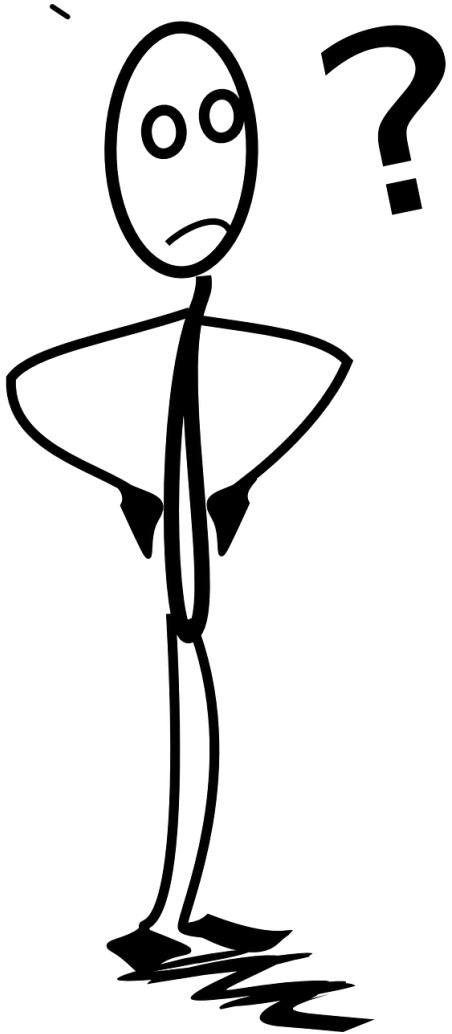
Is It A Spam ?

by Lamiaa Dakir & Jocelyn Dunkley



From: tomking@something.com

Good morning!

The snow/ice storm is creating an ice buildup and conditions where the Grounds crew cannot plow to ground level or apply melting agents until the precipitation stops.

All college-related activities – including classes – are cancelled today (Tuesday 3-14-17). No classes or other activities will be held, including the Graduate School of Social Work. Those employees designated as non-essential should not report to work.

The Tri-College Transportation system will stay closed throughout the day, and resume on a normal schedule tomorrow (Wednesday). Please go to the Transportation section of storm.brynmawr.edu for run times.

Please check the Hotline and website throughout the day for updates.

Weather Hotline: 610-526-7310

Website storm.brynmawr.edu

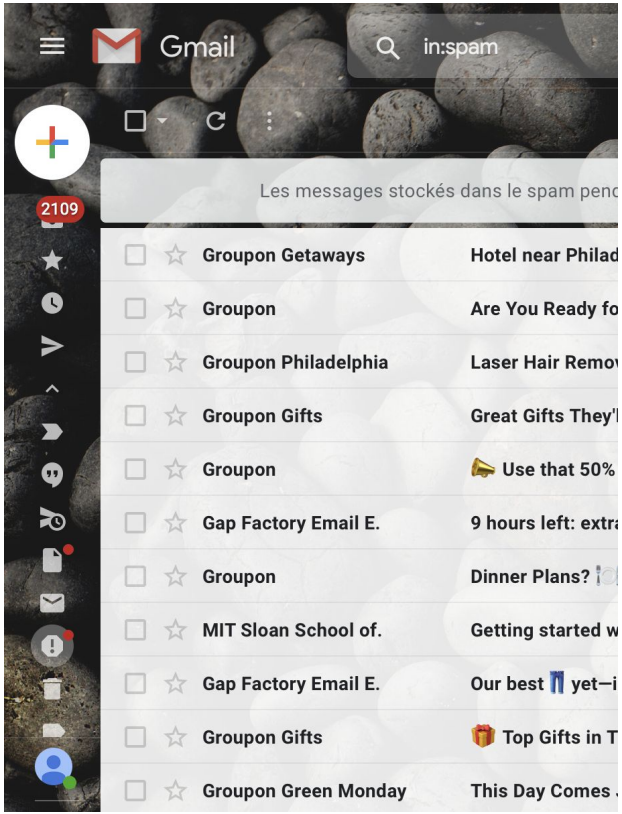
Students – please check the website for updates on Dining, the Library, Fitness Center, Transportation, etc. at storm.brynmawr.edu

The Township has declared a Snow Emergency. Please use caution if you must go out - icy conditions and gusty winds are expected throughout the day.

Thanks

Tom

How is Spam Detected ?



