Relatório de pesquisa

04 de Junho de 2022 | Desafio Data Science

1 Introdução

Um problema comum a usuários conectados à redes de telecomunicações, são contatos indesejados, conhecidos popularmente como SPAM (acrônimo derivado da expressão em inglês “Sending and Posting Advertisement in Mass”, que em tradução literal significa “enviar e postar publicidade em massa”). Tal incômodo, afeta tanto clientes domiciliares, quanto corporativos e organizacionais.

Um estudo realizado no período de outubro de 2020 à setembro de 2021, avaliou globalmente o volume de SPAM circulando via e-mail. Os dados assustam: a média diária de tráfego desse tipo de mensagem é de aproximadamente 165,13 bilhões [5]. A mesma pesquisa aponta que no mês de julho houve um recorde de quase 283 bilhões de mensagens diárias, sendo os Estados Unidos o país com maior veiculação.

A maior preocupação em relação aos SPAM são possíveis consequências, como fraudes, ataques virtuais e clonagens de aparelhos. Uma das possíveis técnicas envolvidas para fins maliciosos é o *phishing*, no qual o Brasil é líder de ocorrências reportadas [6], seguido de França e Portugal. Além dos e-mails, outras fonte de disseminação de SPAM e fraudes são os aplicativos de mensagens/redes sociais [1], bem como o contato via linhas telefônicas [2].

Diante deste contexto, pesquisadores reúnem esforços para evitar que tais ameaças prossigam se alastrando através das redes. Uma opção possível para neutralizar infortúnios decorrentes de SPAM, é filtrar as comunicações recebidas por um usuário, classificando-as como SPAM ou mensagens/ligações comuns. Considerando-se a taxa de acurácia dos sistemas de segurança mais robustos e modernos do mercado, esta estratégia tem se mostrado promissora na redução de riscos e danos aos usuários destas redes [3].

Assim, este artigo visa apresentar uma breve análise exploratória a respeito de um conjunto de dados fornecido, relacionado à mensagens textuais. Além disso, é relatado o desenvolvimento de um modelo de classificação das mensagens, cujo principal objetivo é filtrá-las entre as classes “SPAM” e “não-SPAM”.

2 Desenvolvimento

É muito relevante conhecermos o conjunto de dados com o qual estamos trabalhando, extraindo características, informações, padrões e *insights*. O arquivo disponibilizado para análise possui 5574 mensagens textuais, sendo 4827 exemplares de mensagens comuns e 747 de SPAM. As mensagens foram submetidas a uma etapa de mineração de texto, onde foram destacadas as frequências das 149 palavras mais recorrentes. Além disso, são disponibilizadas colunas contendo a quantidade total de palavras frequentes na mensagem, a quantidade total de palavras da mensagem, a data de recebimento da mensagem e a classe da mensagem (se é ou não SPAM).

Dessa forma, com o objetivo de expressar visualmente as palavras mais frequentes em toda a base de dados, criou-se uma nuvem de palavras (Figura 1). A primeira etapa para sua construção, foi obter a frequência total de cada termo. Depois, carregou-se uma máscara de imagem em formato de carta (simbolizando e-mail) e através do módulo WordCloud, da linguagem de programação Python, gerou-se a figura.

