Data Science Challenge - Senior Labs

Autora: Açucena Rodrigues

1. Introdução

Os Spams são mensagens de conteúdo indesejado que fazem parte do dia-a-dia de quem utiliza a internet ou recebe mensagens no celular. Essas mensagens são enviadas para grandes números de pessoas e geralmente contém algum conteúdo malicioso ou aquela promoção imperdível. É possível utilizar ferramentas de inteligência que sejam capazes de identificar automaticamente essas mensagens de forma automática de forma que o destinatário não chegue sequer a vê-las.

Esse artigo trata dos resultados obtidos para o desafio proposto pelo Senior Labs. O desafio é composto por duas etapas e o objetivo do mesmo é analisar e entender padrões em mensagens de texto e construir um modelo que permita a classificação dessas mensagens entre spam e não spam. Esse artigo será limitado apenas aos requisitos solicitados na descrição do challenge, entretanto, no notebook de código é possível encontrar algumas outras informações interessantes sobre o dataset e algumas outras alternativas de modelagem.

2. Desenvolvimento

Nas subseções a seguir são apresentados os melhores resultados obtidos a partir das análises feitas para as duas etapas do desafio.

2.1. Primeira Etapa: Análise Exploratória

Para a primeira etapa foi proposta uma exploração nos dados, a fim de entender alguns padrões nas mensagens spam e normais e apresentar as informações mais relevantes. A primeira delas é a frequência das palavras nas mensagens. Abaixo, na Figura 1, é apresentada uma nuvem de palavras baseada na frequência dessas palavras em todas as mensagens do dataset. No notebook, é possível encontrar nuvens de palavras divididas para spams e não spams.

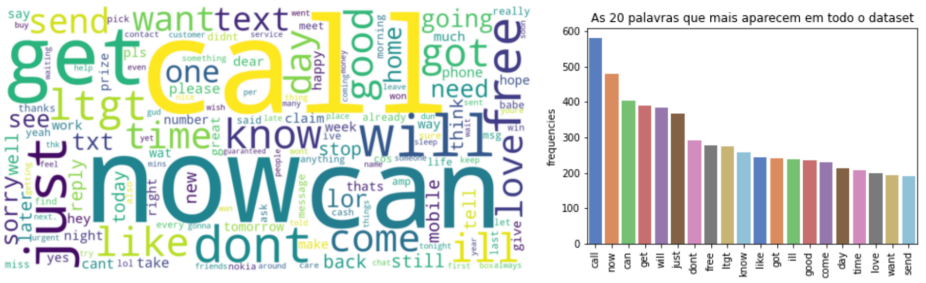
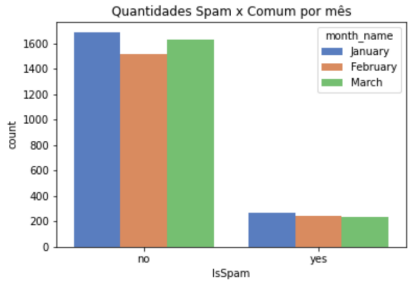


Figura 1. Frequência das palavras no dataset.

Na imagem é possível identificar a palavra "call" como a mais frequente das palavras, seguida da palavra “now” e da palavra “can”. Isso também pode ser visto no gráfico com a frequência das 20 palavras que mais aparecem em todo o dataset. Para mensagens spam, as principais palavras foram “call”, “free” e “now”

Uma outra informação importante a se analisar é a quantidade de mensagens entre spam e não spam. A Figura 2 traz a distribuição de mensagens spam e não spam por mês. A partir dessa imagem é possível perceber que não há muita diferença entre as quantidades de mensagens por mês. O mês de janeiro foi o mês em que mais foram recebidas mensagens spam, entretanto, em uma diferença bem pequena dos outros meses.



Quantidade de mensagens Spam x Comum por mês

month_name	IsSpam	
February	no	1512
	yes	244
January	no	1687
	yes	266
March	no	1628
	yes	237

Figura 2. Quantidade de mensagens spam e não spam por mês.

Um ponto que pode ser destacado a partir dessa imagem é que a classe de mensagens não spam é muito maior que a de mensagens spam, indicando um grande desbalanceamento entre as classes. Entretanto, isso não se mostrou um problema na etapa de modelagem.