Gmail Spam Filtering Factors:

1. **Content :**
   - Grammar and spelling errors
   - Spam trigger words
2. **Domain Name**
   - Blacklisted domain names
3. **Email Header**
4. **Attachments**
5. **Get Whitelisted**
   - List of approved email addresses

Credit: https://clean.email/gmail-spam-filter

Spam Database

Dataset:

- 4601 Examples
  - 1813 Spam (39.4%)
  - 2788 Non-spam (60.6%)
- 58 Attributes
  - 57 Features (continuous)
  - Label
    - 1 means the email is spam
    - 0 means the email is not spam

Spam Database

Features:

- Frequency of some words

```
word_freq_make:          continuous.
word_freq_address:       continuous.
word_freq_all:           continuous.
word_freq_3d:            continuous.
```

- Frequency of some characters

```
char_freq_;:             continuous.
char_freq_(:             continuous.
char_freq_[:             continuous.
char_freq_!:             continuous.
```

- Capital Letters

```
capital_run_length_average: continuous.
capital_run_length_longest: continuous.
capital_run_length_total:   continuous.
```

Spam Database

Pre-processing:

- Built Data object to store X (features) and y (label)

- Split spam database into
  - 75% Training Data
  - 25% Testing Data

Note : To avoid bias when training our models, we made sure to randomly chose points from the dataset to split the data into training and testing

Detect Spam Using Machine Learning

- Decision Trees
- Naive Bayes
- Random Forests
- AdaBoost
- Neural Networks

Detect Spam Using Machine Learning

- Decision Trees
- Naive Bayes
- Random Forests
- AdaBoost
- Neural Networks

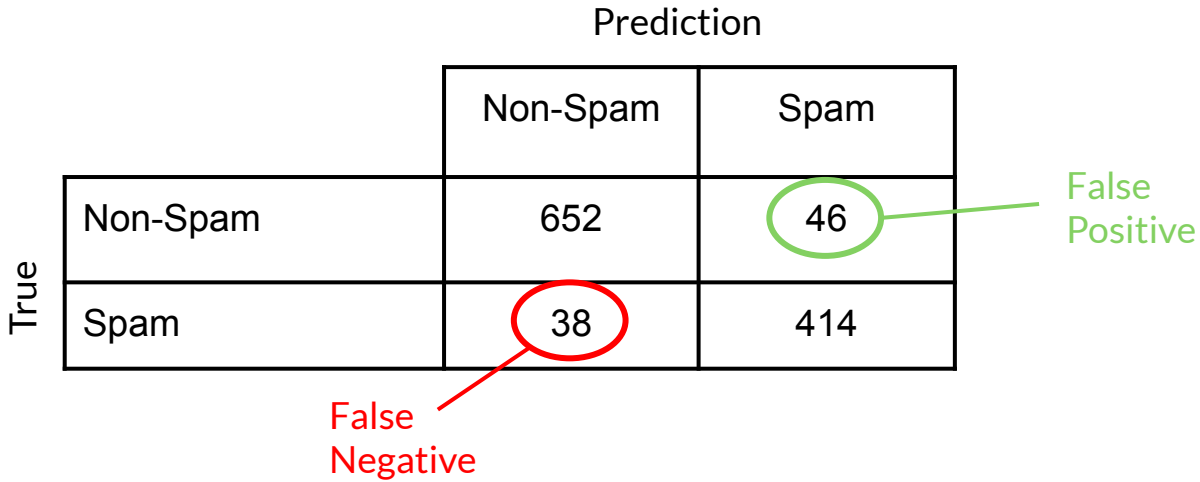
Decision Trees

- Used `sklearn.tree.` **`DecisionTreeClassifier`**(*criterion='**entropy**', splitter='**best**'*)
- No maximum depth
- Relied on entropy to find the best feature

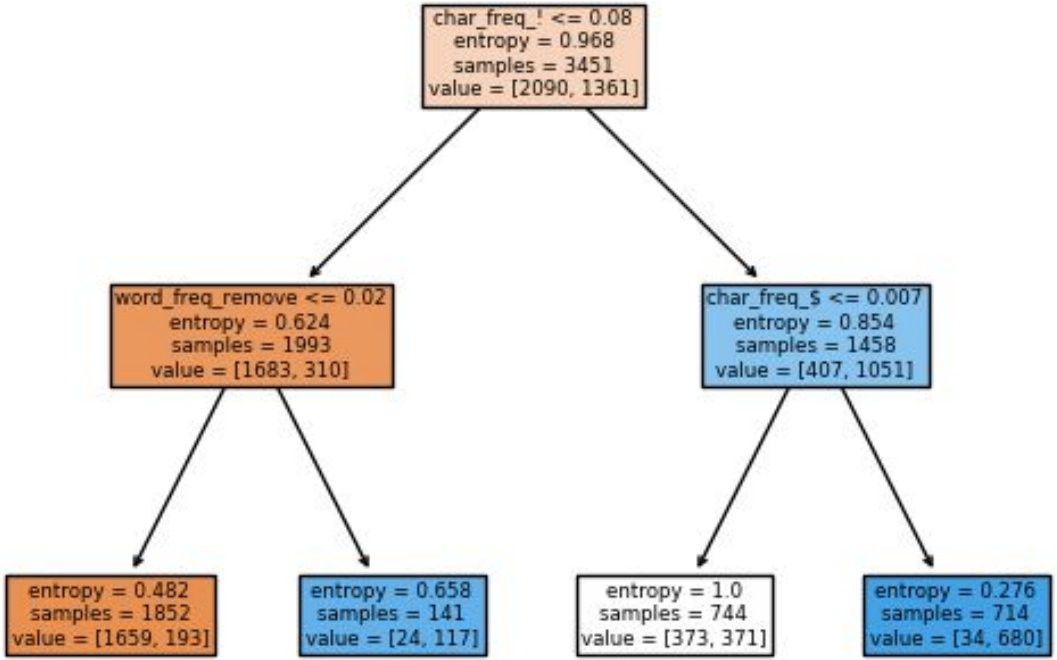
		Prediction: Non-Spam	Prediction: Spam
True	Non-Spam	652	46
True	Spam	38	414

