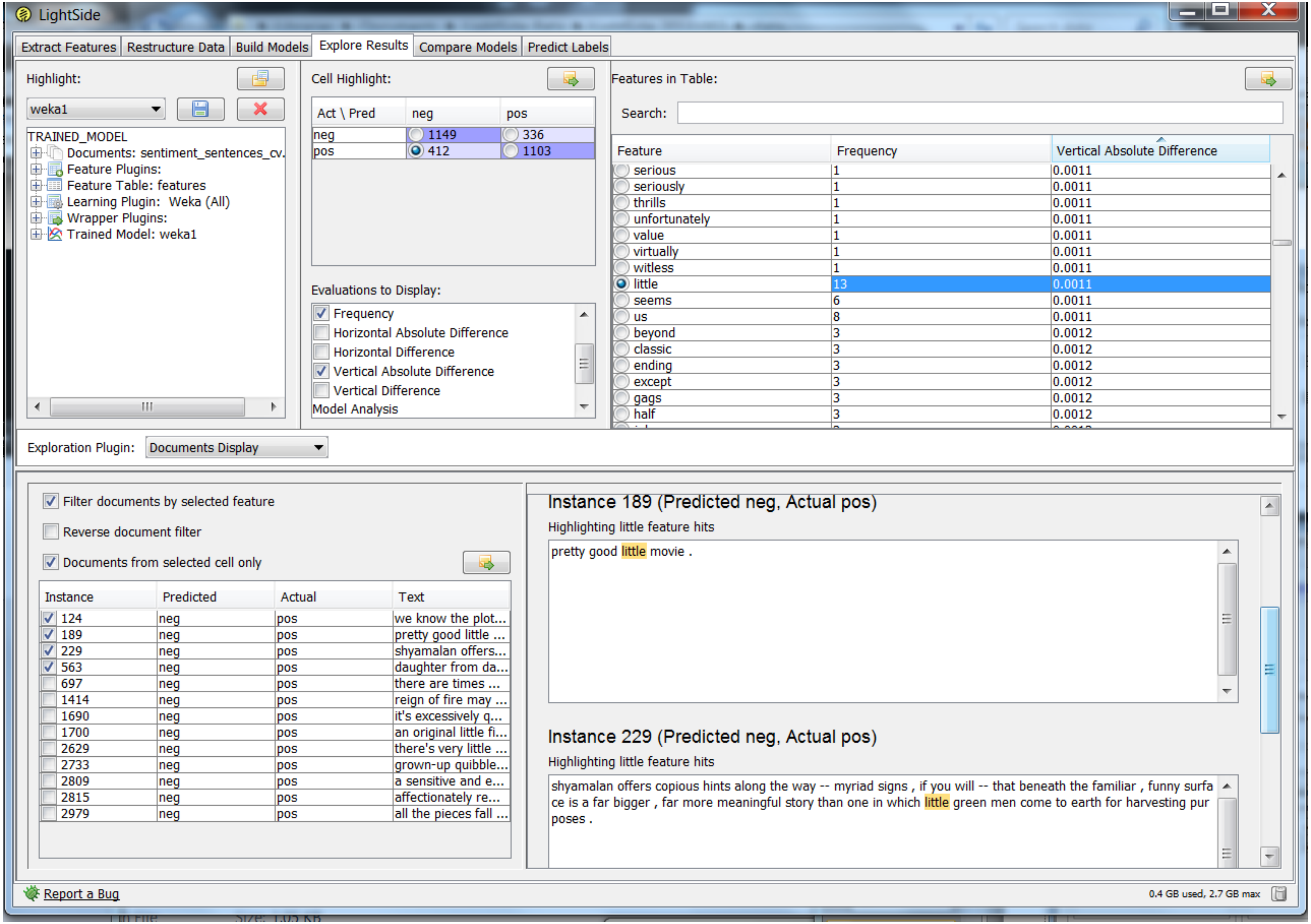

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Configure Weka (All)
Choose
SMO -C 1.0 -L 0.001 -P 1.0
Cross-Validation
Supplied Test Set
No Evaluation
Test Set (CSV):
sentiment_sentence...
DOCUMENT_LIST
Documents: sentiment_senten
Train
Name:
weka
Trained Models:
weka
TRAINED_MODEL
Documents: sentiment_sente
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Open
Look in:
data
Recent Items
Desktop
My Documents
Computer
essay
Gallup
MovieReviews
NewsgroupTopic
sentiment_documents
sentiment_sentences
sentiment_sentences_cv
sentiment_sentences_dev
File name:
sentiment_sentences_dev.csv
Files of type:
CSV (Excel)
Open
Cancel
Report a Bug
0.8 GB used, 2.7 GB max

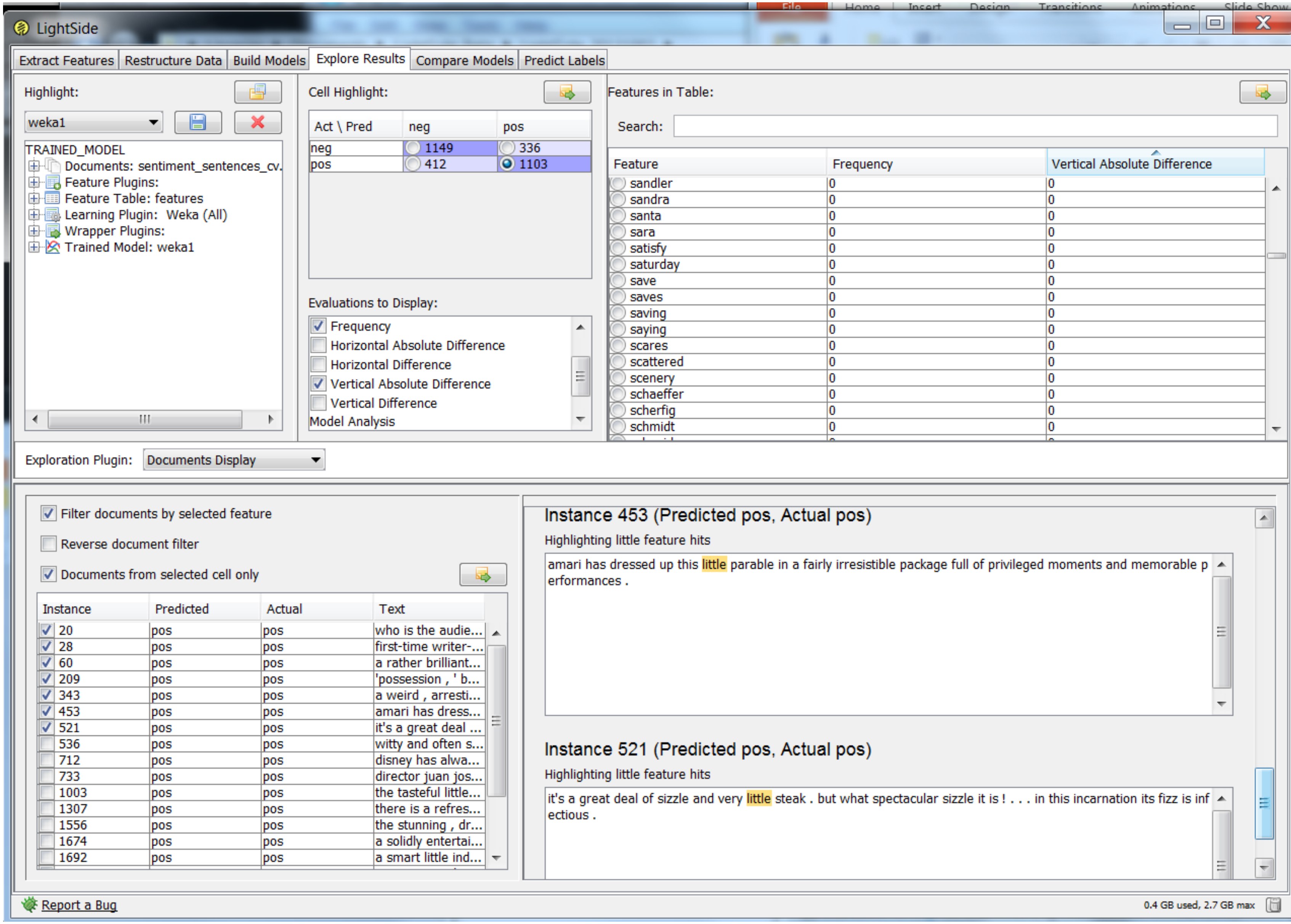

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Configure Weka (All)
Choose
SMO -C 1.0 -L 0.001 -P 1.0
Cross-Validation
Supplied Test Set
No Evaluation
Test Set (CSV):
sentiment_sentence...
DOCUMENT_LIST
Documents: sentiment_senten
Train
Name:
weka2
Feature Selection
#:
1000
Trained Models:
weka1
TRAINED_MODEL
Documents: sentiment_sentences
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Model Evaluation Metrics:
Metric
Value
Accuracy
0.7507
Kappa
0.5015
Model Confusion Matrix:
Act \ Pred
neg
pos
neg
1149
336
pos
412
1103
Report a Bug
0.7 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
weka1
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka1
Cell Highlight:
Act \ Pred
neg
pos
neg 1149 336
pos 412 1103
Evaluations to Display:
Feature Confusion Ranking
Average Value
Frequency
Horizontal Absolute Difference
Horizontal Difference
Vertical Absolute Difference
Features in Table:
Search:
Feature
'60s
'70s
'd
'em
'll
'm
're
's
've
--
-lsb-
-rsb-
1
10
100
101
Exploration Plugin:
Model Output (Text)
+ -0.5744 * (normalized) witless
+ 1 * (normalized) witty
+ 1.1405 * (normalized) wonderful
+ 1.0185 * (normalized) wonderfully
+ -0.516 * (normalized) wooden
+ -1 * (normalized) woody
+ 0.1701 * (normalized) work
+ 1.1733 * (normalized) works
+ 0.6076 * (normalized) world
+ -0.3756 * (normalized) worse
+ -1.3245 * (normalized) worst
+ 1.06 * (normalized) worth
+ -0.2368 * (normalized) would
+ -1.0217 * (normalized) writers
+ -0.9047 * (normalized) wrong
+ 1 * (normalized) wry
+ 1.1511 * (normalized) y
+ 0.6517 * (normalized) year
+ 0.4901 * (normalized) years
+ 0.4529 * (normalized) yet
- 0.2855
Number of kernel evaluations: 144356432 (53.759% cached)
Report a Bug
0.8 GB used, 2.7 GB max

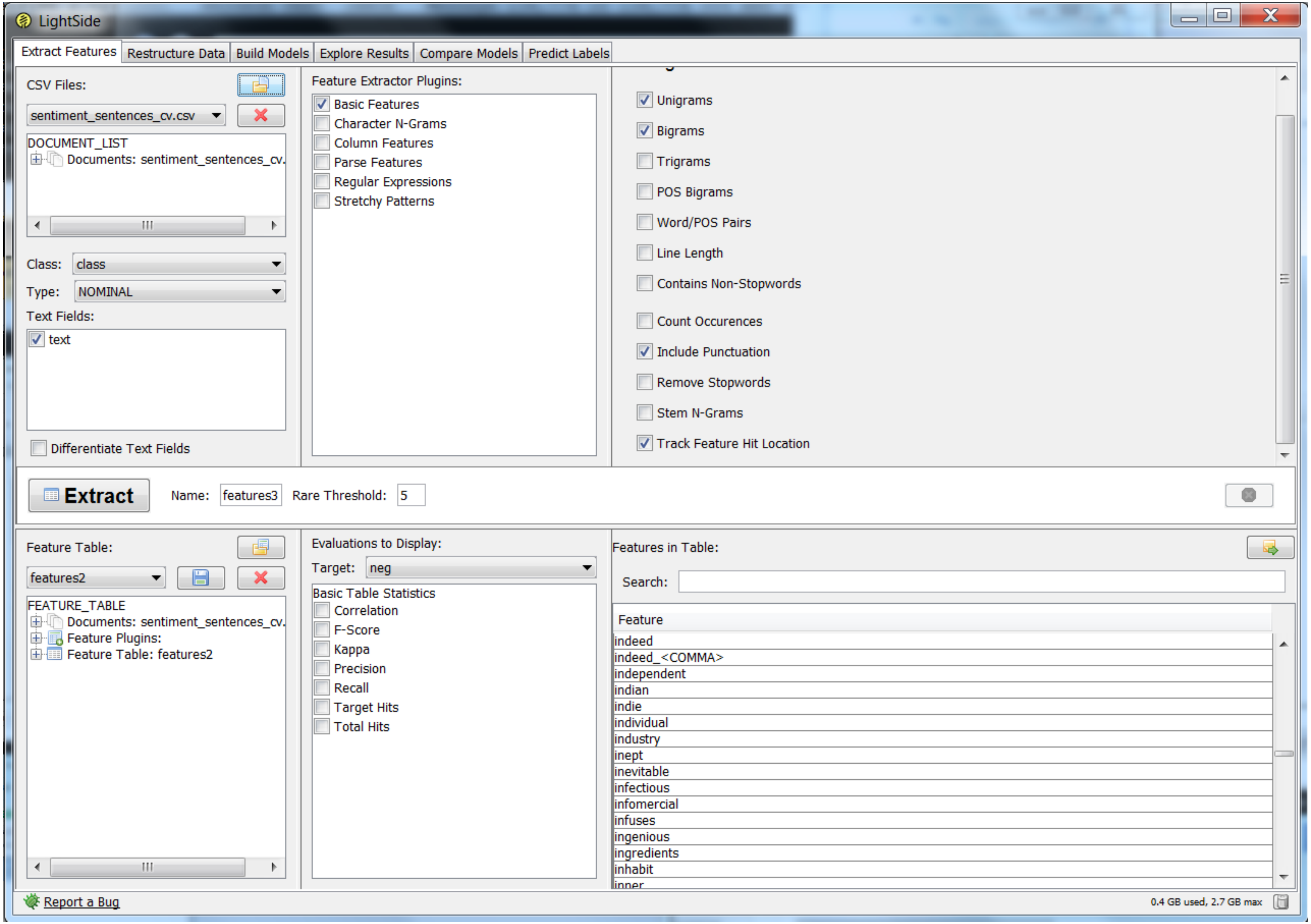

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
weka1
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka1
Cell Highlight:
Act \ Pred
neg
pos
neg 1149 336
pos 412 1103
Evaluations to Display:
Frequency
Horizontal Absolute Difference
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Features in Table:
Search:
Feature
Frequency
Vertical Absolute Difference
serious 1 0.0011
seriously 1 0.0011
thrills 1 0.0011
unfortunately 1 0.0011
value 1 0.0011
virtually 1 0.0011
witless 1 0.0011
little 13 0.0011
seems 6 0.0011
us 8 0.0011
beyond 3 0.0012
classic 3 0.0012
ending 3 0.0012
except 3 0.0012
gags 3 0.0012
half 3 0.0012
Exploration Plugin:
Model Output (Text)
+ 0.3548 * (normalized) lee
+ -0.2567 * (normalized) left
+ 0.4641 * (normalized) legacy
+ 0.6121 * (normalized) leigh
+ -0.4566 * (normalized) less
+ -0.5215 * (normalized) let
+ 0.4781 * (normalized) life
+ -1.1239 * (normalized) lifeless
+ -0.2275 * (normalized) like
+ 0.7136 * (normalized) lips
+ -0.5328 * (normalized) list
+ -1 * (normalized) listless
+ 0.5816 * (normalized) literary
+ 0.939 * (normalized) literate
+ -0.4418 * (normalized) little
+ 1.1512 * (normalized) lively
+ 0.0912 * (normalized) lives
+ -0.0151 * (normalized) long
+ 0.4274 * (normalized) longing
+ 0.2859 * (normalized) look
+ -0.5328 * (normalized) looking
+ -0.9967 * (normalized) loses
+ -0.9503 * (normalized) loud
+ -0.6698 * (normalized) lousy
+ 0.3711 * (normalized) love
Report a Bug
0.3 GB used, 2.7 GB max

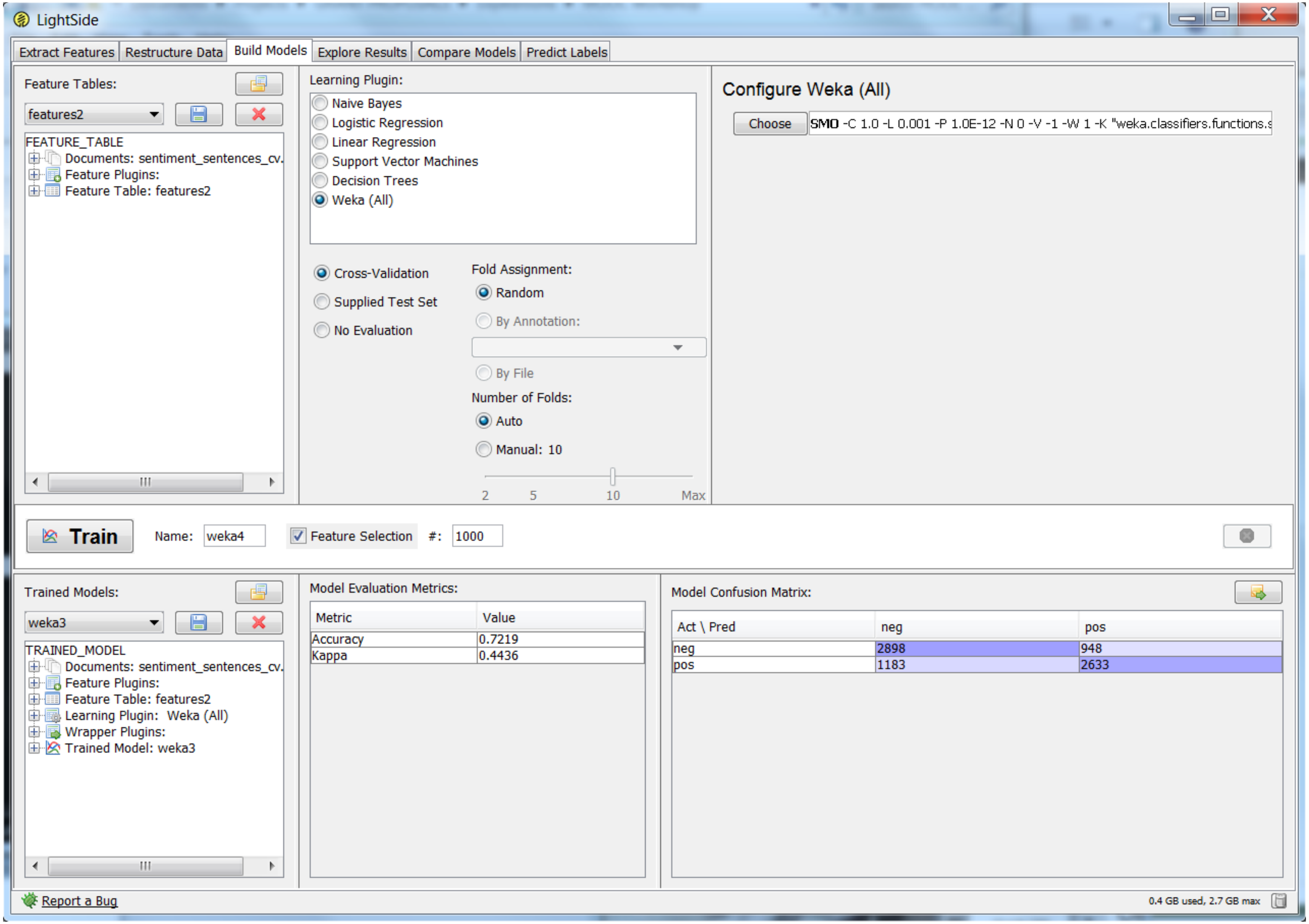

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
weka1
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka1
Cell Highlight:
Act \ Pred
neg
pos
neg 1149 336
pos 412 1103
Evaluations to Display:
Frequency
Horizontal Absolute Difference
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Features in Table:
Search:
Feature
Frequency
Vertical Absolute Difference
serious 1 0.0011
seriously 1 0.0011
thrills 1 0.0011
unfortunately 1 0.0011
value 1 0.0011
virtually 1 0.0011
witless 1 0.0011
little 13 0.0011
seems 6 0.0011
us 8 0.0011
beyond 3 0.0012
classic 3 0.0012
ending 3 0.0012
except 3 0.0012
gags 3 0.0012
half 3 0.0012
Exploration Plugin:
Highlighted Feature Details
Frequency
Model Confusion Matrix:
Act \ Pred
neg
pos
neg 35 12
pos 13 20
Vertical Absolute Difference
Model Confusion Matrix:
Act \ Pred
neg
pos
neg 0 0.018
pos 0.001 0
Report a Bug
0.3 GB used, 2.7 GB max

