Senior Labs - Análise de Spams

Gabriela Yukari Kimura
gykimura10@gmail.com

Introdução

A simplicidade e o baixo custo de utilização dos e-mails permitiram que empresas e organizações envolvessem os seus trabalhos na Internet, utilizando o e-mail como forma de comunicação com os seus clientes e parceiros, realização de processos de recrutamento, e a utilização deste ambiente para o recebimento de reclamações e/ou sugestões. No entanto, o spam, caracterizado como mensagens comerciais e de conteúdo indevido, tem sido uma ferramenta importante para criminosos realizarem atividades ilegais na Internet, como o furto de informações confidenciais, venda de produtos falsificados e a distribuição de malwares. Atualmente, a quantidade de *spams* recebidas diariamente é tamanha que tornou a sua seleção e análise manual impraticável.

Para melhorar a segurança na Internet e auxiliar os usuários que a utilizam, muitos métodos para detecção, análise e investigação de *spams* foram propostos. F. Li et al. [1], em seu artigo, propõe a clusterização de e-mails baseados nas informações de URLs e a quantia em dinheiro mencionada no conteúdo das mensagens. Já Drucker[2], propôs a utilização de uma SVM para a classificação do conteúdo das mensagens relacionados à sua utilização. Ambos apresentaram bom desempenho, mas continuam sujeitos às técnicas maliciosas, que visam dificultar esta detecção.

Neste artigo, são analisados a base de mensagens dos meses de janeiro, fevereiro e março, fornecidas pela empresa Senior Labs. Uma análise exploratória é registrada, e métodos para a classificação e detecção de spams são avaliados.

Metodologia

Para a análise de spams, foi utilizada uma base de dados já fornecida pela Senior Labs. O arquivo csv continha um total de 5574 mensagens, sendo 4827 unidades de mensagens comuns e 747 unidades de mensagens indevidas, os *spams*. Além disso, as mensagens já haviam sido submetidas à uma etapa de mineração de texto, apresentando as principais palavras de cada mensagem, em colunas individuais que apontam quando foram utilizadas, e outras três colunas, com as informações de quantidade total de palavras por mensagem, a data de recebimento e a coluna *IsSpam*, indicando se a mensagem é considerada comum ou um *spam*. Todo o código e análises foram realizadas utilizando *python* versão 3, *jupyter notebook* e visualização interativa com *streamlit*, framework open-source.

Inicialmente, foi realizada a análise das palavras mais frequentes da base de dados. Para isso, utilizando a biblioteca *pandas*, todas as linhas de cada coluna de palavra foram somadas, e as palavras foram ordenadas de maneira decrescente, apresentando as palavras mais utilizadas até as menos frequentes.

Em seguida, foi realizada levantado a quantidade de mensagens comuns e spams recebidas pela Senior Labs, mensalmente. A coluna *Date* foi transformada para *datetime,* e a coluna *IsSpam* foi utilizada para realizar a contagem dos tipos de mensagens, uma vez que apresenta o valor *no* caso seja uma mensagem comum, e *yes*, caso contrário. Além disso, a coluna *Word_Count* foi utilizada para

analisar as estatísticas (mínimo, máximo, média, mediana, desvio padrão e variância) da quantidade de palavras presentes na base de dados.

Por fim, foram levantados os dias de cada mês com maior sequência de mensagens comuns (não *spam*). Para tal, as informações de mês e dia foram extraídas da coluna *Date* e foram geradas duas novas colunas. Isso facilitou o agrupamento por mês e dia, e a contagem das instâncias da coluna *IsSpam* que apresentavam valor *no*, ou seja, a contagem das instâncias de mensagens comuns.

Após realizar estas análises, o dataset foi filtrado, permanecendo somente as informações relevantes para uma modelagem de classificação de mensagens. Foram removidas as colunas contendo o conteúdo completo das mensagens (*Full_Text*) e a Data que foi recebida (coluna *Date*). A coluna IsSpam foi codificada utilizando a método *LabelEncoder* da biblioteca *sklearn*, e foi utilizada como valor de predição, 0 indicando uma mensagem comum, e 1, indicando um *spam*.

