

# A Taste of Applied Machine Learning

Carolyn Penstein Rosé Language Technologies Institute/ Human-Computer Interaction Institute

#### Carolyn Rosé



- Joint appointment between Language Technologies and Human-Computer Interaction
- President of the International Society of the Learning Sciences

PhD in Language and Information Technologies, 1998

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

- Co-Chair of the CSCL community committee
- Associate Editor of the International Journal of Computer Supported Collaborative Learning

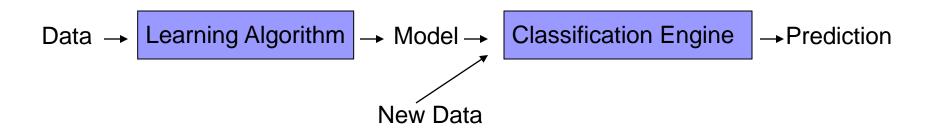
Machine Learning: Conceptual Overview

#### How does machine learning work?

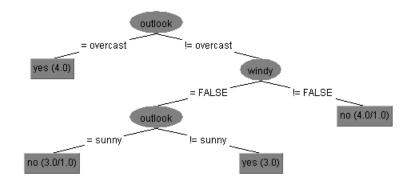
outlook	temperature	humidity	wir	ndy	<u>play</u>	
sunny	hot	high	F	ALSE	no	
sunny	hot	hiah	T	RUF	no	
A slight	Outlook:	<u> - '- (' ( -</u>	U	le le	arner	
+		ny -> No		ne m		
informa	<b>\</b> +	ercast -> Ye	-5		What	
do you	tl Rainy-> Yes			his case?		
sunr Clas				ı E	yes	
	<feature na<="" td=""><td>ame&gt;:</td><td></td><td>E</td><td>yes</td></feature>	ame>:		E	yes	
4	<value< td=""><td>e&gt; -&gt; <predict< td=""><td>ion&gt;</td><td></td><td>yes</td></predict<></td></value<>	e> -> <predict< td=""><td>ion&gt;</td><td></td><td>yes</td></predict<>	ion>		yes	
What v	<value> -&gt; <prediction <value=""> -&gt; <prediction< td=""><td>ion&gt;</td><td>se?</td><td>yes</td></prediction<></prediction></value>		ion>	se?	yes	
overcast				ALSE	yes	
rainy	•••		•	RUE	no	

#### What is machine learning?

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



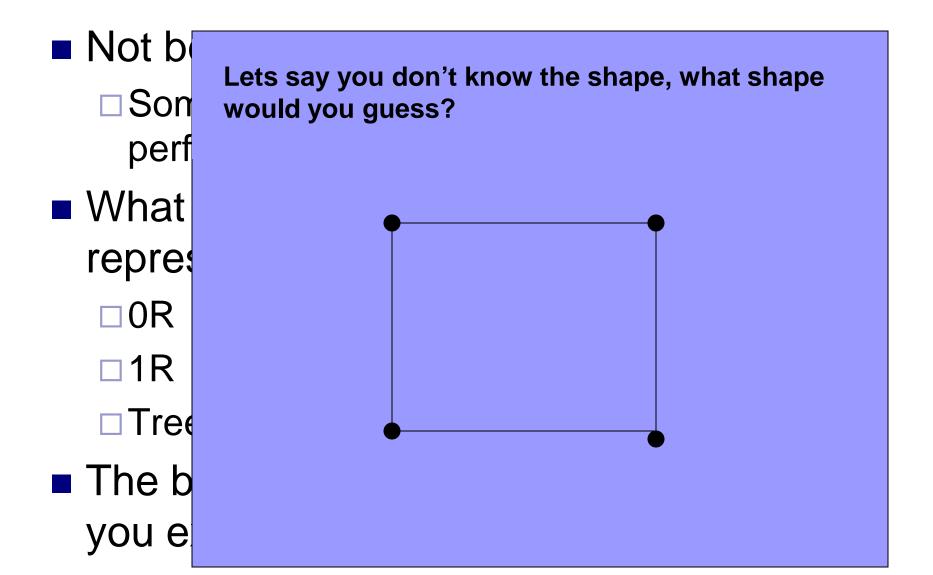
#### More Complex Algorithm...

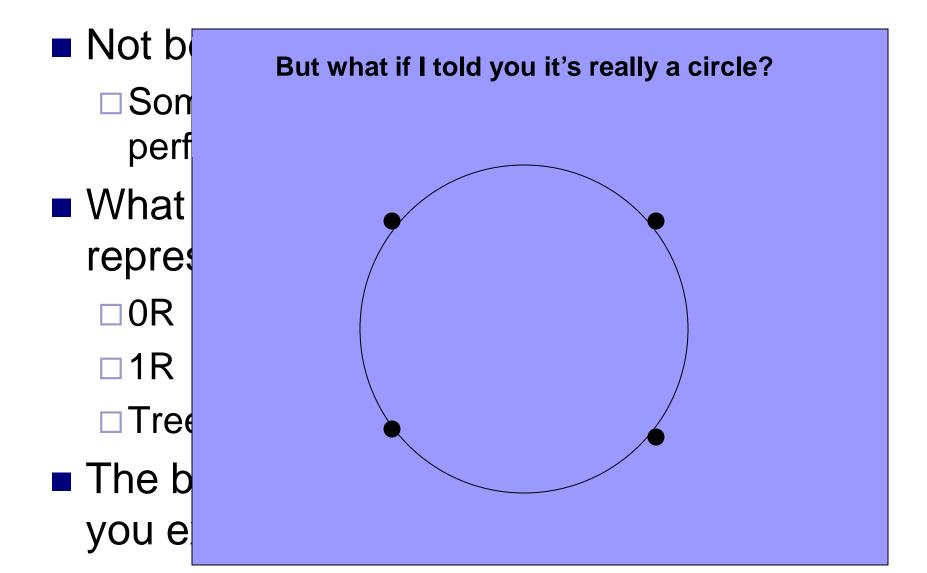


\* Only makes 2 mistakes!

- 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?
   OR
  - □1R
  - □Trees
- The best algorithm for your data will give you exactly the power you need

Not b Lets say you don't know the shape, what shape Son would you guess? perf What repres  $\Box 0R$  $\Box 1R$ The b vou e





Not b □Son perf What repres  $\Box 0R$  $\Box 1R$ The b

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

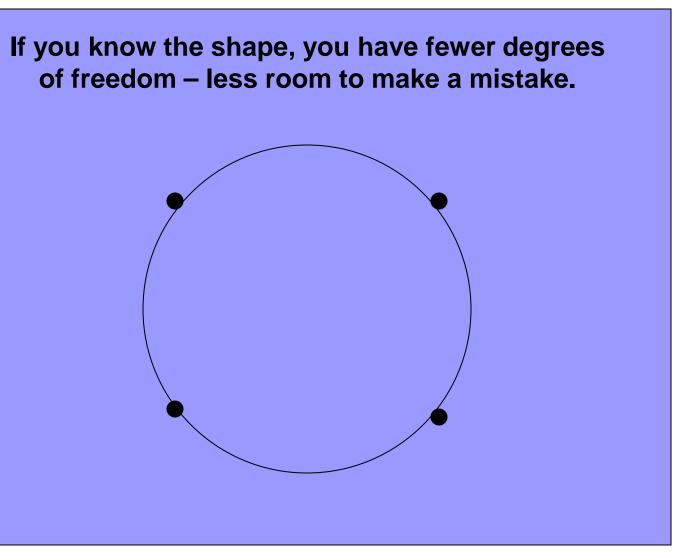
Not b □Son perf What repres  $\Box 0R$  $\Box 1R$ The b

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

Not b □Son perf What repres  $\Box 0R$  $\Box 1R$ The b

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

Not b □ Son perf What repres  $\Box 0R$  $\Box 1R$ The b

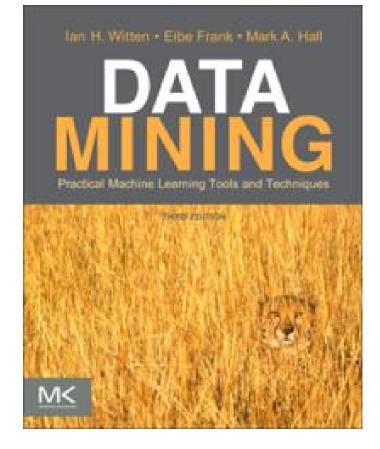


- Not because it is more complex
  - Sometimes more complexity makes performance worse
- What is different in what the three rule representations assume about your data?
   OR
  - □1R
  - □Trees
- The best algorithm for your data will give you exactly the power you need

## 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/)

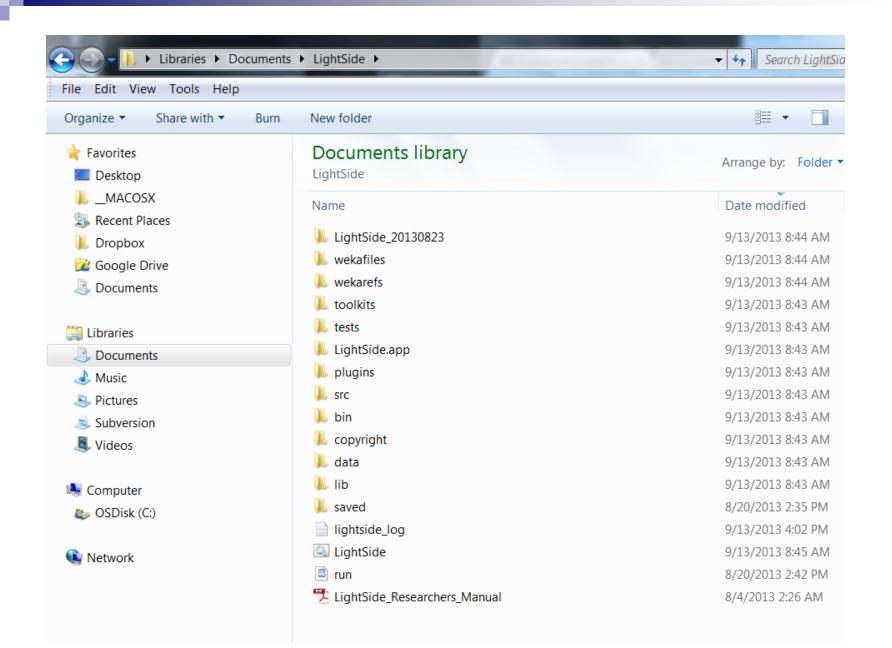
- Open source machine learning toolkit
- Includes Java API

LightSIDE (http://lightsidelabs.com/research)

- Weka add-on for text processing
- □ Developed at CMU!

#### lightsidelabs.com/research/





#### 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.

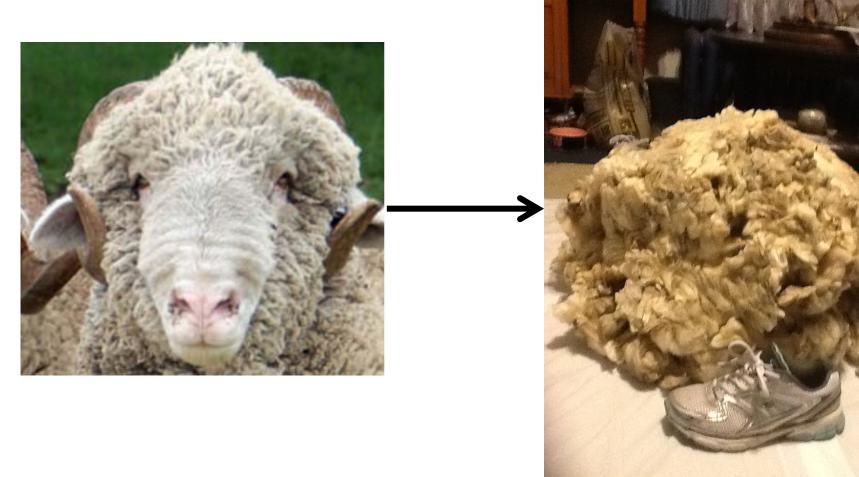
2. And ALL Wrong Side Rows. Sc in each stitch across. Ch.2, turn. 3. \*Work hdc in 4 sc, work post dc as follows: yo, insert hook right to left under post of hdc below next sc, draw up loop, (yo and draw through 2 loops) twice to complete dc. Skp sc behind post dc, work hdc in next sc, work post dc in 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...

