# Análise Exploratória e Classificador para Textos com potencialidade de *Spam*.

Daniel Longhi Fernandes Pedro
longhi88@hotmail.com

**Introdução**

Na era digital, muito conteúdo é gerado, muito texto e informações são trocadas diariamente. Neste contexto, muitos textos podem ser de considerados de desinformação ou são utilizados para a disseminação de má notícias, logo, esses textos cunham o neologismo “fakes news" [1,2]. Diante disto, ferramentas que baseiam-se na estatística e até mesmos modelos de Inteligência Artificial [3] são amplamente utilizados para veracidade e coerência das informações geradas no universo digital.

Já em grandes corporações a situação não é diferente. Com os mesmos problemas e um alto volume de troca de e-mails, é necessário a checagem e validação, uma vez que *spam* tendem fazer grande estrago nas caixas de e-mail. O exagerado volume visa por metodologias e abordagens que consigam identificar e barrar, se necessário, o recebimento de conteúdos que contenham essas características.

**Metodologia**

A análise teve origem com um arquivo referência corporativo em troca de texto. Deste modo, foi possível a construção e detecção visual de palavras que contêm maior recorrência em contextos de *spam*. Todas as informações foram geradas utilizando a linguagem Python 3.7 de uso gráfico e exploratório a IDE Jupyter.

Em um primeiro ponto foi trabalho com o texto para identificar palavras que tinham maior utilização, assim, gerou-se um grande *Word Cloud*. Em segundo momento foi explorado a quantidade de mensagens comum e de *spam* trocadas durante os meses do ano de 2017. Visando quantativamente extrair de forma estatística, foram feitas análises de quantidade de palavras mínimas, máximas, média, mediana, desvio padrão e variância destes textos. Por fim, foi construído um modelo de IA que separasse textos com potencialidades de ser comum ou *spam*.

**Resultados**

Na primeira etapa temos alguns verbos e palavras auxiliares que compõe os textos trocados. Além disto, são palavras que, no geral, não denotam contextos corporativos de

forma direta. Porém, é uma análise que traz uma visão geral do que está sendo tratado nestas mensagens.

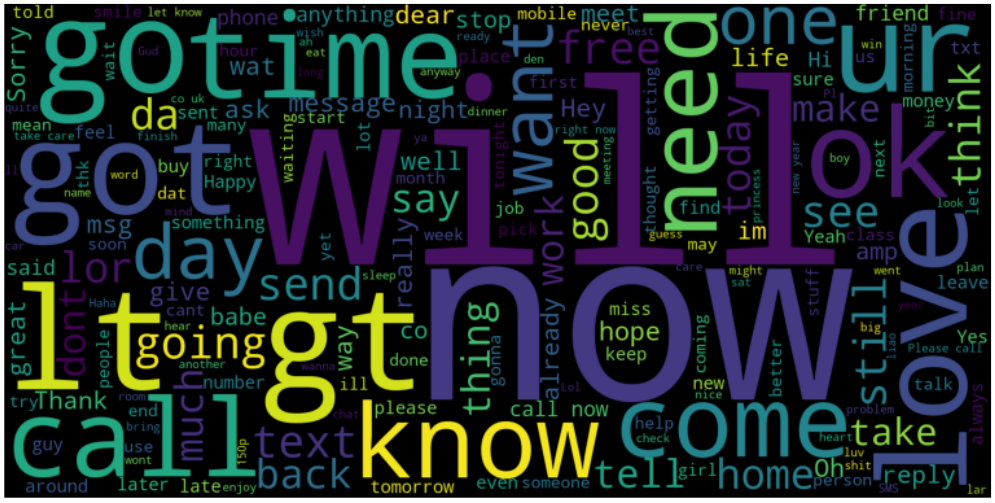


Figura 1 – *Word cloud.*

Na análise de recorrência de palavras, identificou que o número de palavras tem-se mantido constante, sendo uma parte de *spam*.