O *dataset* apresentava um alto desbalanceamento de classes, com 4827 mensagens comuns e somente 747 mensagens de *spam*. Dependendo da complexidade do problema, o desbalanceamento pode prejudicar o treinamento e a qualidade do classificador. Neste projeto, foram analisados a classificação utilizando o *dataset* desbalanceado e balanceado, após aplicado a técnica de *DownSampling*, que iguala a quantidade de instâncias em relação à classe com menor quantidade. Foram avaliados a precisão e tempo de processamento de quatro métodos de classificação (*Random Forest Classifier, K Nearest Neighbors Classifier, Decision Tree Classifier* e *SVM*), já implementados e disponibilizados pela biblioteca *sklearn*. Foi utilizado 80% das instâncias da base para treinamento do modelo, e o restante, para teste.

Resultados

Através da geração do gráfico de frequência das palavras utilizadas nas mensagens recebidas pela Senior Labs, é possível observar a predominância das palavras *call* e *now* em relação às demais. Entretanto, é necessário considerar que na linguagem inglesa, algumas palavras podem apresentar múltiplos sentidos. Por isso, além da análise de frequência de palavras, seria interessante o estudo do contexto das mensagens em que estas palavras estão inseridas. A figura 1 apresenta as 10 palavras mais frequentes no base de dados de mensagens.

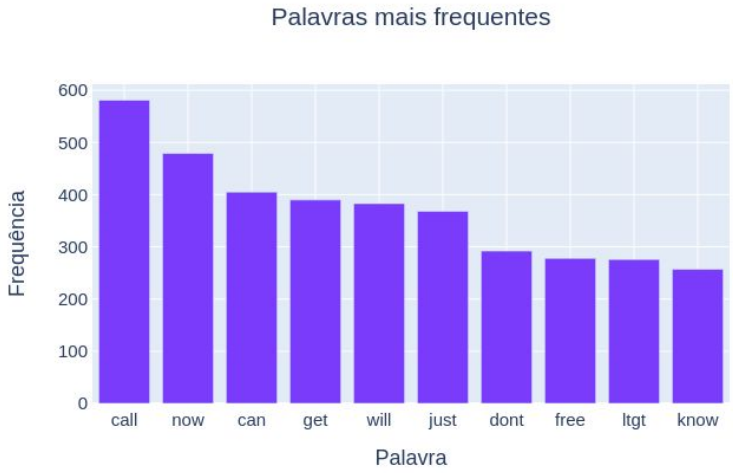

Figura 1. Gráfico com as 10 palavras mais frequentes.

Analisando a quantidade de mensagens comuns e de spam nos meses de janeiro, fevereiro e março, apresentados na figura 2 , é evidente que a maioria delas pertence à classe de mensagens comuns, e que a quantidade de spams é muito próxima entre os meses, representando cerca de 13% do total de mensagens de cada mês. Além disso, a quantidade total de palavras de cada mês também são bastante similares, com uma média de 13 palavras por mês, com o mês de janeiro apresentando uma variância um pouco maior, mas não significativa, em relação ao restante dos meses.

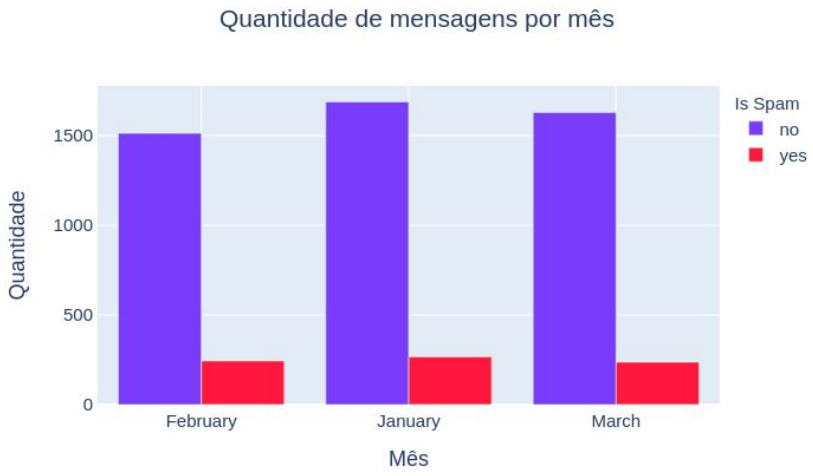

Figura 2. Quantidade de mensagens comuns e spam por mês

A tabela 1 demonstra as estatísticas da quantidade total de palavras utilizadas em cada mês, e a tabela 2 apresenta os dias de cada mês, com uma maior quantidade de mensagens comuns.