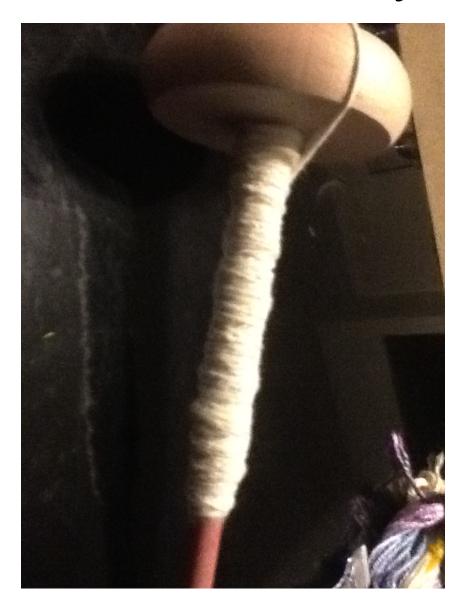


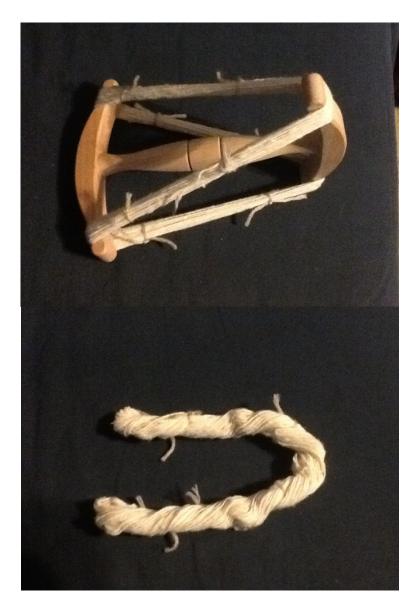




The rectangular batts (like csv files with raw texts) that come out of the drum carder are still not ready to make something out of

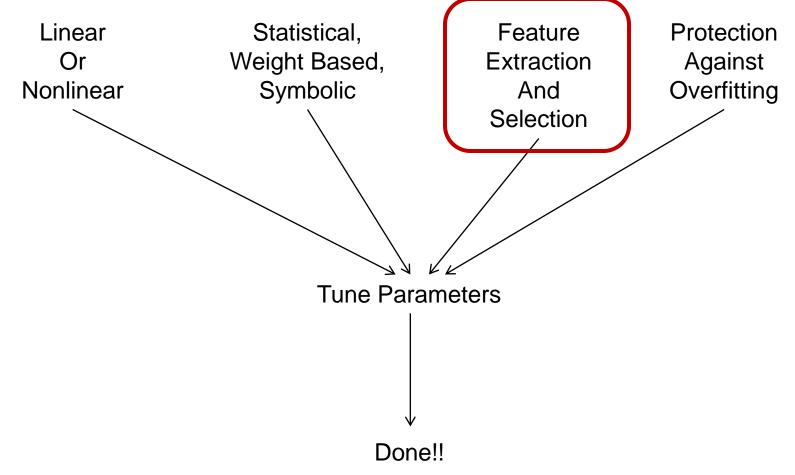
#### Here's where you come in!





#### Text Teaser

### Decisions about Machine Learning Methods



#### Consider this simple example...

SimpleExample.xls				
	A	В		
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				
	А	В		
1	Code	text		
2	Question	Tell me what your favorite color i		
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				
	А	В		
1	Code	text		
2	Question	Tell me what your favorite color is.		
3	Statement	Fell you my favorite color is blue.		
4	Question	Where do you live?		
5	Statement	l ive 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 Dake?		
9	Statement	You should take my applied machine learning course.		
10	Question	Tell me when you get up in the morning.		
11	Statement	Lget up early.		

What distinguishes Questions and Statements?

I versus you is not a reliable predictor

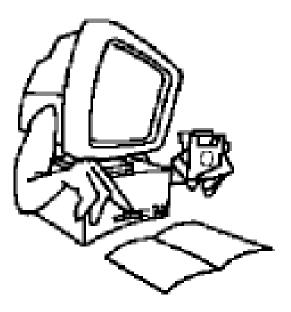
# What are good features for text categorization?

SimpleExample.xls				
	A	В		
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 vour data



## LightSIDE: A quick tour

Edtract Features       Restructure Data       Build Models       Explore Results       Compare Models       Predict Labels         CSV Files:	LightSIDE	Sec. (3. 4)	
Image: State of the state	Extract Features Restructure Data Build Mode	s Explore Results Compare Models Predict Labels	
Stretchy Patterns     POS Bigrams     POS Bigrams     Word/POS Pairs   Line Length   Contains Non-Stopwords   Binary N-grams?   Include Punctuation?   Remove Stopwords?   Stem N-grams?     Feature Table:     Target Hits   Peatures in Table:     Target Hits   Provide Functuation   Paster Table Statistics   Recall   Target Hits   Paster Table Statistics		Basic Features Character N-Grams Column Features	✓ Unigrams
Class: •   Type: IOMINAL   Text Fields: © Contains Non-Stopwords   © Binary N-grams? © Include Punctuation?   © Include Punctuation? Remove Stopwords?   © Inferentiate Text Fields © Stem N-grams?   Image: Image:   Image: Image: <t< td=""><td></td><td></td><td>POS Bigrams</td></t<>			POS Bigrams
Differentiate Text Fields     Differentiate Text Fields     Extract     Name:   features   Rare Threshold:     5     Feature Table:     Target:     Target:     Basic Table Statistics     Recall   Target Hits   Precision   Total Hits   Correlation   F-Score   Kappa     Kappa	Type: NOMINAL 🔻		Line Length
Differentiate Text Fields     Extract     Name:     features     Recall     Target Hits   Precision   Total Hits   Precision   Total Hits   Precision   Correlation   F-Score   Kappa     Kappa     Stem N-grams?     Stem N-grams?     Stem N-grams?     Stem N-grams?     Features in Table:     Search:			✓ Include Punctuation?
Feature Table:     Target:     Target Hits   Precision   Total Hits   Correlation   F-Score   Kappa	Differentiate Text Fields		
Target:     Basic Table Statistics   Recall   Target Hits   Precision   Total Hits   Correlation   F-Score   Kappa	Extract Name: features R	are Threshold: 5	
Basic Table Statistics   Recall   Target Hits   Precision   Total Hits   Correlation   F-Score   Kappa	Feature Table:		Features in Table:
Report a Bug     0.1 GB used, 2.7 GB max		Basic Table Statistics          Recall         Target Hits         Precision         Total Hits         Correlation         F-Score	Search:

LightSIDE	$x \in [\infty]$		*** 1 1 🖏 👦			
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 I		
	Ø Open			Colling advances	×	
Class:	Look in:	📜 data		▼ 🤌 📂 🛄 -		
Type: NOMINA Text Fields:	Recent Items Desktop My Documents Computer	Callup Callup Callup Cale Newsgro Calesentime Calesentime	oupTopic nt_documents			
	Network	File name: Files of type:	sentiment_sentences.csv CSV	T	Open Cancel	
W Report a Bug			Total Hits Correlation F-Score Kappa			0.1 GB used, 2.7 GB max

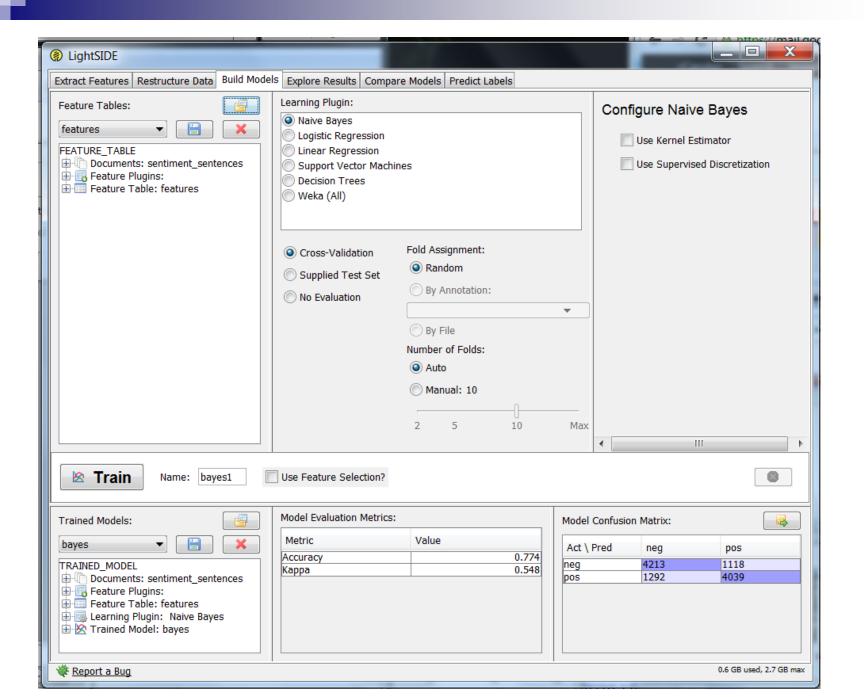
Extract Features Restructure Data Build Models Explore Results Compare Model	Predict Labels
CSV Files: sentiment_sentences DOCUMENT_LIST Documents: sentiment_sentences Class: class Type: NOMINAL Text Fields: Differentiate Text Fields Feature Extractor Plugins: Basic Features Character N-Grams Column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

Extract Features Restructure Data Build Models Explore Results Compare Models Pred	
CSV Files: sentiment_sentences DOCUMENT_LIST Documents: sentiment_sentences Class: class Type: NOMINAL V text text	dict Labels
Differentiate Text Fields  Extract Name: [Features1] Rare Threshold: 5	Remove Stopwords?  Stem N-grams?
Feature Table:   features   FEATURE_TABLE   Documents: sentiment_sentences   Feature Plugins:   Feature Table: features	Features in Table:         Search:         Feature         '60s         '70s         'd         'em         'III         'm         're         's         've            0.3 GB used, 2.7 GB max

LightSIDE						
Extract Features Rest	tructure Data	Build Models	Explore Results Compare Models Predict Labels			
CSV Files:	•		Feature Extractor Plugins:	Configure Basic	Features	
DOCUMENT_LIST			Character N-Grams	🔽 Unigrams		
Documen	Save Featur	e Table	Paras Pastures Pagain Egeneration	C. Agenera	×	
	Look in:	📔 saved		•	🦻 📂 🛄 •	
Class: class	Recent					
Type: NOMIN	Items					
Text Fields:						
✓ text	Desktop					
	My					
	Documents					
Differentiate						
Extrac	Computer					0
Feature Table:		File name:	features.arff		Save Feature Table	
features	Network	Files of type:		<b>•</b>	Cancel	
FEATURE_TABLE	_		LightSIDE Feature Table			
Documents: se	entiment_sente s:	ences	ARFF (Weka)	003		
E Feature Table:			Precision     Total Hits	'70s		^_
				'd 'em		
			F-Score	'll		
			🔲 Карра	'm 're		
				's		
				've 		-
🔆 <u>Report a Bug</u>				·	0.6 GB	used, 2.7 GB max

LightSIDE     Extract Features Restructure Data Build Mode	Company Deputts	na Mandala - Danadista La kala		
Feature Tables:         features         features         FEATURE_TABLE         Documents: sentiment_sentences         Feature Plugins:         Feature Table: features	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machin Decision Trees Weka (All)			Configure Naive Bayes Use Kernel Estimator Use Supervised Discretization
	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random By Annotation: By File Number of Folds: Auto Manual: 10 2 5 10	▼ Max	
Name: bayes	Use Feature Selection?	>		
Trained Models:	Model Evaluation Metrics:		Model	Confusion Matrix:
🔆 <u>Report a Bug</u>				0.3 GB used, 2.7 GB max

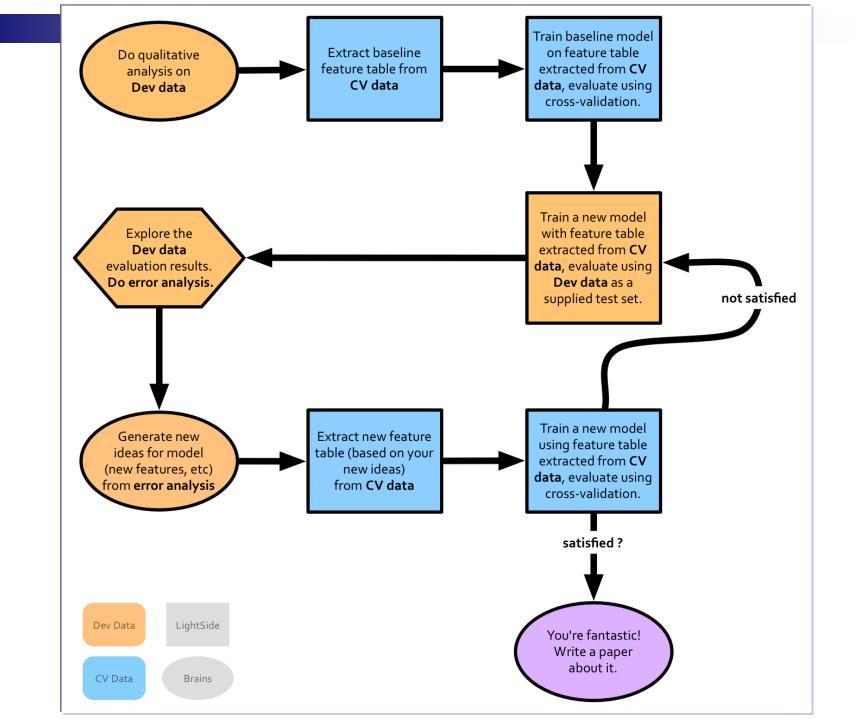
(



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!



