Desafio Senior - Exercício Pesquisador 2020

Nome: Bruna Fortunato

1 O PROBLEMA

A quantidade de *spams* (mensagens não solicitadas) que recebemos diariamente, não para de crescer. Os tipos de spam são diversos: anúncios de produtos / web sites, esquemas para ganhar dinheiro rápido, correntes, pornografia e etc.

O arquivo `sms_senior.csv` contém vários exemplos de mensagens comuns (4827 unidades) e mensagens *spams* (747 unidades). As mensagens foram submetidas a uma etapa de mineração de texto, com o objetivo de identificar as palavras mais frequentes na base de dados.

2 Breve Análise Exploratória

Para conhecer o dataset, vamos plotar a *Wordcloud* baseado no conteúdo do texto sem tratamento: Full_text. A *Wordcloud* pode ser vista na Figura a seguir.

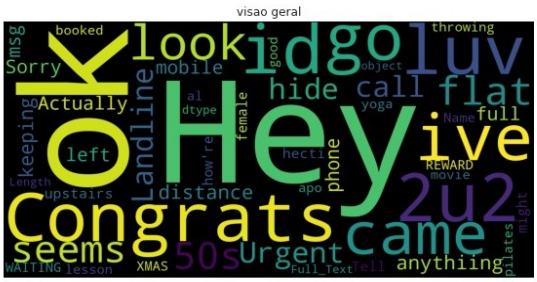

Figura 1: Visão geral do dataset.

Observa-se que existem inúmeras contrações e palavras informais com alta frequência no texto.

2.1 Pré-processamento dos dados

Nesta etapa, realizou-se a filtragem de stopwords utilizando as bibliotecas NLTK e Gensim, disponíveis em Python. Após, realizou-se a implementação de um laço *loop* no qual foram contabilizadas a soma da frequência absoluta de cada palavra. O resultado e exibido conforme a seguir.

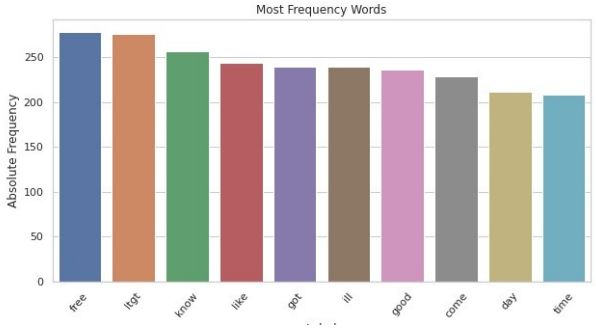

Figura 2: Palavras frequentes.

Observa-se que as labels free, ltgt, know estão entre as mais citadas em todo o dataset. O próximo passo será exibir graficamente o comportamento dessas palavras durante cada mês no qual os dados foram coletados.

2.2 Quantidades de mensagens comuns e spams para cada mês

Nesta etapa, primeiramente foram contados os valores únicos de cada data e desta forma, concluiu-se que que as datas presentes no dataset variam entre janeiro a marco de 2017. Após, realizou-se a plotagem de cada mês de acordo com as Fig.3-5.

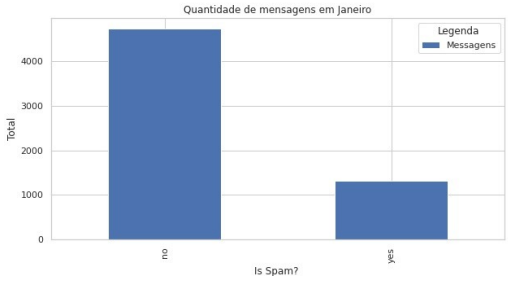

Figura 3: Spams e não Spams em Janeiro.

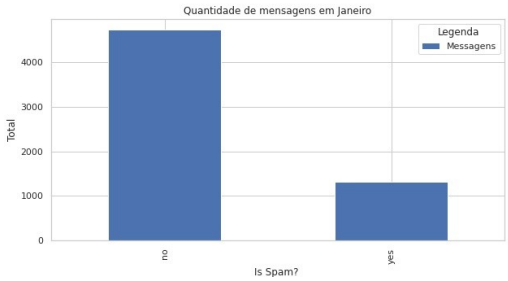

Figura 4: Spams e Não Spams em Fevereiro.

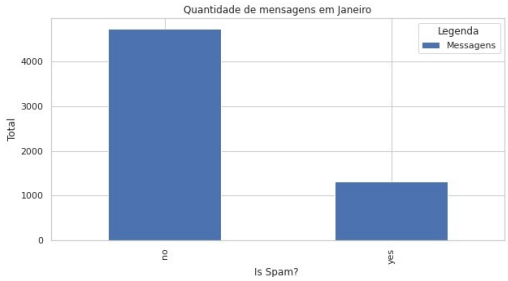

Figura 5: Spams e Não Spams em Março.