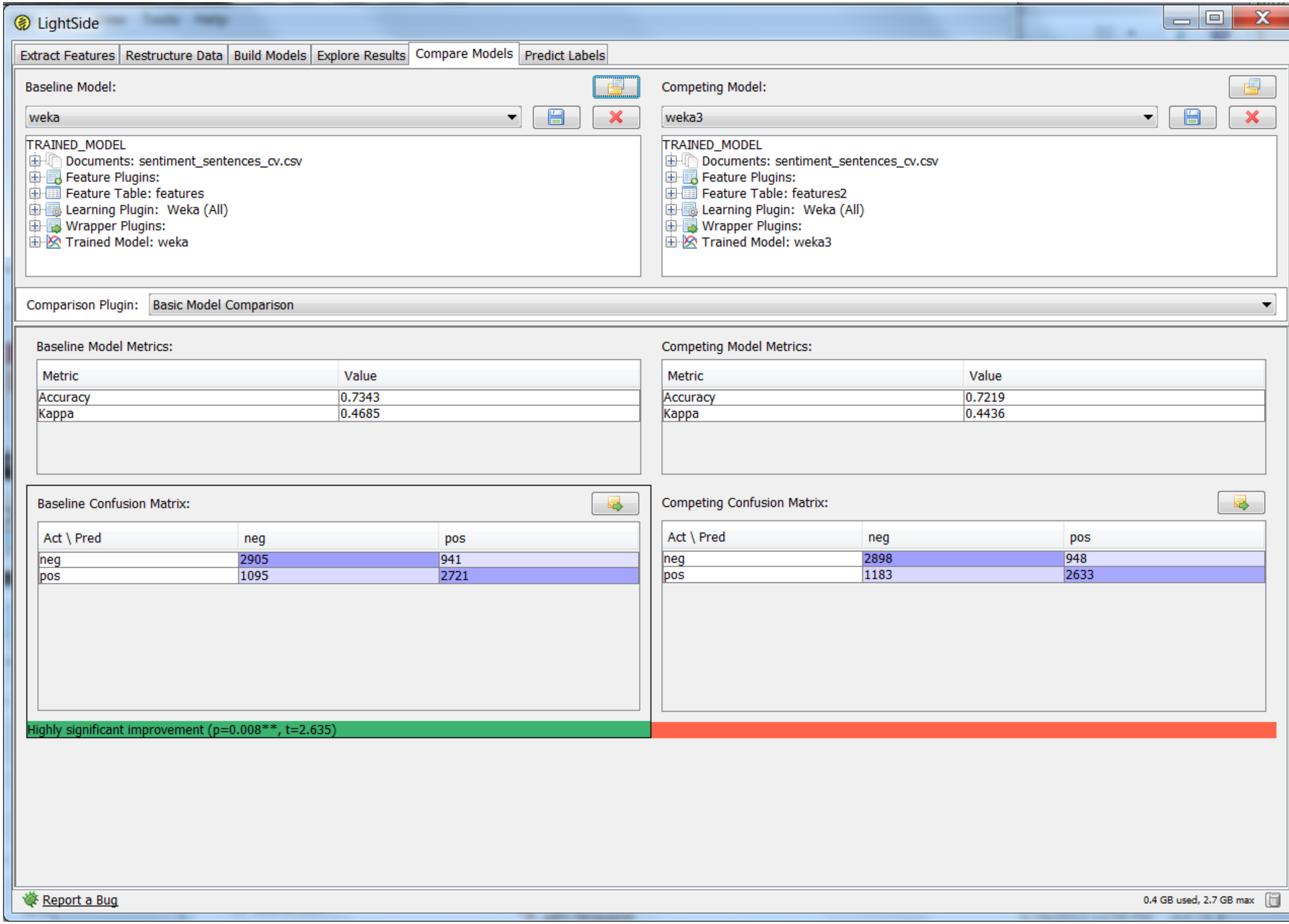

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
weka1
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka1
Cell Highlight:
Act \ Pred
neg
pos
neg 1149 336
pos 412 1103
Evaluations to Display:
Frequency
Horizontal Absolute Difference
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Features in Table:
Search:
Feature
Frequency
Vertical Absolute Difference
serious 1 0.0011
seriously 1 0.0011
thrills 1 0.0011
unfortunately 1 0.0011
value 1 0.0011
virtually 1 0.0011
witless 1 0.0011
little 13 0.0011
seems 6 0.0011
us 8 0.0011
beyond 3 0.0012
classic 3 0.0012
ending 3 0.0012
except 3 0.0012
gags 3 0.0012
half 3 0.0012
Exploration Plugin:
Documents Display
Filter documents by selected feature
Reverse document filter
Documents from selected cell only
Instance
Predicted
Actual
Text
124 neg pos we know the plot...
189 neg pos pretty good little ...
229 neg pos shyamalan offers...
563 neg pos daughter from da...
697 neg pos there are times ...
1414 neg pos reign of fire may ...
1690 neg pos it's excessively q...
1700 neg pos an original little fi...
2629 neg pos there's very little ...
2733 neg pos grown-up quibble...
2809 neg pos a sensitive and e...
2815 neg pos affectionately re...
2979 neg pos all the pieces fall ...
Instance 189 (Predicted neg, Actual pos)
Highlighting little feature hits
pretty good little movie .
Instance 229 (Predicted neg, Actual pos)
Highlighting little feature hits
shyamalan offers copious hints along the way -- myriad signs , if you will -- that beneath the familiar , funny surfa
ce is a far bigger , far more meaningful story than one in which little green men come to earth for harvesting pur
poses .
Report a Bug
0.4 GB used, 2.7 GB max

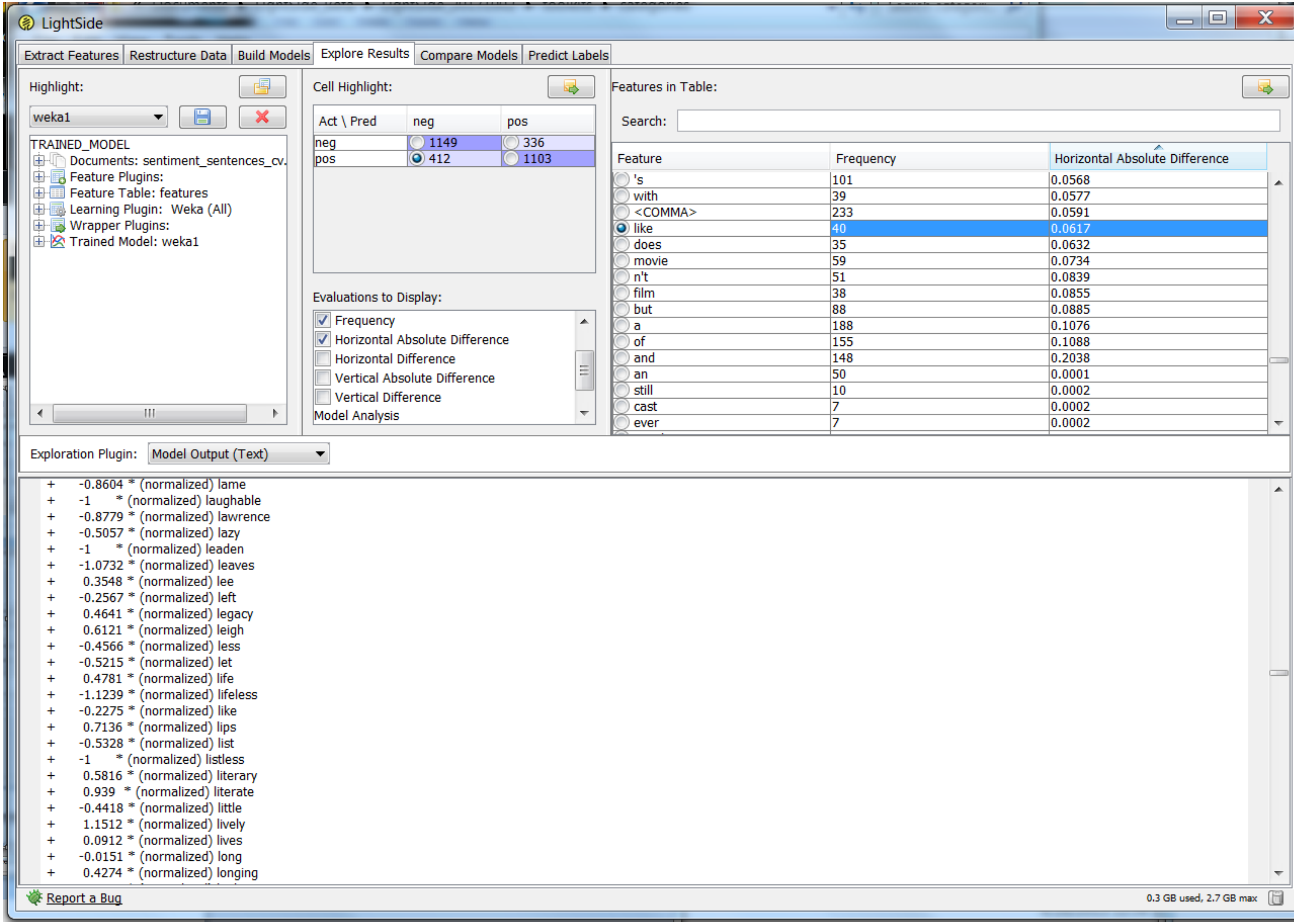

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
weka1
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka1
Cell Highlight:
Act \ Pred
neg
pos
neg 1149 336
pos 412 1103
Evaluations to Display:
Frequency
Horizontal Absolute Difference
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Features in Table:
Search:
Feature
Frequency
Vertical Absolute Difference
sandler 0 0
sandra 0 0
santa 0 0
sara 0 0
satisfy 0 0
saturday 0 0
save 0 0
saves 0 0
saving 0 0
saying 0 0
scares 0 0
scattered 0 0
scenery 0 0
schaeffer 0 0
scherfig 0 0
schmidt 0 0
Exploration Plugin:
Documents Display
Filter documents by selected feature
Reverse document filter
Documents from selected cell only
Instance
Predicted
Actual
Text
20 pos pos who is the audie...
28 pos pos first-time writer-...
60 pos pos a rather brilliant...
209 pos pos 'possession , ' b...
343 pos pos a weird , arresti...
453 pos pos amari has dress...
521 pos pos it's a great deal ...
536 pos pos witty and often s...
712 pos pos disney has alwa...
733 pos pos director juan jos...
1003 pos pos the tasteful little...
1307 pos pos there is a refres...
1556 pos pos the stunning , dr...
1674 pos pos a solidly entertai...
1692 pos pos a smart little ind...
Instance 453 (Predicted pos, Actual pos)
Highlighting little feature hits
amari has dressed up this little parable in a fairly irresistible package full of privileged moments and memorable p
erformances .
Instance 521 (Predicted pos, Actual pos)
Highlighting little feature hits
it's a great deal of sizzle and very little steak . but what spectacular sizzle it is ! . . . in this incarnation its fizz is inf
ectious .
Report a Bug
0.4 GB used, 2.7 GB max

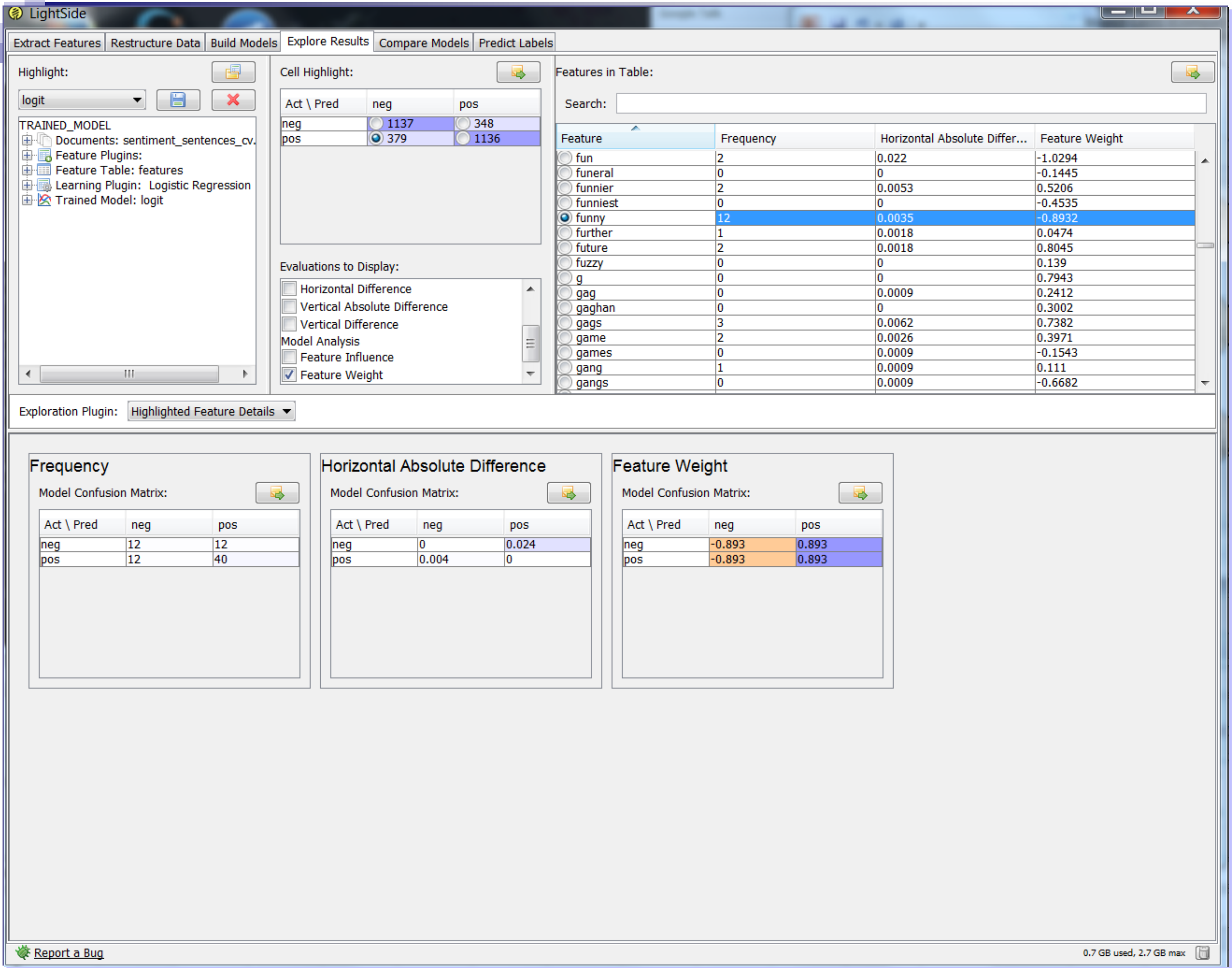
* Testing bigrams as an alternative….