Decision Trees

- Used `sklearn.tree.` **`DecisionTreeClassifier`**(*criterion='**entropy**', splitter='**best**'*)
- No maximum depth
- Relied on entropy to find the best feature



Decision Trees



Detect Spam Using Machine Learning

- Decision Trees
- Naive Bayes
- Random Forests
- AdaBoost
- Neural Networks

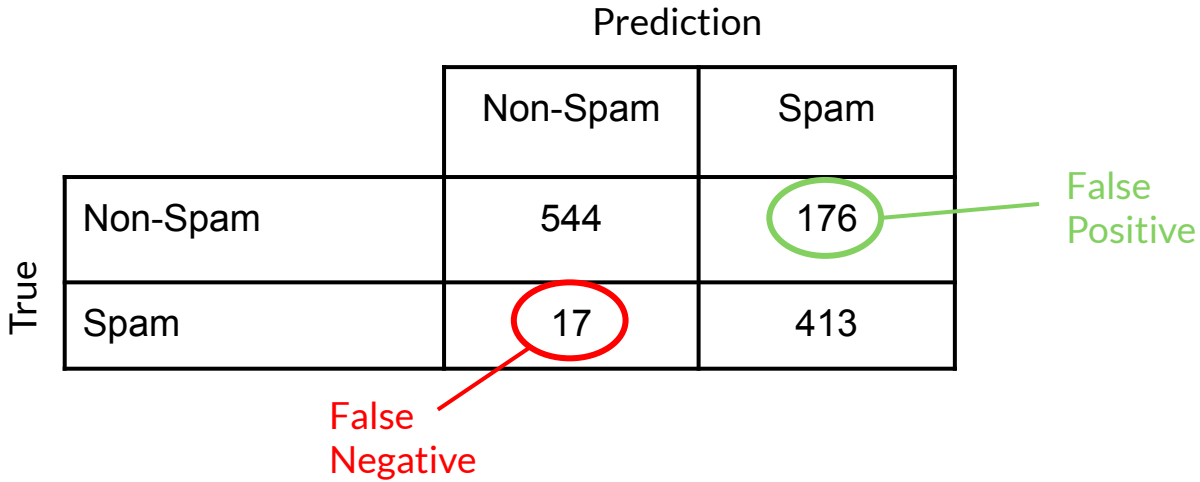
Naive Bayes

- Used `sklearn.naive_bayes.` **`GaussianNB`**(*alpha=1.0*)
- Assumes features are independent
- GaussianNB classifies features that are continuous

		Prediction	
		Non-Spam	Spam
True	Non-Spam	544	176
	Spam	17	413

Naive Bayes

- Used `sklearn.naive_bayes.` **`GaussianNB`**(*alpha=1.0*)
- Assumes features are independent
- GaussianNB classifies features that are continuous



