01

Project Scope

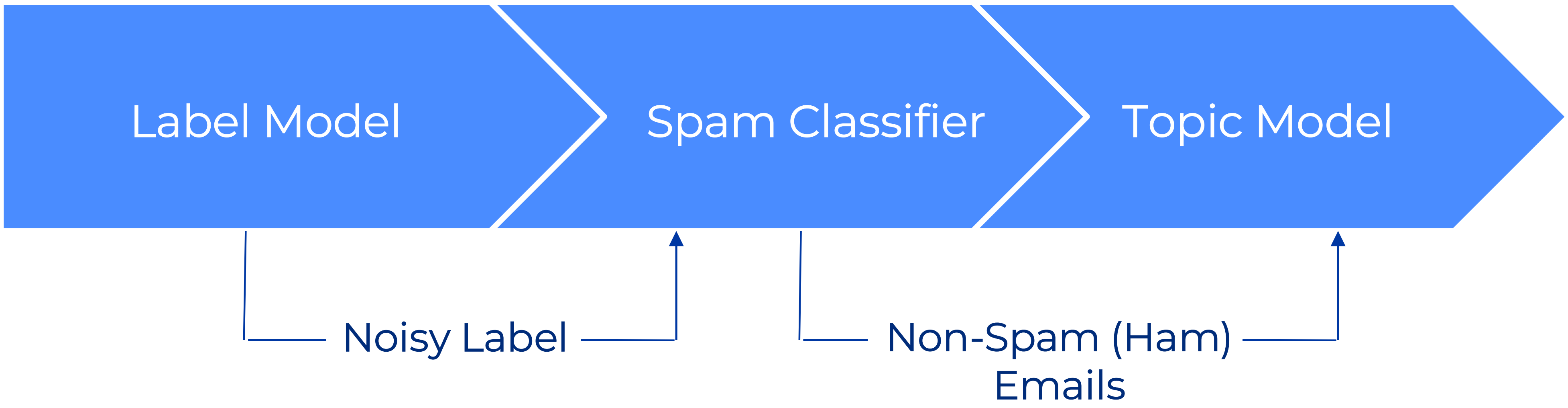
Benefits of the Project

Saving Labeling Cost
(reduce ~80% of
labeled data needed)

Broad Use Cases
(support development of
other ML projects)

Improve Model
Maintenance Process

Model Pipeline



- Label the large dataset with small amount of hand-labeled data
- Classify unwanted spam emails
- Extract the topics of remaining emails for future use

02
Model Development

Spam Classifier - Planning

Dataset

Enron Emails Dataset – Jeff Dasovich's Email

26,371 emails

63 folders

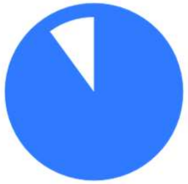
From 1999 to 2002

	Date	From	To	Subject	body_msg	X-Folder_Category
65772	2000-12-13 08:51:00-08:00	rochmanm@spurr.org	tomb@crossborderenergy.com, michael.alexander...	RE: Change in Wood All-party Meeting\nCc: ask...	\n\n?That is extremely bush league, to make a ...	Notes inbox
59957	2002-01-02 16:35:52-08:00	paul.kaufman@enron.com	jeff.dasovich@enron.com	RE:	\n\nIf you pursue my idea--don't refer to the ...	Deleted Items