Algumas estatísticas foram obtidas a partir dos dados de contagens de palavras nas mensagens recebidas. A média da quantidade de palavras nas mensagens para os 3 meses foi de aproximadamente 16 palavras, com uma mediana de 12 ou 13. Janeiro foi o mês em que foi recebida a maior mensagem, com 190 palavras e a menor mensagem recebida em todos os meses possuía duas palavras. A Figura 3 mostra esses valores.

month_name	February	January	March
min	2.000000	2.000000	2.000000
max	100.000000	190.000000	115.000000
mean	16.029043	16.336918	16.285255
median	13.000000	13.000000	12.000000
var	121.935908	157.682535	134.008715

Figura 3. Estatísticas obtidas para a quantidade de palavras em uma mensagem calculadas por mês.

Abaixo, a Figura 4 mostra o dia do mês com o maior número de mensagens recebidas, sendo fevereiro o maior com 72 mensagens no dia 13. Janeiro e março obtiveram um máximo de 69 mensagens em um dia nos dias 01 e 08, respectivamente.

Dia do mês com mais mensagens

	Mês	Dia	Quantidade
0	February	2017-02-13	72
1	January	2017-01-01	69
2	March	2017-03-08	69

Figura 4. Dia do mês com o maior número de mensagens.

2.2. Segunda Etapa: Modelagem

Para a etapa de modelagem, visando aprender um pouco mais sobre NLP, foram utilizadas algumas técnicas específicas relacionadas a essa área para a criação desse modelo. Foram realizados 3 experimentos diferentes, sendo um utilizando as 149 colunas originais do dataset que continham as frequências das palavras, uma segunda abordagem utilizando TF-IDF, esses dois utilizando o Decision Tree como classificador. E um terceiro experimento utilizando Word2Vec e uma rede neural LSTM como classificador. Aqui serão apresentadas as metodologias aplicadas no terceiro experimento. Esses experimentos estão completos no notebook de análise.

Para a criação dos datasets aqui foram utilizadas apenas as colunas com o texto das mensagens completo e a label utilizada foi a coluna “IsSpam”, como proposto no desafio. A primeira etapa foi o pré-processamento onde os textos passaram por etapas de tokenização, que trata da divisão das frases em suas palavras constituintes, remoção de stopwords, remoção de pontuações, todas as palavras foram passadas para minúsculo, remoção de caracteres alfanuméricos e por fim, por uma etapa de stemming que consiste na remoção da parte final das palavras com o objetivo reduzir essa palavras a uma forma mais primitiva.

Após o pré-processamento, as mensagens foram aplicadas ao Word2Vec para a criação das word embeddings. O Word2Vec cria vetores de tamanho n para representar cada uma das palavras no vocabulário para o problema de forma que palavras semelhantes possuem valores próximos nos vetores, com essa semelhança podendo ser calculado por métricas matemáticas. Esse algoritmo resulta em uma matriz de pesos (word embeddings) para cada palavra.

Esses vetores de peso, juntamente com o conjunto de treino foram aplicados a uma rede neural LSTM para a classificação. Uma rede LSTM é um tipo especial de rede neural pois é capaz de aprender conexões de longo prazo, tendo assim um grande poder de predição. Isso ocorre porque o modelo tenta entender quais são as informações de curto prazo que são importantes e devem ser lembradas e como uma memória de longo prazo afeta no entendimento do texto.

Abaixo são apresentados os resultados obtidos com o modelo utilizando Word2Vec + LSTM para o conjunto de teste:

Acurácia	Acurácia Balanceada	F1	Precisão	Recall	ROC AUC
0.9832	0.9582	0.9366	0.9495	0.9241	0.9582

No notebook de análise é possível ver como o modelo foi construído e os parâmetros de treino, como foi avaliado, as métricas para as classes 0 e 1 e outras informações pertinentes a essa análise.

3. Conclusão

Neste artigo são apresentados os resultados obtidos a partir da análise de um dataset contendo 5574 mensagens SMS rotuladas entre spam e não spam. Os resultados obtidos

foram bem satisfatórios. Inicialmente foi realizada uma análise exploratória a fim de entender o comportamento dos dados.

Na sequência foi desenvolvido um modelo baseado em uma combinação de Word2Vec com uma rede neural Long Short-Term Memory (LSTM) obtendo uma acurácia de 0.98 e um F1-score de 0.93. Os resultados obtidos demonstram que além das boas métricas, a rede consegue classificar bem os valores das classes positiva e negativa, demonstrando que o modelo final obtido é um bom modelo para o problema em questão.