Além disso, foram extraídos do dataset os seguintes dados estatísticos :

- Valor máximo de frequência absoluta: 190;
- Valor minemo de frequência absoluta: 2 ;
- Mediana: 16.22;
- Media: 13;
- Desvio Padrão: 11.76.

A maior quantidade de palavras repetidas representa 11.11 , que corresponde a mensagem:

"For me toe love should start with attraction.i should feel that I need her every time around me.she should be the first thing which comes in my thoughts.I would start the day and end it with her.she should be there every time I dream.love will be then when my every breath has her name.my life should happen around her.my life will be named to her.I would cry for her.will give all my happiness and take all her sorrows.I will be ready to fight with anyone for her.I will be in love when I will be doing the craziest things for her.love will be when I don't have to proove anyone that my girl is the most beautiful lady on the whole planet.I will always be singing praises for her.love will be when I start up making chicken curry and end up makiing sambar.life will be the most beautiful then.will get every morning and thank god for the day because she is with me.I would like to say a lot..will tell later.." (18/01/2017).

Para uma melhor visualização, consultar o notebook senio_desafio.ipynb.

2.3 Classificação automática sobre as mensagens como "comum" e "spam"

Nesta etapa, primeiramente eliminou-se as features com informações de classificação e que poderiam causar ruídos para o algoritmo classificador, a saber : *'Date', 'month', 'perc_common_words', 'Common_Word_Count','Word_Count'* . Após, foram aplicados os algoritmos de aprendizagem de máquina para a classificação automática de textos entre Spam ou Não Spam. Os algoritmos escolhidos foram:

- **Naive Bayes :** Classificador probabilístico muito utilizado, baseado no "Teorema de Bayes";
- **Regressão logística:** Classificador que utiliza a função logística para classificação em grupos;
- **Árvore de decisão:** Algoritmo de classificação não linear, amplamente utilizado na aprendizagem supervisionada.

Através dos algoritmos citados, realizou-se a implementação e treinamento dos modelos no qual os resultados são apresentados na sessão a seguir.

Em todos os procedimentos, foram realizados a divisão dos conjuntos de treino e teste entre 70% da base para treino e 30% da base para teste. Além disso, as métricas para avaliação foram:

- **Acurácia:** indica uma performance geral do modelo. Dentre **todas** as classificações, quantas o modelo classificou corretamente;
- **Precisão:** dentre todas as classificações de classe Positivo **que o modelo fez**, quantas estão corretas;
- ***Recall*/Revocação/Sensibilidade:** dentre todas as situações de classe Positivo **como valor esperado**, quantas estão corretamente classificadas;
- **F1-Score:** média harmônica entre precisão e *recall*.

Resultados

Os resultados obtidos estão dados pelas Tabelas 1 – 3.

Naive Bayes	Precisão	Recall	F1-Score	Positivos e Falso Positivos	Negativos e Falso negativos
Não Spam	0.97	0.99	0.98	1432/15	180/46
Spam	0.92	0.80	0.86		
Acurácia	0.96				

Tabela 1: Resultado do algoritmo Naive Bayes.

Regressão Logística	Precisão	Recall	F1-Score	Positivos e Falso Positivos	Negativos e Falso negativos
Não Spam	0.96	0.99	0.98	1437/10	172/54
Spam	0.95	0.76	0.84		
Acurácia	0.96				

Tabela 2: Resultado do algoritmo Regressão Logística.

Arvore de Decisão	Precisão	Recall	F1-Score	Positivos e Falso Positivos	Negativos e Falso negativos
Não Spam	0.97	0.97	0.97	1409/38	177/49
Spam	0.82	0.78	0.80		
Acurácia	0.95				

Tabela 3: Resultado do algoritmo Arvore de Decisão.

Analisando a acurácia, observa-se que os algoritmos Naive Bayes e Regressão Logística apresentaram mesmo valor, sendo a arvore de decisão inferior em comparação aos demais. Já a precisão, *Recall*, F1-score e quantidade de falso positivos e falsos negativos, observa-se que existem divergências entre tais valores. Analisando então a aplicação do modelo no negocio, chega-se a conclusão que e-mails importantes poderiam ser classificados como não importantes em uma maior probabilidade segundo o *Recall* do algoritmo Regressão Logística. Desta forma, o modelo Naive Bayes deve ser o mais adequado para este problema