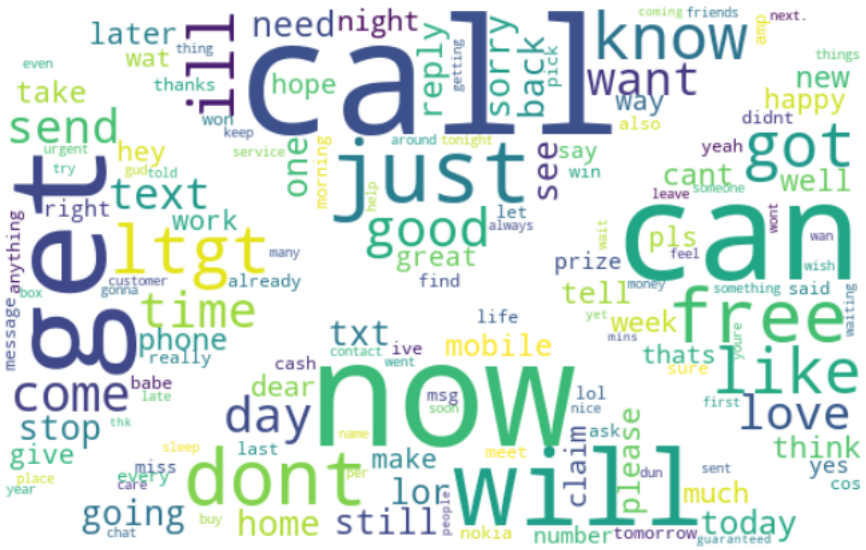

Figura 1: Nuvem de palavras com os termos mais frequentes em toda a base de dados

Também, traçaram-se gráficos para expressar a relação entre as quantidades de mensagens SPAM e não-SPAM mensalmente. O documento contempla registros de janeiro, fevereiro e março, com percentuais de SPAM de 13.3%, 13.9% e 12.7%, respectivamente. Além do gráfico de barras agrupadas anexado à este relatório (Figura 2), também foram gerados três gráficos de setores (pizza/pie), onde se observa com clareza a relação entre os volumes de mensagens de cada classe. Todas estas quatro visualizações gráficas foram dadas através da biblioteca Matplotlib, da linguagem de programação Python.

Em uma terceira etapa de análise, calcularam-se medidas de estatística descritiva (o máximo, o mínimo, a média, a mediana, o desvio padrão e a variância) acerca da quantidade total de palavras para cada um dos meses. Os resultados estão resumidos na Tabela 1. Em uma etapa final das ações analíticas, determinou-se o dia de cada mês que possui a maior sequência de mensagens não-SPAM, sendo em janeiro o dia 26, em fevereiro o dia 04 e em março o dia 31.

Finalmente, desenvolveu-se uma metodologia capaz de classificar automaticamente as mensagens como ”SPAM”ou ”não-SPAM”. Para tanto, utilizaram-se como ferramentas a linguagem de programação Python e a biblioteca scikit-learn. Uma vez que o conjunto de dados já estava minerado, estruturado e as palavras mais frequentes contadas, optou-se pela implementação do algoritmo de

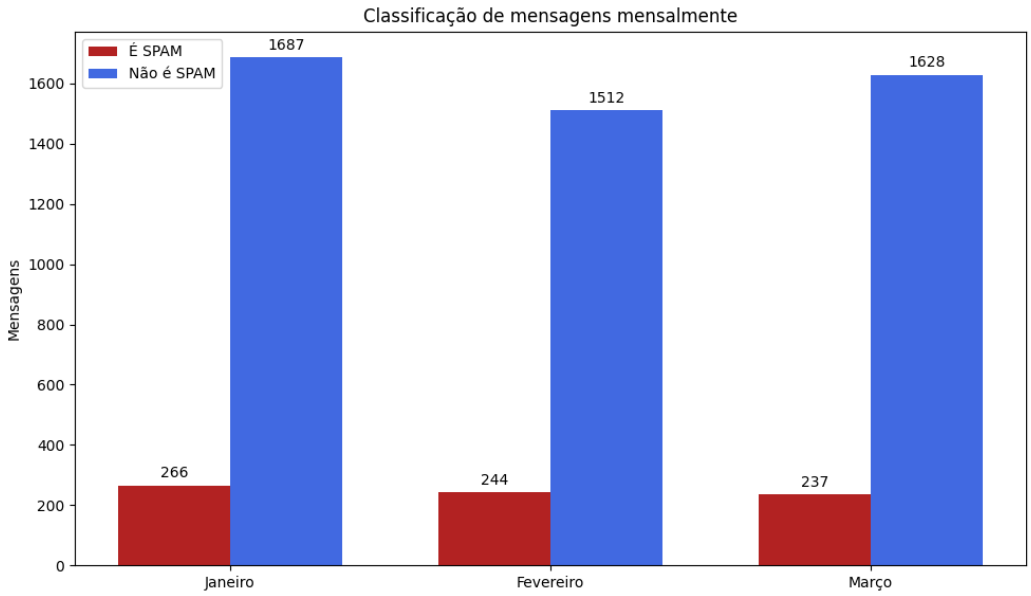

Figura 2: Classificação de mensagens ao longo dos meses

classificação Naïve-Bayes com modelo Multinomial.

