How Your Security Aggressiveness Affects the Accuracy of Threat Detection

Suppose a firm hires you to create a new spam filter.
Parameters to the problem: You classify e-mail messages into 4 types:

80% are "suspicious", which have a 95% chance of being spam
5% are "questionable", which have a 50% chance of being spam
5% are "possible", which have a 10% chance of being spam
10% are "safe", which have a 0.1% chance of being spam

Then, you have 3 choices of an aggressiveness policy based on your risk thresholds:

"Least aggressive"	flag only the suspicious
"Medium aggressive"	flag the suspicious and questionable
"Most aggressive"	flag everything except the safe

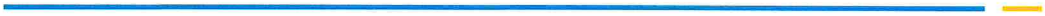
The firm receives 1 million messages a day. What is the effect of the classification & aggressiveness levels on the number of false positives and false negatives perceived by the company?

False positive =	Good traffic stopped
False negative =	Bad traffic let in

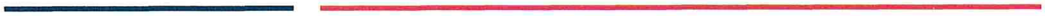
Notice how the proportions increase or decrease depending on the aggressiveness.
More aggressive --> more false positives, fewer false negatives

Least aggressive:

1.3% False positives

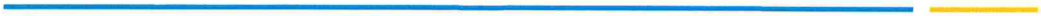


24.2% False negatives

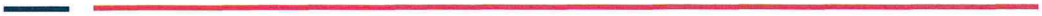


Medium aggressive:

8.0% False positives

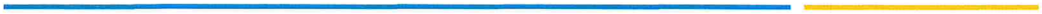


4.2% False negatives

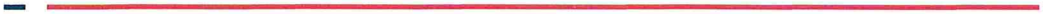


Most aggressive:

20.0% False positives



0.2% False negatives



Perfect system would have:

0% False positives

0% False negatives

INPUT

Total number of cases	1,000,000

The 4 categories of varying degrees of suspicion.	Change first row of values to vary incidence.			
Category	Suspicious	Questionable	Possible	Safe
Proportion classified as such	0.2	0.1	0.1	0.6
Proportion within category that are indeed bad	0.95	0.5	0.1	0.001

CALCULATIONS

	Suspicious	Questionable	Possible	Safe	TOTAL
Total in category	200,000	100,000	100,000	600,000	1,000,000
Those actually bad	190,000	50,000	10,000	600	250,600
Those actually good	10,000	50,000	90,000	599,400	749,400

Least aggressive	Flagged	Not Flagged
Actually bad	190,000	60,600
Actually good	10,000	739,400

Medium aggressive	Flagged	Not Flagged
Actually bad	240,000	10,600
Actually good	60,000	689,400

Most aggressive	Flagged	Not Flagged
Actually bad	250,000	600
Actually good	150,000	599,400

RESULTS

What proportion of legitimate (good) cases were flagged?				**"False positive"**
Least aggressive	10,000	out of	749,400 =	**0.0133**
Medium aggressive	60,000	out of	749,400 =	**0.0801**
Most aggressive	150,000	out of	749,400 =	**0.2002**
What proportion of harmful (bad) cases escaped detection?				**"False negative"**
Least aggressive	60,600	out of	250,600 =	**0.2418**
Medium aggressive	10,600	out of	250,600 =	**0.0423**
Most aggressive	600	out of	250,600 =	**0.0024**
What proportion of the flagged cases were not harmful?				**similar to false positive**
Least aggressive	10,000	out of	200,000 =	**0.0500**
Medium aggressive	60,000	out of	300,000 =	**0.2000**
Most aggressive	150,000	out of	400,000 =	**0.3750**
What proportion of the nonflagged cases were in fact harmful?				**similar to false negative**
Least aggressive	60,600	out of	800,000 =	**0.0758**
Medium aggressive	10,600	out of	700,000 =	**0.0151**
Most aggressive	600	out of	600,000 =	**0.0010**
What proportion of legitimate (good) cases survived scrutiny?				**true negative: opposite of false positive**
Least aggressive	739,400	out of	749,400 =	**0.9867**
Medium aggressive	689,400	out of	749,400 =	**0.9199**
Most aggressive	599,400	out of	749,400 =	**0.7998**
What proportion of harmful (bad) cases were flagged?				**true positive: opposite of false negative**
Least aggressive	190,000	out of	250,600 =	**0.7582**
Medium aggressive	240,000	out of	250,600 =	**0.9577**
Most aggressive	250,000	out of	250,600 =	**0.9976**