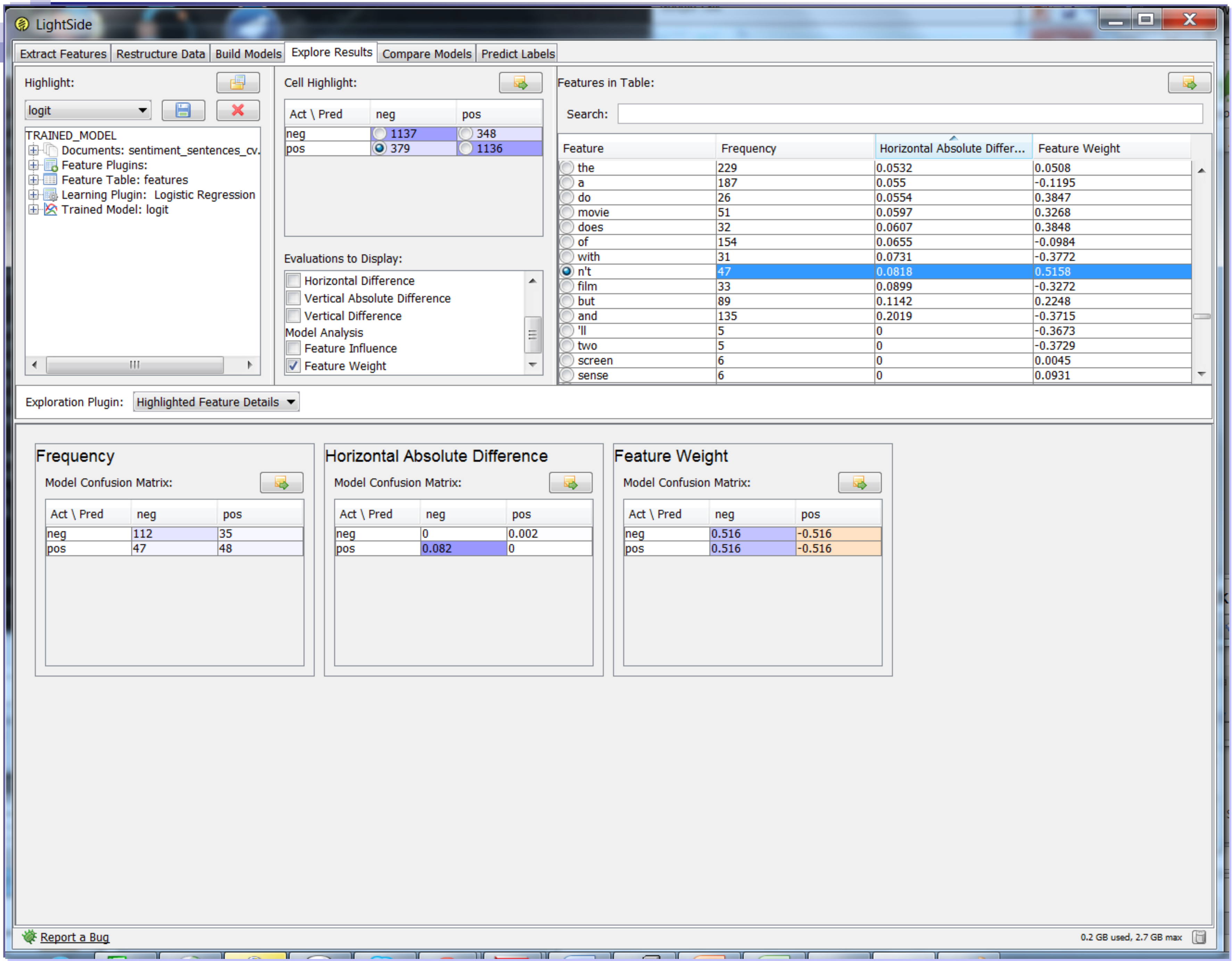


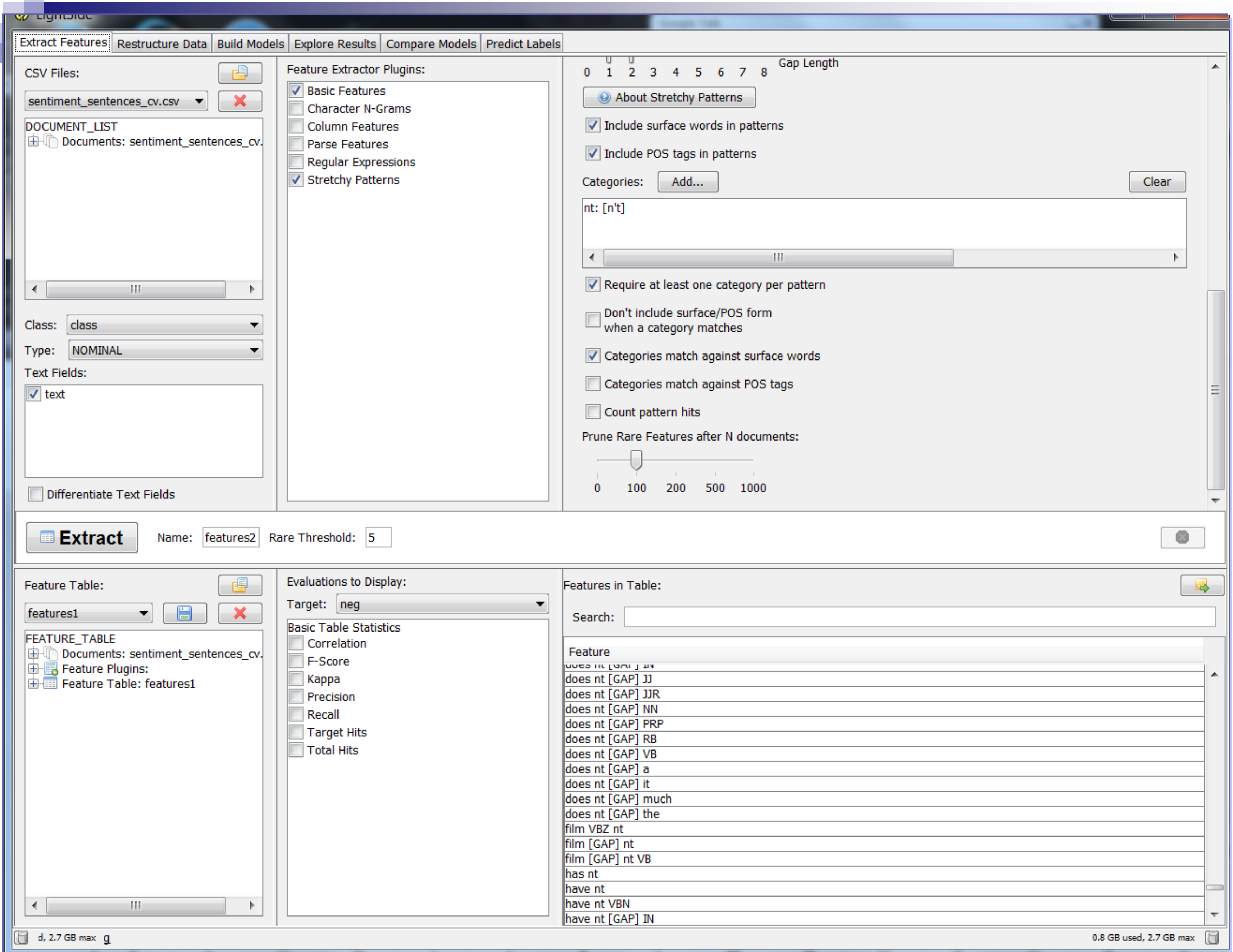

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features2
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features2
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Configure Weka (All)
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.s
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Train
Name:
weka4
Feature Selection
#:
1000
Trained Models:
weka3
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features2
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka3
Model Evaluation Metrics:
Metric
Value
Accuracy
0.7219
Kappa
0.4436
Model Confusion Matrix:
Act \ Pred
neg
pos
neg
2898
948
pos
1183
2633
Report a Bug
0.4 GB used, 2.7 GB max

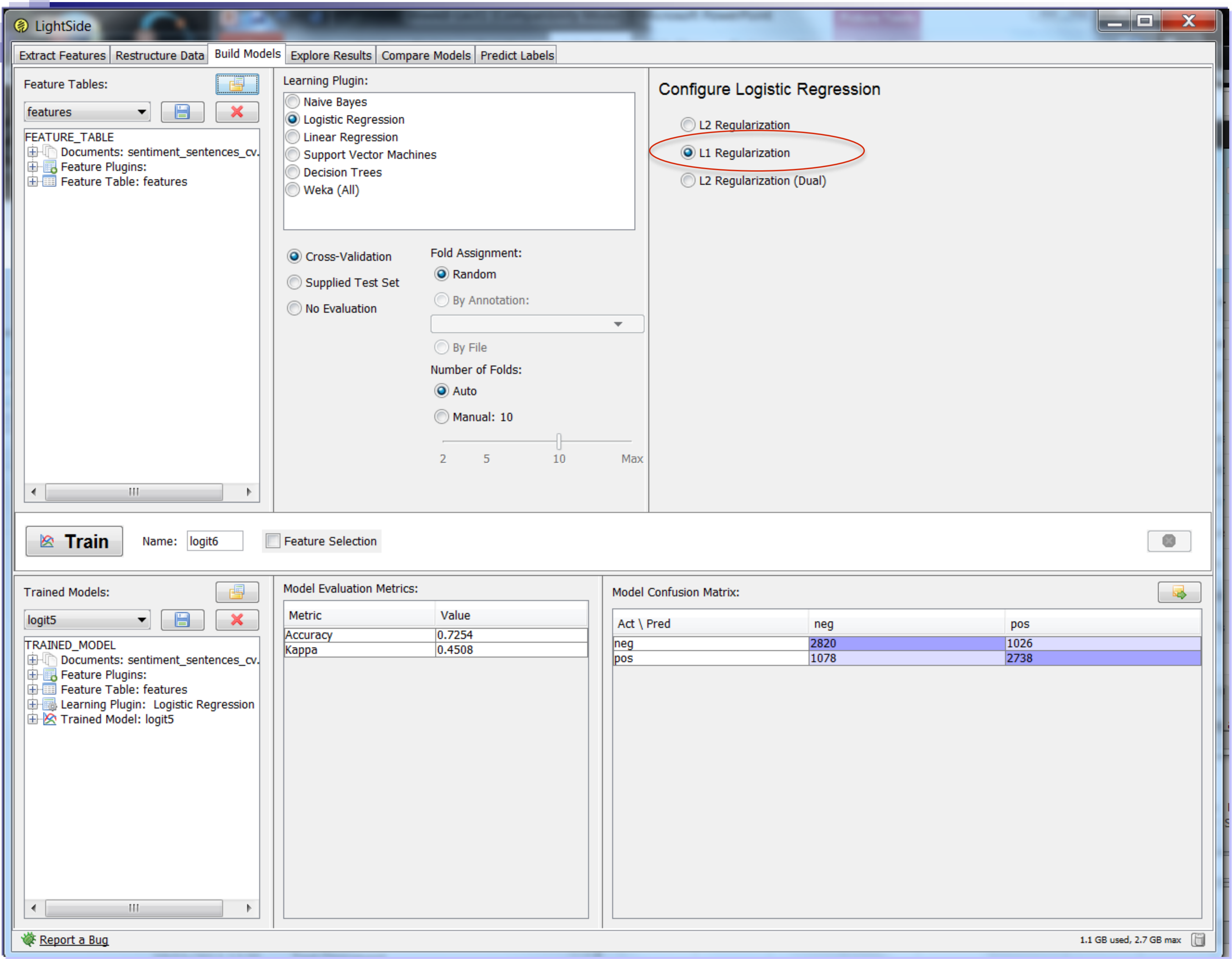

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Baseline Model:
weka
TRAINED_MODEL
Documents: sentiment_sentences_cv.csv
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka
Competing Model:
weka3
TRAINED_MODEL
Documents: sentiment_sentences_cv.csv
Feature Plugins:
Feature Table: features2
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka3
Comparison Plugin: Basic Model Comparison
Baseline Model Metrics:
Metric Value
Accuracy 0.7343
Kappa 0.4685
Competing Model Metrics:
Metric Value
Accuracy 0.7219
Kappa 0.4436
Baseline Confusion Matrix:
Act \ Pred neg pos
neg 2905 941
pos 1095 2721
Competing Confusion Matrix:
Act \ Pred neg pos
neg 2898 948
pos 1183 2633
Highly significant improvement (p=0.008**, t=2.635)
Report a Bug
0.4 GB used, 2.7 GB max

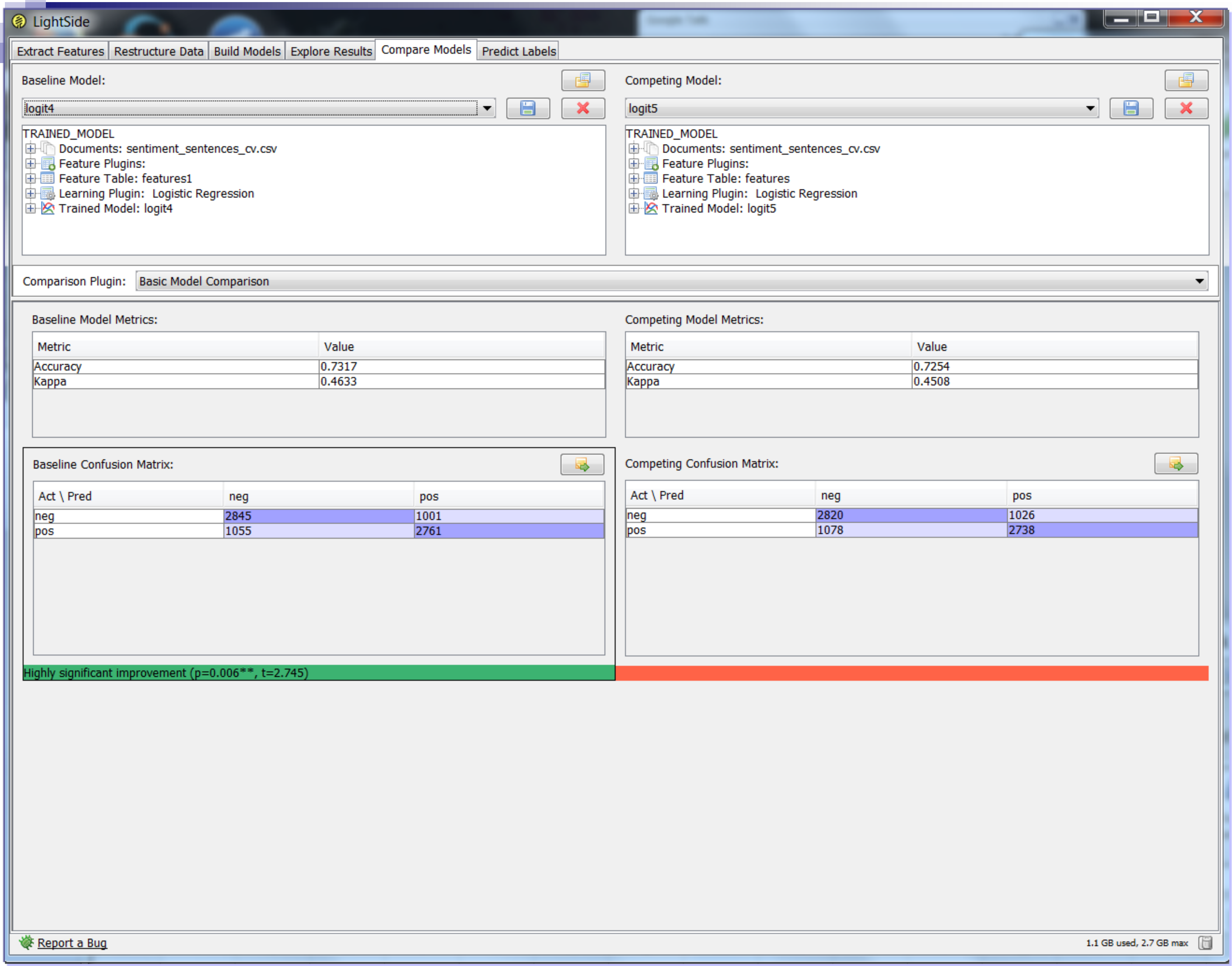

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
weka1
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Weka (All)
Wrapper Plugins:
Trained Model: weka1
Cell Highlight:
Act \ Pred
neg
pos
neg 1149 336
pos 412 1103
Evaluations to Display:
Frequency
Horizontal Absolute Difference
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Features in Table:
Search:
Feature
Frequency
Horizontal Absolute Difference
's 101 0.0568
with 39 0.0577
<COMMA> 233 0.0591
like 40 0.0617
does 35 0.0632
movie 59 0.0734
n't 51 0.0839
film 38 0.0855
but 88 0.0885
a 188 0.1076
of 155 0.1088
and 148 0.2038
an 50 0.0001
still 10 0.0002
cast 7 0.0002
ever 7 0.0002
Exploration Plugin:
Model Output (Text)
+ -0.8604 * (normalized) lame
+ -1 * (normalized) laughable
+ -0.8779 * (normalized) lawrence
+ -0.5057 * (normalized) lazy
+ -1 * (normalized) leaden
+ -1.0732 * (normalized) leaves
+ 0.3548 * (normalized) lee
+ -0.2567 * (normalized) left
+ 0.4641 * (normalized) legacy
+ 0.6121 * (normalized) leigh
+ -0.4566 * (normalized) less
+ -0.5215 * (normalized) let
+ 0.4781 * (normalized) life
+ -1.1239 * (normalized) lifeless
+ -0.2275 * (normalized) like
+ 0.7136 * (normalized) lips
+ -0.5328 * (normalized) list
+ -1 * (normalized) listless
+ 0.5816 * (normalized) literary
+ 0.939 * (normalized) literate
+ -0.4418 * (normalized) little
+ 1.1512 * (normalized) lively
+ 0.0912 * (normalized) lives
+ -0.0151 * (normalized) long
+ 0.4274 * (normalized) longing
Report a Bug
0.3 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
logit
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit
Cell Highlight:
Act \ Pred | neg | pos
neg | 1137 | 348
pos | 379 | 1136
Evaluations to Display:
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Feature Influence
Feature Weight
Features in Table:
Search:
Feature | Frequency | Horizontal Absolute Differ... | Feature Weight
fun | 2 | 0.022 | -1.0294
funeral | 0 | 0 | -0.1445
funnier | 2 | 0.0053 | 0.5206
funniest | 0 | 0 | -0.4535
funny | 12 | 0.0035 | -0.8932
further | 1 | 0.0018 | 0.0474
future | 2 | 0.0018 | 0.8045
fuzzy | 0 | 0 | 0.139
g | 0 | 0 | 0.7943
gag | 0 | 0.0009 | 0.2412
gaghan | 0 | 0 | 0.3002
gags | 3 | 0.0062 | 0.7382
game | 2 | 0.0026 | 0.3971
games | 0 | 0.0009 | -0.1543
gang | 1 | 0.0009 | 0.111
gangs | 0 | 0.0009 | -0.6682
Exploration Plugin:
Highlighted Feature Details
Frequency
Model Confusion Matrix:
Act \ Pred | neg | pos
neg | 12 | 12
pos | 12 | 40
Horizontal Absolute Difference
Model Confusion Matrix:
Act \ Pred | neg | pos
neg | 0 | 0.024
pos | 0.004 | 0
Feature Weight
Model Confusion Matrix:
Act \ Pred | neg | pos
neg | -0.893 | 0.893
pos | -0.893 | 0.893
Report a Bug
0.7 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Highlight:
logit
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit
Cell Highlight:
Act \ Pred | neg | pos
neg | 1137 | 348
pos | 379 | 1136
Evaluations to Display:
Horizontal Difference
Vertical Absolute Difference
Vertical Difference
Model Analysis
Feature Influence
Feature Weight
Features in Table:
Search:
Feature | Frequency | Horizontal Absolute Differ... | Feature Weight
the | 229 | 0.0532 | 0.0508
a | 187 | 0.055 | -0.1195
do | 26 | 0.0554 | 0.3847
movie | 51 | 0.0597 | 0.3268
does | 32 | 0.0607 | 0.3848
of | 154 | 0.0655 | -0.0984
with | 31 | 0.0731 | -0.3772
n't | 47 | 0.0818 | 0.5158
film | 33 | 0.0899 | -0.3272
but | 89 | 0.1142 | 0.2248
and | 135 | 0.2019 | -0.3715
'll | 5 | 0 | -0.3673
two | 5 | 0 | -0.3729
screen | 6 | 0 | 0.0045
sense | 6 | 0 | 0.0931
Exploration Plugin:
Highlighted Feature Details
Frequency
Model Confusion Matrix:
Act \ Pred | neg | pos
neg | 112 | 35
pos | 47 | 48
Horizontal Absolute Difference
Model Confusion Matrix:
Act \ Pred | neg | pos
neg | 0 | 0.002
pos | 0.082 | 0
Feature Weight
Model Confusion Matrix:
Act \ Pred | neg | pos
neg | 0.516 | -0.516
pos | 0.516 | -0.516
Report a Bug
0.2 GB used, 2.7 GB max


Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences_cv.csv
DOCUMENT_LIST
Documents: sentiment_sentences_cv.
Class:
class
Type:
NOMINAL
Text Fields:
text
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
0 1 2 3 4 5 6 7 8
Gap Length
About Stretchy Patterns
Include surface words in patterns
Include POS tags in patterns
Categories:
Add...
Clear
nt: [n't]
Require at least one category per pattern
Don't include surface/POS form when a category matches
Categories match against surface words
Categories match against POS tags
Count pattern hits
Prune Rare Features after N documents:
0 100 200 500 1000
Extract
Name:
features2
Rare Threshold:
5
Feature Table:
features1
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features1
Evaluations to Display:
Target:
neg
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Feature
does nt [GAP] JJ
does nt [GAP] JJR
does nt [GAP] NN
does nt [GAP] PRP
does nt [GAP] RB
does nt [GAP] VB
does nt [GAP] a
does nt [GAP] it
does nt [GAP] much
does nt [GAP] the
film VBZ nt
film [GAP] nt
film [GAP] nt VB
has nt
have nt
have nt VBN
have nt [GAP] IN
0.8 GB used, 2.7 GB max