	Word_Count					
	min	max	mean	median	std	var
Date						
February	2	100	16.029043	13	11.042459	121.935908
January	2	190	16.336918	13	12.557171	157.682535
March	2	115	16.285255	12	11.576213	134.008715

Tabela1. Quantidade de palavras por mês

	Month	Day	Common_Msg
12	Feb	13	72
28	Jan	1	69
66	Mar	8	69

Tabela2. Dias com predominância comum.

Para a modelagem dos dados para classificação, foi analisado o balanceamento de classes. A figura 3 demonstra o alto desbalanceamento apresentado inicialmente pela base de dados, e a figura 4 apresenta as classes já balanceadas, utilizando a técnica de *Downsampling*. Neste caso, ambas as classes apresentam 747 instâncias, número total de instâncias da classe de *spam*.

Figura3. Distribuição inicial das classes.

Figura4. Distribuição das classes balanceadas.

A tabela 3 apresenta os valores de precisão e tempo de processamento apresentados pelos modelos classificadores, utilizando a base de dados desbalanceada e balanceada, respectivamente.

Modelo	**Base de dados balanceado**		**Base de dados desbalanceado**	
	Precisão	**Tempo(s)**	**Precisão**	**Tempo(s)**
KNN k=5	0.93	0.049	0.78	0.113
SVM	0.93	0.899	0.83	0.010
KNN uniform k=3	0.93	0.044	0.83	0.010
KNN distance k=3	0.94	0.049	0.84	0.010
Decision Tree	0.94	0.041	0.86	0.008
Random Forest	0.96	0.474	0.92	0.175

Tabela 3. Desempenho dos modelos classificadores.

É possível observar que utilizando a base de dados desbalanceada, todos os métodos classificadores apresentaram um alto grau de precisão, uma vez que a maioria das instâncias eram da classe normal. Já na avaliação com uma base de dados balanceada, alguns métodos classificadores tiveram maior dificuldade em aprender o padrão das mensagens e realizar a detecção de spam, apresentando uma redução de quase 10% no valor de precisão. O único método que permaneceu com um score acima de 90% foi o classificador Random Forest, apresentando melhor precisão, mas com um tempo de processamento acima dos demais, mas que não teve influência significativa devido ao tamanho relativamente pequeno da base de dados.

Conclusão

A detecção de spams têm sido objeto de estudo há muito tempo. A complexidade de contextualizar o conteúdo das mensagens e detectar todas as possibilidades de fraude existentes são algumas das dificuldades que envolvem essa área de segurança. No projeto proposto, é possível observar que os métodos classificadores utilizados obtiveram um desempenho relativamente bom. Entretanto, é necessário analisar a complexidade das mensagens envolvidas nesta análise, e realizar novos testes com outros conjuntos de mensagens consideradas como *spam*.

Comparando os resultados apresentados pela classificação utilizando a base de dados balanceada e desbalanceada, é possível observar a influência do balanceamento dos dados à precisão do modelo. Na primeira situação, o treinamento do modelo com uma grande quantidade de mensagens comuns tornou o modelo enviesado, passando a classificar a maioria das mensagens como comum. Com a base balanceada, o modelo, apesar de apresentar uma precisão de classificação reduzida, conseguiu aprender a identificar as palavras que representam um spam, e a classificar as mensagens, como pode ser visto nas porcentagens de erros mais equilibradas.

Como continuidade deste projeto, é sugerido a utilização de ferramentas como PCA para redução de dimensionalidade dos dados (seleção das features mais relevantes), e a hiperparametrização dos modelos de classificação, visando a otimização dos mesmos.

Referências

[1] Alishahi, Mina Sheikh, Mohamed Mejri, and Nadia Tawbi. "Clustering spam emails into campaigns." *2015 International Conference on Information Systems Security and Privacy (ICISSP)*. IEEE, 2015

[2] Drucker, H., D. Wu, and V.N. Vapnik, Support vector machines for spam categorization.Neural Networks, IEEE Transactions on, 1999. 10(5): p. 1048-1054.