### 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

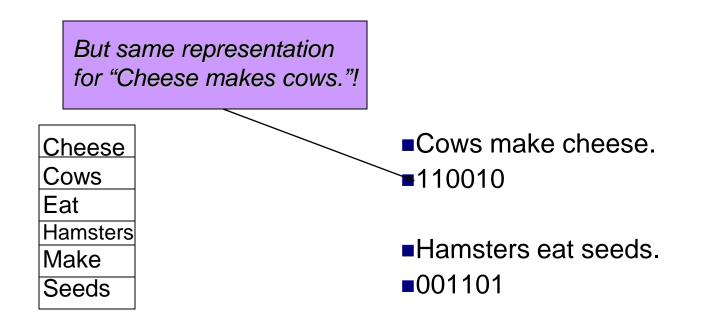
Cheese Cows Eat Hamsters Make Seeds

- 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





LightSIDE	· rise first time from range	
Extract Features Restructure Data Build Models Exp	plore Results Compare Models Predict Labels	
CSV Files:	ature 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:		<ul> <li>FOS Bigranis</li> <li>Word/POS Pairs</li> <li>Line Length</li> <li>Contains Non-Stopwords</li> <li>Binary N-grams?</li> <li>Include Punctuation?</li> <li>Remove Stopwords?</li> </ul>
Differentiate Text Fields  Extract Name: features Rare Th	Threshold: 5	Stem N-grams?
Extract Name: Teatures Rate In	mesnou. 5	
Targ		Features in Table:
Report a Bug	L	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



#### 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

③ LightSide	7		
Extract Features Restructure Data Build Mode	Is Explore Results Compare Models Predict Labels		
CSV Files:	Feature Extractor Plugins:	Configure Basic Features	Â
Gallup.csv	Character N Crams	🔽 Unigrams	
⊕ () Documents: Gallup.csv	Regular Expressions	🔲 Bigrams	
	Stretchy Patterns	Trigrams	Ξ
		POS Bigrams	
Class: Vote		Word/POS Pairs	
Type: NOMINAL		Line Length Contains Non-Stopwords	
Age			
Gender State		Count Occurences	
✓ text		Include Punctuation           Remove Stopwords         Include Punctuation	
Differentiate Text Fields		Stem N-Grams	
Extract Name: features1 R	are Threshold: 5		
Feature Table:	Evaluations to Display:	Features in Table:	
features 🔹 📔 🗙	Target: Negative   Basic Table Statistics	Search:	
FEATURE_TABLE	Correlation F-Score	Feature	
🕀 🧾 Feature Table: features	Rappa Precision	" 'm 're	^
	Recall	's 've	
	Total Hits	<comma> <ellipse></ellipse></comma>	
		<period> <questionmark></questionmark></period>	
		<semicolon></semicolon>	-
👋 <u>Report a Bug</u>		0.2 GB used, 2.7	7 GB max 📋

"Because the cost of healthcare is just outta sight

crazy"

Configure Basic Features
Vnigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Count Occurences
Include Punctuation
Remove Stopwords
Stem N-Grams

"Because the cost of healthcare is just outta sight

crazy"

Configure Basic Features
Vnigrams
🔲 Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Count Occurences
Include Punctuation
Remove Stopwords
Stem N-Grams

"Because the cost of healthcare is just outta sight

crazy"

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

#### "the cost of healthcare"

DT NN PRP NN

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

### 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, superlative10.LS List item marker11.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

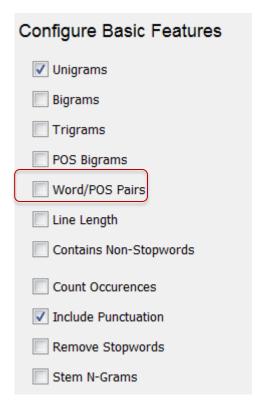
### Part of Speech Tagging

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 whpronoun
36.WRB wh-adverb

## "the cost of healthcare"



4

### "the cost of healthcare"

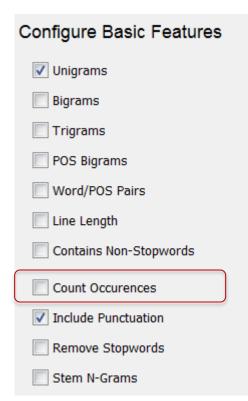
Configure Basic Features Unigrams Bigrams Trigrams POS Bigrams Word/POS Pairs Line Length Contains Non-Stopwords Count Occurences Include Punctuation Remove Stopwords Stem N-Grams

### "the cost of healthcare"

YES

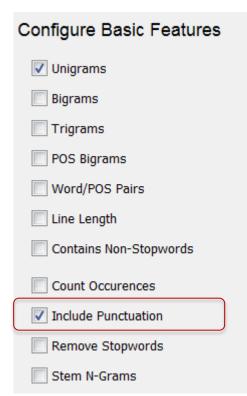
Configure Basic Features
🗸 Unigrams
Bigrams
Trigrams
POS Bigrams
Word/POS Pairs
Line Length
Contains Non-Stopwords
Count Occurences
Include Punctuation
Remove Stopwords
Stem N-Grams

"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.

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

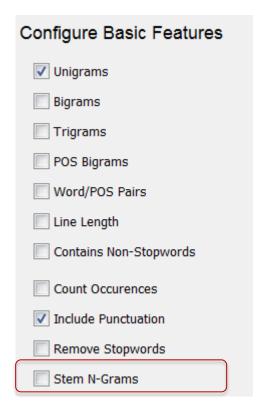


If you uncheck this, punctuation will be ignored and stripped out of the representation.

### "the cost healthcare"

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

"healthcare cost<u>s</u>"  $\rightarrow$  "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

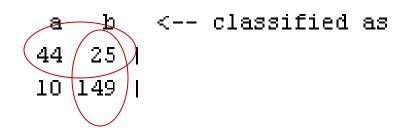
Think like a computer!

- 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 ===

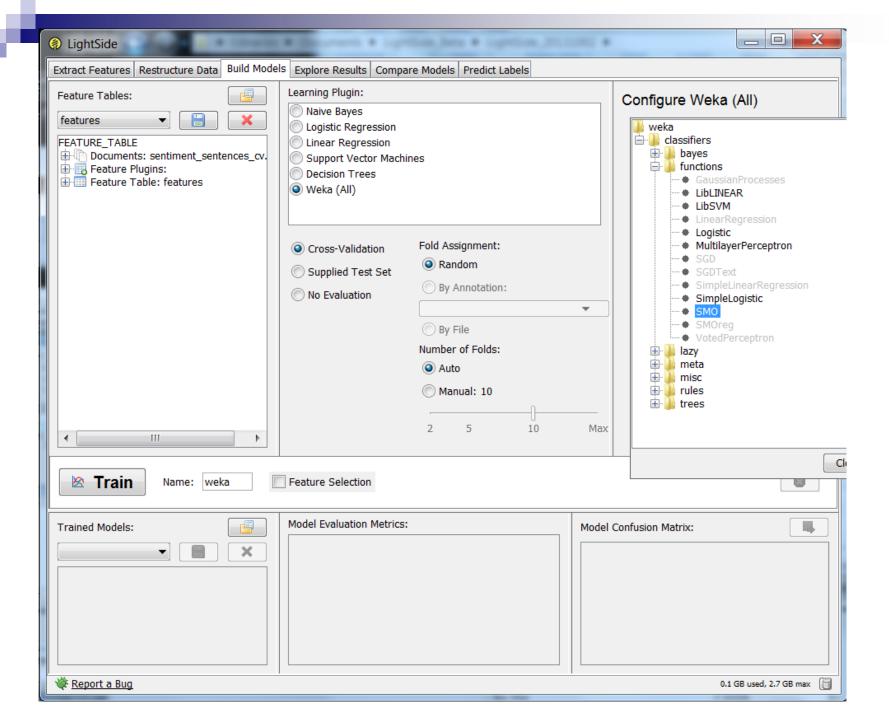


**Goal:** We want to discover how to rerepresent the data so that instances with the same class value look more similar to one another and instances with different class values look more different

- 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)

LightSide			
Extract Features Restructure Data Build Models Explore Results Compare Models Predict Labels			
CSV Files:			
Class: Open	× E		
Type: NOMINAL   Text Fields:     Recent   Recent   Items     A essay   Gallup   MovieReviews   MovieReviews   Name:   fe   Mame:   My   Documents   My   Documents   My   Documents   My   Computer     Sentiment_sentences_cv			
Target Hits	Open Cancel		
Report a Bug 0.1 GB used, 2.7 GB max			

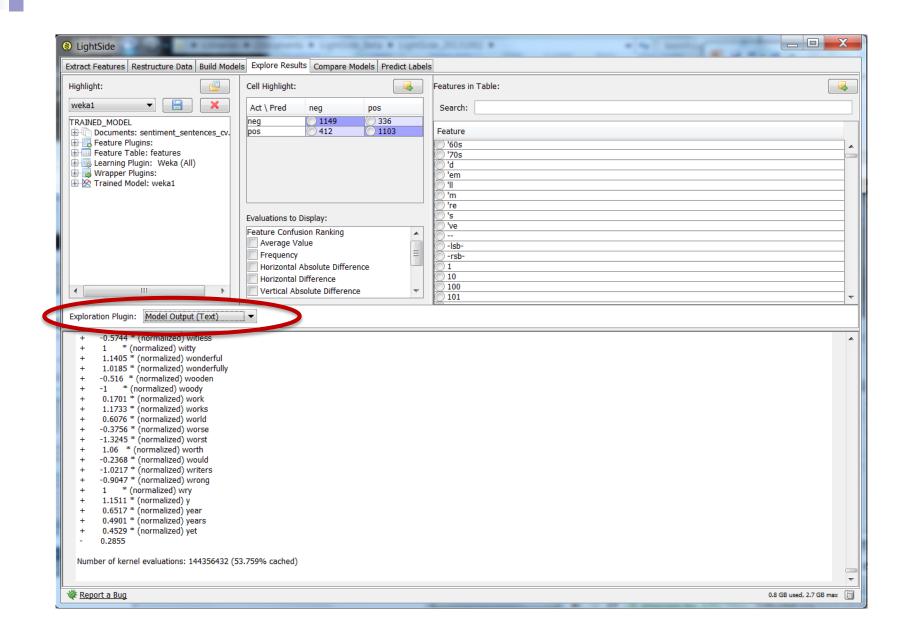
LightSide	· Courses + Lyrian has + Lyria	* (01100) *	
Extract Features Restructure Data Build Mode CSV Files: Sentiment_sentences_cv.csv  X DOCUMENT_LIST Documents: sentiment_sentences_cv.	Is Explore Results Compare Models Predict Labels Feature Extractor Plugins: Basic Features Character N-Grams Column Features Parse Features	Configure Basic Features	
<ul> <li></li> <li></li> <li>Class: class</li> <li>Type: NOMINAL</li> </ul>	Regular Expressions	<ul> <li>Trigrams</li> <li>POS Bigrams</li> <li>Word/POS Pairs</li> <li>Line Length</li> </ul>	Ħ
Text Fields:		<ul> <li>Contains Non-Stopwords</li> <li>Count Occurences</li> <li>Include Punctuation</li> <li>Remove Stopwords</li> <li>Stem N-Grams</li> </ul>	
	Care 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	• • • • • • • • • • • • • • • • • • •



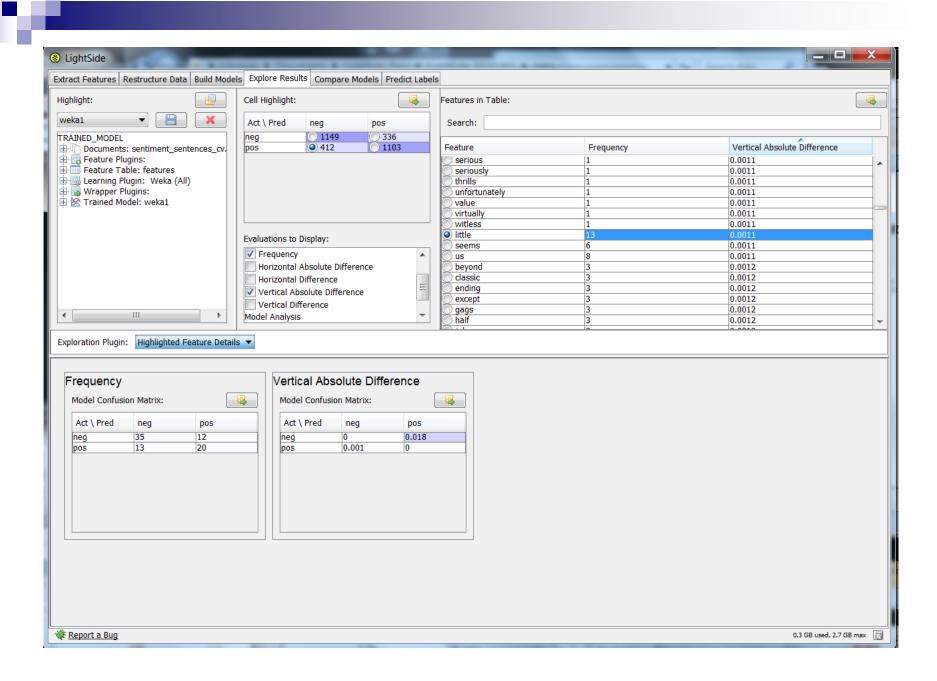
Ø LightSide	a Disease a Light	tion, here & capition, it				
Extract Features Restructure Data Build Mode	s Explore Results Compa	re Models Predict Labels				
Feature Tables:       Image: Constraint of the sector of the	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machin Decision Trees Weka (All)	nes			gure Wek	( <b>a (All)</b> -C 1.0 -L 0.001 -P 1.0
4 111	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random     By Annotation:     By File Number of Folds:     Auto     Manual: 10     2 5 10	▼ Max			
Name: weka1	Feature Selection #: 1	1000				
Trained Models:	Model Evaluation Metrics:		Model	Confusion	Matrix:	
weka	Metric Accuracy Kappa	Value 0.7343 0.4685	Act \ neg pos	Pred	neg 2905 1095	pos 941 2721
🔆 <u>Report a Bug</u>						0.8 GB used, 2.7 GB max 🛛 📋