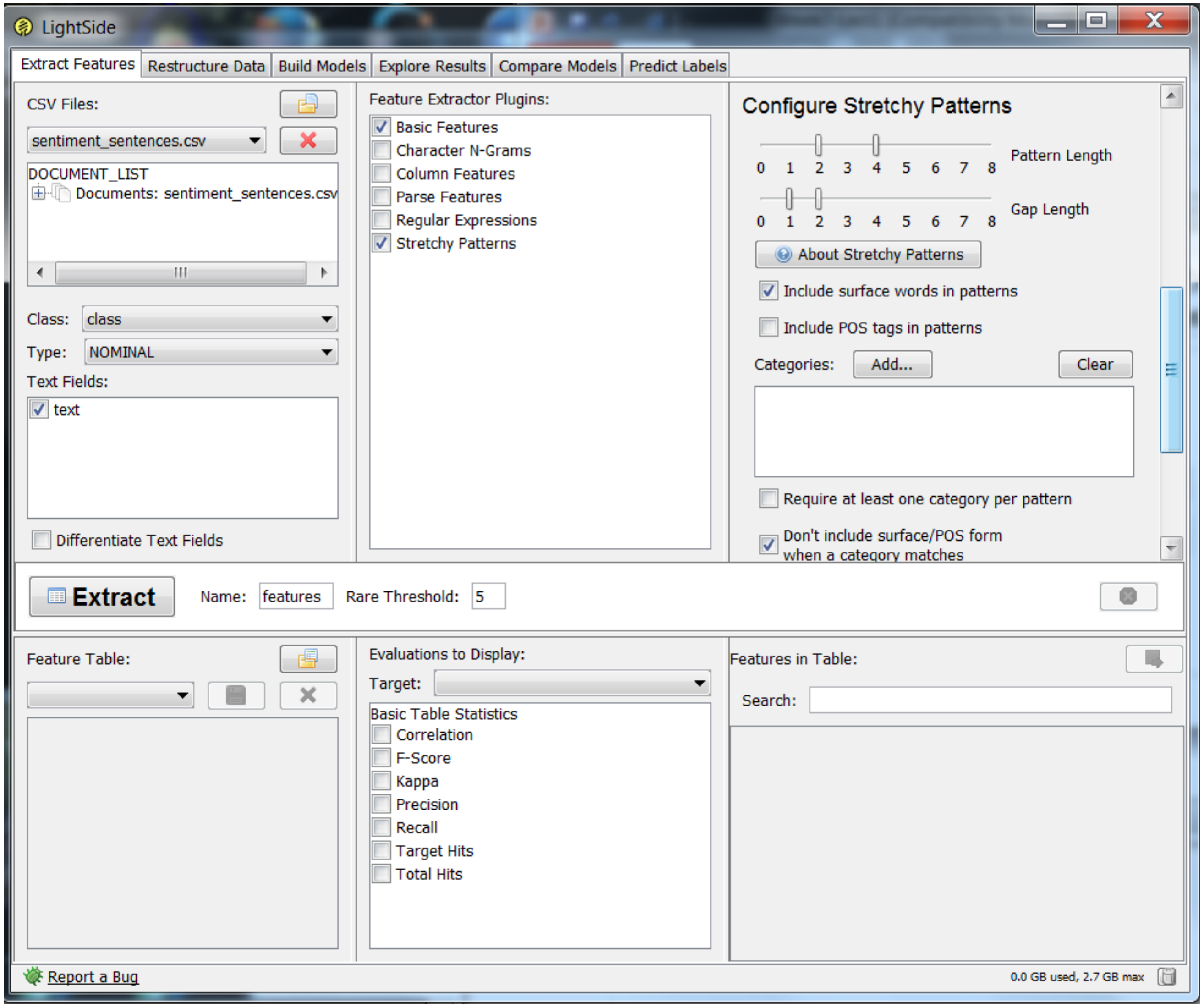

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Configure Logistic Regression
L2 Regularization
L1 Regularization
L2 Regularization (Dual)
Train
Name:
logit6
Feature Selection
Trained Models:
logit5
TRAINED_MODEL
Documents: sentiment_sentences_cv.
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit5
Model Evaluation Metrics:
Metric
Value
Accuracy
0.7254
Kappa
0.4508
Model Confusion Matrix:
Act \ Pred
neg
pos
neg
2820
1026
pos
1078
2738
Report a Bug
1.1 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Baseline Model:
logit4
TRAINED_MODEL
Documents: sentiment_sentences_cv.csv
Feature Plugins:
Feature Table: features1
Learning Plugin: Logistic Regression
Trained Model: logit4
Competing Model:
logit5
TRAINED_MODEL
Documents: sentiment_sentences_cv.csv
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit5
Comparison Plugin:
Basic Model Comparison
Baseline Model Metrics:
Metric
Value
Accuracy
0.7317
Kappa
0.4633
Competing Model Metrics:
Metric
Value
Accuracy
0.7254
Kappa
0.4508
Baseline Confusion Matrix:
Act \ Pred
neg
pos
neg
2845
1001
pos
1055
2761
Competing Confusion Matrix:
Act \ Pred
neg
pos
neg
2820
1026
pos
1078
2738
Highly significant improvement (p=0.006**, t=2.745)
Report a Bug
1.1 GB used, 2.7 GB max

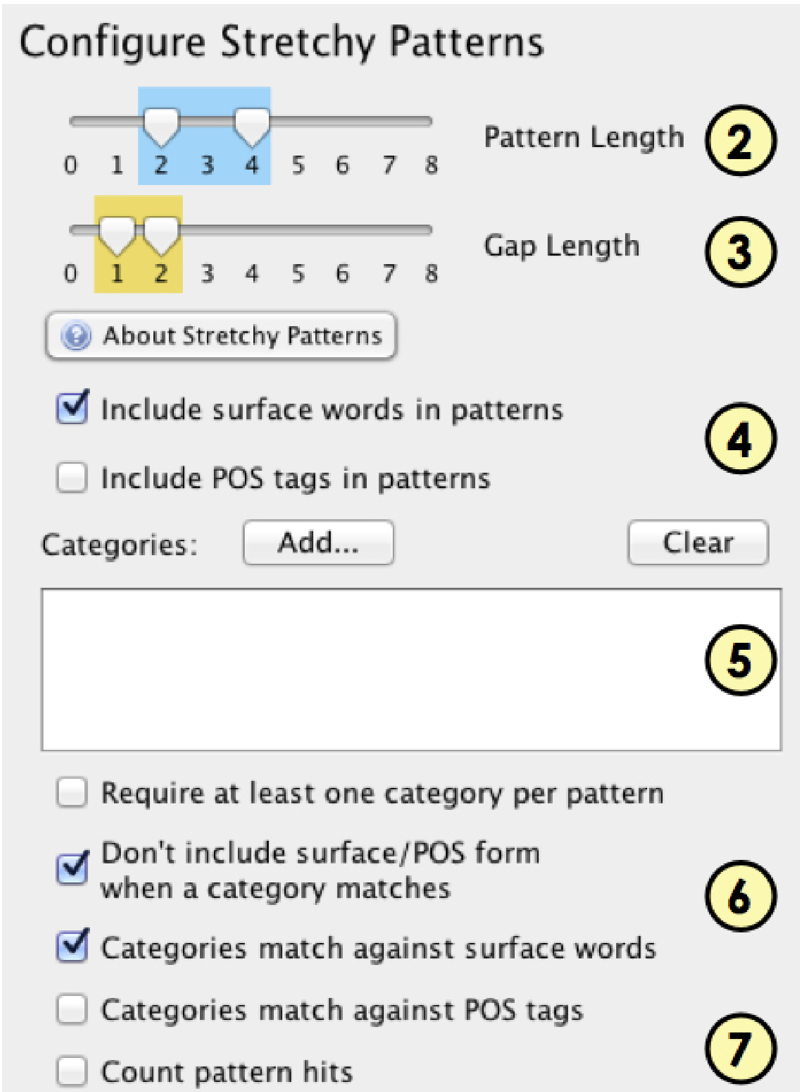
Special Text Features

Stretchy Patterns in LightSIDE

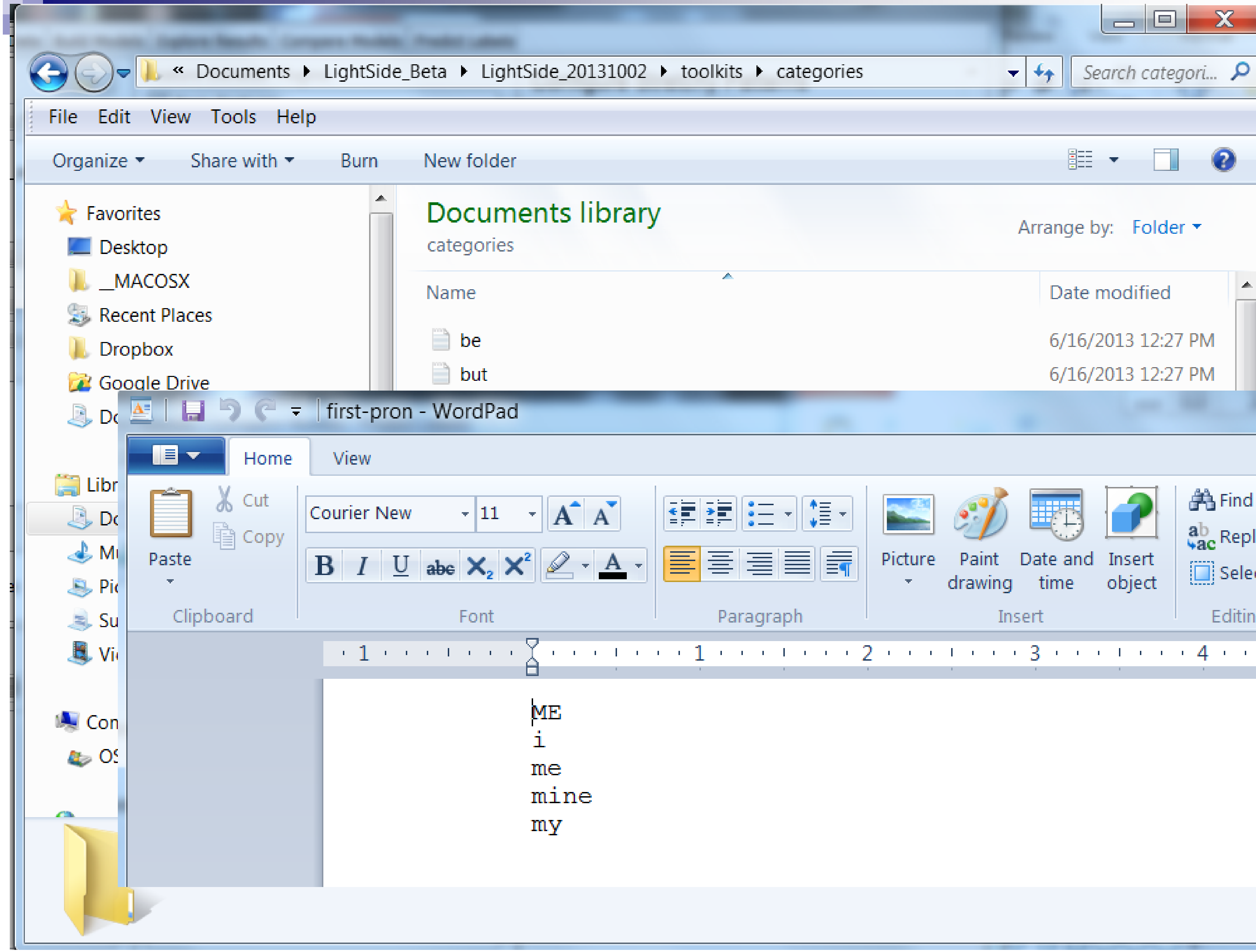
Looking at sentiment_sentences.csv

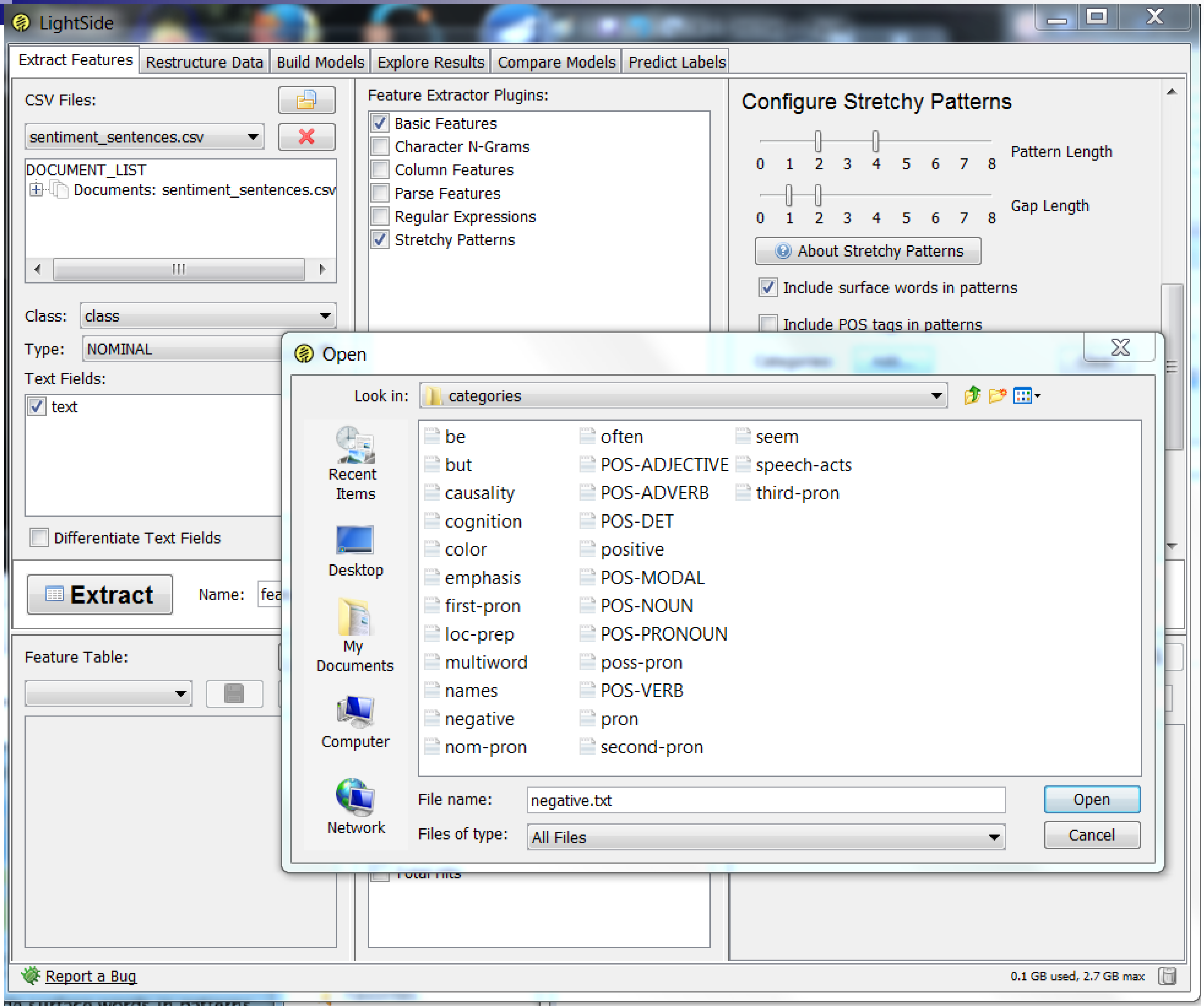


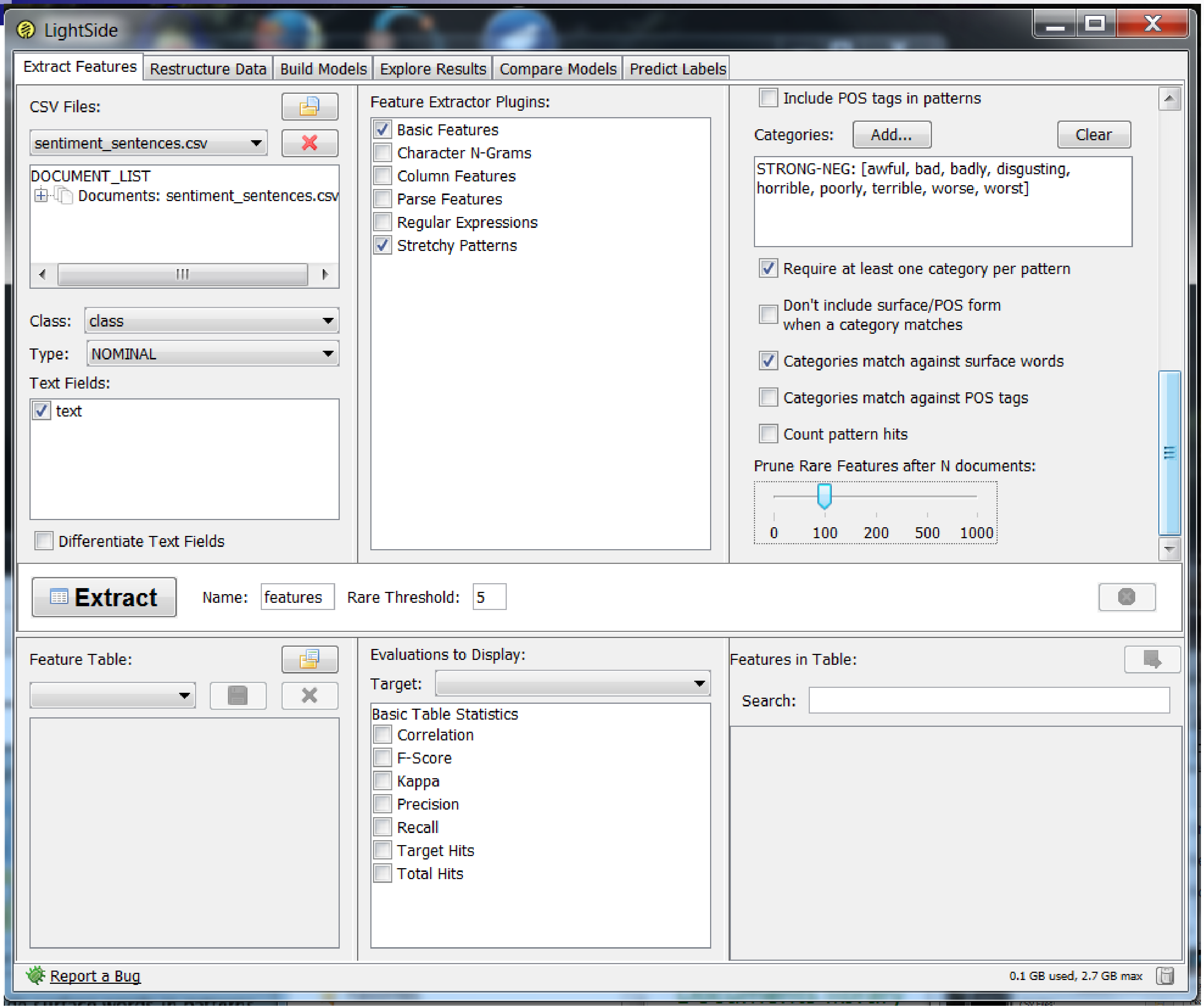
Configuring Stretchy Patterns



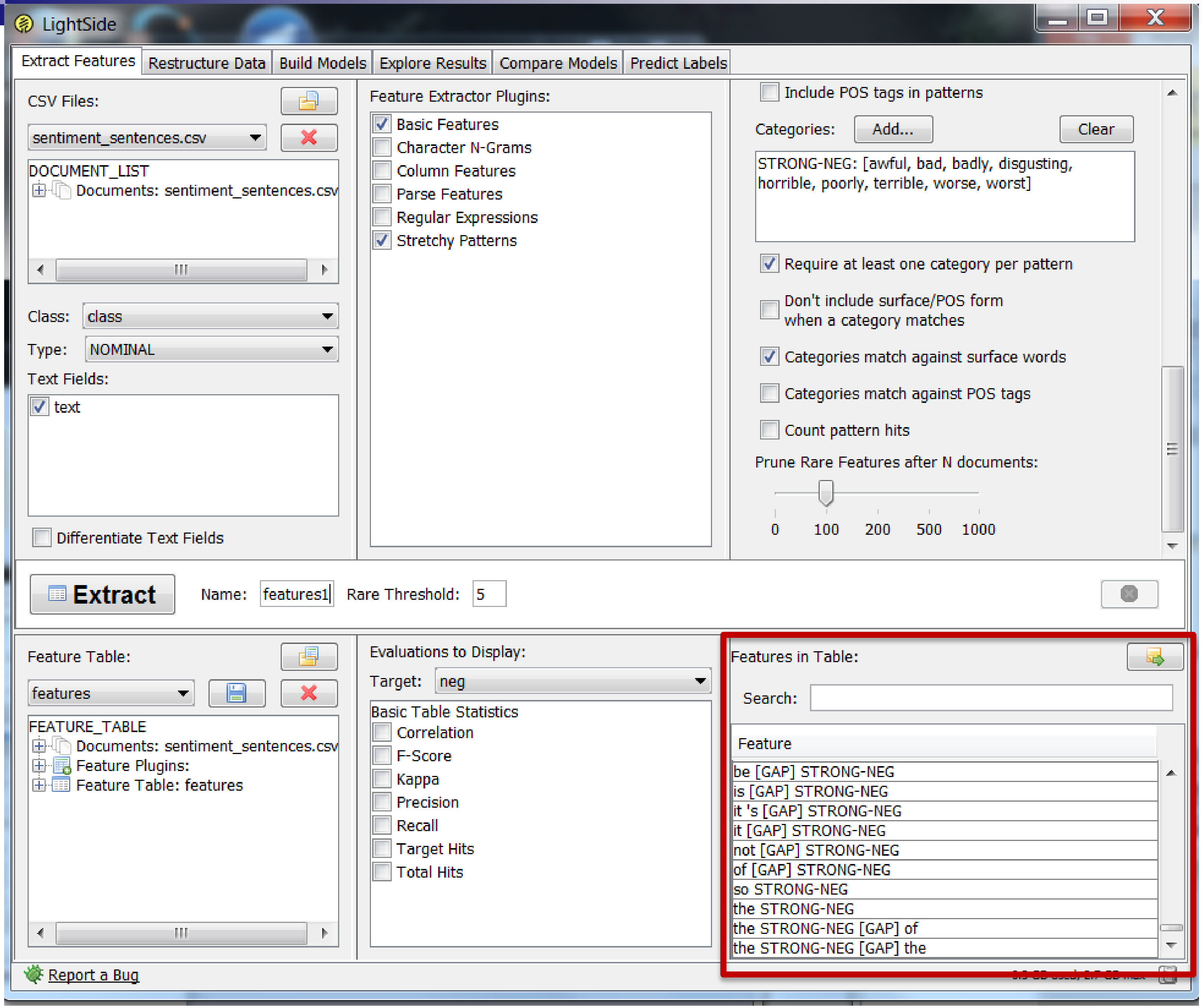
- Longer patterns and longer gaps lead to larger numbers of features
- Categories are useful both for abstraction and for anchoring the patterns