Detect Spam Using Machine Learning

- Decision Trees
- Naive Bayes
- Random Forests
- AdaBoost
- Neural Networks

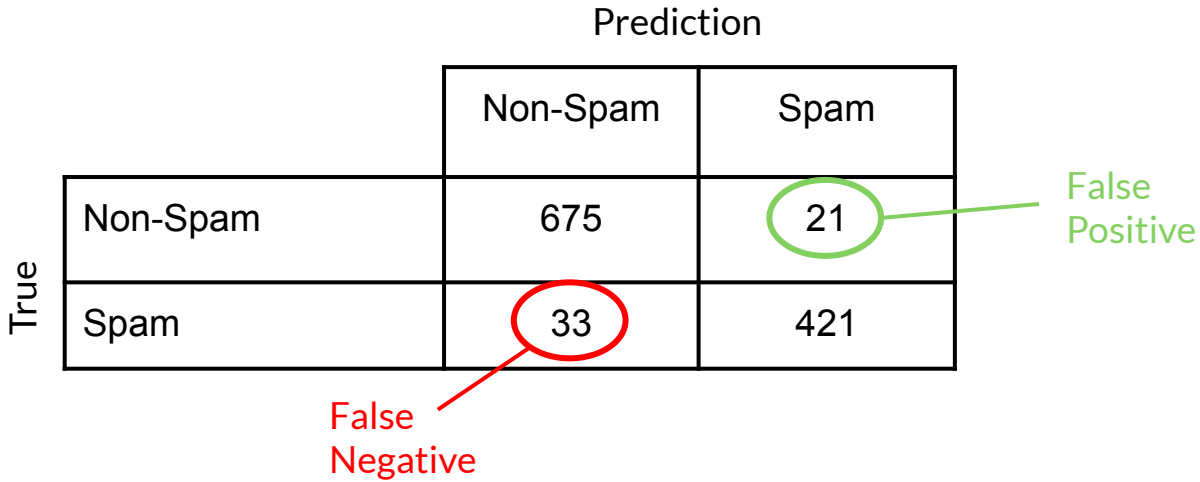
Random Forest

- Used `sklearn.ensemble.` **`RandomForestClassifier`**(*n_estimators=**20***)
- Used bootstrap samples
- Tried Gini and Entropy as criterion (almost similar results)

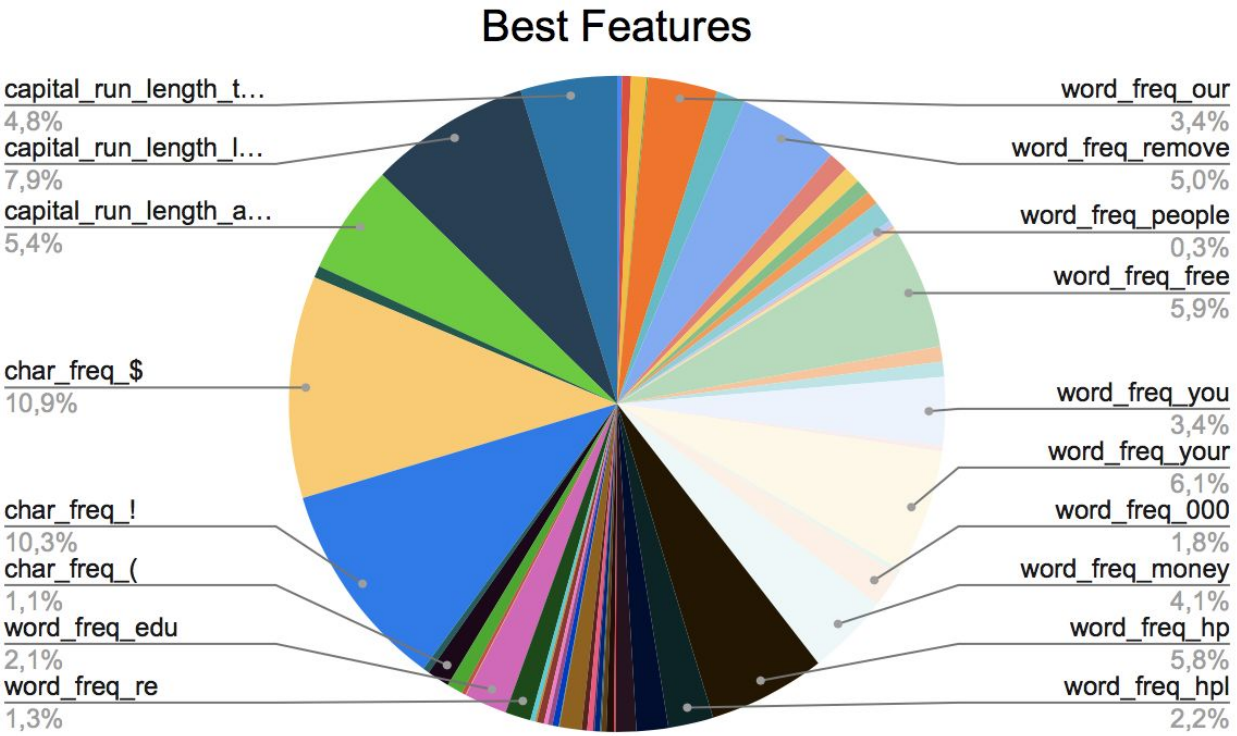
		Prediction	
		Non-Spam	Spam
True	Non-Spam	675	21
	Spam	33	421

Random Forest

- Used `sklearn.ensemble.` **`RandomForestClassifier`**(*n_estimators=**20***)
- Used bootstrap samples
- Tried Gini and Entropy as criterion (almost similar results)