LightSide	R Deserve A species law 4 of	
Extract Features Restructure Data Build Mo	Explore Results Compare Models Predict La	abels
Feature Tables:       Image: Constraint of the sector of the	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machines Decision Trees Weka (All) Cross-Validation Cross-Validation Supplied Test Set No Evaluation DOCUMENT_LIST Document	ence V
Image: Train       Name: weka         Trained Models:       D         Image: Weka       Image: Weka         Image: TRAINED_MODEL       Documents: sentiment_senter         Image: Trained Models:       Image: Weka         Image: Trained Models:       Image: Trained Models:         Image: Trained Models:       Image: Trained	Look in:       Image: data         Look in:       Image: data         Image: data       Image: data </td <td>dev.csv Open Cancel</td>	dev.csv Open Cancel
🖗 <u>Report a Bug</u>		0.8 GB used, 2.7 GB max

LightSide	* (b		•			
Extract Features Restructure Data Build Mode	S Explore Results Compare I	Models Predict Labels				
Feature Tables:       Image: Constraint of the sector of the	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machines Decision Trees Weka (All)			Configur Choose	re Weka (A SMD -C 1,	<b>VII)</b> 0 -L 0.001 -P 1.0
	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>		senter			
4 III >		•	•			
Name: weka2	Feature Selection #: 100	0				
Trained Models:	Model Evaluation Metrics:		Model C	onfusion Mat	trix:	
weka1   TRAINED_MODEL  Documents: sentiment_sentences  Feature Plugins:  Feature Table: features  Learning Plugin: Weka (All)  Wrapper Plugins:  IIII		Value 0.7507 0.5015	Act \ Pr neg pos	red ne 114 412	49	pos 336 1103
🔆 <u>Report a Bug</u>					0.7 GE	used, 2.7 GB max 📋



LightSide	Is Explore Results Compare Models Predict Lab	ale		
Highlight:	Cell Highlight:	Features in Table:		
weka1 🔻 📙 🗙	Act \ Pred neg pos	Search:		
TRAINED MODEL	neg 1149 336			
Documents: sentiment_sentences_cv.	pos 0 412 1103	Feature	Frequency	Vertical Absolute Differen
🖶 🖥 Feature Plugins:		🔘 serious	1	0.0011
🕀 🛄 Feature Table: features		seriously	1	0.0011
🖶 🔜 Learning Plugin: Weka (All)		() thrills	1	0.0011
Wrapper Plugins:		unfortunately	1	0.0011
🗄 🖄 Trained Model: weka1		value	1	0.0011
		virtually witless	1	0.0011 0.0011
		Iittle	13	0.0011
	Evaluations to Display:	, seems	6	0.0011
	Frequency	us	8	0.0011
	Horizontal Absolute Difference	beyond	3	0.0012
	Horizontal Difference	Classic	3	0.0012
	Vertical Absolute Difference	ending	3	0.0012
	Vertical Difference	except	3	0.0012
4	Model Analysis	O gags	3	0.0012
	inoder / indigits	half	3	0.0012
+ 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) litterate + -0.4418 * (normalized) little + 1.1512 * (normalized) lively + 0.0912 * (normalized) lives + -0.0151 * (normalized) long				



ghlight:		re Models Predict Labe	ls			_
	Cell Highlight:		Features in Table:			
eka1 🔻 📔 🗙	nde (Trea neg	pos	Search:			
AINED_MODEL	cv. pos • 412	336	Feature	Frequency	Vertical Absolute Difference	
Feature Plugins:			🔘 serious	1	0.0011	
Feature Table: features			seriously	1	0.0011	
🔤 Learning Plugin: Weka (All)			🔘 thrills	1	0.0011	
Wrapper Plugins:			unfortunately	1	0.0011	
🖄 Trained Model: weka1			Value	1	0.0011	
			Virtually	1	0.0011	
			witless     little	1	0.0011	
	Evaluations to Display:		seems	6	0.0011	
	Frequency		US	8	0.0011	
	Horizontal Absolute Dif		o beyond	3	0.0012	
	Horizontal Difference		Classic	3	0.0012	
	Vertical Absolute Diffe	rence E	ending	3	0.0012	
	Vertical Difference		except	3	0.0012	
		-	🔘 gags	3	0.0012	
			half	3	0.0012	
ploration Plugin: Documents Display	Model Analysis					
		Instan	ce 189 (Predicted neg	g, Actual pos)	10.004D	
ploration Plugin: Documents Display				g, Actual pos)		[
Ploration Plugin:       Documents Display         Image: Plane pla		Highlight	ce 189 (Predicted neg	g, Actual pos)		[
Oloration Plugin: Documents Display  Filter documents by selected featur Reverse document filter  Documents from selected cell only Instance Predicted	e Actual Text	Highlight	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		[
Oloration Plugin: Documents Display  Filter documents by selected featur Reverse document filter  Documents from selected cell only Instance Predicted I 124 neg	e Actual Text pos we know the	e plot	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		[
Ioration Plugin:       Documents Display         Image: Provide the state of the s	e Actual Text pos we know th pos pretty good	e plot	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		[
Ioration Plugin:       Documents Display         Image: Predicted description       Predicted         Instance       Predicted         I24       neg         I29       neg	e Actual Text pos we know th pos pretty good pos shyamalan	e plot offers	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		
Plogin: Documents Display     Filter documents by selected featur     Reverse document filter     Documents from selected cell only     Instance Predicted     124 neg     189 neg     229 neg     563 neg	Actual Text pos we know th pos pretty good pos shyamalan o pos daughter fro	e plot offers	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		
Ploration Plugin: Documents Display     Filter documents by selected feature     Reverse document filter     Documents from selected cell only     Instance Predicted     124 neg     129 neg     229 neg     563 neg     697 neg	e Actual Text pos we know th pos pretty good pos shyamalan pos daughter fro pos there are tir	e plot inttle offers mes	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		
Ioration Plugin:       Documents Display         Image: Provide the second s	e Actual Text pos we know th pos pretty good pos shyamalan o pos daughter fro pos there are tir pos reign of fire	e plot Iittle om da mes may	ce 189 (Predicted neg ing little feature hits	g, Actual pos)		
Ioration Plugin:       Documents Display         ✓       Filter documents by selected feature         Reverse document filter         ✓       Documents from selected cell only         Instance       Predicted         ✓       124         189       neg         ✓       229         189       neg         ✓       1690         1690       neg	e Actual Text pos we know th pos pretty good pos shyamalan of pos daughter fro pos there are tin pos reign of fire pos it's excessiv	e plot Iittle offers mes rely q	ce 189 (Predicted neg ing little feature hits ood little movie .			
Image: system of the system	e Actual Text pos we know th pos pretty good pos shyamalan pos daughter fro pos there are tir pos reign of fire pos it's excessiv	e plot ittile e may rely q ittile f ittile f ittile f	ce 189 (Predicted neg ing little feature hits ood little movie .			
All Andrew Article Andrew Articl	e Actual Text pos we know th pos pretty good pos shyamalan pos daughter fro pos there are tir pos reign of fire pos it's excessiv pos an original pos there's very pos grown-up q	e plot ittle rmay rely q ittle fi uibble Highlight Pretty ga Pretty ga Instan Highlight	ce 189 (Predicted neg ing little feature hits ood little movie .			
✓       Filter documents by selected feature         ✓       Filter documents by selected feature         ✓       Reverse document filter         ✓       Documents from selected cell only         Instance       Predicted         ✓       124         neg       229         ✓       1630         1414       neg         1590       neg         1700       neg         2629       neg         2733       neg         2809       neg	e Actual Text pos we know the pos pretty good pos shyamalan pos daughter fro pos there are tir pos reign of fire pos it's excessiv pos an original pos there's very pos grown-up q pos a sensitive a	e plot Inttle rmay rittle fin rittle fin	ce 189 (Predicted neg ing little feature hits ood little movie . ce 229 (Predicted neg ing little feature hits	g, Actual pos)	that beneath the familiar , funny surfa	
✓       Filter documents by selected feature         ✓       Filter documents by selected feature         ✓       Reverse document filter         ✓       Documents from selected cell only         Instance       Predicted         ✓       124         ✓       189         ✓       563         ✓       97         1414       neg         1690       neg         1700       neg         2629       neg         2809       neg         2809       neg         2815       neg	e Actual Text pos we know th pos pretty good pos shyamalan of pos daughter fro pos there are tin pos reign of fire pos it's excessiv pos an original 1 pos there's very pos a sensitive a pos a sensitive a pos a fifectionate	e plot offers orm da mes e may little fi little fi little fi little s little s li	ce 189 (Predicted neg ing little feature hits ood little movie . ce 229 (Predicted neg ing little feature hits lan offers copious hints along t	g, Actual pos) the way myriad signs , if you will	I that beneath the familiar , funny surfa en men come to earth for harvesting pur	
<ul> <li>✓ Filter documents by selected featur</li> <li>✓ Filter documents by selected featur</li> <li>✓ Reverse document filter</li> <li>✓ Documents from selected cell only</li> <li>Instance</li> <li>✓ Predicted</li> <li>✓ 124 neg</li> <li>✓ 189 neg</li> <li>✓ 229 neg</li> <li>✓ 563 neg</li> <li>697 neg</li> <li>1414 neg</li> <li>1690 neg</li> <li>1700 neg</li> <li>1700 neg</li> <li>2629 neg</li> <li>2733 neg</li> <li>2809 neg</li> </ul>	e Actual Text pos we know the pos pretty good pos shyamalan pos daughter fro pos there are tir pos reign of fire pos it's excessiv pos an original pos there's very pos grown-up q pos a sensitive a	e plot offers orm da mes e may little fi little fi little fi little s little s li	ce 189 (Predicted neg ing little feature hits ood little movie . ce 229 (Predicted neg ing little feature hits lan offers copious hints along t	g, Actual pos) the way myriad signs , if you will		