« Documents ▸ LightSide_Beta ▸ LightSide_20131002 ▸ toolkits ▸ categories
Search categori...
File Edit View Tools Help
Organize Share with Burn New folder
Favorites
Desktop
_MACOSX
Recent Places
Dropbox
Google Drive
Documents library
categories
Arrange by: Folder
Name
Date modified
be 6/16/2013 12:27 PM
but 6/16/2013 12:27 PM
first-pron - WordPad
Home View
Paste
Cut
Copy
Clipboard
Courier New 11
Font
Paragraph
Picture
Paint drawing
Date and time
Insert object
Insert
Find
Replace
Select all
Editing
ME
i
me
mine
my

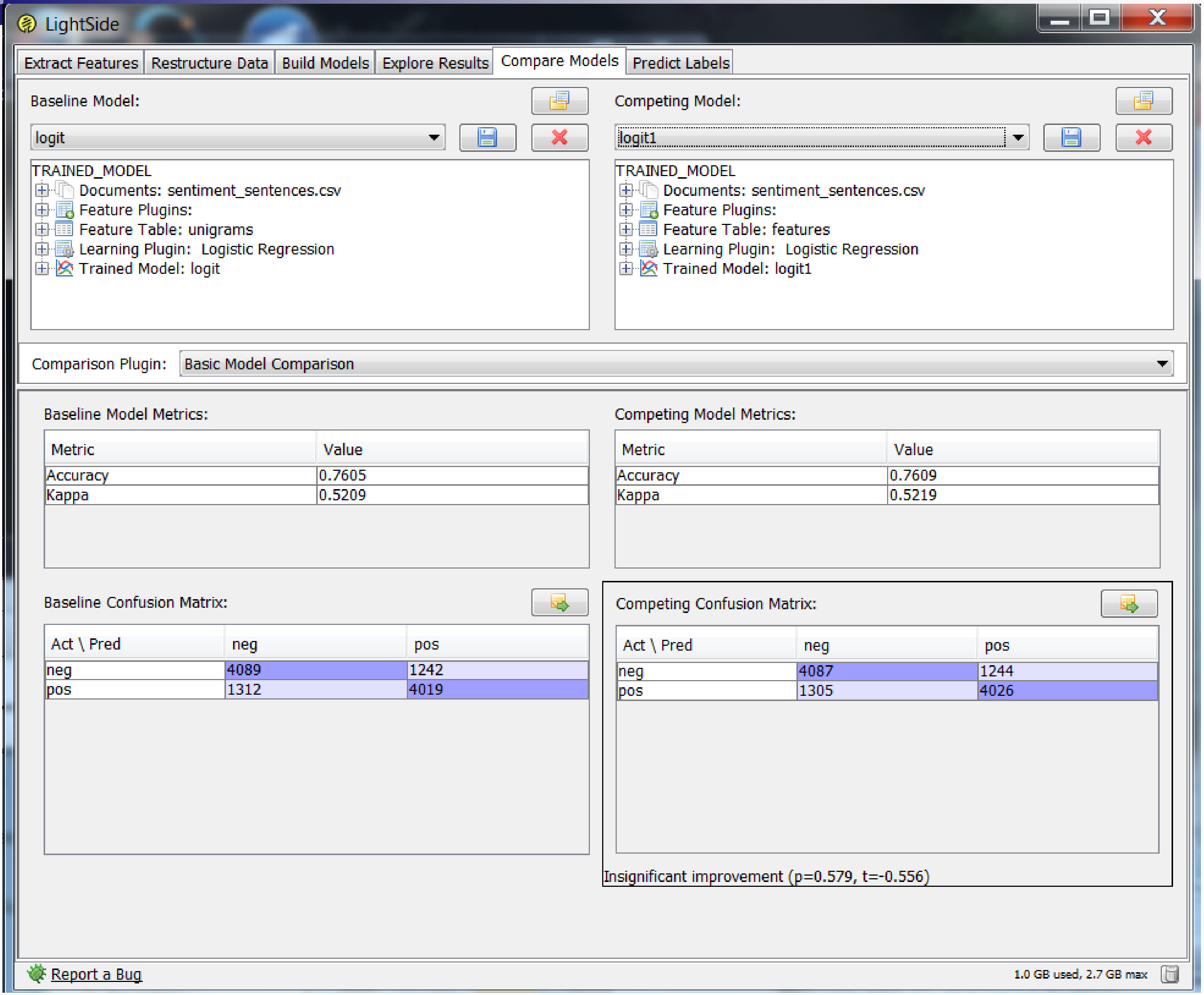

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences.csv
DOCUMENT_LIST
Documents: sentiment_sentences.csv
Class:
class
Type:
NOMINAL
Text Fields:
text
Differentiate Text Fields
Extract
Name:
Feature Table:
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Stretchy Patterns
Pattern Length
Gap Length
0 1 2 3 4 5 6 7 8
About Stretchy Patterns
Include surface words in patterns
Include POS tags in patterns
Open
Look in:
categories
Recent Items
Desktop
My Documents
Computer
Network
be
but
causality
cognition
color
emphasis
first-pron
loc-prep
multiword
names
negative
nom-pron
often
POS-ADJECTIVE
POS-ADVERB
POS-DET
positive
POS-MODAL
POS-NOUN
POS-PRONOUN
poss-pron
POS-VERB
pron
second-pron
seem
speech-acts
third-pron
File name:
negative.txt
Files of type:
All Files
Open
Cancel
Report a Bug
0.1 GB used, 2.7 GB max



LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences.csv
DOCUMENT_LIST
Documents: sentiment_sentences.csv
Class:
class
Type:
NOMINAL
Text Fields:
text
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Include POS tags in patterns
Categories:
Add...
Clear
STRONG-NEG: [awful, bad, badly, disgusting, horrible, poorly, terrible, worse, worst]
Require at least one category per pattern
Don't include surface/POS form when a category matches
Categories match against surface words
Categories match against POS tags
Count pattern hits
Prune Rare Features after N documents:
0
100
200
500
1000
Extract
Name:
features
Rare Threshold:
5
Feature Table:
Evaluations to Display:
Target:
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Report a Bug
0.1 GB used, 2.7 GB max

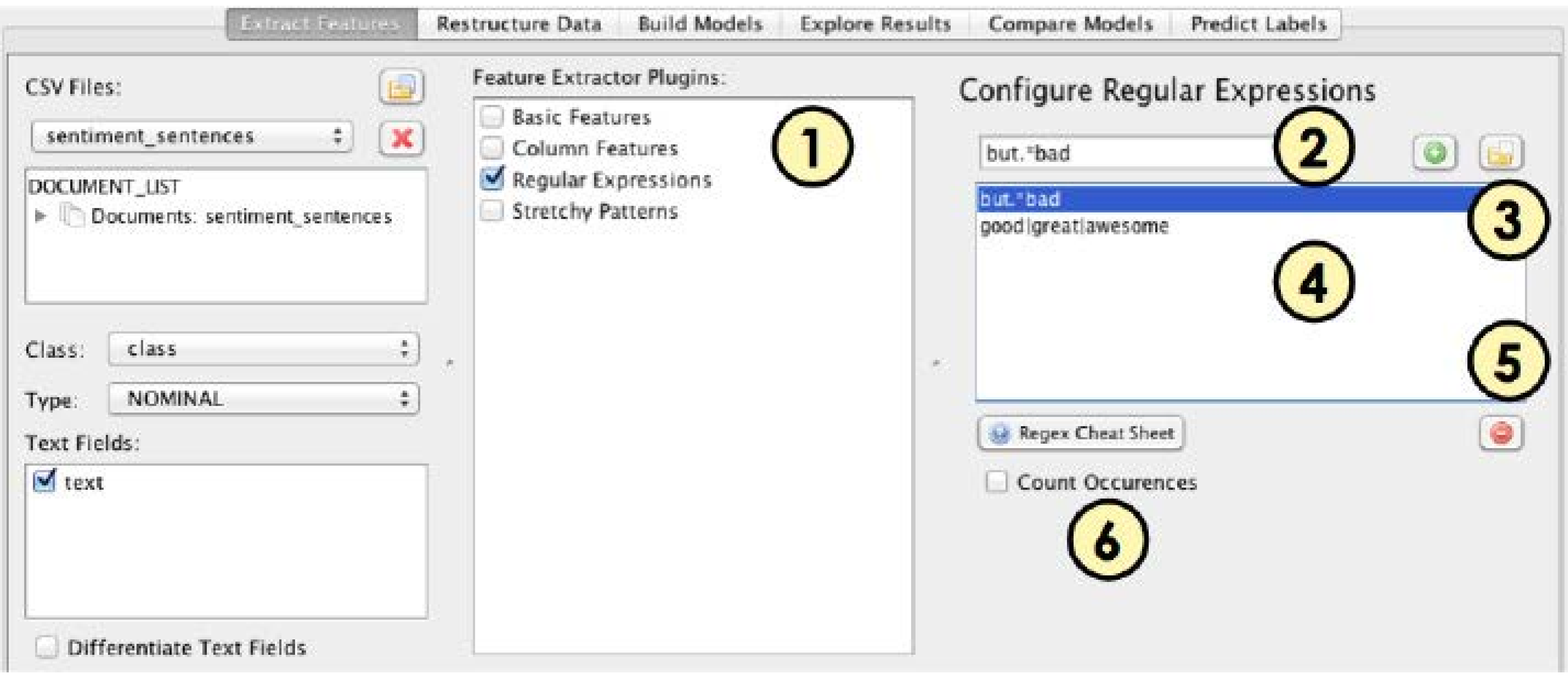


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
sentiment_sentences.csv
DOCUMENT_LIST
Documents: sentiment_sentences.csv
Class:
class
Type:
NOMINAL
Text Fields:
text
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Include POS tags in patterns
Categories:
Add...
Clear
STRONG-NEG: [awful, bad, badly, disgusting, horrible, poorly, terrible, worse, worst]
Require at least one category per pattern
Don't include surface/POS form when a category matches
Categories match against surface words
Categories match against POS tags
Count pattern hits
Prune Rare Features after N documents:
0
100
200
500
1000
Extract
Name:
features1
Rare Threshold:
5
Feature Table:
features
FEATURE_TABLE
Documents: sentiment_sentences.csv
Feature Plugins:
Feature Table: features
Evaluations to Display:
Target:
neg
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Feature
be [GAP] STRONG-NEG
is [GAP] STRONG-NEG
it 's [GAP] STRONG-NEG
it [GAP] STRONG-NEG
not [GAP] STRONG-NEG
of [GAP] STRONG-NEG
so STRONG-NEG
the STRONG-NEG
the STRONG-NEG [GAP] of
the STRONG-NEG [GAP] the
Report a Bug


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Baseline Model:
logit
TRAINED_MODEL
Documents: sentiment_sentences.csv
Feature Plugins:
Feature Table: unigrams
Learning Plugin: Logistic Regression
Trained Model: logit
Competing Model:
logit1
TRAINED_MODEL
Documents: sentiment_sentences.csv
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit1
Comparison Plugin:
Basic Model Comparison
Baseline Model Metrics:
Metric
Value
Accuracy
0.7605
Kappa
0.5209
Competing Model Metrics:
Metric
Value
Accuracy
0.7609
Kappa
0.5219
Baseline Confusion Matrix:
Act \ Pred
neg
pos
neg
4089
1242
pos
1312
4019
Competing Confusion Matrix:
Act \ Pred
neg
pos
neg
4087
1244
pos
1305
4026
Insignificant improvement (p=0.579, t=-0.556)
Report a Bug
1.0 GB used, 2.7 GB max

Regular Expressions

- ***** allows the previous part of the regex to re-peat, but it is not necessary.
- **+** is the same, but requires the previous part to match at least once.
- **?** allows the previous part to happen either once or not at all, but does not match further.
- **.** is a wildcard, matching any one character.
- Certain character classes are predefined, like \w (any character A-Z), \d (any digit 0-9), and \s (any type of space character).



American Street Gangs

Predict gang affiliation from posts

- **Crips, Bloods, Hoovers**
  - crips started in South Central LA
  - Pirus, Bloods, Hoovers from crips
- Chicago based
  - People Nation
    - **vice lords, latin kings, stones**
  - Folk nation
    - **gangster disciples**
- **Trinitarios**
  - hispanic gang based in NYC

Graffiti Based Style Features

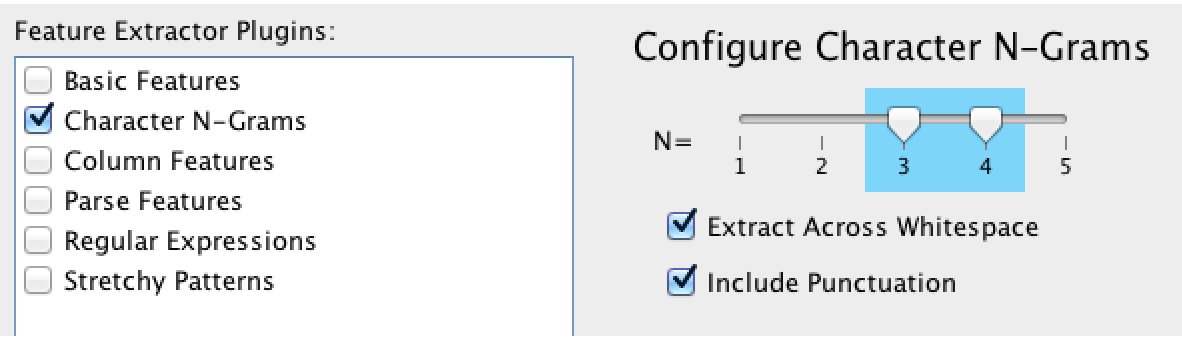
Graffiti

- Social messages
- Stylistic writing
- crossing out other gangs

On the board

c	ck	ckrab, ckome
ck	cc	fucc, blocc
p	pk	pkut, ...
h	hk	whky, hkappens
b	bk	bk1, bkang
e	3	3ast
s	5	5hit
c	c^	c^rime, c^uh

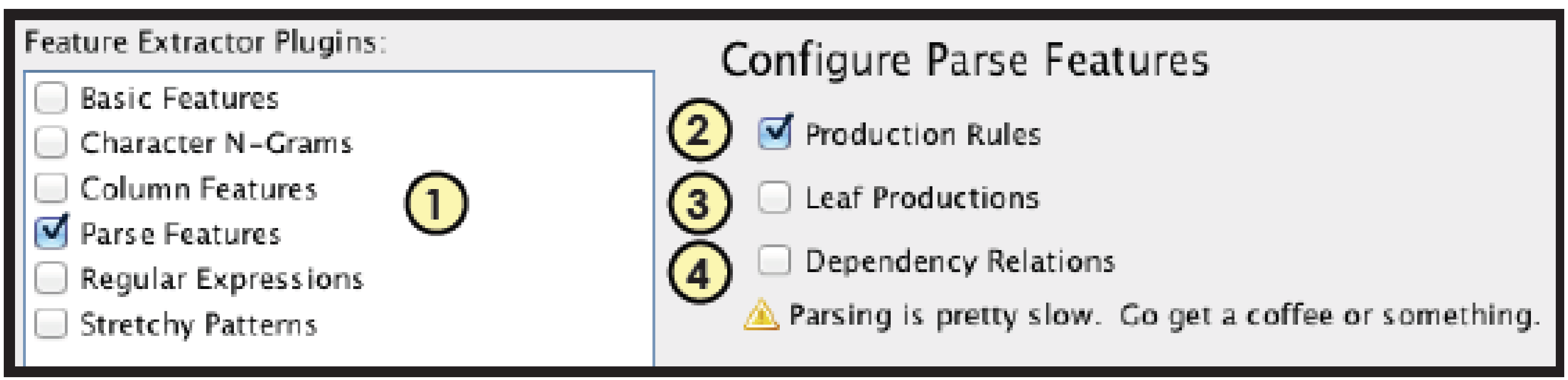
Character N-grams



- Character bigrams can detect graffiti style features
- Could also be used to identify consistent endings on words (i.e., that indicate formality or gender)

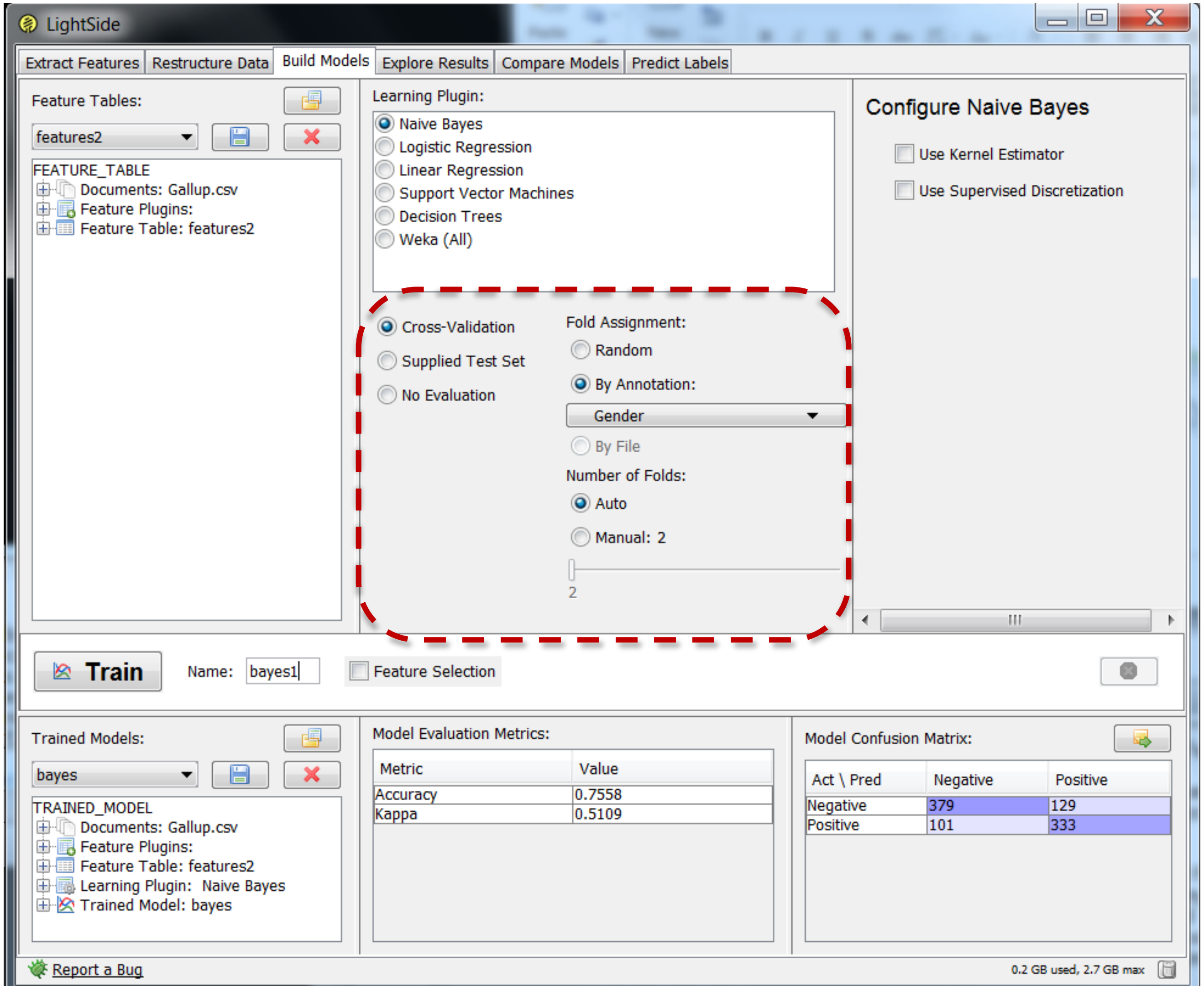
Parse Features

- Word based features lose all structure and order within sentences
- Parse features can capture that
- But they are SLOW!!