Random Forest



Detect Spam Using Machine Learning

- Decision Trees
- Naive Bayes
- Random Forests
- AdaBoost
- Neural Networks

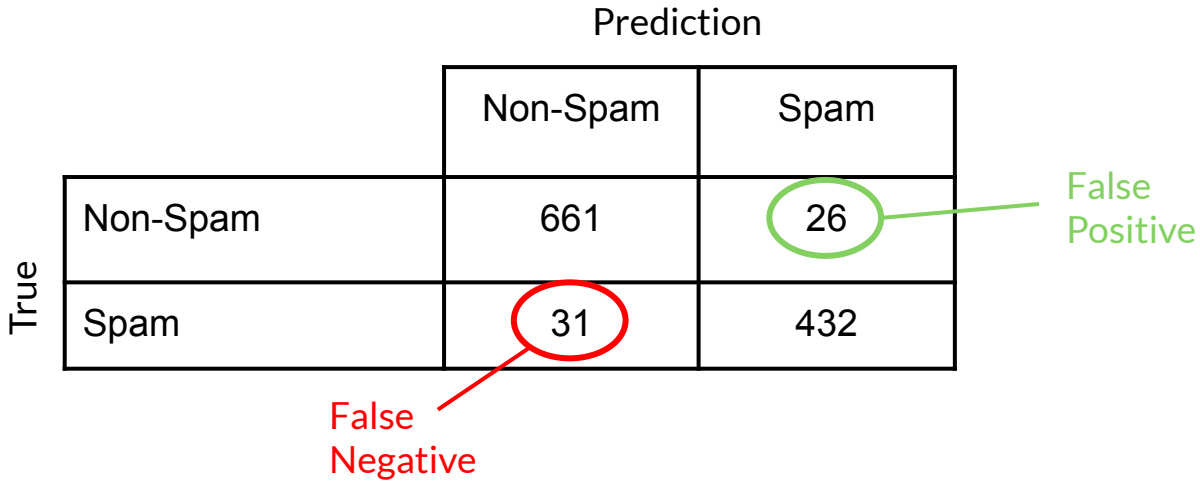
AdaBoost

- Used `sklearn.ensemble.` **`AdaBoostClassifier`**`(` *n_estimators=**200***`)`
- The base classifier is **`DecisionTreeClassifier`**`(max_depth=1)` (i.e Decision Stump)

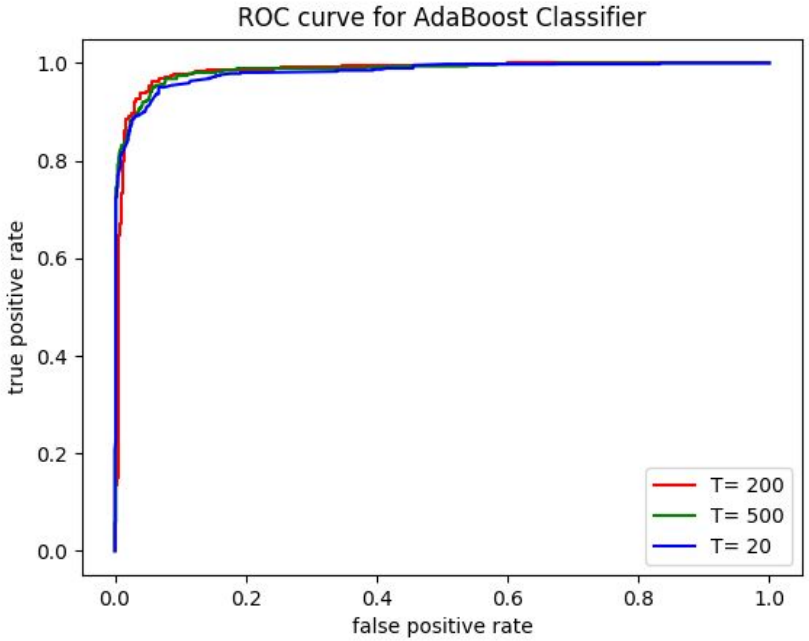
		Prediction	
		Non-Spam	Spam
True	Non-Spam	661	26
	Spam	31	432

AdaBoost

- Used `sklearn.ensemble.` **`AdaBoostClassifier`**`(` *n_estimators=**200***`)`
- The base classifier is **`DecisionTreeClassifier`**`(max_depth=1)` (i.e Decision Stump)