Goal of the Spam Classifier

Precision

90%+

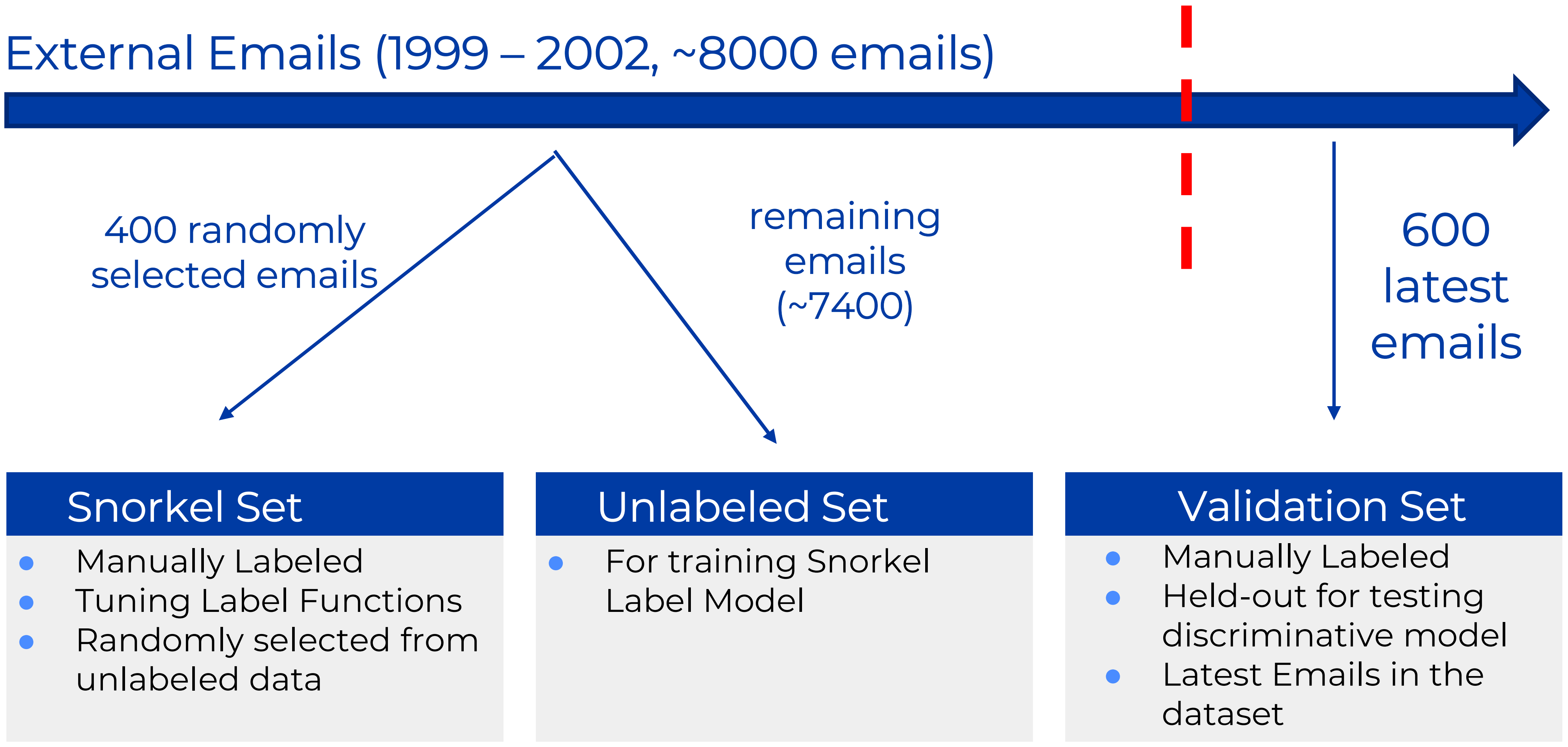


Coverage

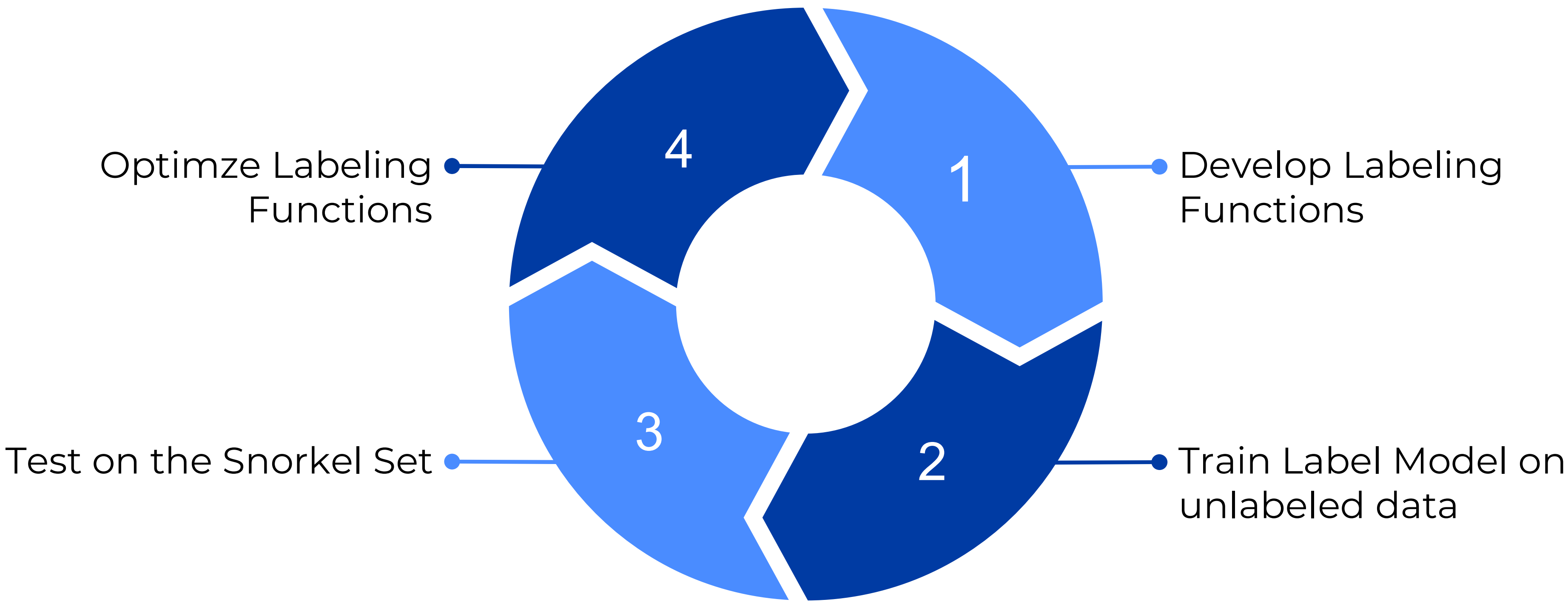
02
Model Development

Labeling Model - Snorkel

Snorkel


External Emails (1999 – 2002, ~8000 emails)
400 randomly selected emails
remaining emails (~7400)
600 latest emails
Snorkel Set
Manually Labeled
Tuning Label Functions
Randomly selected from unlabeled data
Unlabeled Set
For training Snorkel Label Model
Validation Set
Manually Labeled
Held-out for testing discriminative model
Latest Emails in the dataset

Developing Label Model



Labeling Functions

	j	Polarity	Coverage	Overlaps	Conflicts	Correct	Incorrect	Emp. Acc.
newswire	0	[1]	0.0175	0.0125	0.0075	7	0	1.000000
career	1	[1]	0.0425	0.0425	0.0100	15	2	0.882353
haas_promotion_address	2	[1]	0.0700	0.0500	0.0075	28	0	1.000000
count_to_enron	15	[0]	0.2875	0.1625	0.0100	113	2	0.982609
reply	16	[0]	0.3325	0.2600	0.0375	119	14	0.894737
jeff_count	17	[0]	0.2025	0.1950	0.0225	77	4	0.950617

Label Model - Performance on Snorkel Set

```
validation value count:  0    242
 1    105
-1     53
Name: label, dtype: int64
{'accuracy': 0.9394812680115274, 'precision': 0.8857142857142857, 'recall': 0.9117647058823529, 'f1':
0.8985507246376812}
```

Precision	**88.57%**
Recall	91.18%
% of Data Labeled	86.75%

Label Model - Performance on Validation Set

```
validation value count:  1    269
 0    242
-1     89
Name: label, dtype: int64
{'accuracy': 0.9256360078277887, 'precision': 0.9479553903345725,
'recall': 0.9139784946236559, 'f1': 0.9306569343065693}
```

