Insper Instituto de Ensino e Pesquisa

# “Fighting spam with statistics”

**Royal Statistical Society**

**Por Joshua Goodman and David Heckerman**

**Disciplina de Inteligência Artificial, Prof. F. Barth**

**Feito por Carolina Hirschheimer, Natália Carreras e Rodrigo Nigri**

Spam é um e-mail que chega ao usuário sem que ele tenha solicitado ou considere a hipótese de recebê-lo. Ele é usado com fins comerciais para fazer pessoas adquirirem algum produto ou serviço, podendo ser também um meio para disseminação de golpes e informações falsas. Como mostrado no artigo "*Fighting spam with statistics",* o spam é um problema de longa data. De acordo com a *Brightmail*, por custar apenas 0.01¢/spam, várias empresas e pessoas optam por utilizar esse recurso, de modo que 50% dos e-mails na internet em 2004 eram spam. Em julho de 2021, cerca de 84% dos e-mails foram relatados como spam. Desses e-mails, 36% promoviam algum produto ou serviço, 32% possuíam conteúdos adultos, 27% eram sobre assuntos financeiros e 3% eram golpes ou fraudes, de acordo com a *Spam Laws*.

Diante disso, empresas começaram a tentar bloquear o spam. O Hotmail, no início dos anos 2000, percebeu que muitos spams eram enviados com assunto "*From Hotmail*" ou, genericamente, "*From Empresa*" e tentou bloquear e-mails que começavam com "*From"*. Porém, no processo, perceberam que estavam bloqueando muitos falsos positivos, como e-mails de aniversário, que tinham como assunto "*From Susan*".

Nesse sentido, para contabilizar não só uma palavra, mas todo o e-mail na análise, pode-se utilizar o classificador Naive-Bayes. Naive-Bayes é um modelo estatístico de machine learning que é capaz de prever se um e-mail é spam, com base na mensagem. Estatisticamente, o algoritmo é escrito como: $P(spam|mensagem) = \frac{P(mensagem|spam) \times P(spam)}{P(mensagem)}$. De modo geral, isso significa que a chance de um e-mail ser spam é igual à chance de uma pessoa receber uma das bilhões de mensagens de spam possíveis vezes a probabilidade do spam ocorrer, sobre a probabilidade da mensagem ocorrer.

Na prática, o Naive-Bayes é implementado em algumas etapas: primeiro, deve-se classificar bases de treinamento e teste manualmente. Em seguida, roda-se o Naive-Bayes para essa base de treinamento e, após ensinar o algoritmo, pode-se testar o classificador em uma base de dados de teste. Então, verifica-se a acurácia e precisão do classificador e são realizadas iterações para melhorar essas métricas. Por fim, pode-se utilizar o classificador para verificar se os e-mails são spam ou não.

O classificador Naive-Bayes, no entanto, é suscetível ao erro em algumas situações. A primeira é quando há frases comuns em e-mails verídicos e de spam, como "clique aqui para se inscrever". Ademais, se determinada palavra acidentalmente tende a aparecer mais em um e-mail de spam, esse efeito pode ser potencializado no modelo preditivo.

Em face dessas questões, outras soluções foram criadas para combater o spam. Uma delas são os computing puzzles. O conceito por trás dos computing puzzles sustenta-se no fato de que, a fim de verificar se os e-mails são legítimos, são enviados puzzles para bloquear a caixa de e-mail do remetente. O programa de e-mail passa ~15 segundos resolvendo o problema e, sem você ver, resolve no fundo o problema em microssegundos, enviando a resposta de volta para o destinatário do e-mail.

Para alguém que está disseminando spams e costuma enviar milhões de e-mails por dia, 15 segundos limitam o envio a ~6 mil e-mails diários, tornando-se essenciais no combate. Um dos problemas do uso de computational puzzles, todavia, é que essa estimativa de 15 segundos está extremamente sujeita a variações, podendo levar, na prática, de 1 segundo a 1 minuto para resolução desses problemas.

Em suma, ambas metodologias possuem benefícios e debilidades. Ainda, é crucial buscar soluções para o spam, que não apenas motiva usuários a deixarem de usar e-mails, mas também é um mecanismo para orquestração de diversos crimes.

# Referências Bibliográficas


**[1]** GOODMAN, Joshua ; HECKERMAN, David. **Fighting spam with statistics.** Significance, v. 1, n. 2, p. 69–72, 2004.

**[2]** MURPHY, Kevin. **Naive Bayes classifiers.** [s.l.: s.n., s.d.]. Disponível em: <https://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf>. Acesso em: 15 nov. 2022.

**[3] Encyclopedia of Bioinformatics and Computational Biology.** Google Books. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=rs51DwAAQBAJ&oi=fnd&pg=PA403&dq=naive+bayes&ots=q__06bPrQT&sig=5nBOdoQVmJ3LCE2JVE5x6akgkJE#v=onepage&q=naive%20bayes&f=false>. Acesso em: 15 nov. 2022.

**[4] Spam e-mail traffic share 2021 | Statista**. Statista. Disponível em: <https://www.statista.com/statistics/420391/spam-e-mail-traffic-share/ >. Acesso em: 15 nov. 2022.

**[5] Global average daily spam volume 2021 | Statista**. Statista. Disponível em: <https://www.statista.com/statistics/1270424/daily-spam-volume-global/ >. Acesso em: 15 nov. 2022.

**[6] Spam Statistics and Facts**. Spamlaws. Disponível em: <https://www.spamlaws.com/spam-stats.html >. Acesso em: 15 nov. 2022.

**[7]** TSANG, Patrick ; SMITH, Sean. **Combating Spam and Denial-of-Service Attacks with Trusted Puzzle Solvers.** [s.l.: s.n., s.d.]. Disponível em: <https://www.cs.dartmouth.edu/~sws/pubs/ts08a.pdf>. Acesso em: 15 nov. 2022.