Email Spam Detection

CS5262 Final Project
Yizhou Guo

Background

Email spam detection has been developing in the past decade.

Popular dataset: Spambase, Spam archive, **Enron-Spam**, Spam Assassin, TREC, PU(1,2,3,A), etc. [1][2]

Popular method: **SVM**, k-NN, Genetic Algorithm, **ANN**, Naïve Bayes, Random Forest, NLP, etc. [2]

Key measurements: accuracy, false-positive rate

Previous results

Multiple Linear Regression: No previous results available.

Support Vector Machine: 80% to 97% (using different datasets). [1]

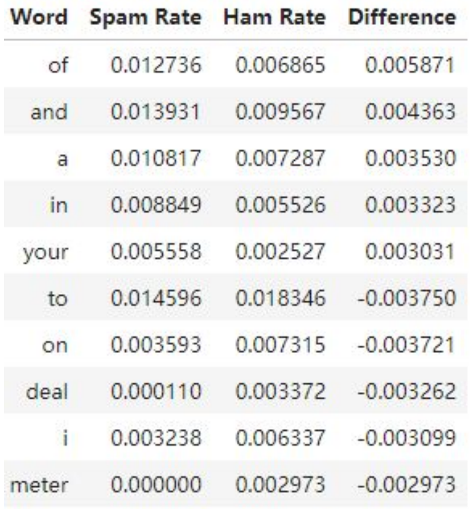
Artificial (Feed-Forward) Neural Network: 90% to 99% (using different datasets). [1]

Feature Selection

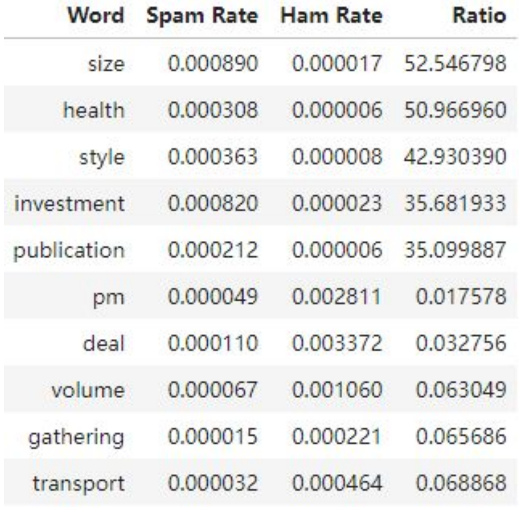
Most significant occurrence rate difference (5 positive + 5 negative)

Most significant occurrence rate ratio for words appeared more than 5 times (5 largest + 5 smallest)

3 Engineered Features:
Special character ratio,
Unrecognized word ratio,
Re-recognized word ratio after removing all special characters



Word	Spam Rate	Ham Rate	Difference
of	0.012736	0.006865	0.005871
and	0.013931	0.009567	0.004363
a	0.010817	0.007287	0.003530
in	0.008849	0.005526	0.003323
your	0.005558	0.002527	0.003031
to	0.014596	0.018346	-0.003750
on	0.003593	0.007315	-0.003721
deal	0.000110	0.003372	-0.003262
i	0.003238	0.006337	-0.003099
meter	0.000000	0.002973	-0.002973



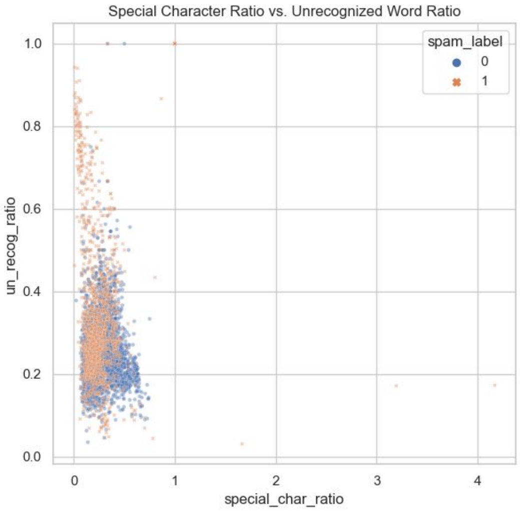
Word	Spam Rate	Ham Rate	Ratio
size	0.000890	0.000017	52.546798
health	0.000308	0.000006	50.966960
style	0.000363	0.000008	42.930390
investment	0.000820	0.000023	35.681933
publication	0.000212	0.000006	35.099887
pm	0.000049	0.002811	0.017578
deal	0.000110	0.003372	0.032756
volume	0.000067	0.001060	0.063049
gathering	0.000015	0.000221	0.065686
transport	0.000032	0.000464	0.068868

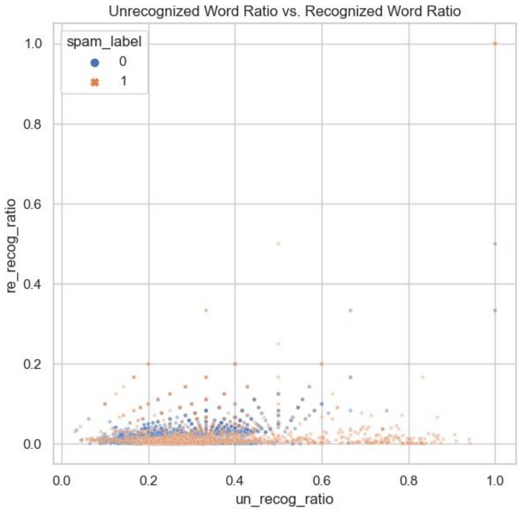
Feature Selection

Most significant occurrence rate difference (5 positive + 5 negative)

Most significant occurrence rate ratio for words appeared more than 5 times
(5 largest + 5 smallest)

3 Engineered Features:
  Special character ratio,
  Unrecognized word ratio,
  Re-recognized word ratio after removing all special characters





Multiple Linear Regression Results

Top appearance rate difference		**Top appearance rate ratio**	
of	-0.00223	size	0.01699
and	0.00906	**health**	**0.22850**
a	0.02359	style	-0.02158
in	0.00217	**investment**	**0.08626**
your	0.03869	**publication**	**0.07130**
to	-0.01231	pm	-0.01542
on	-0.02590	volume	-0.02231
deal	**-9.24216**	**gathering**	**-0.08056**
in	-0.00248	transport	-0.01160
meter	-0.00218	**deal (duplicate)**	**-9.24216**

Engineered specs		
	special_char_ratio	**-0.22852**
	un_recog_ratio	**0.92008**
	re_recog_ratio	**0.19200**

Results

Threshold	0.5	0.6
Accuracy	77.9%	75.9%
False Positive Rate	20.5%	12.5%

SVM Results

Activation function	Linear	Poly	RBF
Accuracy	78.7%	72.6%	79.5%
False Positive Rate	23.8%	22.2%	18.6%

Neural Network results

Threshold: 0.5

(Layer, Cells)	(2, [512,256])	(2, [128,64])	(1, [512])
Accuracy	85.7%	85.2%	84.5%
False Positive Rate	20.7%	21.3%	18.3%

Threshold: 0.6

(Layer, Cells)	(2, [512,256])	(2, [128,64])	(1, [512])
Accuracy	86.8%	85.6%	84.9%
False Positive Rate	17.2%	17.5%	20.1%

Future possibilities

Use more features to achieve higher performance

Implement NLP methods

Reference

[1] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019

[2] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa, Machine learning for email spam filtering: review, approaches and open research problems, Heliyon, Volume 5, Issue 6, 2019