Naïve Bayes and Logistic Regression for Ham and Spam

Assignment Report

Machine Learning

To the

The University Of Texas at Dallas

Submitted by:

Prashant Prakash

(Net id: pxp141730)

The University of Texas at Dallas

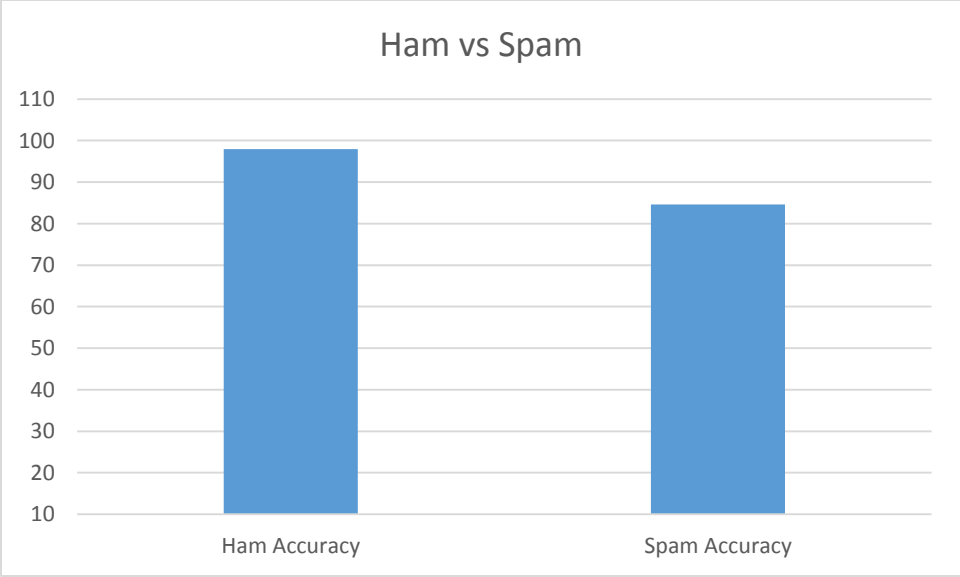
Dallas – 75252

February, 2015

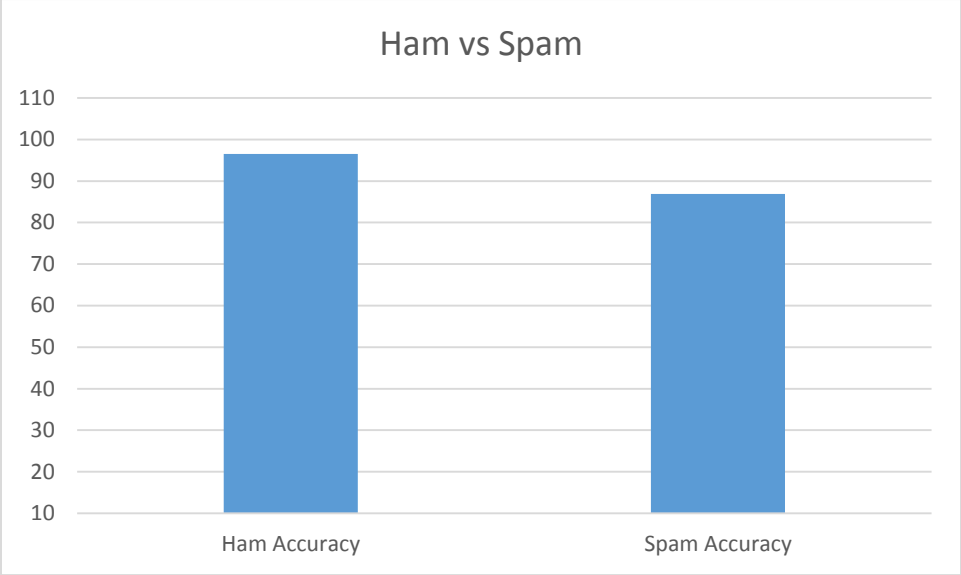
Naïve Bayes

- **Setup:**
  Training Data:
  Number of Spam Files: 123
  Number of Ham Files: 340
  Test Data:
  Number of Spam Files: 130
  Number of Ham Files: 348