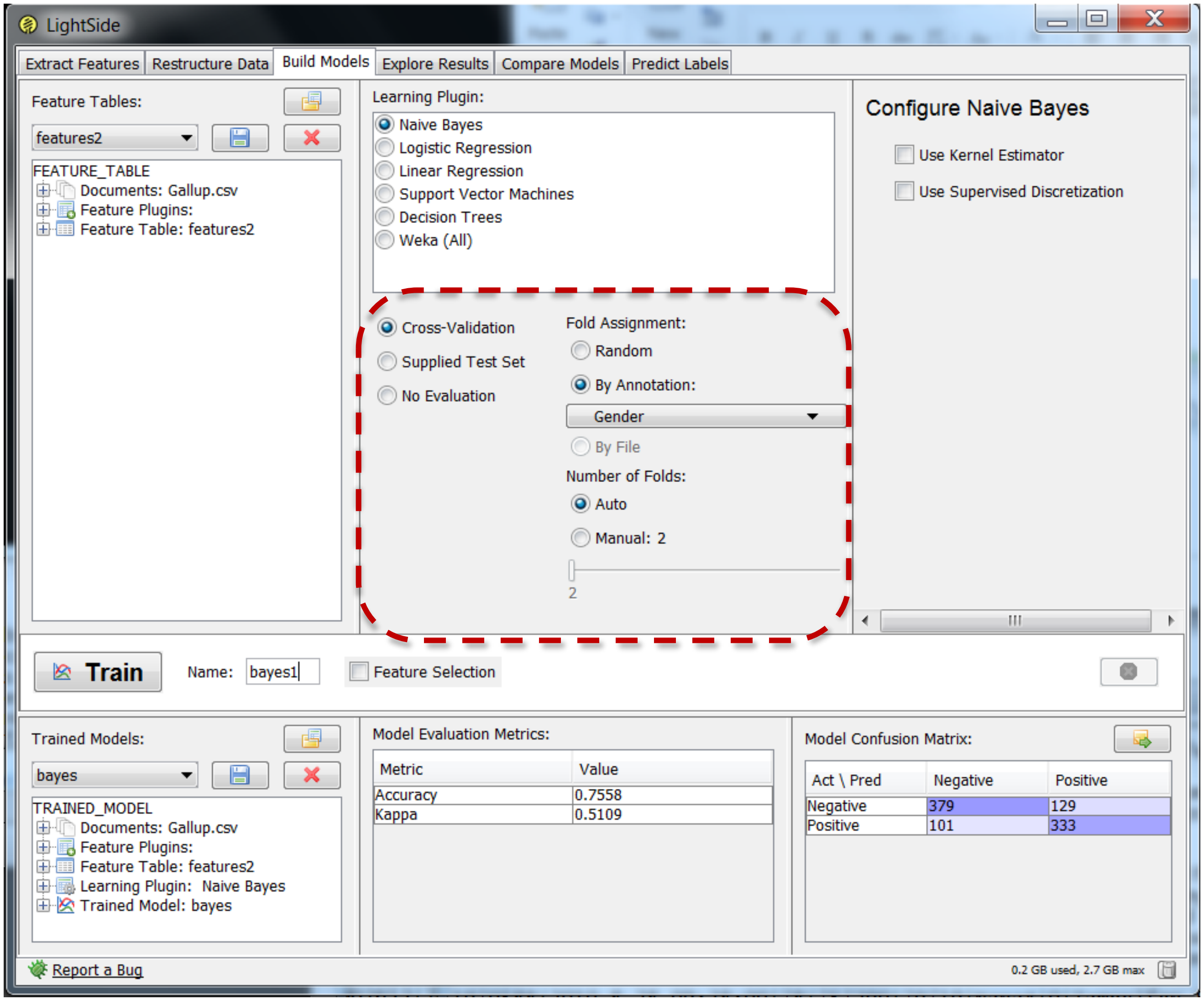


Leveraging Subpopulations through Multi-Level Modeling

Evaluation

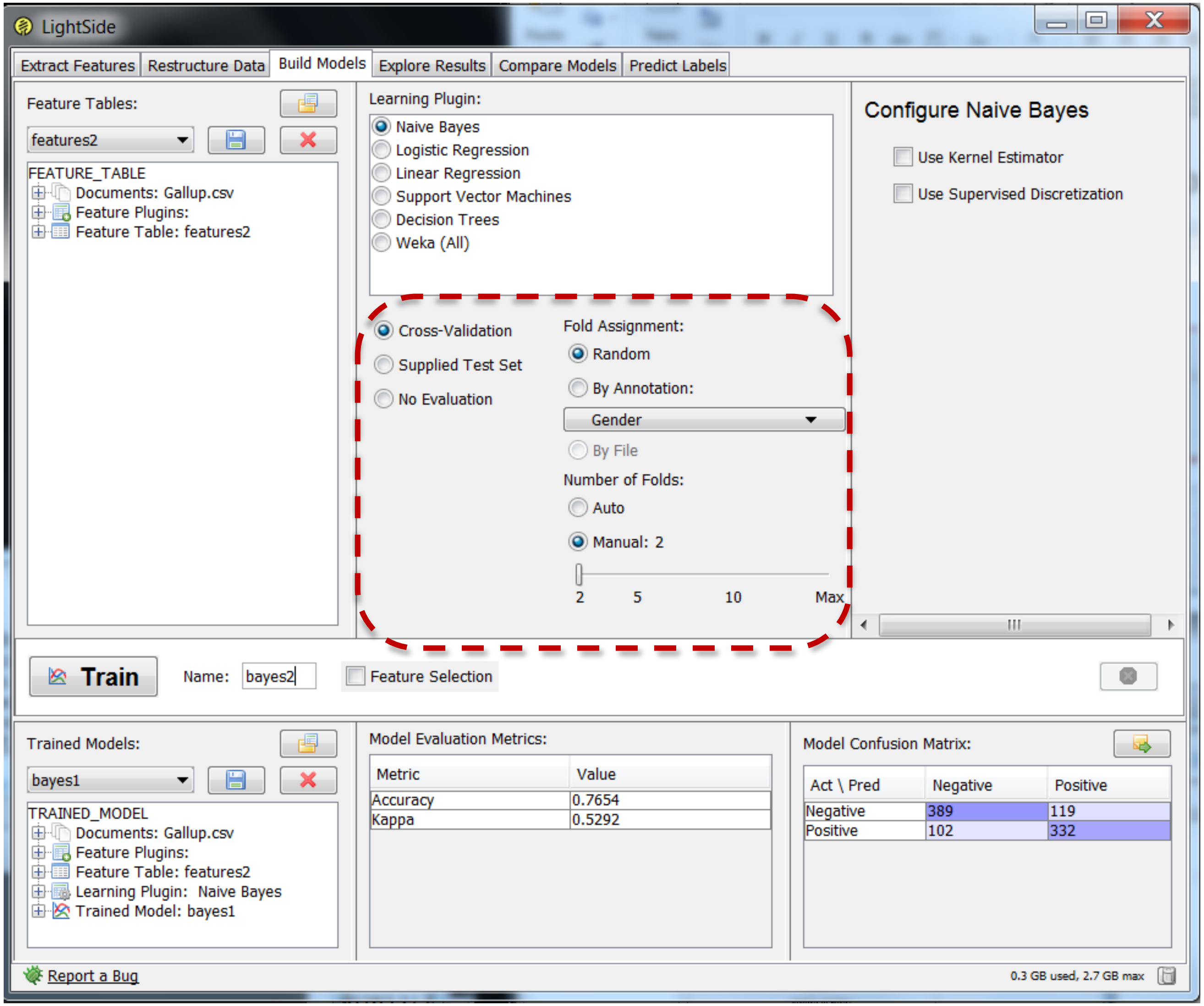

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features2
FEATURE_TABLE
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Configure Naive Bayes
Use Kernel Estimator
Use Supervised Discretization
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
Gender
By File
Number of Folds:
Auto
Manual: 2
2
Train
Name: bayes1
Feature Selection
Trained Models:
bayes
TRAINED_MODEL
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin: Naive Bayes
Trained Model: bayes
Model Evaluation Metrics:
Metric
Value
Accuracy
0.7558
Kappa
0.5109
Model Confusion Matrix:
Act \ Pred
Negative
Positive
Negative
379
129
Positive
101
333
Report a Bug
0.2 GB used, 2.7 GB max

Evaluation



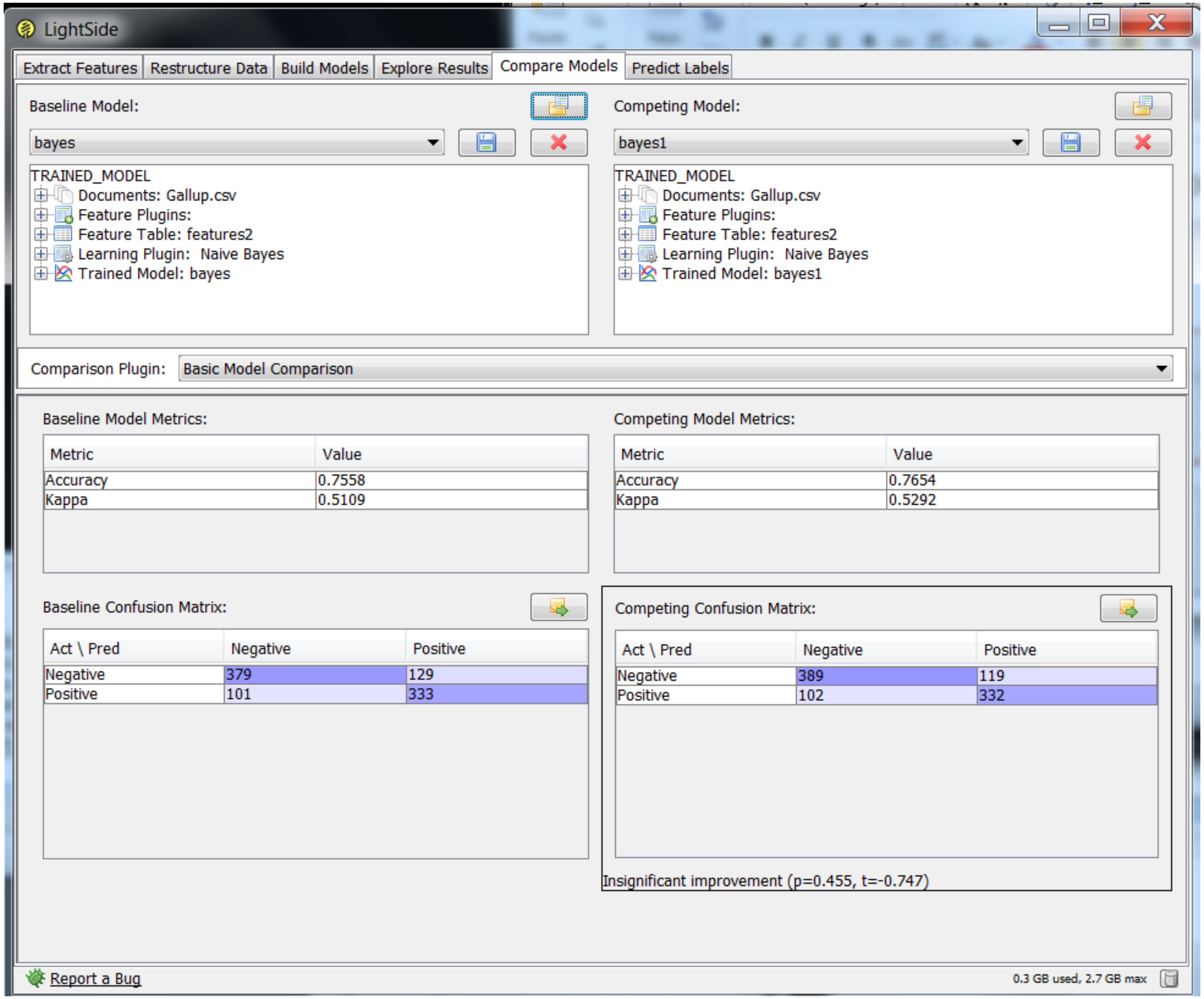
LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features2
FEATURE_TABLE
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Configure Naive Bayes
Use Kernel Estimator
Use Supervised Discretization
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
Gender
By File
Number of Folds:
Auto
Manual: 2
2
Train
Name: bayes1
Feature Selection
Trained Models:
bayes
TRAINED_MODEL
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin: Naive Bayes
Trained Model: bayes
Model Evaluation Metrics:
Metric | Value
Accuracy | 0.7558
Kappa | 0.5109
Model Confusion Matrix:
Act \ Pred | Negative | Positive
Negative | 379 | 129
Positive | 101 | 333
Report a Bug
0.2 GB used, 2.7 GB max

Evaluation



LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features2
FEATURE_TABLE
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Configure Naive Bayes
Use Kernel Estimator
Use Supervised Discretization
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
Gender
By File
Number of Folds:
Auto
Manual: 2
2
5
10
Max
Train
Name:
bayes2
Feature Selection
Trained Models:
bayes1
TRAINED_MODEL
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin: Naive Bayes
Trained Model: bayes1
Model Evaluation Metrics:
Metric
Value
Accuracy
0.7654
Kappa
0.5292
Model Confusion Matrix:
Act \ Pred
Negative
Positive
Negative
389
119
Positive
102
332
Report a Bug
0.3 GB used, 2.7 GB max

Evaluation


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Baseline Model:
bayes
TRAINED_MODEL
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin: Naive Bayes
Trained Model: bayes
Competing Model:
bayes1
TRAINED_MODEL
Documents: Gallup.csv
Feature Plugins:
Feature Table: features2
Learning Plugin: Naive Bayes
Trained Model: bayes1
Comparison Plugin:
Basic Model Comparison
Baseline Model Metrics:
Metric
Value
Accuracy
0.7558
Kappa
0.5109
Competing Model Metrics:
Metric
Value
Accuracy
0.7654
Kappa
0.5292
Baseline Confusion Matrix:
Act \ Pred
Negative
Positive
Negative
379
129
Positive
101
333
Competing Confusion Matrix:
Act \ Pred
Negative
Positive
Negative
389
119
Positive
102
332
Insignificant improvement (p=0.455, t=-0.747)
Report a Bug
0.3 GB used, 2.7 GB max

Why is performance different?

- Men and women used language differently
- Different focus
  - Women had a more personal focus
  - Men had a more national/objective focus

What is different in how men and women talk?

- Word-based features capture more content than style, and are thus vulnerable to domain specificity.

male	female
linux	shopping
microsoft	mom
gaming	cried
server	freaked
software	pink

(Schler 2006)

What is different in how men and women talk?

- Women's language as "deviant" - *Lakoff (1975)* or "more varied" - *Chambers (1992)*
- Extrathematic details in conversational storytelling - time and location (male), people and speech acts (female). *Johnstone (1993)*
  "...after a full three years..."
  "...he would sit and talk to my mother..."

What is different in how men and women talk?

- Hedging, qualifiers, and intensifiers -
  "I think I might have said ..."
  "So he brought to me..."
  "I'm sometimes so jealous of people"

- "like" particle - gender variations in placement and usage
  Iyeiri, Yaguchi, Okabe (2005)
  "...and then, we asked ***like*** *four and one..."*
  *"****Like****, instead of advanced, basic, proficient, and whatever..."*

Confounded with other variables

- Men sound older and women sound younger (Argamon et al., 2007)

- Men sound more like non-fiction and women sound more like fiction (Argamon et al., 2003)

Why do low level features overfit?

- In a linear model, positive weights push the decision towards one class while negative weights push the decision towards the other class

- The magnitude of the weight indicates how much of a push that feature gives

Why do low level features overfit?

- What happens if the same feature predicts age, gender, and social class?
  - If you are predicting gender, then the average value for each feature assumes the mix of age and social class in the data set you trained for
    - The weights normalize for this mix
    - If the mix changes, then the normalization will be wrong
  - So the weights won't predict gender correctly anymore on datasets where the mix of those other factors is different

Never saw MOH in train, so trained model will overpredict extent of swearing among males on test set

Train

FYH MYL
MYH MOL MYL
MYL MYL
MYH MOL FOH
MYL
MOL
MYH MOL MYL MYL
MOL
MYH MYL
MOL FYH MYL
FYH MYH MOL FYH
MYL MYH MOL FYH FYH
MYH FOH FOH
MYH FOH FOH
FYH FOH FOH FOH

Test

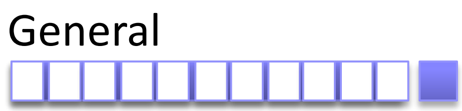
FYL MOH
MOH MOH FOL MOH
FYL FOL
FYL MOL FOL FOL
FYL MOL
MOH FOL
MOH MYH MOL
FYL MOH MYH
FYL MYH MOH
FYL MYH
MOH

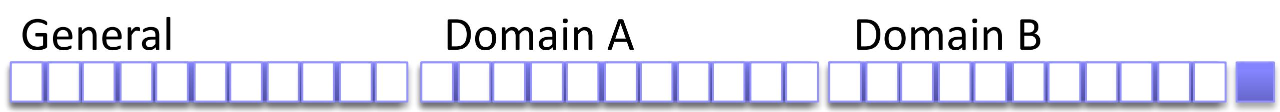
Evaluation of Domain Generality

Occupation	Unigram	Unigram + Bigram	POS Bigram	Stretchy Patterns
Engineering	49.5 (-.01)	53 (.06)	49 (-.02)	50.5 (.01)
Education	49 (-.02)	52 (.04)	54.5 (.09)	51 (.02)
Internet	55.5 (.11)	47.5 (-.05)	55.5 (.11)	56.5 (.13)
Law	51.5 (.03)	46.5 (-.07)	46.5 (-.07)	50.5 (.01)
Non-Profit	50 (0)	54 (.08)	49 (-.02)	51. (.02)
Technology	50 (0)	53.5 (.07)	50 (0)	51.5 (.03)
Arts	48 (-.04)	46.5 (-.07)	51 (.02)	55.4 (.11)
Media	53 (.06)	50 (0)	45 (-.1)	51.5 (.02)
Science	52 (.04)	48 (-.04)	40.5 (-.19)	59.5 (.19)
Student	51 (.02)	46 (-.09)	55 (.10)	62 (.24)
Average	50.95 (.002)	49.7 (-.007)	49.6 (.01)	53.94 (.08)
Random CV	61.05 (.22)	59.65 (.19)	57.95 (.16)	62.8 (.26)

- Contrast random CV and leave-one-occupation-out CV
- All feature space representations show significant drop between random CV and leave-one-occupation-out CV
- Only stretchy patterns remain significantly above random performance

Feature Splitting (Daumé III, 2007)



