Classificando spams em uma pequena base de dados de mensagens SMS

Renzo A. Viloche Morales
renvmorales@gmail.com

5 de novembro de 2017

1 Introdução

Mineração de texto (processamento de linguagem natural PLN?) é comumente usada com a finalidade de analisar e obter uma descrição geral em dados não-estruturados (por exemplo, um texto único, ou mensagens enviadas a partir de vários usuários). Este tipo de operação não é apenas importante para fins de uma análise exploratória, mas também necessário durante a etapa de pré-processamento que fornecerá dados de entrada para treinar um modelo (algoritmo) de classificação através de uma técnica de aprendizado de máquina.

Neste trabalho, uma pequena base de dados de 5.574 mensagens SMS contendo spams (mensagens consideradas fora do interesse do destinatário) é analisada brevemente e convertida numericamente usando a técnica de tf-idf (?). Uma vez concluído o processamento, a base de dados é dividida em duas partes: uma para "teste" e outra para "treinamento". O objetivo aqui é de avaliar rapidamente a capacidade de classificação para quatro modelos de aprendizado: *naïve-Bayes*, regressão logística, SVM e *Multi-layer Perceptron*. As análises, teste e validação de performance dos modelos são todos realizados usando a linguagem de código aberto Python, especificamente os módulos: *numpy, pandas, scikit-learn, wordcloud, nltk, string*.

2 Metodologia

3 Resultados

3.1 Descrição da base de dados

A base de dados contida no arquivo `SMS.csv` é composta por diversas mensagens comuns (4.827) e spams (747) em inglês dispostos na forma de 5.574 linhas e 154 colunas. Os atributos (cada coluna) da base de dados são descritos a seguir:

- 1 coluna contendo a mensagem original (`Full_Text`);
- 149 colunas com valores inteiros que indicam a frequência de uma determinada palavra na mensagem (`got ... wan`);
- 1 coluna contendo a quantidade de palavras frequentes na mensagem (`Common_Words_Count`);
- 1 coluna contendo a quantidade total de palavras da mensagem (`Word_Count`);
- 1 coluna contendo a data de recebimento da mensagem (`Date`);

- 1 coluna que identifica se a mensagem é spam ou não (`IsSpam`).

Sempre que possível, a base de dados neste trabalho será referida como `SMS`.

3.2 Análise exploratória

Inicialmente foi realizado uma análise das palavras mais frequentes em toda a base. Para isto, as frequências totais de cada uma das 149 palavras mais comuns foram calculadas. O seguinte gráfico de barras na figura 1 exibe cada uma das frequência para palavras com frequência de no mínimo igual a 150.

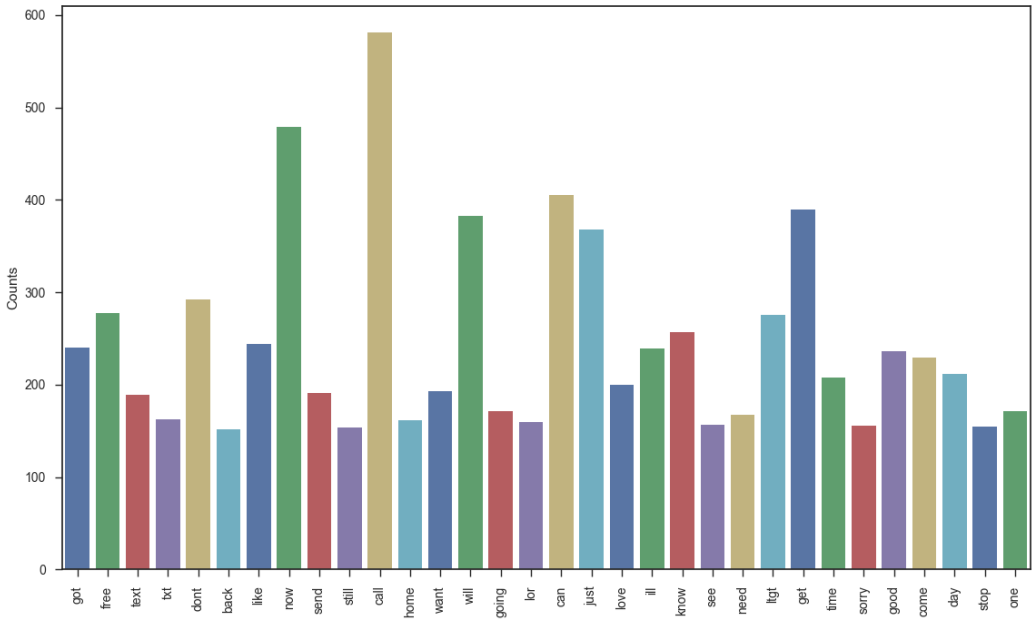

Figura 1: Gráfico de barras com frequências de palavras mais comuns de `SMS` superiores a 150.

