DESAFIO SENIORLABS

Victor Accete Nicácio Placido

1 Introdução

A quantidade de spams que recebemos é enorme. Estima-se que cerca de 122.33 bilhões de emails são spams (Cvetićanin), isso sem contar os spams via mensagens de texto.

Para combater esse problema, são empregadas técnicas de filtragem de spam (spam filtering). Filtros de spam detectam mensagens indesejadas e automaticamente a classificam como spam, para que o sistema em que o filtro está inserido decida o que fazer com ele. Por exemplo, no caso de emails, há uma caixa específica para spams.

Diante disso, foi me proposto o desafio de analisar um dataset de mensagens de texto (contendo spams e não spams) para analisar e para desenvolver um modelo de classificação para filtrar as mensagens de spam.

Foram propostas as seguintes tarefas:

- 1. Exibir gráfico as palavras mais frequentes em toda a base de dados
- 2. Exibir gráfico com as quantidades de mensagens comuns e spams para cada mês;
- 3. Calcular o máximo, o mínimo, a média, a mediana, o desvio padrão e a variância da quantidade total de palavras para cada mês;
- 4 Exibir o dia de cada mês que possui a maior sequência de mensagens comuns (não spam).

Além disso, no final, foi proposto que eu aplicasse um modelo de machine learning para classificar automaticmente as mensagems como "comum" ou "spam" e dissertar brevemente sobre os resultados.

2 Desenvolvimento

O ambiente de desenvolvimento escolhido para esse desafio foi o Colab, devido à celeridade de já ter tudo previamente instalado, tornando um ambiente adequado para prototipações. O Colab pode ser acessado através deste link: https://colab.research.google.com/drive/1RV54KdQLA_kOv9PXXahRRI_rfRBRBtvU?usp=sharing.

Para ler o arquivo, foi necessário mudar a codificação para utf-8, antes de subir para o Colab. Para isso, abri o arquivo .csv fornecido no editor de textos Sublime Text 3 e então

salvei com a codificação utf-8. Dessa forma, consegui abrir o arquivo sem configurações adicionais no código.

Após uma breve exploração inicial do arquivo, foi dado início à execução dos desafios propostos. Pudemos analisar as palavras mais frequentes no dataset e verificamos conforme a tabela 1.

Um comparativo mais detalhado das palavras mais frequentes pode ser conferido na imagem 1.

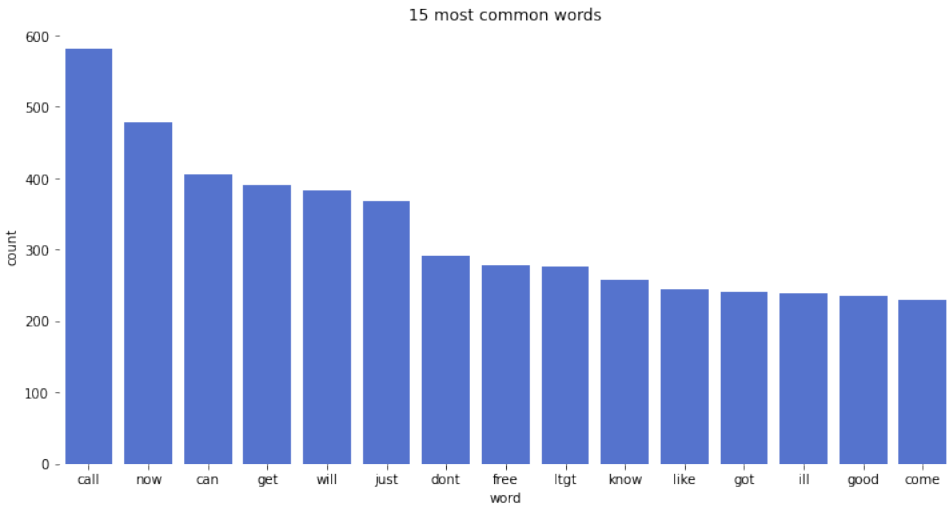


Figure 1: 15 palavras mais comuns no dataset.

Table 1: 5 palavras mais frequentes.