ightSide		_		_		Ph. 4		
act Features	Restructure Data E	Build Models Expl	lore Results Compare Mode	s Predict Labe	els			
hlight:	[	Cell H	Highlight:		Features in Table:			
eka1	▼ 📄	X Act V		005	Search:			
AINED_MODEL		nces cv. pos	0 1149 0 412	) 336 ) 1103	Feature	Frequency	Vertical Absolute Difference	
- Feature P	ts: sentiment_senter lugins:	ices_cv. pos	112	1105	( ) sandler	0	0	
	able: features				sandra	0	0	
	Plugin: Weka (All)				Santa Santa	0	0	
🐻 Wrapper I					🔘 sara	0	0	
🖄 Trained M	1odel: weka1				Satisfy	0	0	
					Saturday	0	0	
					Save	0	0	
		Evalu	ations to Display:		saves	0	0	
					saving	0	0	
			requency	*	🔘 saying	0	0	
		H	Iorizontal Absolute Difference		Scares	0	0	
		Н	lorizontal Difference		scattered	0	0	
		V	ertical Absolute Difference	=	scenery	0	0	
			ertical Difference		Schaeffer	0	0	
						0	0	
		Mode	Analysis	-	Scherfig			
	III	olay 🔻	el Analysis	-	schmidt	0	0	
oloration Plugi	n: Documents Disp	olay 🔻	el Analysis		nce 453 (Predicted pos	0		
Filter docu	in: Documents Disp	eature	el Analysis	Highlight	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		
✓ Filter docu ✓ Filter docu	n: Documents Disp ments by selected fe	eature		Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		
Ioration Plugi  Filter docu Reverse do Documents Instance 20	n: Documents Disp ments by selected fe ocument filter s from selected cell o	eature		Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		
Ioration Plugi  Filter docu Reverse do Documents Instance 20 20 28	n: Documents Disp iments by selected fe ocument filter s from selected cell o Predicted	eature Actual	Text	Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		
Ioration Plugi Filter docu Reverse do Documents Instance 20 20 28 20 60	n: Documents Disp ments by selected fe ocument filter s from selected cell o Predicted pos	eature Actual pos	Text who is the audie	Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		
Ioration Plugi         Ioration Plugi         Filter docu         Reverse do         Documents         Instance         22         28         20         28         200         200         200         200         200         200         200         200         200         200         200         200         200         200	n: Documents Disp ments by selected for ocument filter s from selected cell o Predicted pos pos	eature Actual pos pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b	Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		_
Ioration Plugi         Ioration Plugi         Filter docu         Reverse do         Documents         Instance         20         28         60         209         343	n: Documents Disp iments by selected fe ocument filter s from selected cell of Predicted pos pos pos pos pos pos	eature Pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arrest	Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		_
<ul> <li>Filter docu</li> <li>Filter docu</li> <li>Reverse do</li> <li>Documents</li> <li>Instance</li> <li>20</li> <li>28</li> <li>60</li> <li>209</li> <li>343</li> <li>453</li> </ul>	n: Documents Disp ments by selected fe ocument filter s from selected cell of Predicted pos pos pos pos pos pos pos	eature only Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress	Highlight amari h	ice 453 (Predicted pos ting little feature hits has dressed up this little parable	, Actual pos)		_
Ioration Plugi           Ioration Plugi           Ioration Plugi           Reverse do           Documents           Instance           20           28           60           209           343           453	n: Documents Disp ments by selected fe ocument filter s from selected cell o Predicted pos pos pos pos pos pos pos pos pos pos	Actual pos pos pos pos pos pos pos pos pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arrest amari has dress it's a great deal	Highlight amari h erforma	ice 453 (Predicted pos ting little feature hits has dressed up this little parable ances .	o , Actual pos) e in a fairly irresistible package ful		_
Ioration Plugi           ✓           Filter docu           Reverse do           ✓           Documents           Instance           ✓           20           28           60           ✓	n: Documents Disp ments by selected fe ocument filter s from selected cell of pos pos pos pos pos pos pos pos pos pos	eature Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress it's a great deal witty and often s	Highlight amari h erforma	ice 453 (Predicted pos ting little feature hits has dressed up this little parable ances .	o , Actual pos) e in a fairly irresistible package ful		-
Ioration Plugi           Ioration Plugi           Instance           2           Documents           Instance           20           28           209           343           453           536           712	n: Documents Disp ments by selected fe ocument filter s from selected cell of pos pos pos pos pos pos pos pos pos pos	eature only Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress it's a great deal witty and often s disney has alwa	Highlighi amari h erforma	nce 453 (Predicted pos ting little feature hits has dressed up this little parable ances .	o , Actual pos) e in a fairly irresistible package ful		_
✓         Filter docu           ✓         Filter docu           Reverse do         ✓           ✓         Documents           Instance         ✓           ✓         20           ✓         28           ✓         60           ✓         209           ✓         2343           ✓         453           ✓         521           536         712           733	n: Documents Disp ments by selected fe ocument filter s from selected cell of pos pos pos pos pos pos pos pos pos pos	eature only Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , 'b a weird , arresti amari has dress it's a great deal witty and often s disney has alwa director juan jos	Highlight amari h erforma Instan Highlight	nce 453 (Predicted pos ting little feature hits has dressed up this little parable ances .	, Actual pos) e in a fairly irresistible package ful	I of privileged moments and memorable p	
<ul> <li>✓ Filter docu</li> <li>✓ Filter docu</li> <li>✓ Reverse do</li> <li>✓ Documents</li> <li>✓ Instance</li> <li>✓ 20</li> <li>✓ 28</li> <li>✓ 60</li> <li>✓ 209</li> <li>✓ 343</li> <li>✓ 453</li> <li>✓ 453</li> <li>✓ 521</li> <li>536</li> <li>712</li> <li>733</li> <li>1003</li> </ul>	n: Documents Disp	eature only Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress it's a great deal witty and often s disney has alwa director juan jos the tasteful little	Highlight amari h erforma Instan Highlight	nce 453 (Predicted pos ting little feature hits has dressed up this little parable ances .	, Actual pos) e in a fairly irresistible package ful		
<ul> <li>✓ Filter docu</li> <li>✓ Filter docu</li> <li>✓ Reverse do</li> <li>✓ Documents</li> <li>✓ Instance</li> <li>✓ 20</li> <li>✓ 28</li> <li>✓ 60</li> <li>✓ 209</li> <li>✓ 343</li> <li>✓ 453</li> <li>✓ 521</li> <li>536</li> <li>७ 712</li> <li>७ 733</li> <li>1003</li> <li>1307</li> </ul>	n: Documents Disp ments by selected fe ocument filter s from selected cell of pos pos pos pos pos pos pos pos	eature only Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress it's a great deal witty and often s disney has alwa director juan jos the tasteful little there is a refres	Highlight amari h erforma Instan Highlight	ting little feature hits ance 453 (Predicted pos ting little feature hits as dressed up this little parable ances .	, Actual pos) e in a fairly irresistible package ful	I of privileged moments and memorable p	
✓         Filter docu           ✓         Filter docu           ✓         Reverse do           ✓         Documents           ✓         Documents           ✓         20           ✓         28           ✓         60           ✓         209           ✓         343           ✓         521           ✓         536           ✓         712           ✓         733           1003         1307           1556         1556	n: Documents Disp ments by selected fe ocument filter s from selected cell of pos pos pos pos pos pos pos pos	Actual pos pos pos pos pos pos pos pos pos pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress it's a great deal witty and often s disney has alwa director juan jos the tasteful little there is a refres the stunning , dr	Highlight amari h erforma Instan Highlight it's a gr	ting little feature hits ance 453 (Predicted pos ting little feature hits as dressed up this little parable ances .	, Actual pos) e in a fairly irresistible package ful	I of privileged moments and memorable p	
✓         Filter docu           ✓         Filter docu           Reverse do         ✓           ✓         Documents           ✓         20           ✓         28           ✓         60           ✓         299           ✓         453           ✓         521           536         712           733         1003           1307	n: Documents Disp ments by selected fe ocument filter s from selected cell of pos pos pos pos pos pos pos pos	eature only Actual pos	Text who is the audie first-time writer a rather brilliant 'possession , ' b a weird , arresti amari has dress it's a great deal witty and often s disney has alwa director juan jos the tasteful little there is a refres	Highlight amari h erforma Instan Highlight it's a gr	ting little feature hits ance 453 (Predicted pos ting little feature hits as dressed up this little parable ances .	, Actual pos) e in a fairly irresistible package ful	I of privileged moments and memorable p	

#### \* Testing bigrams as an alternative....

Ednat Fabries Restructure Data Build Models Explore Results   CSV Fles: Feature Extractor Flagins:   Windgrame Unigrame   Retire Extractor Flagins:   Character Hortons   Character Hortons <th>SV Files:       Feature Edractor Plugins:       V Ingrams         Serdiment_sentences_cv.cv.       V       Base Features         DOCUMENT_LIST       Same Features       V Bigrams         Closs:       Closs:       Beguint Expressions       V Brave Features         Closs:       Closs:       Stretchy Paterns       V Brave Features         Type:       NOMDAL       Inc. Length       Contact Plugins:       V Word/POS Pairs         V text       Contact Plugins:       Contact Plugins:       Contact Plugins:       Contact Plugins:         V text       Inc. Length       Contact Plugins:       Contact Plugins:       Contact Plugins:       Ene Length         Differentiate Text Fields       Count Occurrences       V Induée Punchadion       Ene M-Grams         Differentiate Text Fields       V       Track Feature Hit Location       Inc.         Feature Table:       V       Features:       Feature In Table:       Inc.         Feature Table:       V       Feature Plugins:       Feature In Table:       Inc.         Feature Table:       V       Search       Feature       Inc.       Inc.         Feature Table:       V       Search       Feature       Inc.       Inc.         Feature Table:       V</th> <th>③ LightSide</th> <th></th> <th></th> <th>٢</th>	SV Files:       Feature Edractor Plugins:       V Ingrams         Serdiment_sentences_cv.cv.       V       Base Features         DOCUMENT_LIST       Same Features       V Bigrams         Closs:       Closs:       Beguint Expressions       V Brave Features         Closs:       Closs:       Stretchy Paterns       V Brave Features         Type:       NOMDAL       Inc. Length       Contact Plugins:       V Word/POS Pairs         V text       Contact Plugins:       Contact Plugins:       Contact Plugins:       Contact Plugins:         V text       Inc. Length       Contact Plugins:       Contact Plugins:       Contact Plugins:       Ene Length         Differentiate Text Fields       Count Occurrences       V Induée Punchadion       Ene M-Grams         Differentiate Text Fields       V       Track Feature Hit Location       Inc.         Feature Table:       V       Features:       Feature In Table:       Inc.         Feature Table:       V       Feature Plugins:       Feature In Table:       Inc.         Feature Table:       V       Search       Feature       Inc.       Inc.         Feature Table:       V       Search       Feature       Inc.       Inc.         Feature Table:       V	③ LightSide			٢
CSV Files:	SV Files:       Feature Ediractor Plugins:       V Intgrams         gentment_sentences_ox.cov       V         DOCUMENT_LIST       Searc Features         Documents: sentiment_sentences_ox       Page Features         Regular Expressions       Stretchy Patienns         Closs:       dass         Type:       NomINAL         Type:       NomINAL         Type:       NomINAL         Total Halds:       Contains Non-Stopwords         V text       Outrothy Patienns         Differentiate Text Fields       Count Courrences         V text       Include Functionion         Feature Table:       Features         Feature Table:       Features         Feature Table:       Features         Pociaments: sentiment_sentences, on the plane;       Feature Table:         Feature Table:	Extract Features Restructure Data Build Mode	Is Explore Results Compare Models Predict Labe	Is	
Differentiate Text Fields     Image: Track Feature Bit Location     Image: Track Feature Bit Loc	□ Differentiate Text Fields       Image: Track Feature Hit Location         Image: Track Features       Rare Threshold: 5         Image: Feature Table:       Image: Target: Ta	sentiment_sentences_cv.csv   DOCUMENT_LIST   Documents: sentiment_sentences_cv.   Image: Class:   Class:   class:   class:   Type:   NOMINAL   Text Fields:	Basic Features     Character N-Grams     Column Features     Parse Features     Regular Expressions	<ul> <li>Øigrams</li> <li>Trigrams</li> <li>POS Bigrams</li> <li>Word/POS Pairs</li> <li>Line Length</li> <li>Contains Non-Stopwords</li> <li>Count Occurences</li> <li>Include Punctuation</li> <li>Remove Stopwords</li> </ul>	• III
features2     FEATURE_TABLE   © Correlation   P Score   Basic Table Statistics   Correlation   P Score   Kappa   Precision   Recall   Target Hits   Total Hits     Indeed   Indeed   Indeed   Indeed   Indeed   Indeed   Indeed   Indeed   Indeed   Indie   Indie   Indie   Indian   Indie   Indian   Indie   Indian   Indie   Indie   Indie   Indian   Indie   Indie   Indie   Indian   Indie   Indian   Indie   Indian   Indie	features2     FEATURE_TABLE   Documents: sentiment_sentences_ov.   Feature Plugins:   Feature Table: features2     Target: neg     Basic Table Statistics   Correlation   Feature Table: features2     Target Hits   Total Hits     Search:     Search:     Feature   indeed_   indeed_   indeed_   indeed_   indian   indian   indial   individual   individual   individual   industry   inevitable		are Threshold: 5	Track Feature Hit Location	•
🖗 Report a Bug 0.4 GB used, 2.7 GB max	infomercial infuses ingenious ingredients inhabit inner	features2     Image: Constraint of the sentences	Target:       neg         Basic Table Statistics         Correlation         F-Score         Kappa         Precision         Recall         Target Hits	Search: Feature indeed indeed_ <comma> independent indian india individual individual individual individual industry inept inept inepttable infectious infomercial infuses ingenious</comma>	

LightSide							- 0 X
Extract Features Restructure Data Build Mode	Is Explore Results Compa	re Models Predict Labels					
Feature Tables:       Image: Constraint of the sector of the	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machi Decision Trees Weka (All)	nes		Configure Weka ( Choose SMD -C 1		N 0 -V -1 -W 1 -K "weka.clas	sifiers.functions.s
< <u> </u>	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random By Annotation: By File Number of Folds: Auto Manual: 10 2 5 10	▼ Max				
Name: weka4	Feature Selection #:	1000					
Trained Models:	Model Evaluation Metrics:		Model C	onfusion Matrix:			
weka3 🗸 📙 🗙	Metric	Value	Act \ P	ed	neg	pos	
TRAINED MODEL	Accuracy	0.7219 0.4436	neg		2898	948	
Documents: sentiment_sentences_cv.     Feature Plugins:     Feature Table: features2     Learning Plugin: Weka (All)     Wrapper Plugins:     Trained Model: weka3	Kappa		pos		1183	2633	
🖗 <u>Report a Bug</u>						0.4 G	3 used, 2.7 GB max   [[]

0							
LightSide						-	
Extract Features Restructure D	ata Build Models Explore R	esults Compare Models Predict L	abels.				
Baseline Model:				Competing Model:			
weka		▼	×	weka3			▼ □ ×
TRAINED_MODEL Documents: sentiment_ Feature Plugins: Feature Table: features Learning Plugin: Weka Wrapper Plugins: Trained Model: weka	_			TRAINED_MODEL Documents: s Feature Plugi Feature Table Learning Plug Wrapper Plug Trained Mode	e: features2 jin: Weka (All) jins:		
Comparison Plugin: Basic Mo	del Comparison						▼]
Baseline Model Metrics:				Competing Model Me	etrics:		
Metric	Value			Metric		alue	
Accuracy Kappa	0.734			Accuracy Kappa		7219 4436	
Baseline Confusion Matrix:				Competing Confusion	n Matrix:		
Act \ Pred	neg	pos		Act \ Pred	neg	pos	
neg	2905	941		neg	2898	948	
pos	1095	2721		pos	1183	2633	
Highly significant improvement	(p=0.008**, t=2.635)						
🔆 <u>Report a Bug</u>							0.4 GB used, 2.7 GB max 📋