Precision	**94.80%**
Recall	91.40%
% of Data Labeled	85.17%

Better than Snorkel Set

Label Model - Model Output

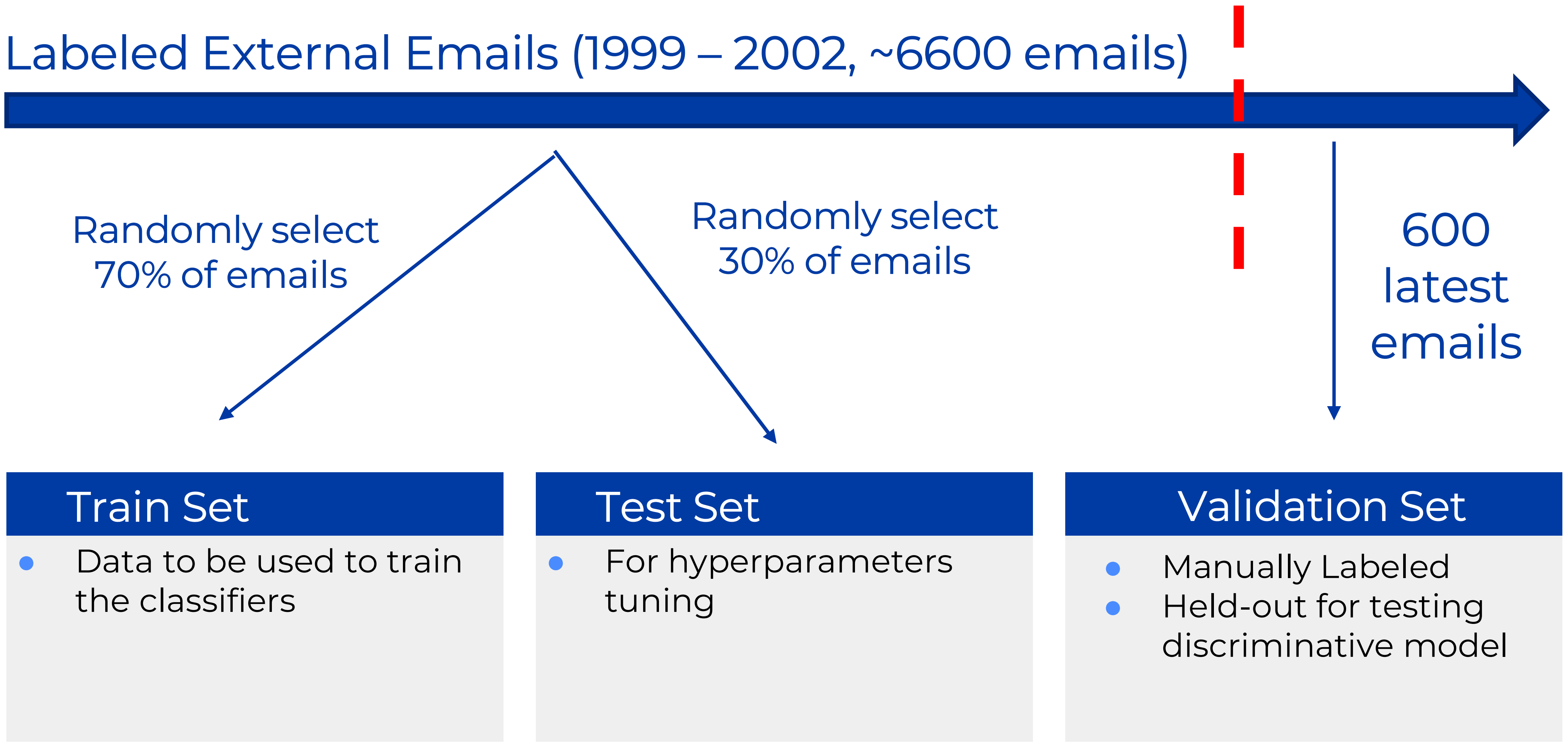
```
labeled_data['label'].value_counts() ...

0    4031
1    1939
```

Output:
4031 Ham Emails
1939 Spam Emails

02

Model Development

Spam Classifier


Labeled External Emails (1999 – 2002, ~6600 emails)
Randomly select 70% of emails
Randomly select 30% of emails
600 latest emails
Train Set
Data to be used to train the classifiers
Test Set
For hyperparameters tuning
Validation Set
Manually Labeled
Held-out for testing discriminative model

Evaluation Methods

Precision

90%+

Coverage

Generalization

Preprocessing Approaches

Pre-trained Word Embeddings

- Google's Pretrained Word2Vec Model
- Vector size: 300

Word Embeddings

- Trained by training set emails
- Vector size: 300

TF-IDF

- Trained by training set emails
- Vocabulary Size: 944

Classifiers

- Decision Tree
- XGBoost
- Random Forest
- Support Vector Machine
- Logistic Regression

Combined into

Final Ensemble Model

Considering the size of the dataset. Deep Learning Algorithms is not used.