Palavra	Contagem
call	581
now	479
can	405
get	390
will	383

Em seguida, verificamos as quantidades de mensagens (spams ou comuns) para cada mês. E, conforme a figura 2, verificamos que não há uma diferença significativa entre mensagens comuns e spams nos três primeiros meses do ano, que foram os meses disponibilizados no dataset.

Em seguida, calculamos algumas métricas em relação à contagem de palavras de cada mensagem. Verificamos as métricas expressas na tabela 2

Por fim, verificamos a maior sequência de mensagens que não era spam para cada mês e

Figure 2: Mensagens por mês. Comuns (azul) e Spam (laranja).

Table 2: Métricas das contagens de palavras das mensagens.

Métrica	Janeiro	Fevereiro	Março
Máximo	190	100	115
Mínimo	2	2	2
Média	16.34	16.03	16.29
Mediana	13	13	12
Desvio padrão	12.56	11.04	11.58
Variância	157.68	121.94	134.01
Total	31906	28147	30372

verificamos o resultado da tabela 3.

Table 3: Maiores sequências de mensagens comuns por mês.

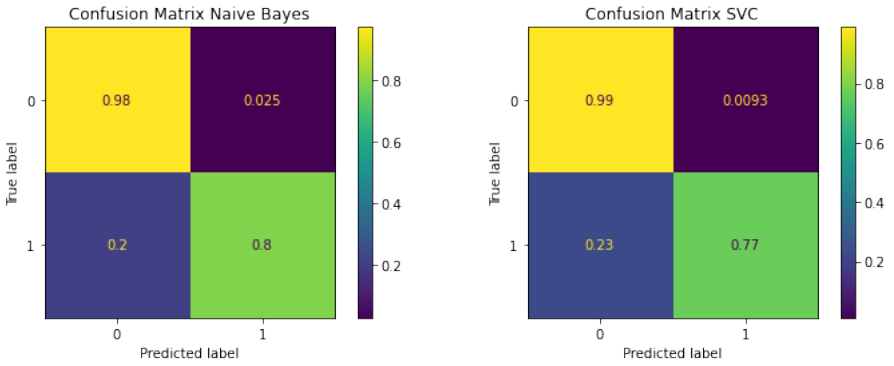
Mês	Sequência	Dia da ocorrência
Janeiro	31 mensagens	26/01
Fevereiro	39 mensagens	04/02
Março	46 mensagens	31/03

Por fim, para a segunda etapa. Realizamos a classificação. Escolhemos dois modelos para testar: o SVM, que com o Kernel linear é conhecido por ser bom em classificação de textos (KOWALCZYK); e o Naive Bayes Multinomial, que apesar de não poder ser usado em problemas de regressão, é útil para classificação. Os resultados com diferentes métricas podem ser conferidos na tabela 4.

Table 4: Avaliação dos modelos.

Algoritmo	Precisão	Acurácia	F1
MNB	0.796	0.951	0.815
SVM	0.769	0.960	0.842

As matrizes de confusão podem ser vistas na imagem 3.



(a) Naive Bayes (b) SVM

Figure 3: Matrizes de confusão.

3 Conclusões

Trabalhar com dados sanitizados de antemão ajudou bastante no processo, principalmente no momento da classificação. Em relação aos resultados obtidos, pudemos observar que não há uma diferença sensível no comportamento das mensagens nos três primeiros meses do ano.

Uma alternative de análise seria verificar quais palavras são mais comuns em mensagens de spam e quais são mais comuns em mensagens comuns. No entanto, isso não foi feito para este trabalho e pode ser sugestão de trabalho futuro.

Em relação à classificação, os dois modelos apresentaram boa acurácia, mas acredito que a melhor métrica nesse caso seria a precisão, para minimizar a quantidade de falsos positivos. A precisão, no entanto, não demonstrou resultados tão bons quanto as demais métricas. No caso de emails e mensagens, um falso positivo pode ser extremamente prejudicial para o cliente. Dessa forma, a prioridade deveria ser minimizar ao máximo o número de falsos positivos.

References

[Cvetićanin] Cvetićanin, N. What's on the other side of your inbox – 20 spam statistics for 2021. *https://dataprot.net/statistics/spam-statistics/*.

[KOWALCZYK] KOWALCZYK, A. Linear kernel: Why is it recommended for text classification? *https://www.svm-tutorial.com/2014/10/svm-linear-kernel-good-text-classification/*.