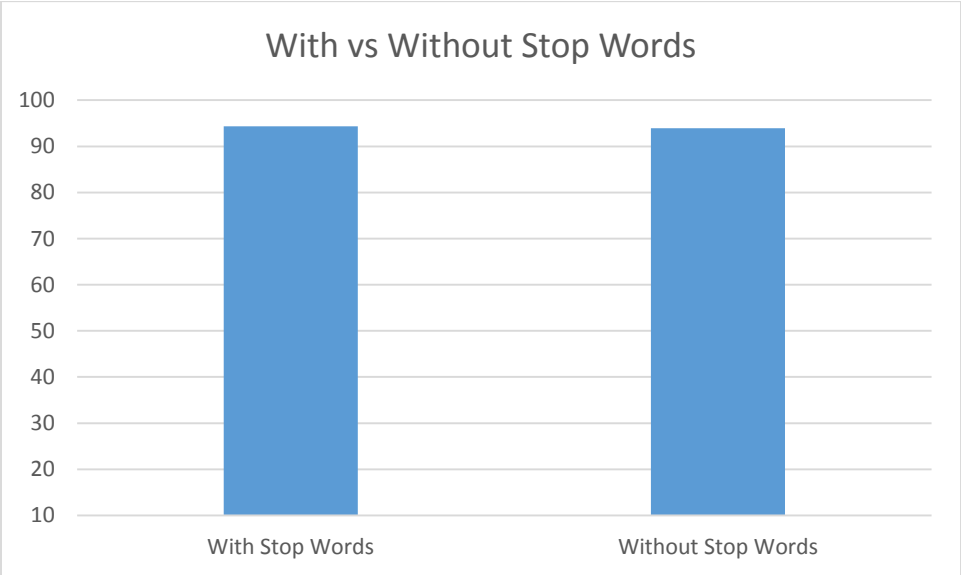
- **Accuracy with Stop Words**
  Number of Tokens Generated: 9192
  Spam Accuracy: 84.61%
  Ham Accuracy: 97.98 %



- Accuracy After Removal of Stop Words
  Spam Accuracy: 86.92%
  Ham Accuracy: 96.55%
  Number of Tokens Generated: 9076



- Overall Accuracy Comparison Before and After removal of Stop words



Explanation: As we can see that there is not big change is number of tokens after removal of stop words. So definitely Naïve Bayes won't show a big change in accuracy because this Algorithm is based on the counts of tokens, count of words occurred in spam mail, count of words occurred in ham mails. There are not many stop words in the training data set provided to us. So there is not much impact of stop words in our number of features. Though you can see there is an increase in accuracy of spam after removal of stop words. It means there are more stop words in spam train data and after removal of stop words we removed the noise from training data.

Logistic Regression

- **Setup:**
  Training Data:
  Number of Spam Files: 123
  Number of Ham Files: 340
  Test Data:
  Number of Spam Files: 130
  Number of Ham Files: 348

- With stop Words

No of iterations	Eta(learning rate)	lambda	Spam Accuracy	Ham Accuracy	Over All Accuracy
100	.01	0	85.38%	94.54%	92.05%
100	.01	1	86.15%	94.82%	92.46%
100	.01	5	12.30%	99.71%	75.99%
100	.01	10	8.46%	99.71%	74.89%
100	.01	15	83.84%	87.93%	86.82%
100	.01	20	5.38%	99.71%	74.05%
100	.01	25	4.61%	99.71%	73.68%

- Without Stop words

No of iterations	Eta(learning rate)	lambda	Spam Accuracy	Ham Accuracy	Over All Accuracy
100	.01	0	86.15%	97.41%	94.35%
100	.01	1	88.46%	96.83%	94.56%
100	.01	5	90.0%	97.41%	95.39%
100	.01	10	60.76%	99.42%	89.12%
100	.01	15	96.92%	18.67%	41.12%
100	.01	20	16.92%	99.71%	77.85%
100	.01	25	17.69%	99.71%	78.23%

- **Explanation:**

  There is an increase in the Ham and Spam Accuracy after removal of Stop words. The reason is reduction in features has removed noise from data. Stop words frequency count was big in numbers. The best practice in Text classification is to always remove stop words from training Data. Removal of Stop words always help in improving accuracy.
  We can see that sometime there is sudden decrease in the spam Accuracy after increasing value of lambda. This is the case when after penalizing weights the decision boundary has included spam data on ham side. This can be the case of under fitting as the decision boundary is not proper.

Smoothing and Feature Selection

- **IDEA:**
  The total training documents are 478 and number of features after removal of stop words are 9076. My main idea for smoothing is reduction in number of features and for that I used "**Porter's Algorithm**" for **Stemming**. After applying Stemming technique the number of features reduced to 7565.
  In this case Behavior of Naïve Bayes:
  Spam Accuracy: 93.07%
  Ham Accuracy: 93.39%
  Spam Accuracy got increased but Ham Accuracy got decreased little bit. The reason is after applying stemming, most of the tokens from Spam train data participated in stemming more than ham data.
  Behavior of Logistic Regression:
  Number of iterations: 100, eta=.01, lambda=0
  Spam Accuracy: 87.69%
  Ham Accuracy: 90.80%
  Spam Accuracy got increased but Ham Accuracy got decreased. The reason is same as explained for Naïve Bayes above.

  For Feature Selection: We have already removed stop words. So I used **TF (Term Frequency Technique)** to eliminate features which have occurred very less number of times. So it participates less in taking decision for a particular mail to be ham or spam. After applying TF the number of features reduced drastically: 2941. Which means there are many words in the documents which have occurred only once.
  Naïve Bayes Behavior:
  Spam Accuracy: 91.53%
  Ham Accuracy: 91.09%
  Logistic Regression Behavior:
  Number of iterations: 100, eta=.01, lambda=0
  Spam Accuracy: 83.84%
  Ham Accuracy: 89.94%
  There is a reduction in the accuracy for Ham as well as Spam. This is the case of under fitting. The number of features are not enough to decide a proper decision boundary between Ham and Spam.