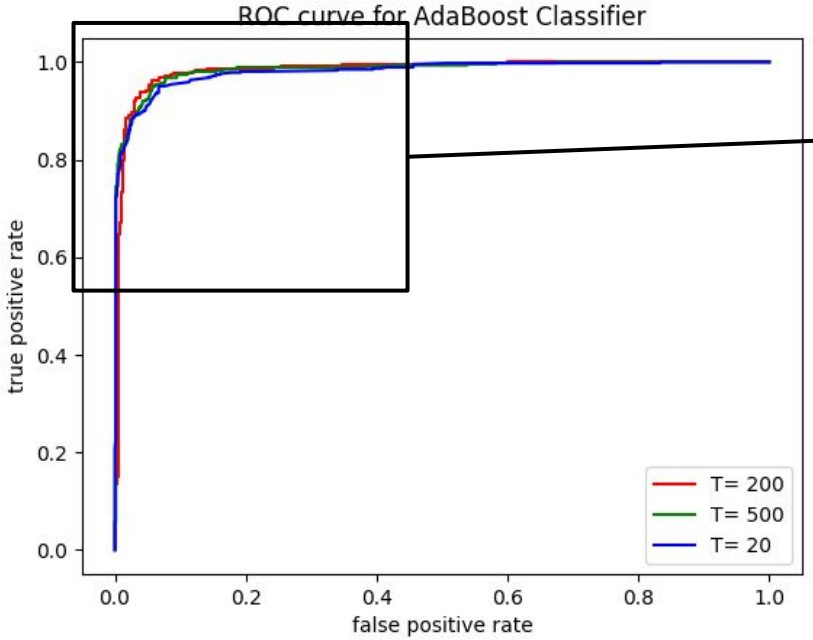


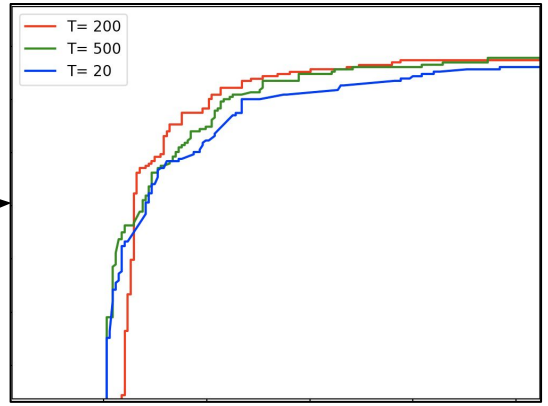
AdaBoost using Decision Stumps



We varied the number of classifiers used to observe how the algorithm behaves.

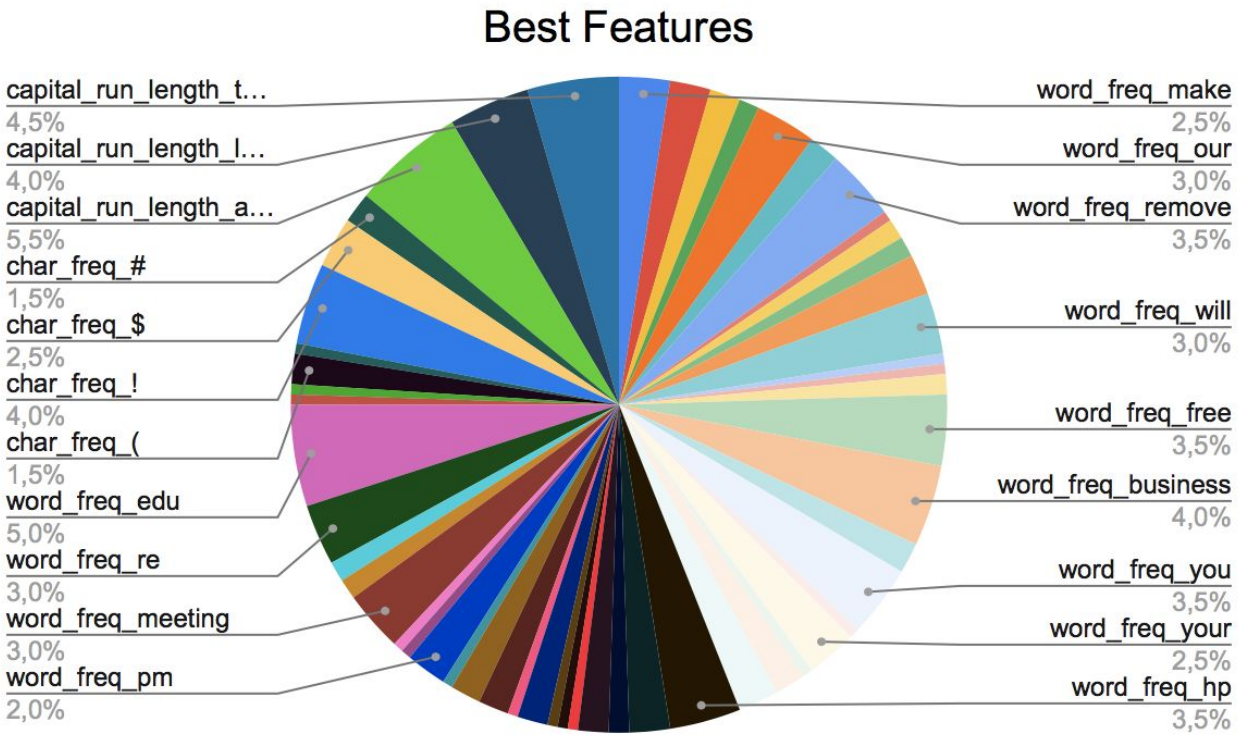
AdaBoost using Decision Stumps





- T = 200 gives the most optimal result
- T = 500 might be overfitting
- T = 20 is underfitting