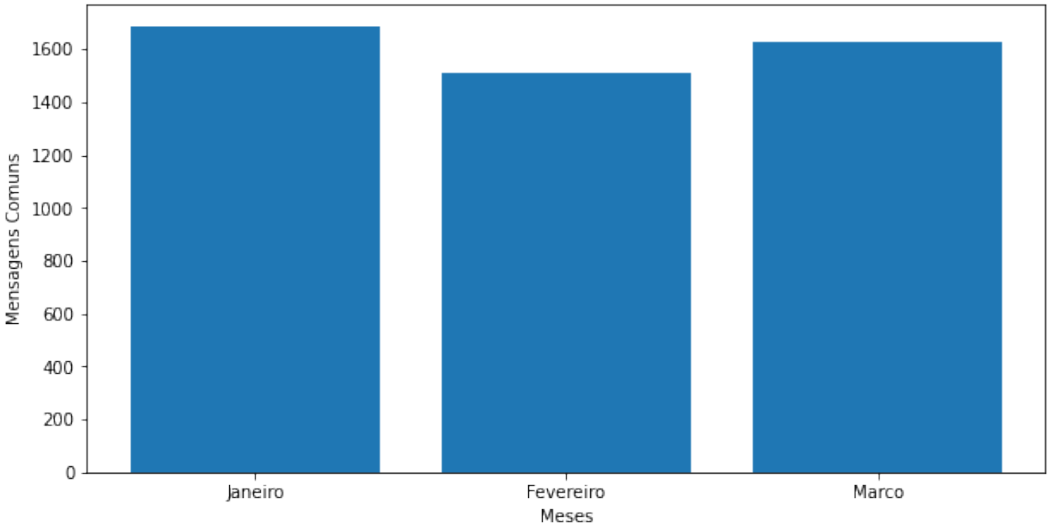


Figura 2 – Quantidade de mensagens comuns por mês.

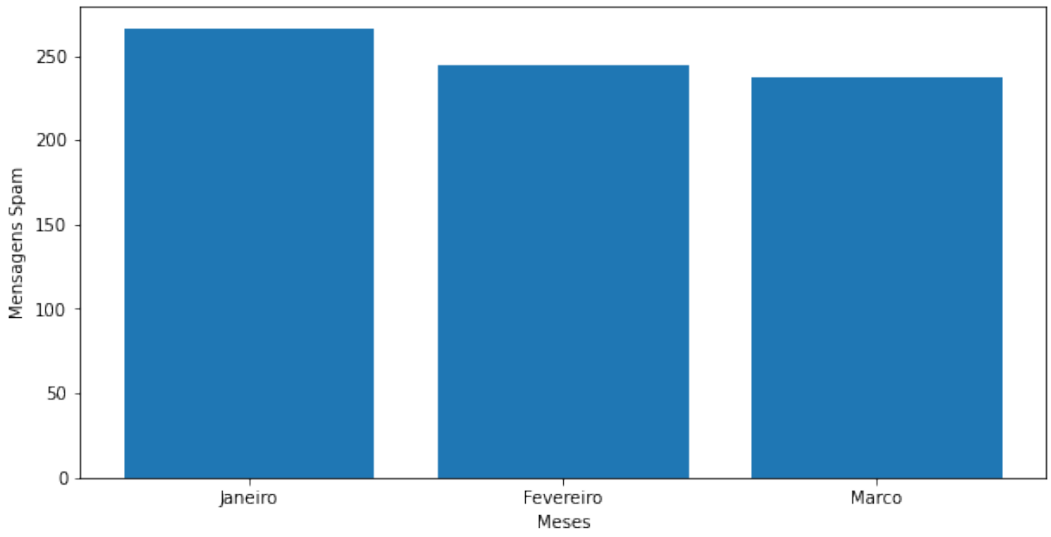


Figura 3 – Quantidade de mensagens *spam* por mês.