LightSide			
Extract Features Restructure Data Build Mod	dels Explore Results Compare Models Predict Labels		
Highlight:	Cell Highlight:	atures in Table:	
weka1 🔹 📔 🔀	Act \ Pred neg pos S	Search:	
TRAINED_MODEL	neg 1149 336 pos 0 412 1103 Fe	eature Frequency	Horizontal Absolute Differen
<ul> <li>Documents: sentiment_sentences_cv.</li> <li>Feature Plugins:</li> </ul>		)'s 101	0.0568
Feature Table: features		) with 39	0.0568
🖶 🔜 Learning Plugin: Weka (All)		<comma> 233</comma>	0.0591
🗄 🐻 Wrapper Plugins:		) like 40	0.0617
🗄 🖄 Trained Model: weka1		does 35	0.0632
		) movie 59	0.0734
		) n't 51	0.0839
		) film 38	0.0855
		) but 88	0.0885
		) a 188 ) of 155	0.1076 0.1088
		) and 148	0.2038
		) an 50	0.0001
		) still 10	0.0002
۰ III ۲	Vertical Difference	cast 7	0.0002
• III P	Model Analysis	ever 7	0.0002
+ -0.8779 * (normalized) lawrence + -0.5057 * (normalized) lazy + -1 * (normalized) lazy + -1 * (normalized) leaden + -0.2567 * (normalized) lee + -0.2567 * (normalized) left + 0.6121 * (normalized) legacy + 0.6121 * (normalized) legacy + -0.4566 * (normalized) less + -0.5215 * (normalized) let + 0.4781 * (normalized) life + -1.1239 * (normalized) life + -0.2267 * (normalized) life + -0.5228 * (normalized) lifs + -0.5328 * (normalized) lifs + -1 * (normalized) list + -1 * (normalized) list + 0.5816 * (normalized) listess			

Extract Features Restructure Data Build Models Explore Results Compare Models Predict Labels	(a) LightSide	and the second se		and the second	
Highlight:       Cell Highlight:       Features in Table:         Image: Cell Highlight:       Feature Search:       Feature Search:         I	Extract Features, Restructure Data, Ruild Mode	La Explore Results Compare Models	Prodict Labola		
Image: Contrast in the sentences of the sentencesentences of the sentences of the sentences of					
Frequency       Horizontal Absolute Differ       Feature Weight         Image: Sentiment_sentiment_sentences_ov       Image: Sentiment_sentences_ov       Image: Sentiment_sentences_ov         Image: Sentiment_sentences_ov       Image: Sentiment_sentences_ov       Image: Sentiment_sentence         Image: Sentiment_sentences_ov       Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence         Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence         Image: Sentiment_sentence       Image: Sentence       Image: Sentiment_sent	Highlight:	Cell Highlight:	Features in Table:		
Frequency       Horizontal Absolute Differ       Feature Weight         Image: Sentiment_sentiment_sentences_ov       Image: Sentiment_sentences_ov       Image: Sentiment_sentences_ov         Image: Sentiment_sentences_ov       Image: Sentiment_sentences_ov       Image: Sentiment_sentence         Image: Sentiment_sentences_ov       Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence         Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence       Image: Sentiment_sentence         Image: Sentiment_sentence       Image: Sentence       Image: Sentiment_sent		Anth David and and			
Peature Plugins:       pos       379       1136         Peature Plugins:       pos       379       1136         Peature Plugins:       Learning Plugin:       Logistic Regression         Peature Table:       features       0       -0.1445         Image: Plugin:       Learning Plugin:       Logistic Regression       0       -0.4435         Peature Table:       features       0       0       -0.4435         Peature Table:       features       0       0       -0.4355         Peature Table:       features       0       0       -0.4355         Peature Table:       features       0       0       -0.4355         Plugin:       Plugin:       Plugin:       1       0.0005       0.0032         Plugin:       Plugin:       Horizontal Difference       0       0       0.018       0.00474         future       2       0.0018       0.0018       0.00474       1       0.00039       0.2412         gegan       0       0       0.00026       0.3371       0       0       0.3002         gegans       0       0.00026       0.3371       0       0       0.00026       0.1543         genes					
Feature Plugins:       0.022       1.0294         Feature Table: features       0       0       -0.1445         Learning Plugin: Logistic Regression       0       0       -0.4335       0.5206         funnier       2       0.0033       0.5206       0         funnier       2       0.0035       0.6932       0         funnier       1       0.0018       0.0474       0       0       0.139       0         further       1       0.0018       0.0474       0       0       0.139       0         gg       0       0       0       0.393       0.5206       0       0       0.393       0         vertical Difference       0       0       0       0.0009       0.2412       0       0       0       0.302       0       0       0.302       0       0       0.302       0       0       0.302       0       0       0.302       0       0       0.302       0				Frequency Horizontal Absolute Differ	Feature Weight
Feature Table: features       0       0       0.1445         Image: Control of the contrel of the control of the control of the cont	Feature Plugins:				<u> </u>
Image: Learning Plugin: Logistic Regression       0       0.0033       0.5206         Image: Learning Plugin: Logistic Regression       0       0       -0.4535         Image: Learning Plugin: Logistic Regression       Image: Learning Plugin: Logistic Regression       0       0         Image: Learning Plugin: Logistic Regression       Image: Learning Plugin: Logistic Regression       0       0       -0.4535         Image: Learning Plugin: Logistic Regression       Image: Learning Plugin: Logistic Regression       0       0       0         Image: Learning Plugin: Logistic Regression       Image: Learning Plugin: Logistic Regression       0       0       0         Image: Learning Plugin: Learning Plugin: Highlighted Feature Details        Image: Learning Plugin: Learning Pl	🕀 🛄 Feature Table: features			0 0	-1.0294
Image: second				2 0.0053	0.5206
Frequency       Model Confusion Matrix:       Att \ Pred meg pos meg 0       0.0024       0.893       0.893	🗄 🖄 I rained Model: logit				
Frequency       future       2       0.0018       0.8045         Frequency       0       0       0.139       0       0       0.7943         gag       0       0       0.0009       0.2412       0       0       0.3002         gaghan       0       0       0       0.3002       0.3002       0.3002       0.3002         gaghan       0       0       0       0.3002       0.3002       0.3002       0.3002         gages       3       0.0056       0.3971       0       0       0.3002       0.3002         game       2       0.0026       0.3971       0       0       0.0009       0.111         gangs       0       0       0.0009       0.111       0       0       0.6682         Exploration Plugin:       Highlighted Feature Details               Model Confusion Matrix:       Image					
Evaluations to Display:       0       0.139         Horizontal Difference       9       0       0       0.7943         Yetrical Absolute Difference       9       0       0       0.7943         Yetrical Absolute Difference       9       0       0       0.7943         Yetrical Absolute Difference       9       0       0       0.3002         Yetrical Absolute Difference       9       3       0.0062       0.7382         Model Analysis       9       3       0.0002       0.7382         Yetrical Difference       9       9       0       0.0002       0.7382         Yetrical Difference       9       9       0       0.0002       0.7382         Yetrical Absolute Difference       9       9       0       0.0009       0.111         Yetrical Absolute Difference       9       0       0.0009       0.6682         Exploration Plugin:       Highlighted Feature Details        Feature Weight       Model Confusion Matrix:       Act \ Pred neg pos					
Horizontal Difference       gag       0       0.0009       0.2412         gaghan       0       0       0.3002         gags       3       0.0062       0.7382         Model Analysis       ame       2       0.0026       0.3971         games       0       0.0009       0.111         gangs       0       0.0009       0.111         gangs       0       0.0009       0.111         gangs       0       0.0009       0.111         gangs       0       0.0009       0.6682		Evaluations to Display:		0 0	
Vertical Absolute Difference       gaghan       0       0       0.0000       0.0002         Vertical Difference       gaghan       0       0       0.0026       0.3971         Vertical Difference       game       2       0.0026       0.3971         games       0       0       0.0009       0.111         gang       1       0.0009       0.111         gangs       0       0.0009       0.0111         gangs       0       0.0009       0.1543         Exploration Plugin:       Highlighted Feature Details           Frequency       Model Confusion Matrix:       Model Confusion Matrix:           Model Confusion Matrix:       Act \ Pred neg pos neg       0       0.0024         neg       12       12       12       0       0.024					
Vertical Difference       gags       3       0.0062       0.7382         Model Analysis       game       2       0.0026       0.3971         games       0       0.0009       -0.1543         gangs       0       0.0009       -0.1543         gangs       0       0.0009       -0.1543         gangs       0       0.0009       -0.1682         Exploration Plugin:       Highlighted Feature Details			999		
Model Analysis       game       2       0.0026       0.3971         Games       0       0.0009       -0.1543         gang       1       0.0009       0.111         gangs       0       0.0009       -0.6682			O gags		
Image: Peature Weight       gang       1       0.0009       0.111         Exploration Plugin:       Highlighted Feature Details       -0.6682         Frequency       Model Confusion Matrix:       Image: Post of the post of			= game	2 0.0026	0.3971
Frequency       Model Confusion Matrix:       Image: pos in reg       0       0.0009       -0.6682         Act \ Pred       neg       pos in reg       0       0.0009       -0.6682			games		
Exploration Plugin: Highlighted Feature Details	4 III •	Feature Weight			
Frequency         Model Confusion Matrix:       Image       Horizontal Absolute Difference       Model Confusion Matrix:       Image       Feature Weight         Act \ Pred       neg       pos       Act \ Pred       neg       pos       Act \ Pred       neg       pos         neg       12       12       0       0.024       Image       pos       Image       pos			2 guilgo		010002
Model Confusion Matrix:       Image: Model Confusion Matrix: <t< th=""><th>Exploration Plugin: Highlighted Feature Detai</th><th>s 🔻</th><th></th><th></th><th></th></t<>	Exploration Plugin: Highlighted Feature Detai	s 🔻			
	Model Confusion Matrix:       Act \ Pred     neg       pos       neg     12	Model Confusion Matrix:	Model Confusion       pos     Act \ Pred       0.024     neg	Matrix:	

② LightSide						
Extract Features Restructure Data Build Mode	Is Explore Results Compare Models Predict Labe	els				
Highlight:	Cell Highlight:	Features in Table:				
TRAINED_MODEL  Documents: sentiment_sentences_cv.  Feature Plugins:	neg 1137 348 pos 379 1136	Feature	Frequency 229	Horizontal Absolute Differ	Feature Weight	
Feature Table: features Learning Plugin: Logistic Regression		a do movie	187 26 51	0.055 0.0554 0.0597	-0.1195 0.3847 0.3268	
	Evaluations to Display:	of with	32 154 31	0.0607 0.0655 0.0731	0.3848 -0.0984 -0.3772	
	Horizontal Difference Vertical Absolute Difference Vertical Difference	n't     film     but     and	47 33 89 135	0.0818 0.0899 0.1142 0.2019	0.5158 -0.3272 0.2248 -0.3715	
	Model Analysis	VIII Vivo screen	5 5 6	0 0 0 0	-0.3673 -0.3729 0.0045	
Exploration Plugin: Highlighted Feature Detai		🔘 sense	6	0	0.0931	-
Frequency         Model Confusion Matrix:         Act \ Pred       neg       pos         neg       112       35         pos       47       48	Image: Second state of the second s		usion Matrix:			
🔆 Report a Bug					0.2 GB used, 2.7	GB max 📋

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

	els Explore Results Compa	re Models Predict Labels				
Feature Tables:	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machi Decision Trees Weka (All)	nes	C L2 Regula			
< <u>III</u> >	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random By Annotation: By File Number of Folds: Auto Manual: 10 2 5 10	▼ Max			
Name: logit6	Feature Selection					[
Trained Models:	Model Evaluation Metrics:		Model Confusion Matrix:			
logit5 👻 📋 🗙	Metric	Value	Act \ Pred	neg	pos	
	Accuracy Kappa	0.7254 0.4508	neg	2820	1026	
TRAINED_MODEL			pos	1078	2738	

LightSid	e				Anapa Tak					X
Extract Feat	ures Restructure Data B	uild Models Explore Res	ults Compare Models Predict Labels	5						
Baseline Mo	odel:				Competing Model:					
logit4			<b></b>	×	logit5				-	×
TRAINED_N	IODEL				TRAINED_MODEL					
Doci	uments: sentiment_senten	ces_cv.csv			Documents: sentime	ent_sentences_cv.csv				
🕂 🗐 Feat	ure Plugins: ure Table: features1				Feature Plugins:	ures				
🗎 🕀 🔜 Lear	nina Pluain: Loaistic Rear	ession			🗄 📾 Learning Plugin: Lo	aistic Rearession				
🗄 🖄 I rai	ned Model: logit4				🗄 🖄 Trained Model: logit	5				
Compariso	n Plugin: Basic Model Co	mparison								•
Baseline I	Model Metrics:				Competing Model Metrics:					
Metric		Value			Metric		Value			
Accuracy		0.731			Accuracy		0.7254			
Карра		0.463	3		Карра		0.4508			
Deseline	Canfunian Matuku				Competing Confusion Matri					
Baseline	Confusion Matrix:	1			Competing Confusion Math	х.				
Act \ Pre	ed	neg	pos		Act \ Pred	neg		pos		
neg		2845 1055	1001 2761		neg pos	2820 1078		1026 2738		
pos		1055	2701		pos	1078		2730		
Highly sign	ificant improvement (p=0.0	006**, t=2.745)								