AdaBoost


Best Features
capital_run_length_t…
4,5%
capital_run_length_l…
4,0%
capital_run_length_a…
5,5%
char_freq_#
1,5%
char_freq_$
2,5%
char_freq_!
4,0%
char_freq_(
1,5%
word_freq_edu
5,0%
word_freq_re
3,0%
word_freq_meeting
3,0%
word_freq_pm
2,0%
word_freq_make
2,5%
word_freq_our
3,0%
word_freq_remove
3,5%
word_freq_will
3,0%
word_freq_free
3,5%
word_freq_business
4,0%
word_freq_you
3,5%
word_freq_your
2,5%
word_freq_hp
3,5%

Detect Spam Using Machine Learning

- Decision Trees
- Naive Bayes
- Random Forests
- AdaBoost
- Neural Networks

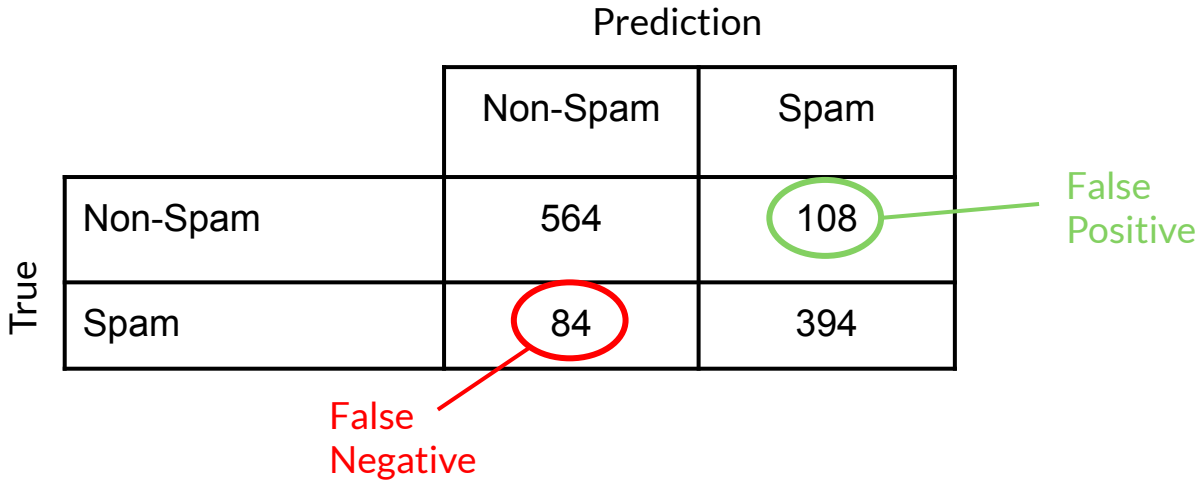
Fully Connected Neural Network

- Architecture was Flatten -> layer with 4000 hidden units and ReLU -> output layer with 2 and softmax

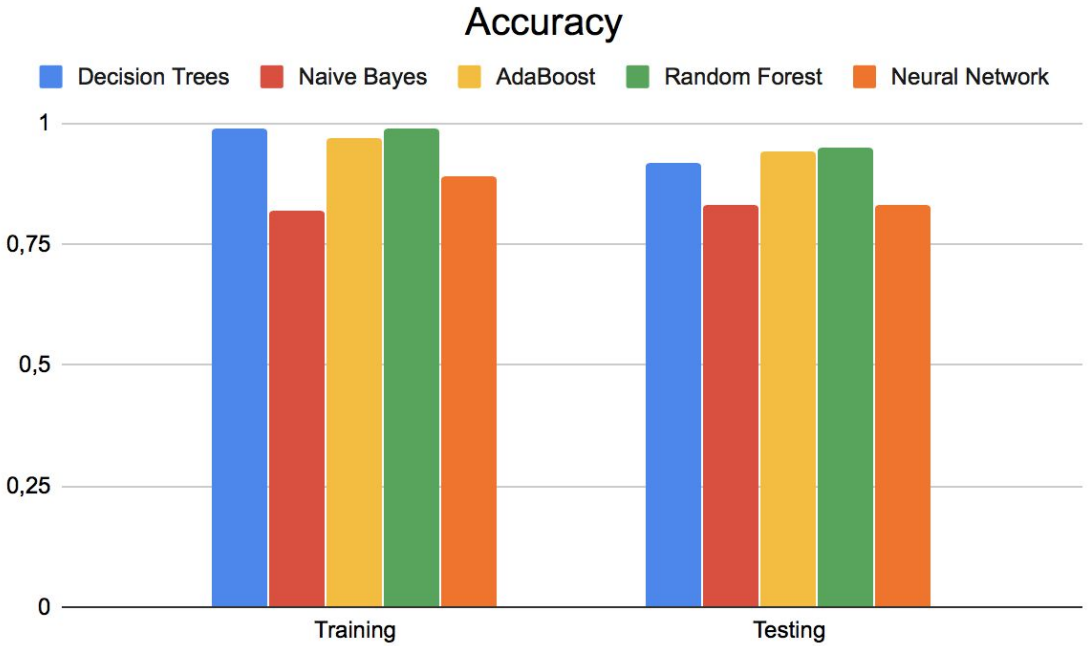
		Prediction	
		Non-Spam	Spam
True	Non-Spam	564	108
	Spam	84	394