Os experimentos foram executados de acordo com a seguinte metodologia: retiveram-se aleatoriamente 30% das observações e classes do conjunto total de dados, para testes, enquanto os outros 70% compuseram um conjunto de treino. O modelo foi treinado com o conjunto de treino e realizou predições a partir do conjunto de testes. As classes preditas pelo modelo foram confrontadas com as rotulações originais do conjunto de testes previamente retido, donde extraíram-se as métricas avaliativas. Para evitar resultados tendenciosos, performaram-se 50 replicatas, com *seeds* de randomização na etapa de retenção variando entre 1 e 51. Dessa forma, todos as configurações de conjuntos treino/teste foram distintas. A acurácia média do modelo entre os 50 testes foi de 95.85%.

Todas as implementações realizadas durante o decorrer deste trabalho, juntamente com os gráficos e os resultados completos estão disponíveis em um repositório do github [4].

3 Conclusões

Ainda há muito o que avançar em relação à privacidade, segurança e confiabilidade em redes de telecomunicações. Ao disponibilizar serviços aos usuários, é de obrigação dos fornecedores que tais quesitos sejam de fato, atendidos. Assim, mais do que nunca, em um contexto global onde informações se espalham intercontinentalmente com velocidades inimagináveis, o investimento em novas tecnologias anti-SPAM torna-se imprescindível.

Tabela 1: Medidas estatísticas em relação à quantidade total de palavras para cada mês

	Janeiro	**Fevereiro**	**Março**
Máximo	190	100	115
Mínimo	2	2	2
Média aritmética	16.3	16.0	16.2
Mediana	13.0	13.0	12.0
Desvio padrão	12.5	11.0	11.5
Variância	157	121	134

Neste breve artigo relatou-se o processo de análise exploratória realizada em um conjunto de mensagens textuais, onde gerou-se uma nuvem de palavras com os termos mais frequentes da base de dados, traçaram-se gráficos comparativos entre as mensagens comum e SPAM para cada mês, calcularam-se medidas estatísticas referentes à quantidade total de palavras em mensagens para cada mês e determinou-se o dia de cada mês em que houve a maior sequência de mensagens comuns. Além disso, desenvolveu-se e avaliou-se um modelo de aprendizado de máquina, capaz de classificar mensagens como SPAM ou não-SPAM com acurácia superior a 95%.

Referências

[1] Dácio Castelo Branco. Golpes de phishing utilizam plataforma de blog do telegram para enganar usuários. https://canaltech.com.br/seguranca/golpes-de-phishing-utilizam-plataforma-de-blog-do-telegram-para-enganar-usuarios-217772/, Jun 2022. Acesso em: 03 de Jun. de 2022.

[2] Nilton Kleina. Brasil foi país mais afetado por ligações de spam em 2021. https://www.tecmundo.com.br/mercado/230768-brasil-pais-afetado-ligacoes-spam-2021.htm, Dec 2021. Acesso em: 03 de Jun. de 2022.

[3] Neil Kumaran. Spam does not bring us joy—ridding gmail of 100 million more spam messages with tensorflow. https://cloud.google.com/blog/products/g-suite/ridding-gmail-of-100-million-more-spam-messages-with-tensorflow, Fev 2019. Acesso em: 03 de Jun. de 2022.

[4] Felipe F. Lorenci. Desafio senior. https://github.com/lorencifelipe/desafioSenior, Jun 2022. Acesso em: 03 de Jun. de 2022.

[5] Statista. Average daily spam volume worldwide from october 2020 to september 2021. https://www.statista.com/statistics/1270424/daily-spam-volume-global/, Apr 2022. Acesso em: 03 de Jun. de 2022.

[6] Statista. Countries most targeted by phishing attacks worldwide in 2021. https://www.statista.com/statistics/266362/phishing-attacks-country/, May 2022. Acesso em: 03 de Jun. de 2022.