# Special Text Features

### Stretchy Patterns in LightSIDE Looking at sentiment\_sentences.csv

LightSide		
Extract Features Restructure Data Build Models	Explore Results Compare Models Predict Labels	
Sentiment_sentences.csv  Sentiment_sentences.csv DOCUMENT_LIST Documents: sentiment_sentences.csv	Feature Extractor Plugins:  Basic Features Character N-Grams Column Features Parse Features Regular Expressions Stretchy Patterns	Configure Stretchy Patterns 0 1 2 3 4 5 6 7 8 0 1 2 3 4 5 6 7 8 Gap Length O 1 2 3 4 5 6 7 8 O 1 2 3 4 5 6 7 8 O 1 2 3 4 5 6 7 8
Class: class  Type: NOMINAL  Text Fields:  text		✓ Include surface words in patterns ☐ Include POS tags in patterns Categories: Add Clear ☐ Require at least one category per pattern
Differentiate Text Fields		✓ Don't include surface/POS form when a category matches
Extract         Name:         features         Rare	e Threshold: 5	
	Evaluations to Display: Target:  Basic Table Statistics Correlation F-Score Kappa Precision Recall Target Hits	Features in Table:
Report a Bug	Total Hits	0.0 GB used, 2.7 GB max

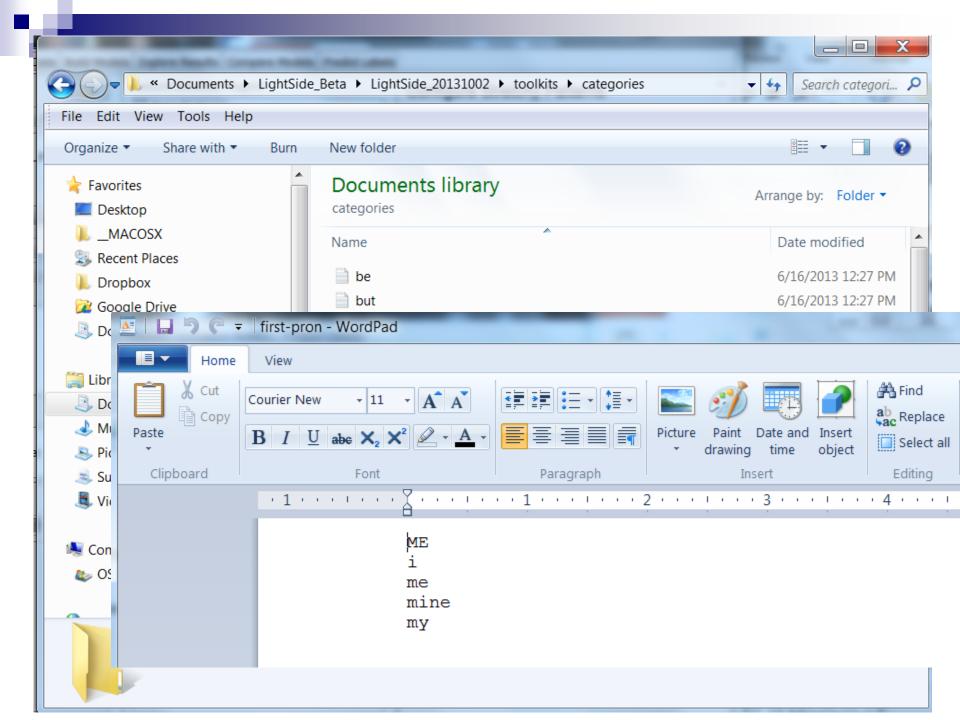
# **Configuring Stretchy Patterns**

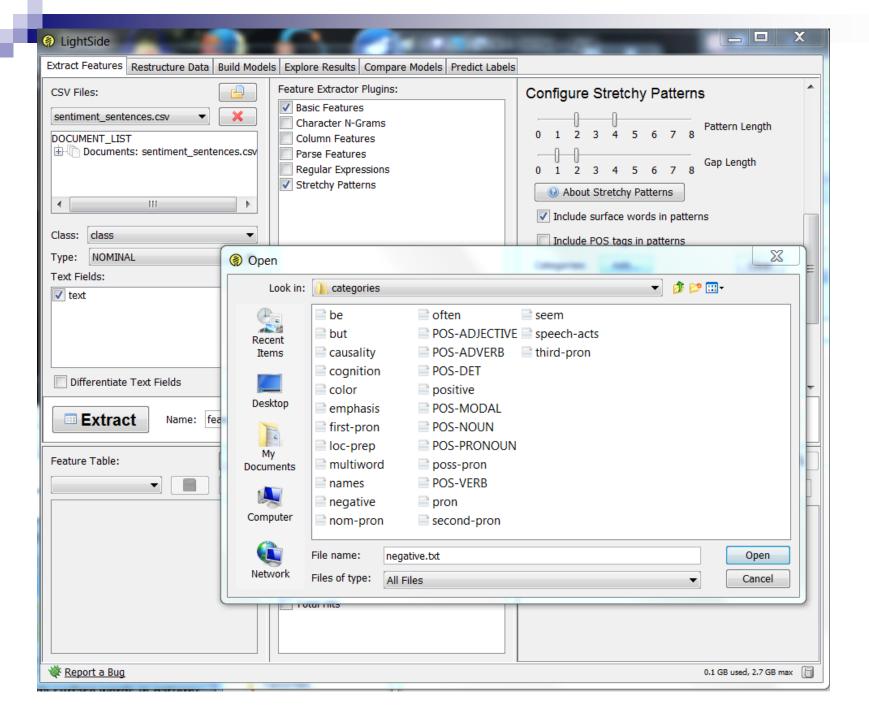
#### **Configure Stretchy Patterns**

0 1 2 3 4 5 6 7 8 Pattern Length	2
0 1 2 3 4 5 6 7 8 Gap Length	3
About Stretchy Patterns	
Include surface words in patterns Include POS tags in patterns	4
Categories: Add	lear
	5
Require at least one category per pattern	5
<ul> <li>Require at least one category per pattern</li> <li>Don't include surface/POS form when a category matches</li> </ul>	5
	<b>5</b>
✓ Don't include surface/POS form when a category matches	<b>5</b>

 Longer patterns and longer gaps lead to larger numbers of features

Categories are useful both for abstraction and for anchoring the patterns





(i) LightSide		
Extract Features Restructure Data Build Model	s Explore Results Compare Models Predict Labels	
CSV Files:	Feature Extractor Plugins:	<ul> <li>Include POS tags in patterns</li> <li>Categories: Add Clear</li> <li>STRONG-NEG: [awful, bad, badly, disgusting, horrible, poorly, terrible, worse, worst]</li> <li>Require at least one category per pattern</li> <li>Don't include surface/POS form when a category matches</li> <li>Categories match against surface words</li> <li>Categories match against POS tags</li> <li>Count pattern hits</li> <li>Prune Rare Features after N documents:</li> </ul>
Differentiate Text Fields		0 100 200 500 1000
Rame: features R	are Threshold: 5	8
Feature Table: <td>Evaluations to Display: Target:  Basic Table Statistics Correlation F-Score Kappa Precision Recall Target Hits Total Hits</td> <td>Features in Table:</td>	Evaluations to Display: Target:  Basic Table Statistics Correlation F-Score Kappa Precision Recall Target Hits Total Hits	Features in Table:
Report a Bug		0.1 GB used, 2.7 GB max

Sentiment_sentences.csv       Image: Construction of the sentences.csv         DOCUMENT_LIST       Image: Construction of the sentences.csv         Image: Construction of the sentences.csv       Image: Construction of the sentences.csv         Image: Construction of the sentences.csv       Image: Construction of the sentences.csv         Image: Construction of the sentences.csv       Image: Construction of the sentences.csv         Image: Construction of the sentences.csv       Image: Construction of the sentences.csv         Image: Construction of the sentences.csv       Image: Construction of the sentences.csv	e Results Compare Models Predict Labels Extractor Plugins: c Features racter N-Grams mn Features e Features ular Expressions tchy 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
Sentiment_sentences.csv  Basic Chara DOCUMENT_LIST Documents: sentiment_sentences.csv Regu Stretu	c Features racter N-Grams mn Features e Features Jlar Expressions	Categories: Add Clear STRONG-NEG: [awful, bad, badly, disgusting, horrible, poorly, terrible, worse, worst]
Class: class  Type: NOMINAL  Text Fields:  text		<ul> <li>Don't include surface/POS form when a category matches</li> <li>Categories match against surface words</li> <li>Categories match against POS tags</li> <li>Count pattern hits</li> <li>Prune Rare Features after N documents:</li> </ul>
Differentiate Text Fields  Extract Name: features1 Rare Thresh	nold: 5	0 100 200 500 1000
features       Image: Control         FEATURE_TABLE       Image: Control         Image: Control       Control	neg  ble Statistics elation ore ba ision all let Hits I Hits	Features in Table:       Image: Constraint of the strong of

LightSide		p-st i				- 🗆 🗙
Extract Features R	estructure Data Build M	odels Explore Results Compar	e Models Predict Labels			
Extract Features Restructure Data Build Models Explore Results Compare Models   Baseline Model:   Baseline Model:     Iogit     Iogit     Indite     Iogit     Indite     Iogit     Indite     Iogit     Indite     Iogit     Indite     Iogit     Indite     Indite						
Comparison Plugin	Basic Model Compari	son				▼]
Baseline Model M	letrics:		Competing Model	Metrics:		
Metric	V	'alue	Metric	Va	alue	
Accuracy	0.	7605	Accuracy	0.7	7609	
Карра		5209	Карра		219	
Baseline Confusio	on Matrix:		Competing Confu	sion Matrix:		
Act \ Pred	neg	pos	Act \ Pred	neg	pos	
neg	4089	1242	neg	4087	1244	
pos	1312	4019	pos	1305	4026	
			Insignificant improv	/ement (p=0.579, t=-0.5	56)	
🔆 <u>Report a Bug</u>					1.0 GB u	sed, 2.7 GB max 🕅

### **Regular Expressions**

- \* allows the previous part of the regex to repeat, 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 o-9), and \s (any type of space character).

Extract Features	Restructure Data Build Models Exp	lore Results Compare Models Predict Labels
CSV Files:	Feature Extractor Plugins:	Configure Regular Expressions
Text Fields:		

### American Street Gangs Predict gang affiliation from posts

#### • Crips, Bloods, Hoovers

- crips started in South Central LA
- Pirus, Bloods, Hoovers from crips
- Chicago based
  - o People Nation
    - vice lords, latin kings, stones
  - o Folk nation
    - gangster disciples

#### Trinitarios

hispanic gang based in NYC



### **Graffiti Based Style Features**



#### Graffiti

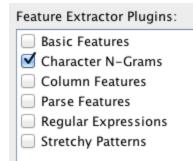
Social messages Stylistic writing crossing out other gangs

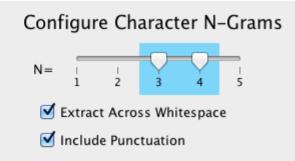


### On the board

- 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!!

Feature Extractor Plugins:	Configure Parse Features
<ul> <li>Basic Features</li> <li>Character N-Grams</li> <li>Column Features</li> <li>Parse Features</li> <li>Regular Expressions</li> <li>Stretchy Patterns</li> </ul>	<ul> <li>2 Production Rules</li> <li>3 Leaf Productions</li> <li>4 Dependency Relations</li> <li>△ Parsing is pretty slow. Go get a coffee or something.</li> </ul>

Leveraging Subpopulations through Multi-Level Modeling

(ightSide)		B				
Extract Features Restructure Data Build Mod	els Explore Results Compa	are Models Predict Labels				
Feature Tables:       Image: Constraint of the seature         features2       Image: Constraint of the seature         FEATURE_TABLE       Image: Constraint of the seature         Image: Feature Plugins:       Image: Feature Table: features2	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Mach Decision Trees Weka (All)	ines			<b>Jure Naive</b> Jse Kernel Esti Jse Supervised	-
	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random Fold Assignment: Fold				
Train Name: bayes1	Feature Selection			•	III	
Trained Models:	Model Evaluation Metrics Metric Accuracy Kappa	: Value 0.7558 0.5109	Act \ Negat Positiv	ive	Matrix: Negative 379 101	Positive 129 333
🔆 <u>Report a Bug</u>					0.2	2 GB used, 2.7 GB max 📋

(ightSide)		B				
Extract Features Restructure Data Build Mod	els Explore Results Compa	are Models Predict Labels				
Feature Tables:       Image: Constraint of the seature         features2       Image: Constraint of the seature         FEATURE_TABLE       Image: Constraint of the seature         Image: Feature Plugins:       Image: Feature Table: features2	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Mach Decision Trees Weka (All)	ines			<b>Jure Naive</b> Jse Kernel Esti Jse Supervised	-
	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random Fold Assignment: Fold				
Train Name: bayes1	Feature Selection			•	III	
Trained Models:	Model Evaluation Metrics Metric Accuracy Kappa	: Value 0.7558 0.5109	Act \ Negat Positiv	ive	Matrix: Negative 379 101	Positive 129 333
🔆 <u>Report a Bug</u>					0.2	2 GB used, 2.7 GB max 📋

③ LightSide		· · · · ·							
Extract Features Restructure Data Build Models Explore Results Compare Models Predict Labels									
Feature Tables:	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machines Decision Trees Weka (All)			Configure Naive Bayes Use Kernel Estimator Use Supervised Discretization					
Train Name: bayes2	<ul> <li>Cross-Validation</li> <li>Supplied Test Set</li> <li>No Evaluation</li> </ul>	Fold Assignment: Random By Annotation: Gender By File Number of Folds: Auto Manual: 2 2 5 10	Max	4	4				
Trained Models:	Model Evaluation Metrics	:	Model	Confusion Matrix:					
bayes1  TRAINED_MODEL  Cocuments: Gallup.csv  Feature Plugins:  Feature Table: features2  Cearning Plugin: Naive Bayes  Trained Model: bayes1	Metric Accuracy Kappa	Value 0.7654 0.5292	Act \ Negat Positiv	ive 389	Positive 119 332				
Image: Weight and Section 2.7 GB max									