Model Result

	Google Word2Vec	Word2Vec	TF-IDF
Train Set	Precision: 94.43% Recall: 87.51%	Precision: 94.39% Recall: 89.30%	Precision: 95.07% Recall: 86.61%
Test Set	Precision: 91.56% Recall: 84.71 %	Precision: 92.25% Recall: 89.04%	Precision: 94.53% Recall: 86.21%
Validation Set	Precision: 94.46% Recall: 84.49%	Precision: 90.60% Recall: 89.10%	Precision: 95.88% Recall: 84.49%
Snorkel Set	Precision: 86.09% Recall: 83.19%	Precision: 90.09% Recall: 84.03%	Precision: 90.57% Recall: 80.67%

Other Model Result

W2V with Bigram & Lemmatization

```
validation set result:
[[251  46]
 [ 39 264]]
precision: 0.8516129032258064
recall: 0.8712871287128713
accuracy: 0.8583333333333333
f1_score: 0.8613376835236541
```

```
snorkel set result:
[[268  13]
 [ 14 105]]
precision: 0.8898305084745762
recall: 0.8823529411764706
accuracy: 0.9325
f1_score: 0.8860759493670887
```

TF-IDF with Bigram & Lemmatization

```
validate set result:
[[290   7]
 [ 51 252]]
precision: 0.972972972972973
recall: 0.8316831683168316
accuracy: 0.9033333333333333
f1_score: 0.8967971530249109
```

```
snorkel set result:
[[263  18]
 [ 22  97]]
precision: 0.8434782608695652
recall: 0.8151260504201681
accuracy: 0.9
f1_score: 0.8290598290598290
```

Spam Classifier - Model Output

Output:

5256 Ham Emails – To be used in Topic Model

2852 Spam Emails

02
Model Development

Topic Model

Model Selection

Latent Dirichlet Allocation (LDA)

```
Coherence Score of 1-topics:  0.3212744020059067
Coherence Score of test data:  0.24797148806455938
Coherence Score of 2-topics:  0.5963888291927538
Coherence Score of test data:  0.5510887949473797
Coherence Score of 3-topics:  0.5830304509603105
Coherence Score of test data:  0.5771009383239412
Coherence Score of 4-topics:  0.6328632041573607
Coherence Score of test data:  0.6048475795126138
Coherence Score of 5-topics:  0.6116075659057547
Coherence Score of test data:  0.5677539480593501
```

LDA Mallet

```
Coherence Score of 1-topics:  0.3203094152523298
Coherence Score of 2-topics:  0.5939300876088229
Coherence Score of 3-topics:  0.5818415781661549
Coherence Score of 4-topics:  0.6009294347324005
Coherence Score of 5-topics:  0.5529872781430463
```

Latent Semantic Analysis (LSA)

```
Coherence Score of 1-topics:  0.23144560306583434
Coherence Score of 2-topics:  0.3007698711546833
Coherence Score of 3-topics:  0.38806370223833747
Coherence Score of 4-topics:  0.4638299192789817
Coherence Score of 5-topics:  0.4545758453201181
```

Number of Topics

```
Coherence Score of 4-topics:  0.6328632041573607

Coherence Score of test data:  0.6048475795126138

Coherence Score of 5-topics:  0.6116075659057547

Coherence Score of test data:  0.5677539480593501
```

```
Coherence Score of 6-topics:  0.5262818742912393

Coherence Score of test data:  0.4827635734183347

Coherence Score of 7-topics:  0.5633251190008167

Coherence Score of test data:  0.5220360879213233

Coherence Score of 8-topics:  0.6005059633081636

Coherence Score of test data:  0.5251150218544316
```

Top 15 words in each topic

Topic: 0

	weight
"com"	0.210
"e-mail"	0.041
"ca"	0.027
"gov"	0.020
"cpuc"	0.013
"net"	0.012
"org"	0.012
"energy"	0.009
"bill"	0.009
"u"	0.009
"bracepatt"	0.006
"williams"	0.006
"john"	0.005
"state"	0.005
"doc"	0.005

Topic: 1

	weight
"say"	0.018
"power"	0.014
"california"	0.012
"state"	0.012
"energy"	0.011
"rate"	0.008
"price"	0.008
"would"	0.007
"electricity"	0.006
"utility"	0.006
"gas"	0.006
"plan"	0.005
"customer"	0.005
"market"	0.005