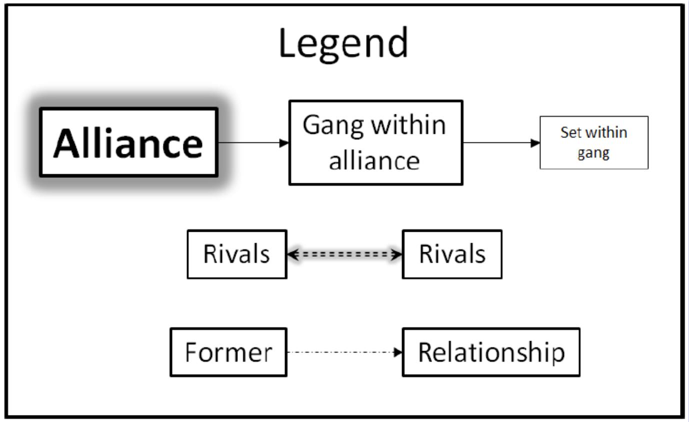


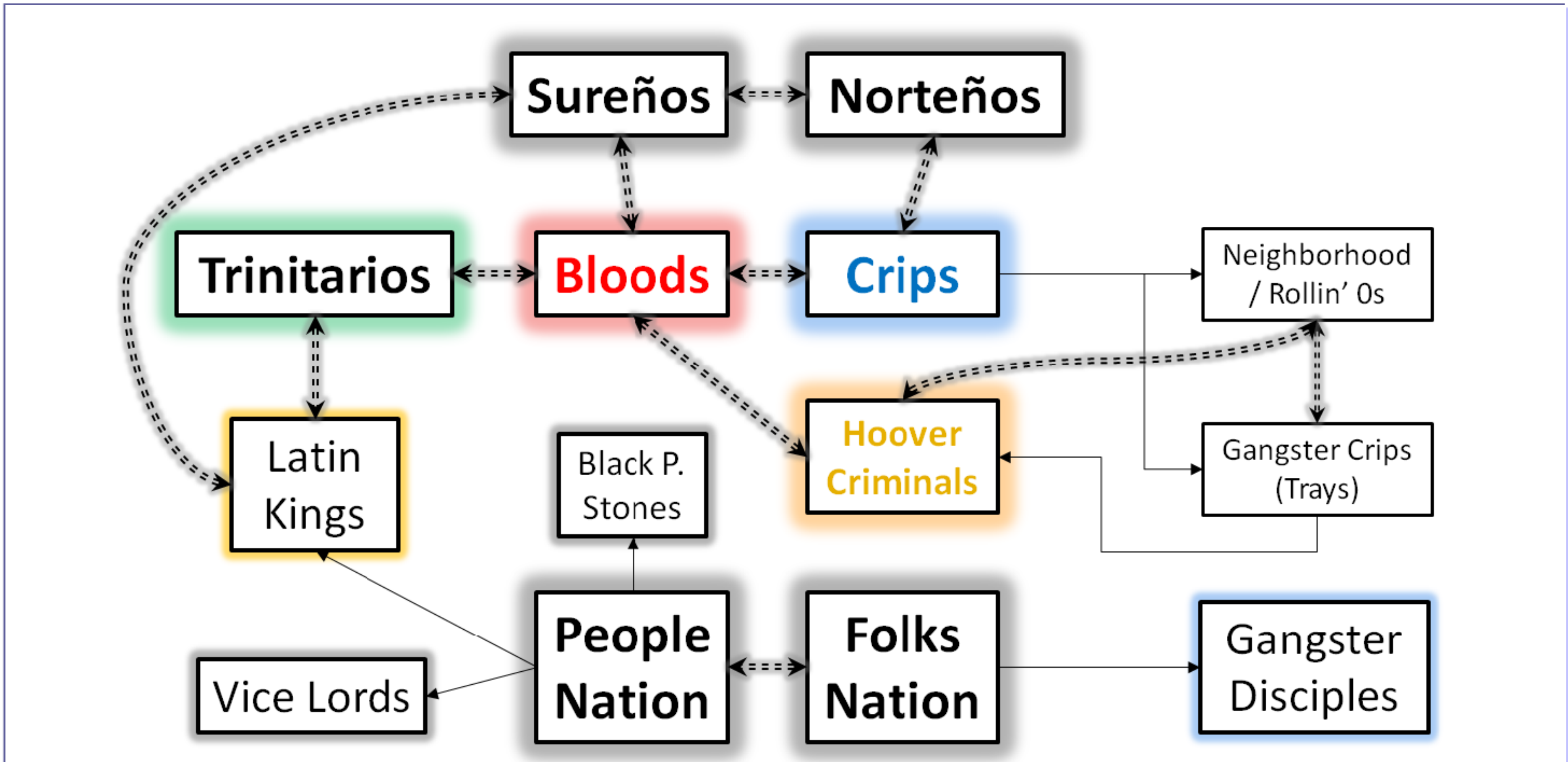
Why is this nonlinear?

It represents the interaction between each feature and the Domain variable

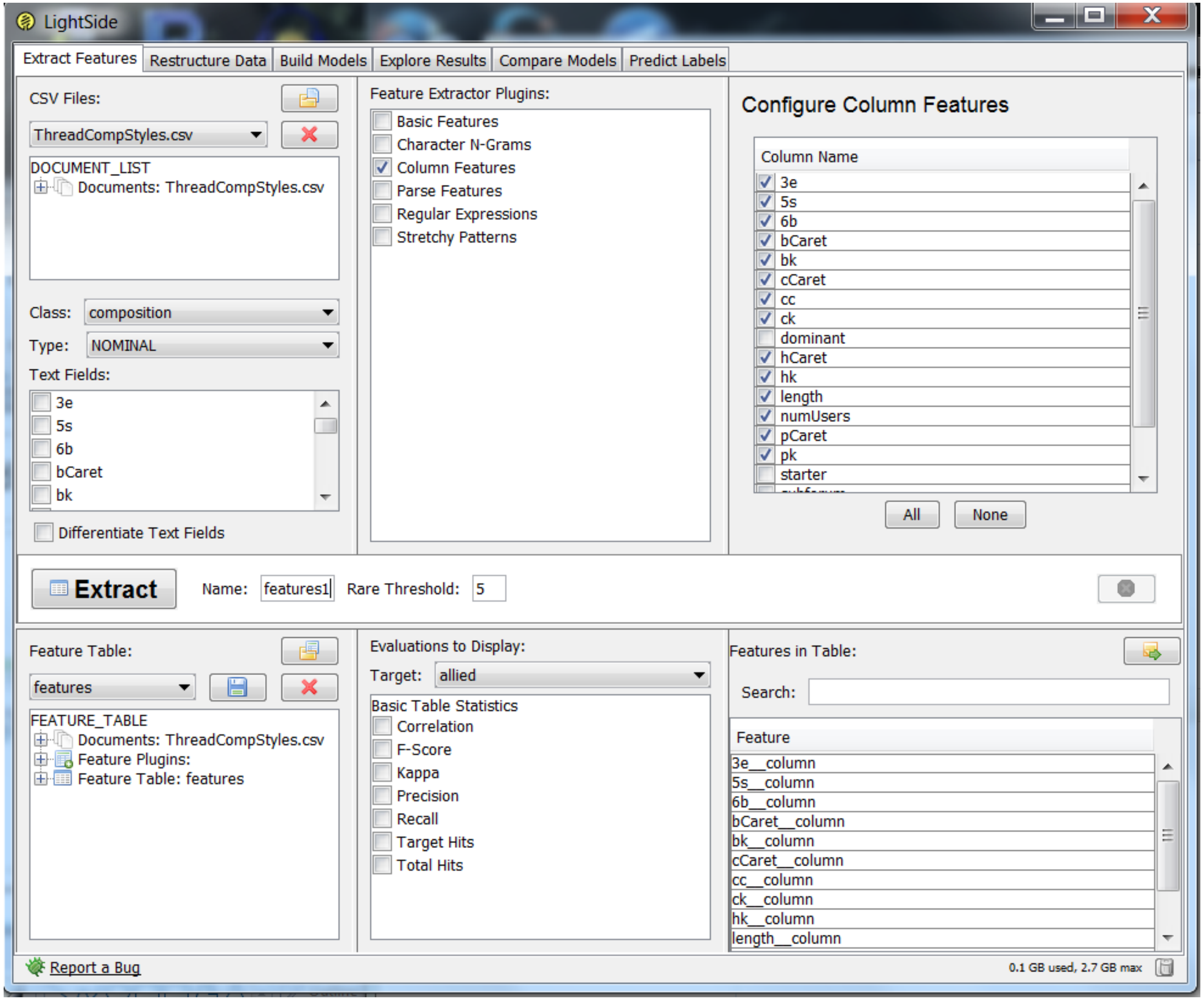
Now that the feature space represents the nonlinearity, the algorithm to train the weights can be linear.

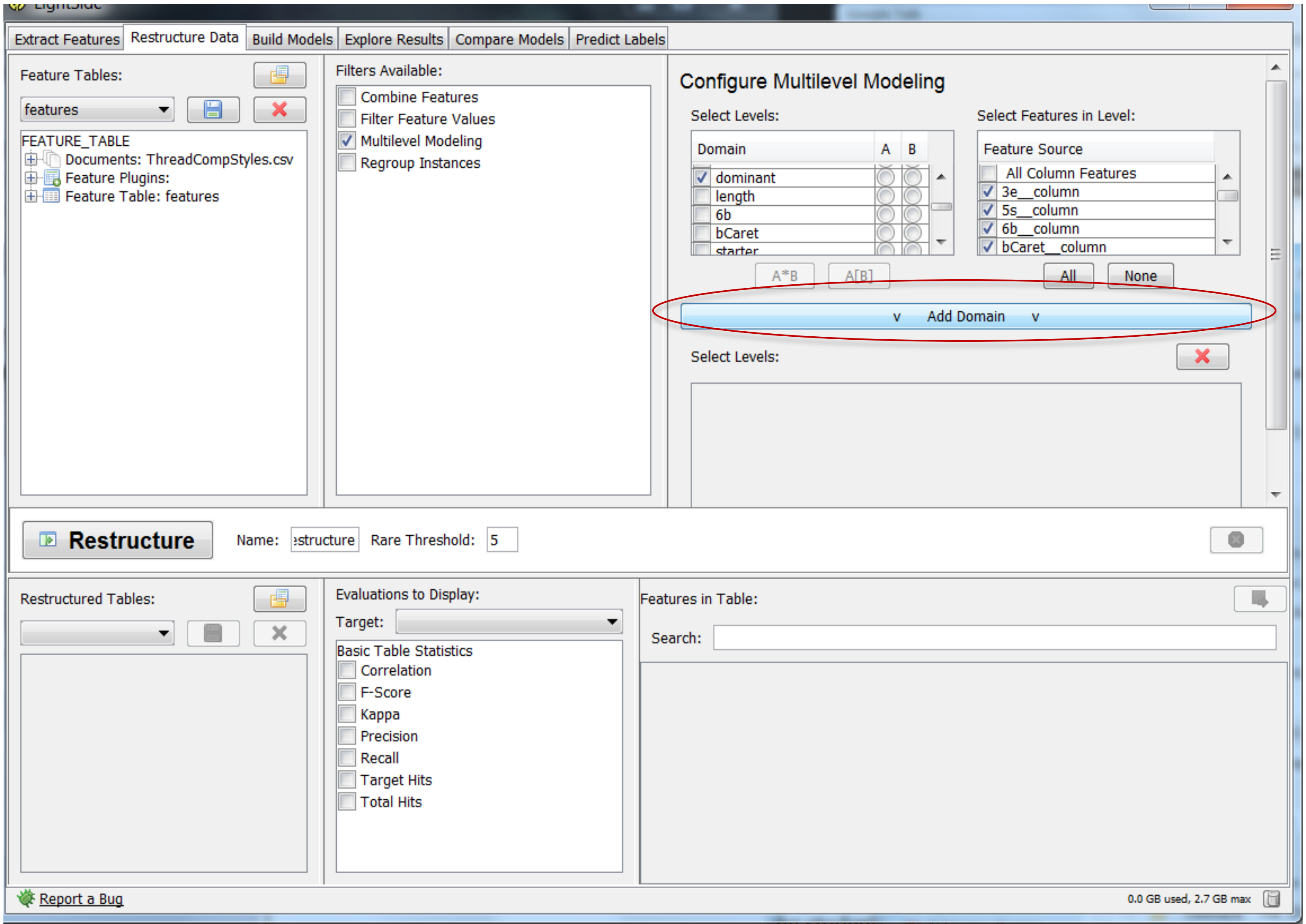
Gang Alliances



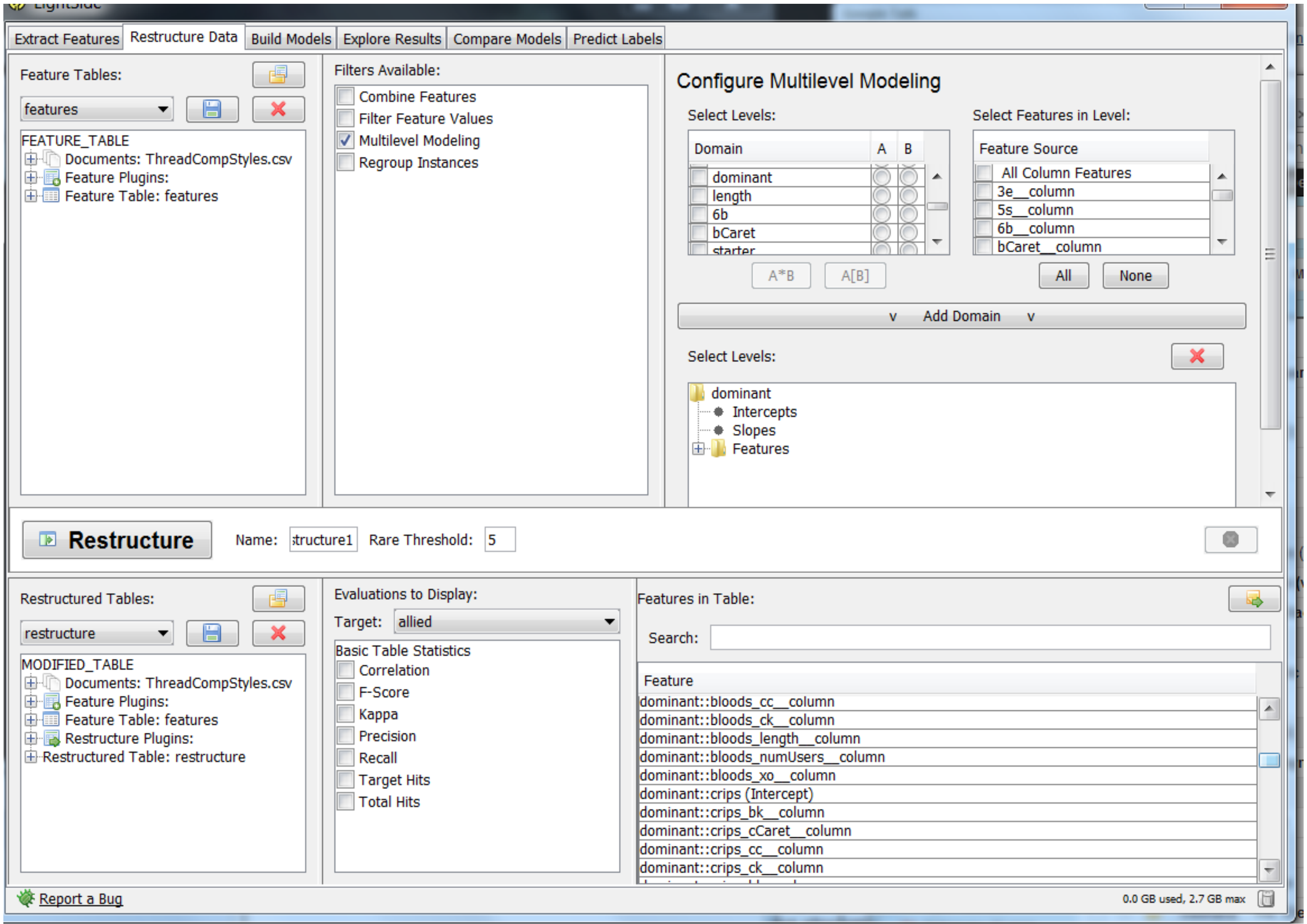


Gangs Data

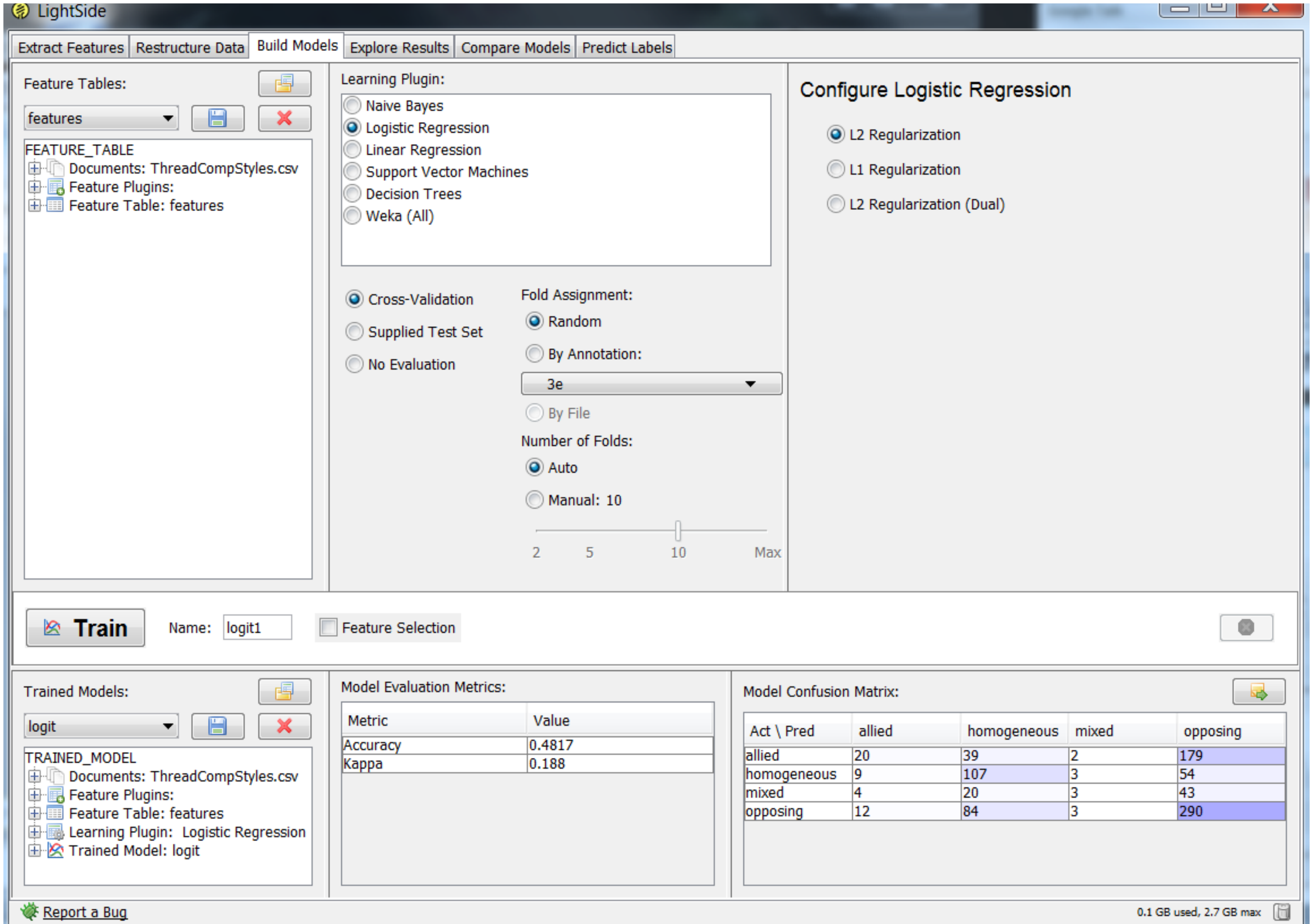

LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
CSV Files:
ThreadCompStyles.csv
DOCUMENT_LIST
Documents: ThreadCompStyles.csv
Class: composition
Type: NOMINAL
Text Fields:
3e
5s
6b
bCaret
bk
Differentiate Text Fields
Feature Extractor Plugins:
Basic Features
Character N-Grams
Column Features
Parse Features
Regular Expressions
Stretchy Patterns
Configure Column Features
Column Name
3e
5s
6b
bCaret
bk
cCaret
cc
ck
dominant
hCaret
hk
length
numUsers
pCaret
pk
starter
All
None
Extract
Name: features1
Rare Threshold: 5
Feature Table:
features
FEATURE_TABLE
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Evaluations to Display:
Target: allied
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Feature
3e__column
5s__column
6b__column
bCaret__column
bk__column
cCaret__column
cc__column
ck__column
hk__column
length__column
Report a Bug
0.1 GB used, 2.7 GB max