Fully Connected Neural Network

- Architecture was Flatten -> layer with 4000 hidden units and ReLU -> output layer with 2 and softmax

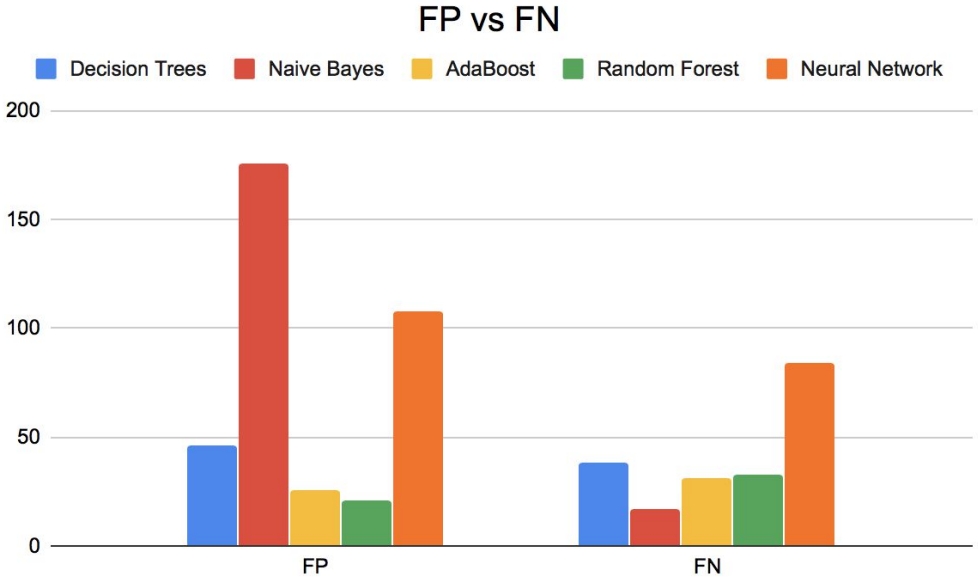


Performance Evaluation



- During the training, Decision Trees and Random Forest methods achieve the highest accuracy. NN performance was in between NB and the others.
- During the testing, AdaBoost and Random Forest show a better performance than the other methods. NN performed almost the same as NB.

Performance Evaluation



- Naive Bayes is very strict. It classified many emails as spam (high FP).
- Random Forest and AdaBoost seem to do well overall.
- In particular Random Forest has a lower FP
- NN had both a high FP and FN

Advantages of Random Forest

- High training and testing accuracy (99% training - 95% testing)
- Takes **less time** because it requires a **lower number of classifiers**
- Random Forest was able to give **reliable importance values** to each features

Testing on Real Email

Spam Email

Subject: Join the thousands who are now sp@m-free

FORGET SPAM BLOCKERS!

Get SMART Spam Control That Always Delivers The Email You Want!

Finally, we discovered the ultimate solution that is guaranteed to stop all spam without losing any of your important email! This revolutionary advanced technology
also protects you 100% against ALL email-borne viruses — both known and unknown.

We didn't believe it either until we actually tried it. So you be the judge and see for yourself.

{LINK}

Sara's Email

Hi Jocelyn and Lamiaa,

Apologies for the delay getting you feedback. I love the dataset and don't foresee any issues. A few comments:

- For your presentation, I think it will be important to explain the features. Are these words, capitalization, etc? It would also be very interesting to do a feature analysis to see which features are most helpful for predicting the output (and if these change across methods).

- I also think it will be important to talk about previous work on this dataset.

- I think your investigation of hyper-parameters is a great idea. For this part, you probably have enough data to do a proper cross-validation, even if the number of folds is low.

-For the NN method, I would probably recommend using a FC architecture unless there are some features that would benefit from a convolutional layer.

- Runtime might also be worth noting/comparing.

Thank you! Looking forward to the results. All the best,
-Sara

Testing on Real Email

	Spam Email (label = 1)	Sara's Email (label = 0)
Decision Trees	0	0
Naive Bayes	1	0
Random Forest	0	0
AdaBoost	1	0
Neural Network	0	0

Conclusion & Future Work

1. We should use machine learning to detect spam !

Future goals:

1. Collect a larger spam dataset (can't have too much data!) to train/test and a larger database of real spam emails to test our model on
2. Tweak the best model (ex. AdaBoost) to get the highest possible accuracy
3. Compare varieties of the same type of algorithm to see if something subtle could help.

Thank you !
Questions ?