Topic: 2

	weight
"enron"	0.182
"com"	0.049
"na"	0.041
"hou"	0.032
"ect"	0.030
"ee"	0.028
"ees"	0.022
"jeff"	0.019
"dasovich"	0.016
"pm"	0.014
"steffes"	0.014
"subject"	0.014
"james"	0.013
"corp"	0.011
"richard"	0.011

Topic: 3

	weight
"com"	0.030
"jeff"	0.025
"enron"	0.021
"call"	0.017
"dasovich"	0.014
"pm"	0.012
"subject"	0.012
"meeting"	0.011
"get"	0.011
"send"	0.010
"please"	0.010
"may"	0.009
"best"	0.008
"message"	0.007
"work"	0.007

LDA topic 1 – California Energy Issues

67694	NEWS: St…	* from A…	0	0.999413…	0	0

* from Associated Press (4/9/01)
* discusses rates and blackout potential for each state
* Louisiana on verge of power crisis, they say...
* I think I might move to Nebraska!

Here's an attachment for easy distribution:

Here's the full article for quick scanning:
Monday April 9 4:23 PM ET
Power Situation by State
By The Associated Press,
A state-by-state look at the electric power situation for summer:

Alabama: Blackouts and rolling brownouts are unlikely as utilities are required to maintain a reserve 15 percent above what is needed to meet peak demand. Residential and business customers will pay rates about 1.1 percent above a year ago.

Alaska: Blackouts and brownouts are unlikely. Hydroelectric power is plentiful, and generating systems are not being taxed. Electric bills are expected to fall slightly.

Arizona: There is little likelihood of a power interruption. Utilities are prepared to handle the summer electricity demand, and price caps prevent the major utilities from increasing...

57358	CERA rep…	a pretty…	0	0.999046…	0	0

a pretty fair analysis of the California mess....

CERA Alert: December 13, 2000

Title: California on the Brink
CERA Knowledge Areas: Western Energy, N. American Power, N. American Gas

CALIFORNIA ON THE BRINK
The California Stalemate
California moved closer to the brink of an outage today as concerns over credit-worthiness of buyers brought the possibility that generators would avoid selling to the California market. While numerous factors have contributed to the high cost of power incurred by California's utilities, the root cause of the current crisis is a lack of new generation. The current credit crisis and its threat to supplies could spark state action to address the situation. The collective efforts of all market participants should be focused on increasing generation capacity as quickly as possible.

Western power prices have skyrocketed well beyond the record levels set this summer. Perhaps because frozen rates insulate the majority of California consumers and companies from...

Topic Model - Model Output

Output:

Emails are sorted into 4 topics

1. government parties
2. California Energy Issue
3. Meetings related
4. others

A Vector indicating the likelihood of the email in each topic

Model Maintenance

Snorkel + Spam Classifier	• Metrics: Drop in Precision (Validation vs New Emails) • Redevelopment: Precision drops > 2.5% • Evaluation Frequency: 1 year
Topic Model	• Metrics: Drop in Coherence Score (Test vs New Emails) • Redevelopment: Coherence score drops > 3% • Evaluation Frequency: 6 months • Remark: Other ML models need to re-tune hyperparameters if the redevelopment is performed

03

Business Use Case

Use Cases in business

Email Classification Engine	• Spam classification • Segmenting emails into topics. • As an input in other email ML models
Snorkel + Spam Classifier	• The development process can be applied to other topic classification (e.g. casual chat, meeting frequency) • Can be applied to imbalance classes problem (requires extra steps, e.g. data augmentation, up-sampling, down-sampling)

Example of using Topic Model as Input

Without Topic Vector

```
validate set result:
[[286  11]
 [ 47 256]]
precision: 0.9588014981273408
recall: 0.8448844884488449
accuracy: 0.9033333333333333
f1_score: 0.8982456140350877
```

```
snorkel set result:
[[271  10]
 [ 23  96]]
precision: 0.9056603773584906
recall: 0.8067226890756303
accuracy: 0.9175
f1_score: 0.8533333333333334
```

With Topic Vector

```
validate set result:
[[282  15]
 [ 41 262]]
precision: 0.9458483754512635
recall: 0.8646864686468647
accuracy: 0.9066666666666666
f1_score: 0.9034482758620689
```

```
snorkel set result:
[[271  10]
 [ 20  99]]
precision: 0.908256880733945
recall: 0.8319327731092437
accuracy: 0.925
f1_score: 0.868421052631579
```

Thank you!