O gráfico mostra 32 palavras com ocorrência superior ao mínimo estabelecido (para efeito de melhor visualização). A lista completa destas palavras é: `got, free, text, txt, dont, back, like, now, send, still, call, home, want, will, going, lor, can, just, love, ill, know, see, need, ltgt, get, time, sorry, good, come, day, stop, one`.

Um outro recurso de mais fácil visualização é a nuvem de palavras. Nela, o tamanho de cada palavra é proporcional a sua frequência de ocorrência. Neste caso uma nuvem de palavras foi aplicada sobre o universo de palavras mais comuns. A figura 2 apresenta a nuvem com as 50 palavras de maior ocorrência. Através do recurso de nuvem fica muito mais fácil identificar, por exemplo, que as palavras mais utilizadas foram: `call, can, now, will, get, just`.

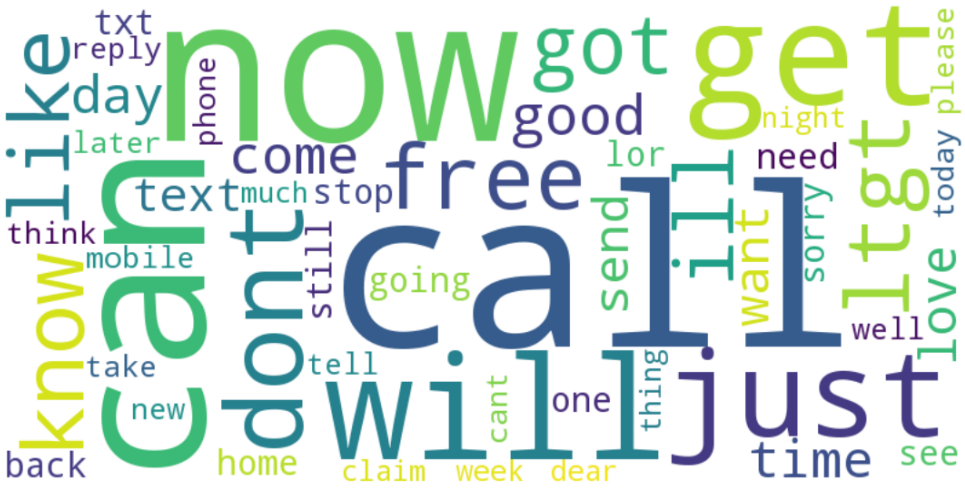

Figura 2: Nuvem com as 50 palavras mais frequentes dentro de um universo de palvras mais comuns de `SMS`.

A seguir, foi feita uma análise da quantidade de mensagens normais e de spam por mês. O gráfico de barras na figura 3 apresenta estas quantidades para os meses de Janeiro, Fevereiro e Março. É possível ver que o número de mensagens classificadas como "spam" é bem reduzido quando comparado ao número de mensagens comuns. O número de spams aparenta posuir mais homogeneidade e uma frequência superior a 200 vezes por mês.

Para o atributo `Word_Count`, um número de estatísticas descritivas foram calculadas: max, min, média, mediana e desvio padrão. A tabela 1 apresenta o cenário encontrado para esta variável.

Tabela 1: Diferentes estatísticas encontradas para o atributo `Word_Count`.

	min.	**max.**	**média**	**mediana**	**desvio padrão**
2017-Jan	2	190	16.34	13	12.56
2017-Fev	2	100	16.03	13	11.04
2017-Mar	2	115	16.28	12	11.58

Observa-se uma característica de dispersão de valores em torno de medidas de centralidade, uma vez que o desvio padrão é comparável aos valores de média/mediana. O fato da média ser um pouco maior que a mediana indica a existência de algumas mensagens muito longas (com muitas palavras). Isto está de acordo pois o valor máximo encontrado para esta variável é sempre superior a 100 palavras, o que acaba deslocando o valor da média nesta direção.