DightSide								
Extract Features Restructure Data Build Models Explore Results Compare Models Predict Labels								
Baseline Model:	Competing Model:							
bayes	bayes1 🔹 🔛							
TRAINED_MODEL Documents: Gallup.csv Feature Plugins: Feature Table: features2 Learning Plugin: Naive Bayes Trained Model: bayes	TRAINED_MODEL							
Comparison Plugin: Basic Model Comparison								
Baseline Model Metrics:		Competing Model M	etrics:					
Metric Value		Metric	Value					
Accuracy 0.7558		Accuracy	0.7654					
Kappa 0.5109		Карра	0.5292	2				
Baseline Confusion Matrix:	Competing Confusion Matrix:							
Act \ Pred Negative	Positive	Act \ Pred	Negative	Positive				
Negative 379 Positive 101	129 333	Negative Positive	389 102	119 332				
			nent (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 lyeiri, 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 to overpredict extent of swearing	
<u>Train</u>	<u>Test</u>
FYH MOL MYL MYH MOL MYL MYH MOL FOH MYH MOL MYL MYH MOL MYL MYH MOL MYL MYH MOL MYL FYH MYH MOL FYH FYH FYH MYH FOH FOH FYH FOH FOH FYH FOH FOH FOH FOH FOH MYH FOH FOH FOH FOH	FYL MOH MOH MOH FOL MOH FYL FOL FOL FYL MOL FOL FYL MOH FYL MOH FYL MYH MOH FYL MYH MOH
FYH FOH FOH FOH	MOH

## Evaluation of Domain Generality

Occupation	Unigr	am	Unig Bigra	ram + nm	POS Bigram	Strete Patter	•
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	i (.19)	57.95 (.16)	62.8	(.26)

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

#### Feature Splitting (Daumé III, 2007)

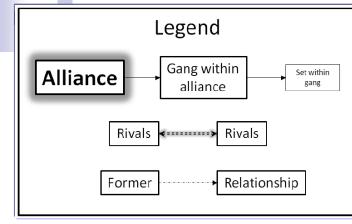




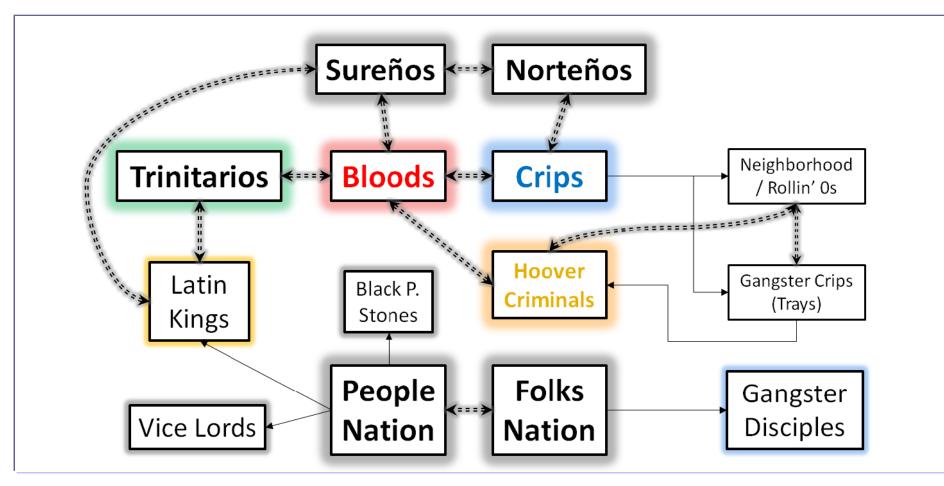
Why is this nonlinear?

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 Mode	els 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 Column Features         Column Name         3e         5s         6b         5c         6b         bk         ccaret         cc         ck         dominant         hk         hk         v hk         v length         v numUsers         v pCaret         v pk         starter         automation
Extract Name: features1 F	Rare Threshold: 5	
Feature Table:         features         FEATURE_TABLE         Documents: ThreadCompStyles.csv         Feature Plugins:         Feature Table: features	Target: allied  Basic Table Statistics Correlation F-Score Kappa Precision Recall Target Hits Total Hits	Features in Table:
🖗 <u>Report a Bug</u>		0.1 GB used, 2.7 GB max

Extract Features Restructure Data Build Models Explore Results Compare Models Predict L	abels
Feature Tables:       Image: Sector Analysis of the sector o	Configure Multilevel Modeling         Select Levels:         Domain         A         B         Column Features         Select Sect         Select Sect         Select Levels:         Select Sect         V         Add Domain         V         Select Levels:
Restructure         Name: Estructure         Rare Threshold:         5	
Restructured Tables:       Evaluations to Display:         Target:       Target:         Basic Table Statistics       Correlation         F-Score       Kappa         Precision       Recall         Target Hits       Total Hits	Features in Table:
🌾 <u>Report a Bug</u>	0.0 GB used, 2.7 GB max 📋

Ŋ8

W LIGHTOIDE		the second se	
Extract Features Restructure Data Build Mode	ls Explore Results Compare Models Predict La	hels	
Extract Features Restructure Data Build Mode Feature Tables: features FEATURE_TABLE Feature Plugins: Feature Table: features Feature Table: features	Is Explore Results Compare Models Predict La 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
Restructure Name: struc	ture1 Rare Threshold: 5		
Restructured Tables:	Evaluations to Display: Target: allied	Features in Table:	
restructure 🔻 📋 🗙	Basic Table Statistics	Search:	
MODIFIED_TABLE Documents: ThreadCompStyles.csv Feature Plugins: Restructure Plugins: Restructure Plugins: Restructure d Table: restructure Restructure	Correlation F-Score Kappa Precision Recall Target Hits Total Hits	Feature dominant::bloods_cccolumn dominant::bloods_ckcolumn dominant::bloods_lengthcolumn dominant::bloods_numUserscolumn dominant::crips (Intercept) dominant::crips_cCaretcolumn dominant::crips_cCaretcolumn dominant::crips_ckcolumn	0.0 GB used, 2.7 GB max
w <u>Report a Bug</u>			0.0 GB used, 2.7 GB max

(\$) LightSide								-	
Extract Features Restructure Data Build Mode	Is Explore Results Compa	re Models Predict Label	s						
Feature Tables:         features         FEATURE_TABLE         Documents: ThreadCompStyles.csv         Feature Plugins:         Feature Table: features	Learning Plugin: Naive Bayes Logistic Regression Support Vector Machin 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	0	L2 Regularizati	on	1	
Name: logit1	Feature Selection								
Trained Models:	Model Evaluation Metrics:			Model (	Confusior	n Matrix:			
Iogit       Image: Complexity of the sector of	Metric Accuracy Kappa	Value 0.4817 0.188		Act \ F allied homog mixed opposit	eneous	allied 20 9 4 12	homogeneous 39 107 20 84	mixed 2 3 3 3 3	opposing 179 54 43 290
W Report a Bug									0.1 GB used, 2.7 GB max  📋
							122		
							122		

Ŋ2

LightSide		_			-			
Extract Features Restructure Data Build Mode	s Explore Results Compare	Models Predict Labels						
Feature Tables:       Image: Constraint of the sector of the	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machine Decision Trees Weka (All) Cross-Validation Supplied Test Set No Evaluation		• •	0	gure Logistic	י ו ו	1	
Name: logit2	Feature Selection							
Trained Models:	Model Evaluation Metrics:		 Model	Confusion	Matrix:			
logit1 👻 📄 🗙	Metric	Value 0.6124	Act \ F	Pred	allied	homogeneous	mixed	opposing
TRAINED_MODEL	Accuracy Kappa	0.4108	allied	eneous	87 21	24 132	3 4	126 16
Documents: ThreadCompStyles.c:			mixed		8	5	21	36
Feature Table: features     Restructure Plugins:     Restructured Table: restructure			opposi	ng	48	40	7	294
🔆 <u>Report a Bug</u>							0.	.1 GB used, 2.7 GB max  📋

xtract Features Restructure Data Build Models Explore Results Compare Models   Baseline Model: Competing Model:   logit Iogit1   TRAINED_MODEL   Courdents: ThreadCompStyles.csv   Feature Plugins:   Feature Table: features   Learning Plugin: Logistic Regression   X	
ogit  RAINED_MODEL  Documents: ThreadCompStyles.csv  Feature Plugins:  Feature Table: features  Earning Plugin: Logistic Regression  K Trained Model: logit  C Trained Model: logit	
RAINED_MODEL  TRAINED_MODEL  Documents: ThreadCompStyles.csv  Feature Plugins: Feature Table: features  Feature Table: features  Feature Table: features  Feature Table: legit  Feature Table: features  Feature Table: features  Feature Table: features  Feature Table: legit  Feature Table: features  Feature Table: legit  Feature Table: features  Feature Table: features  Feature Table: features  Feature Table: legit  Feature Table: features  Feat	
Image: Documents: ThreadCompStyles.csv	
H K Trained Model: logit1	
omparison Plugin: Basic Model Comparison	
Baseline Model Metrics: Competing Model Metrics:	
Metric Value Metric Value	
Accuracy 0.4817 Accuracy 0.6124	
Kappa 0.188 Kappa 0.4108	
Baseline Confusion Matrix:	
Act \ Pred allied homogeneous mixed opposing Act \ Pred allied homogeneous mixed	d opposing
allied 20 39 2 179 allied 87 24 3	126
homogeneous 9 107 3 54 homogeneous 21 132 4	16
mixed 4 20 3 43 mixed 8 5 21	36
opposing 12 84 3 290 opposing 48 40 7	294

opposing				
mixed		_		
homogeneous				
allied				
⊿ Frequencies				-
Level		Count	F	Prob
allied		240	0.27	523
homogeneous	5	173	0.19	839
mixed			0.07	
opposing			0.45	
Total		872	1.00	000
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.

opposing				
mixed		_		
homogeneous				
allied				
⊿ Frequencies				-
Level		Count	F	Prob
allied		240	0.27	523
homogeneous	s	173	0.19	839
mixed			0.07	
opposing			0.45	
Total		872	1.00	000
N Missing	0			

## **Feature Analysis**

 Style features that distinguish Allied from Opposing differ by dominant gang

Crips:

- □ Allied: bCaret
- □ Opposing: CC, PK, cCaret gang is in an

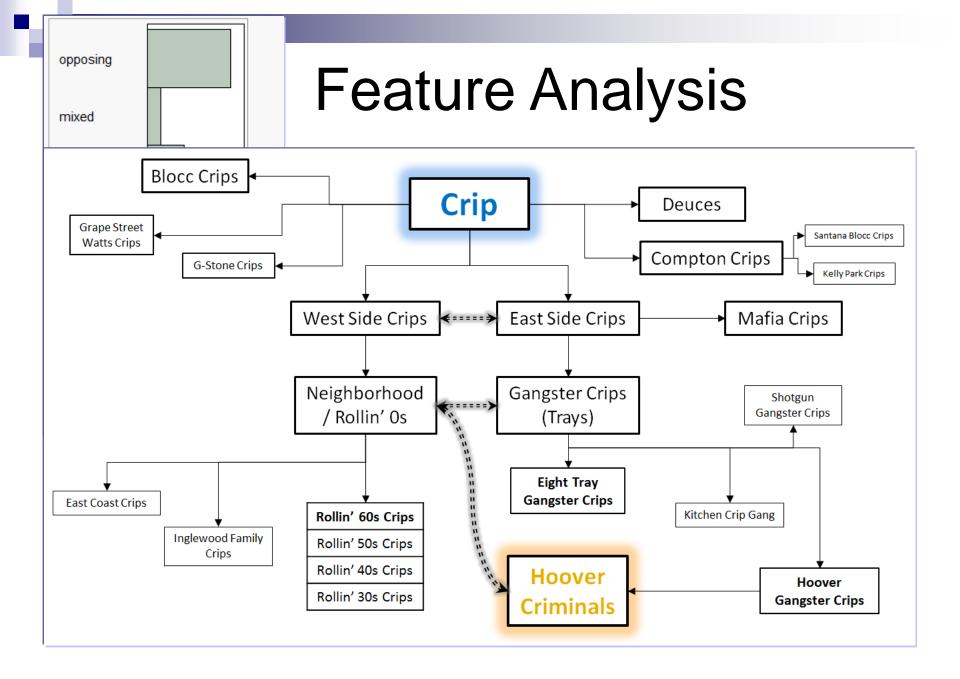
#### Bloods:

- □ 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.



opposing				
mixed				
homogeneous				
allied				
⊿ Frequencies				
Level		Count		Prob
allied		240	0.27	523
homogeneous	;	173	0.19	839
mixed			0.07	
opposing			0.45	
Total N Missing	0	872	1.00	000

## Feature Analysis

 Unigram features that distinguish Allied from Opposing don't differ by dominant gang as much as style features

#### Universal:

- □ Allied: Imao, you, crew
- □ Opposing: forever, wtf, where w

#### Crips:

- $\Box$  Allied: Iol
- Opposing: know, about

#### Bloods:

- □ Allied: niggas, the
- Opposing: at

We see relationship words, but not gang identity words.