Para a parte estatística de composição por palavra para grupos comuns, há certa variabilidade, principalmente em números máximos. Já para as *spams* esse número tende a variar pouco.

| | min | max | mean | median | std | var |
|---|---|---|---|---|---|---|
| **Date** | | | | | | |
| **1** | 2 | 190 | 14.867220 | 11 | 12.728394 | 162.012015 |
| **2** | 2 | 100 | 14.562831 | 11 | 10.966891 | 120.272688 |
| **3** | 2 | 115 | 14.917076 | 11 | 11.532744 | 133.004183 |

Figura 4 – Estatística de palavras por texto em textos comuns.

| | min | max | mean | median | std | var |
|---|---|---|---|---|---|---|
| **Date** | | | | | | |
| **1** | 3 | 36 | 25.657895 | 27 | 5.456472 | 29.773088 |
| **2** | 5 | 36 | 25.114754 | 26 | 6.042642 | 36.513526 |
| **3** | 2 | 39 | 25.683544 | 27 | 6.315593 | 39.886720 |

Figura 5 – Estatística de palavras por texto em textos *spams*.

Ainda, há um agrupamento exploratório do maior dia por mês onde houve o máximo de troca de mensagens comuns.

| | Mes | Dia | Qtd |
|---|---|---|---|
| **0** | Fevereiro | 13 | 72 |
| **1** | Janeiro | 1 | 69 |
| **2** | Marco | 8 | 69 |

Figura 6 – Dia em que houve maior troca de textos comuns.

| | Mes | Dia | Qtd |
|---|---|---|---|
| **0** | Fevereiro | 18 | 14 |
| **1** | Marco | 10 | 14 |
| **2** | Janeiro | 12 | 13 |

Figura 6 – Dia em que houve maior troca de textos *spams*.

Como parte final desta análise exploratória e a identificação rotulada do que são comuns e *spams*, foi criado com I.A. dois algoritmos para classificar ser utilizado para descartar e-mails que tenham conteúdo de *spam*. Utilizando *Multilayer Perceptron*, foram alcançados resultados de acurácia em quase 96%. Na matriz de confusão alguns textos foram classificados de maneira errônea. Deste modo, há uma necessidade maior de tratamento e transformação dos dados para obtenção de melhores resultados.

```
MSE: 0.04
Matrix de Confusão
[[945  25]
 [ 23 122]]
Acc
0.95695067264574
```

Figura 7 – *Scores* para modelo Multilayer Perceptron.

Para fins de comparação, um outro modelo utilizando Árvore de Decisão foi utilizado, houve a utilização da técnica de *cross validation*, que divide o conjunto de dados para um treinamento por partes. No entanto houve resultados parecidos.

```
MSE: 0.04
Matrix de Confusão
[[957  13]
 [ 33 112]]
Acc
0.9587443946188341
```

Figura 8 – *Scores* para modelo Multilayer Perceptron.

Vale ressaltar, que quanto mais informações textuais catalogadas como *spam* e texto habitual trocado pela corporação via e-mail, melhores classificadores teremos. Ainda, ressalto a necessidade de balanceamento do *dataset*, item importante e que impacta nos resultados finais dos modelos.

**Conclusão**

A pesquisa, análise exploratória e construção de um modelo primário para a classificação de textos comuns e *spams* nos permite traçar técnicas e aprimorar conhecimento tanto na aquisição dos dados, como em testes de novas metodologias. Ainda, precisa-se levar em consideração a possibilidade da definição de um dicionário comum na troca de e-mails da empresa. Com isso, ao saber os vocábulos mais utilizados internamente, torna-se mais compreensível e facilita na classificação de textos que não são daquele contexto. Recomenda-se ainda o uso de metodologias de linguagem de processamento natural, com isso, ganha-se também em sintaxe, o que pode vir ser uma nova *feature* qualquer modelo de Inteligência Artificial.

**Referência**

1 - The science of fake news, David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, Jonathan L. Zittrain. Science, 2018. 10.1126/science.aao2998

2 - Fake news on online social media: propagation and reactions to misinformation in search of clicks. C. Delmazol e Jonas C.L. Valente. 2018. http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S2183-54622018000100012

3 - Automatic Detection of Fake News. V. Pérez-Rosas, B.Kleinberg, A. Lefevre, R. Mihalcea. 2017. https://arxiv.org/abs/1708.07104