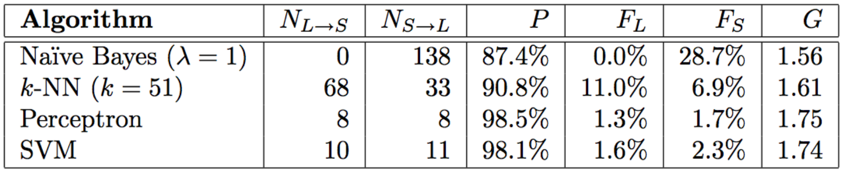
Introduction

It is important to study what people did and create our study on top of it. In this section, we focus on summarize previous studies of spam/ham e-mail filtering and their machine learning methods. At the end of this section, we point out the new methods and new data we use in this project to show our understanding.

Previous Studies

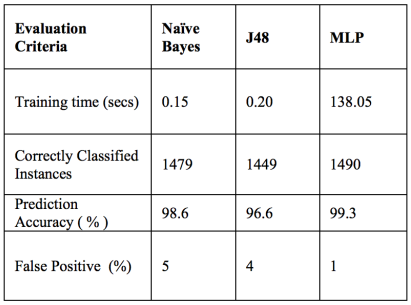
People already came up with the idea of spam/ham e-mail filtering before 2004. In the paper *Machine Learning Techniques in Spam Filtering* written by Konstantin in 2004, the experiment used four main methods : Naive Bayes, K-NN, Perceptron, and SVM. And compare accuracy rate of each methods. In this basic practice, they found the Perceptron method has the highest accuracy rate, 98.5% with a corpus of 1099 messages.



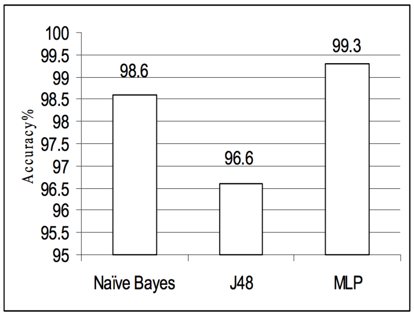
Algorithm	$N_{L \to S}$	$N_{S \to L}$	P	F_L	F_S	G
Naïve Bayes ($\lambda = 1$)	0	138	87.4%	0.0%	28.7%	1.56
k-NN ($k = 51$)	68	33	90.8%	11.0%	6.9%	1.61
Perceptron	8	8	98.5%	1.3%	1.7%	1.75
SVM	10	11	98.1%	1.6%	2.3%	1.74

In 2010, *Email Spam Filtering using Supervised Machine Learning Techniques* written by V.Christina, they used Naive Bayes, J-48(Decision Tree) and Multilayer Perceptron. And they found out MLP performed the best with 99.3% accuracy rate when experimenting with a corpus of 1500 messages.

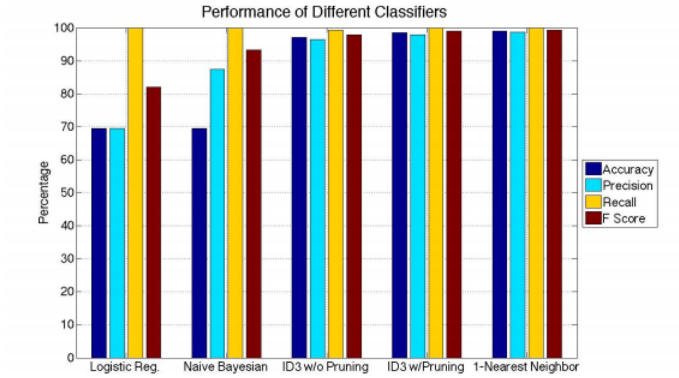
COMPARATIVE RESULTS OF THE CLASSIFIERS



Evaluation Criteria	**Naïve Bayes**	**J48**	**MLP**
Training time (secs)	0.15	0.20	138.05
Correctly Classified Instances	1479	1449	1490
Prediction Accuracy (%)	98.6	96.6	99.3
False Positive (%)	5	4	1



In 2016, *Spam Mail Detection using Classification* written by Parhat and Gambhir used Naive Bayes, SVM and J-48(Decision Tree). And they found out Naive Bayes performed the best with 76% accuracy rate in their experiment. And *Email Spam Detection* written by Ge and Lauren, used the corpus from TREC 2007 with 1000 messages. They tried logistic regression, Naive Bayesian, Decision Tree and K-NN. The finally found KNN with highest 99% accuracy rate.



Our works

1. **Use multiple data source**: In each paper, they mainly use a single year of corpus data. In our project, we tried to source different e-mail and integrate them. The format of each data source is different thus hard to clean. And we successfully got to manage a huge data set.

2. **Try 6 methods at the same time**: Previous studies compare accuracy rate with different methods, but they didn't compare them all at a time. So we studied the methods from 2004 to 2016, and apply all of possible methods with adequate tuning parameters to compare them

3. **Apply 1-Gram and 2-Gram**: Each paper marked that data processing step is important to a good result. Here, we introduced bag of words of 1-Gram and 2-Gram methods in the feature engineering part. And we can see different result of accuracy rate in the following section.