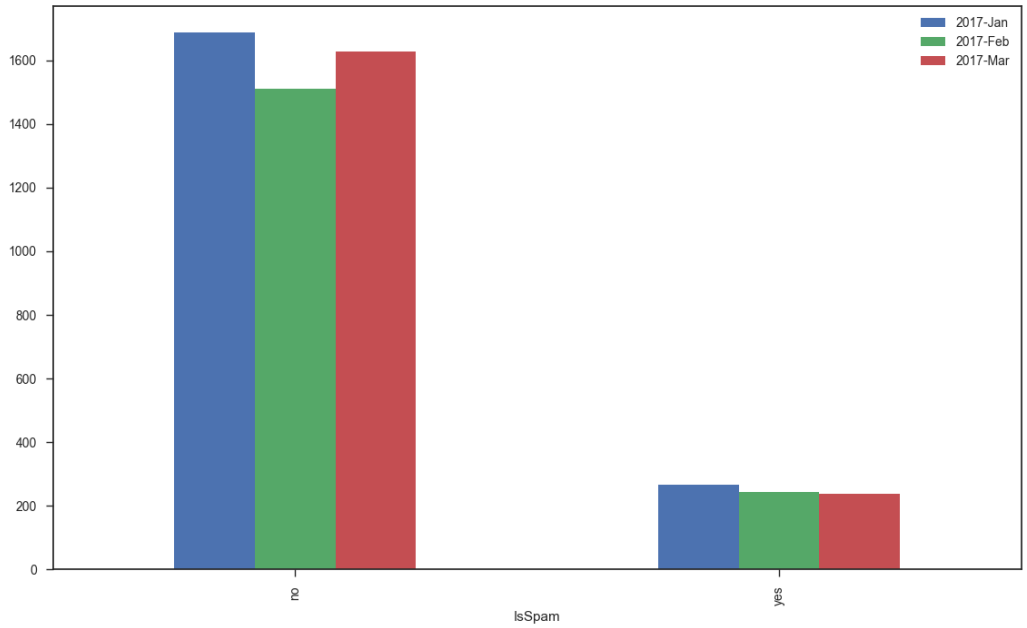

Figura 3: Número mensal de mensagens comuns vs. spams para `SMS`.

3.3 Validação cruzada usando "k-pastas"

A tabela 2 apresenta os valores médios e erro associado das métricas de avaliação de desempenho f1-score, *precision* (precisão) e *recall* (revocação) encontrados aplicando validação cruzada com 10 pastas. A tabela 3 indica os tempos registrados em cada etapa realizado de forma independente. É possível ver que os melhores desempenhos gerais ocorrem para modelos mais complexos (SVM e MLP) com valores do f1-score comparáveis dentro da margem de erro. A complexidade de cada modelo está normalmente associado ao custo computacional que reflete no tempo de cada processo. No entanto, não necessariamente o modelo mais complexo irá apresentar desempenho superior. Isto é visível para o tempo de resposta do modelo de SVM da validação cruzada é muito menor (próximo a 16 segundos) quando comparado ao modelo MLP (acima de 1 minuto).

Tabela 2: Estimativas de métricas de desempenho para os modelos de classificação usados.

	f1-score	**precisão**	**revocação**
Naïve-Bayes	0,840 (± 0,072)	**1,00** (± 0,000)	0.726 (± 0,104)
Regressão logística	0,805 (± 0,069)	0,989 (± 0,031)	0,680 (± 0,095)
SVM	0,942 (± 0,048)	0,987 (± 0,032)	0,901 (± 0,070)
MLP	**0,943** (± 0,027)	0,963 (± 0,046)	**0,924** (± 0,052)

Tabela 3: Tempos registrados para o processo de validação cruzada com 10 pastas.

	f1-score	**precisão**	**revocação**
Naïve-Bayes	8,912 s	8,971 s	8,949 s
Regressão logística	9,228 s	9,076 s	9,073 s
SVM	16,295 s	16,283 s	16,293 s
MLP	**73,721 s**	**74,789 s**	**72,395 s**

4 Conclusão