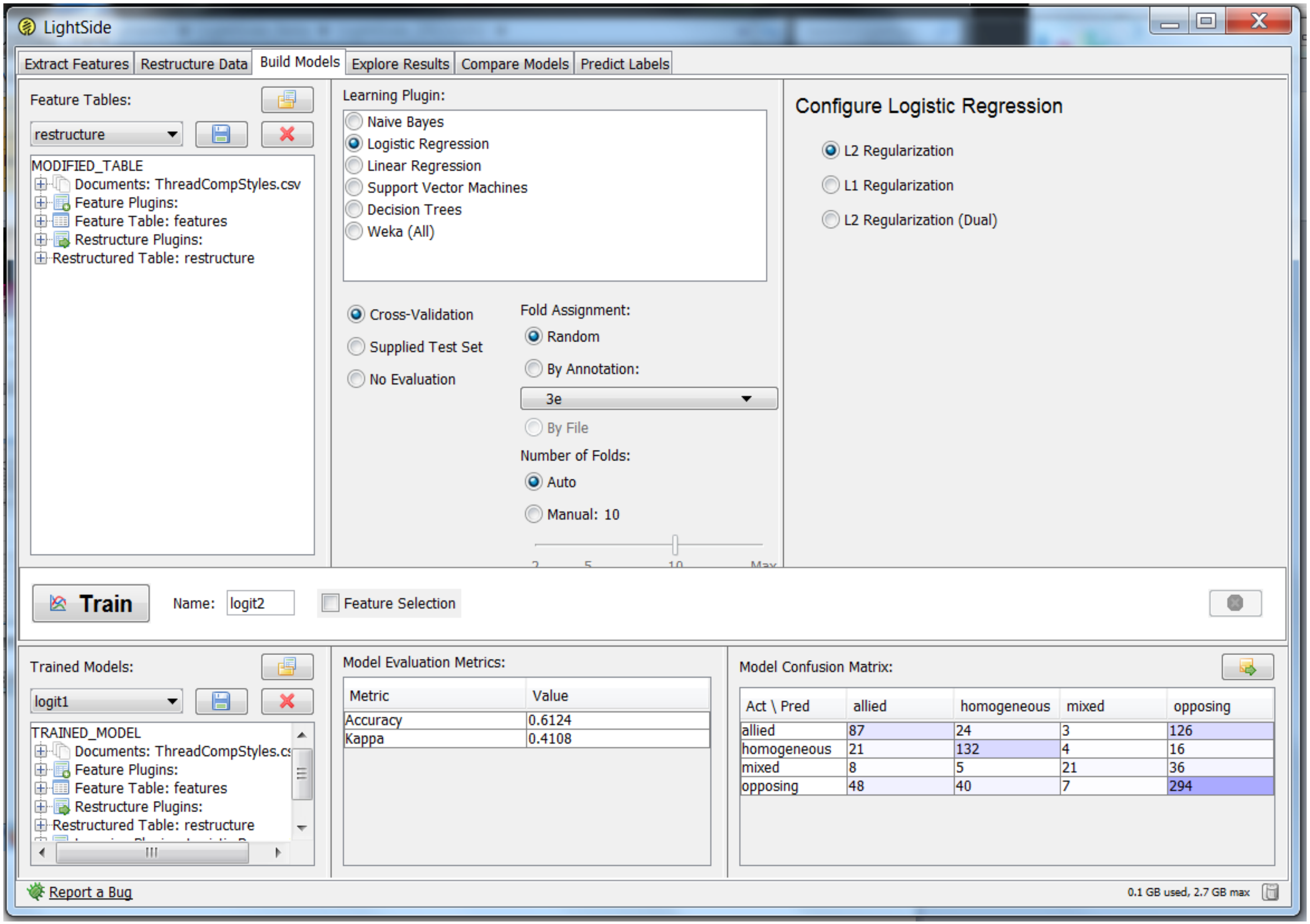


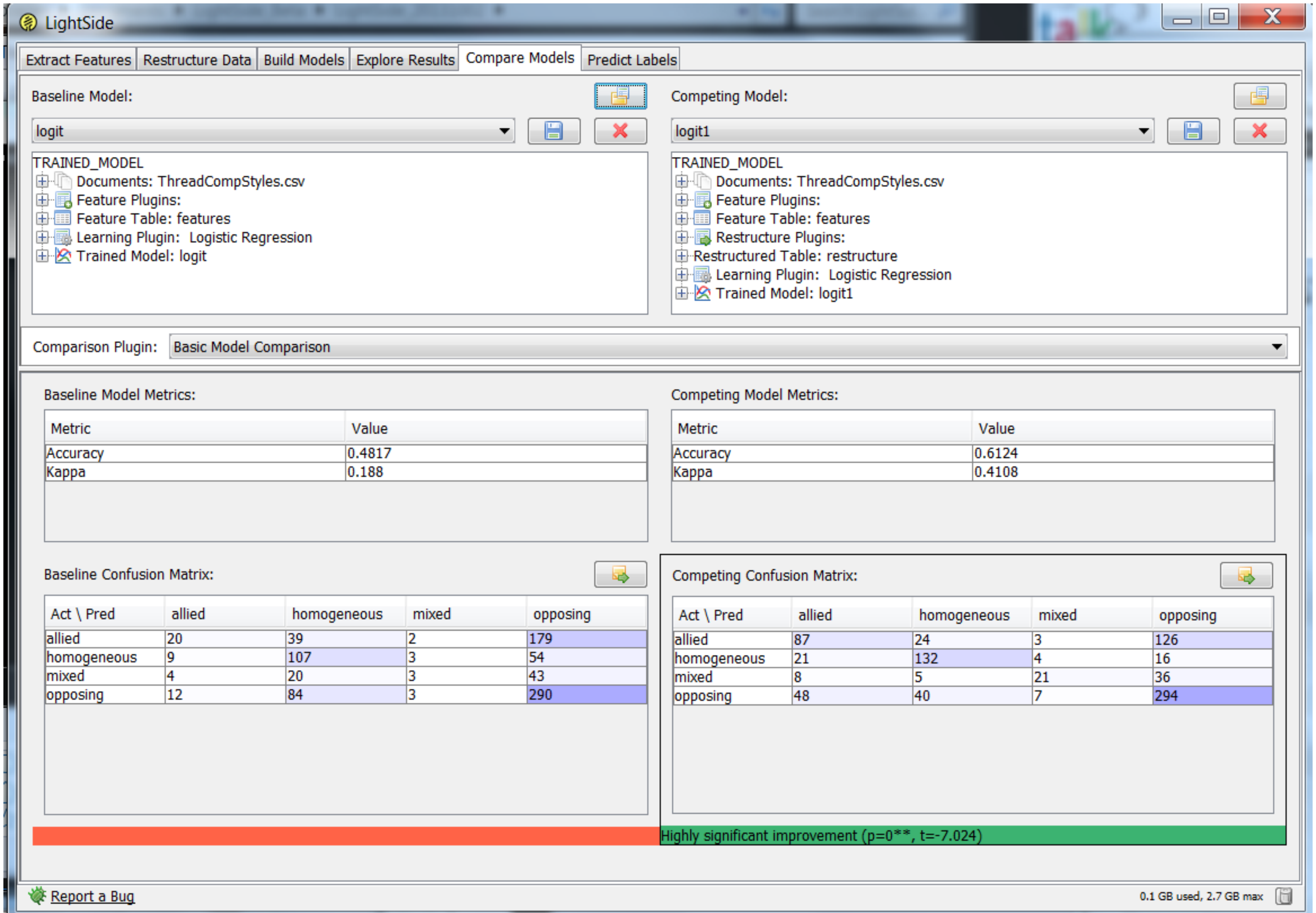
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Filters Available:
Combine Features
Filter Feature Values
Multilevel Modeling
Regroup Instances
Configure Multilevel Modeling
Select Levels:
Domain
A
B
dominant
length
6b
bCaret
starter
Select Features in Level:
Feature Source
All Column Features
3e__column
5s__column
6b__column
bCaret__column
A*B
A[B]
All
None
v Add Domain v
Select Levels:
Restructure
Name:
estructure
Rare Threshold:
5
Restructured Tables:
Evaluations to Display:
Target:
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Report a Bug
0.0 GB used, 2.7 GB max

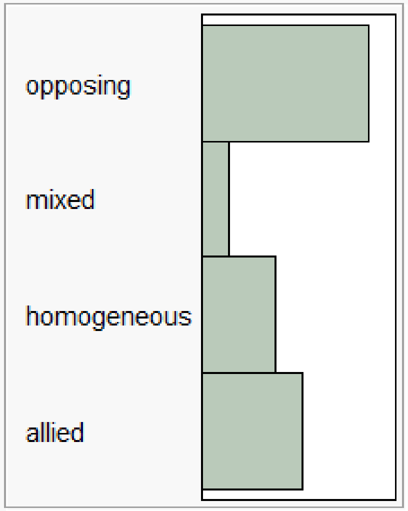


Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Filters Available:
Combine Features
Filter Feature Values
Multilevel Modeling
Regroup Instances
Configure Multilevel Modeling
Select Levels:
Domain
A
B
dominant
length
6b
bCaret
starter
A*B
A[B]
Select Features in Level:
Feature Source
All Column Features
3e__column
5s__column
6b__column
bCaret__column
All
None
Add Domain
Select Levels:
dominant
Intercepts
Slopes
Features
Restructure
Name:
structure1
Rare Threshold:
5
Restructured Tables:
restructure
MODIFIED_TABLE
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Restructure Plugins:
Restructured Table: restructure
Evaluations to Display:
Target:
allied
Basic Table Statistics
Correlation
F-Score
Kappa
Precision
Recall
Target Hits
Total Hits
Features in Table:
Search:
Feature
dominant::bloods_cc__column
dominant::bloods_ck__column
dominant::bloods_length__column
dominant::bloods_numUsers__column
dominant::bloods_xo__column
dominant::crips (Intercept)
dominant::crips_bk__column
dominant::crips_cCaret__column
dominant::crips_cc__column
dominant::crips_ck__column
Report a Bug
0.0 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
features
FEATURE_TABLE
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
3e
By File
Number of Folds:
Auto
Manual: 10
2
5
10
Max
Configure Logistic Regression
L2 Regularization
L1 Regularization
L2 Regularization (Dual)
Train
Name:
logit1
Feature Selection
Trained Models:
logit
TRAINED_MODEL
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit
Model Evaluation Metrics:
Metric Value
Accuracy 0.4817
Kappa 0.188
Model Confusion Matrix:
Act \ Pred allied homogeneous mixed opposing
allied 20 39 2 179
homogeneous 9 107 3 54
mixed 4 20 3 43
opposing 12 84 3 290
Report a Bug
0.1 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Feature Tables:
restructure
MODIFIED_TABLE
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Restructure Plugins:
Restructured Table: restructure
Learning Plugin:
Naive Bayes
Logistic Regression
Linear Regression
Support Vector Machines
Decision Trees
Weka (All)
Cross-Validation
Supplied Test Set
No Evaluation
Fold Assignment:
Random
By Annotation:
3e
By File
Number of Folds:
Auto
Manual: 10
Configure Logistic Regression
L2 Regularization
L1 Regularization
L2 Regularization (Dual)
Train
Name: logit2
Feature Selection
Trained Models:
logit1
TRAINED_MODEL
Documents: ThreadCompStyles.cs
Feature Plugins:
Feature Table: features
Restructure Plugins:
Restructured Table: restructure
Model Evaluation Metrics:
Metric Value
Accuracy 0.6124
Kappa 0.4108
Model Confusion Matrix:
Act \ Pred allied homogeneous mixed opposing
allied 87 24 3 126
homogeneous 21 132 4 16
mixed 8 5 21 36
opposing 48 40 7 294
Report a Bug
0.1 GB used, 2.7 GB max


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict Labels
Baseline Model:
logit
TRAINED_MODEL
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Learning Plugin: Logistic Regression
Trained Model: logit
Competing Model:
logit1
TRAINED_MODEL
Documents: ThreadCompStyles.csv
Feature Plugins:
Feature Table: features
Restructure Plugins:
Restructured Table: restructure
Learning Plugin: Logistic Regression
Trained Model: logit1
Comparison Plugin:
Basic Model Comparison
Baseline Model Metrics:
Metric Value
Accuracy 0.4817
Kappa 0.188
Competing Model Metrics:
Metric Value
Accuracy 0.6124
Kappa 0.4108
Baseline Confusion Matrix:
Act \ Pred allied homogeneous mixed opposing
allied 20 39 2 179
homogeneous 9 107 3 54
mixed 4 20 3 43
opposing 12 84 3 290
Competing Confusion Matrix:
Act \ Pred allied homogeneous mixed opposing
allied 87 24 3 126
homogeneous 21 132 4 16
mixed 8 5 21 36
opposing 48 40 7 294
Highly significant improvement (p=0**, t=-7.024)
Report a Bug
0.1 GB used, 2.7 GB max



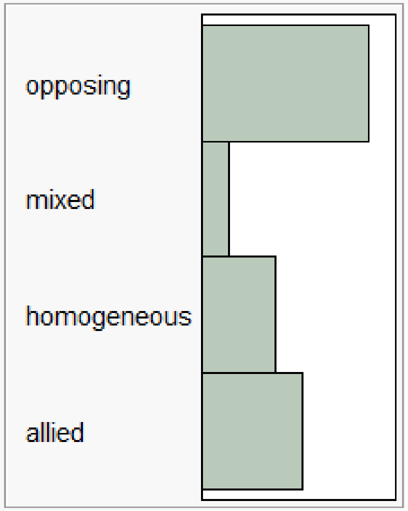
Frequencies

Level	Count	Prob
allied	240	0.27523
homogeneous	173	0.19839
mixed	63	0.07225
opposing	396	0.45413
Total	872	1.00000
N Missing	0	

Feature Analysis

- Style features that distinguish Allied from Opposing differ by dominant gang
- **Crips:**
  - Allied: **bCaret**
  - Opposing: CC, PK, cCaret
- **Bloods:**
  - Allied: **XO, CC**
  - Opposing: hCaret, BK
- **Latin Kings:**
  - Allied: CC, **XO**
  - Opposing: 5S

When the dominant gang is in an allied thread, we see style features that unite them against opposing gangs.



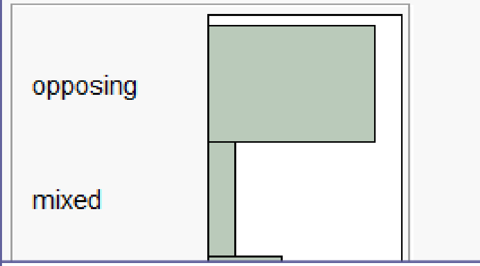
Frequencies

Level	Count	Prob
allied	240	0.27523
homogeneous	173	0.19839
mixed	63	0.07225
opposing	396	0.45413
Total	872	1.00000
N Missing	0	

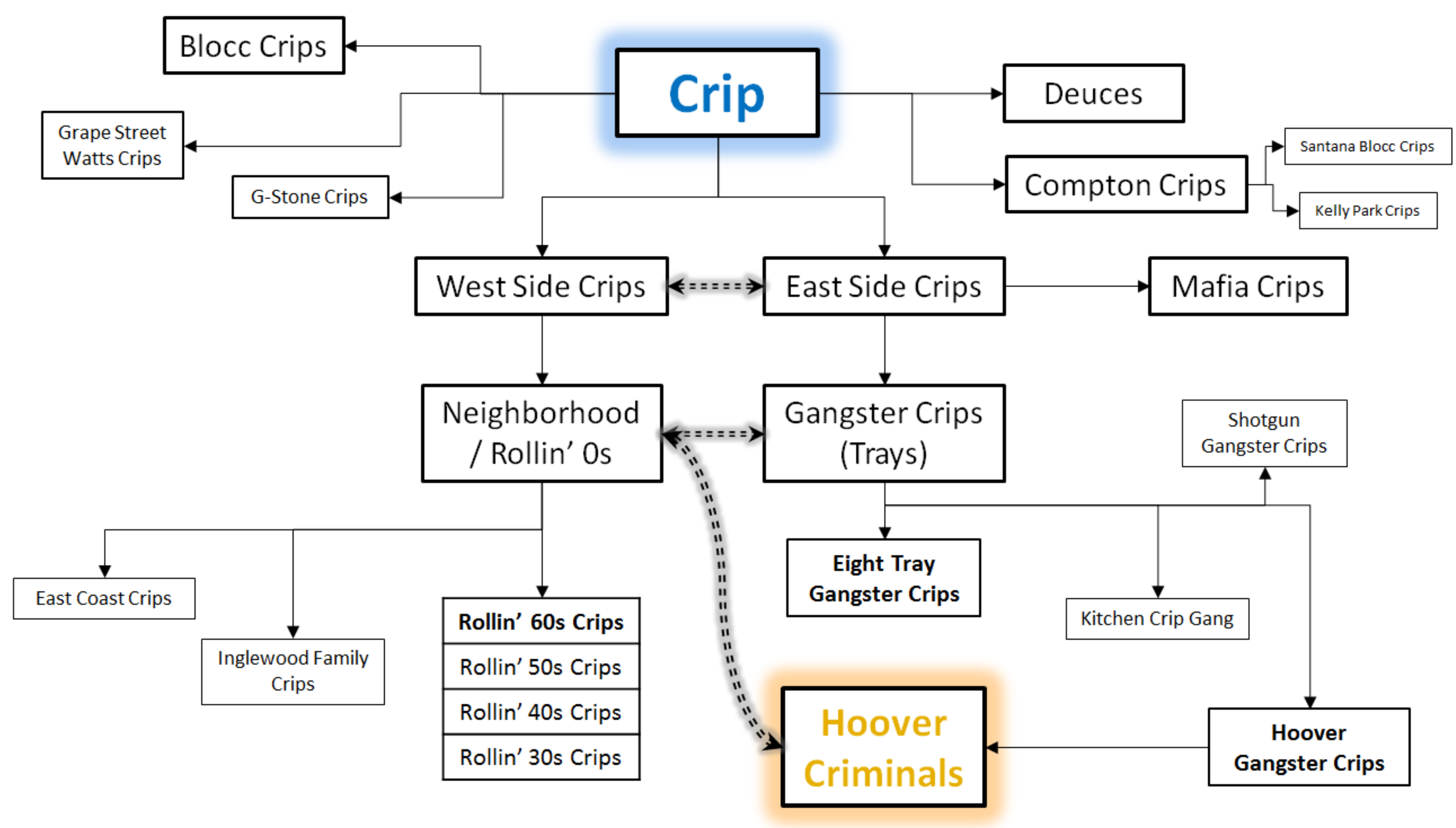
Feature Analysis

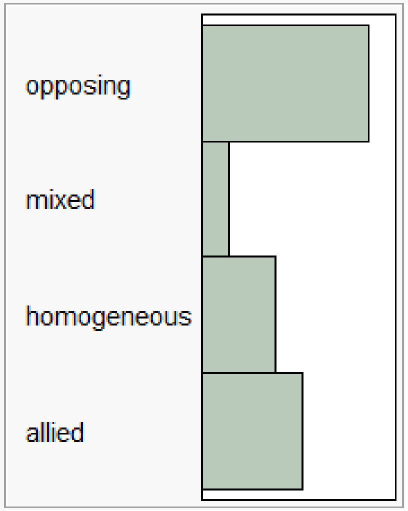
- Style features that distinguish Allied from Opposing differ by dominant gang
- **Crips:**
  - Allied: bCaret
  - Opposing: **CC**, **PK**, **cCaret**
- **Bloods:**
  - Allied: XO, CC
  - Opposing: hCaret, **BK**
- **Latin Kings:**
  - Allied: CC, XO
  - Opposing: **5S**

When the dominant gang is in an opposing thread, we also see features that unite the opposing gangs against them.



Feature Analysis





Frequencies

Level	Count	Prob
allied	240	0.27523
homogeneous	173	0.19839
mixed	63	0.07225
opposing	396	0.45413
Total	872	1.00000
N Missing	0	

Feature Analysis

- Unigram features that distinguish Allied from Opposing don't differ by dominant gang as much as style features
- **Universal:**
  - Allied: lmao, you, crew
  - Opposing: forever, wtf, where
- **Crips:**
  - Allied: lol
  - Opposing: know, about
- **Bloods:**
  - Allied: niggas, the
  - Opposing: at

We see relationship words, but not gang identity words.