
INSIGHT
Data Science Laboratory
Federal University of Ceará

# PRIMEIROS PASSOS NO MUNDO DE DATA SCIENCE & MACHINE LEARNING

INSIGHT

DEPARTAMENTO DE COMPUTAÇÃO

UNIVERSIDADE FEDERAL DO CEARÁ

# AGENDA

1. MÉTRICAS & AVALIAÇÃO DE MODELOS
2. CONSTRUINDO UM DETECTOR DE FAKE NEWS ACERCA DO COVID-19
3. HANDS-ON
4. BIBLIOGRAFIA

# 1. REVIEW

# 2. MÉTRICAS & AVALIAÇÃO DE MODELOS

# COMO EU SEI QUE UM MODELO É MELHOR QUE OUTRO?

# HOJE IREMOS CONHECER ALGUMAS MÉTRICAS!

# REGRESSÃO

# MÉTRICAS PARA REGRESSÃO

- Variações básicas sobre a diferença entre o que você previu e os valores reais.
- **Erro médio absoluto (MAE)**
  - Erro médio

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

test set — predicted vaue — actual value

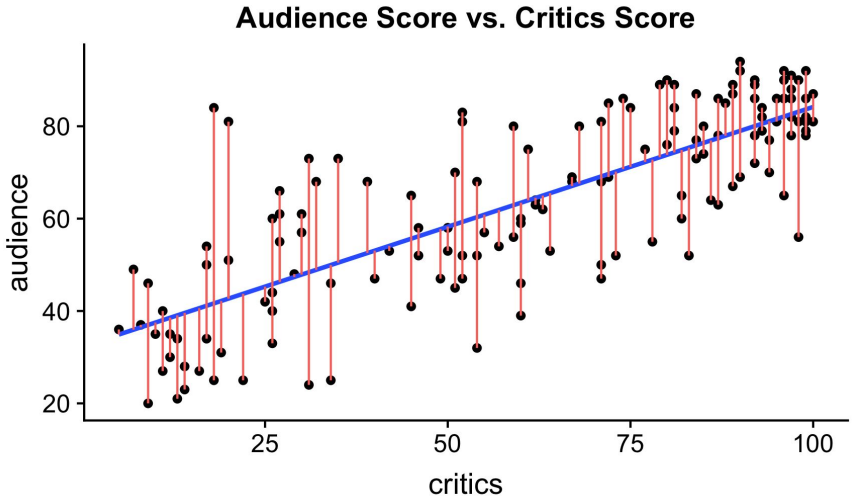
# MÉTRICAS PARA REGRESSÃO

- **Erro quadrático médio (RMSE)**
  - Média das diferenças entre as predições e observações reais ao quadrado
  - Pense na distância Euclidiana entre o vetor de valores corretos e o vetor de valores previstos mediados por ’n’, onde n é o número de pontos.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

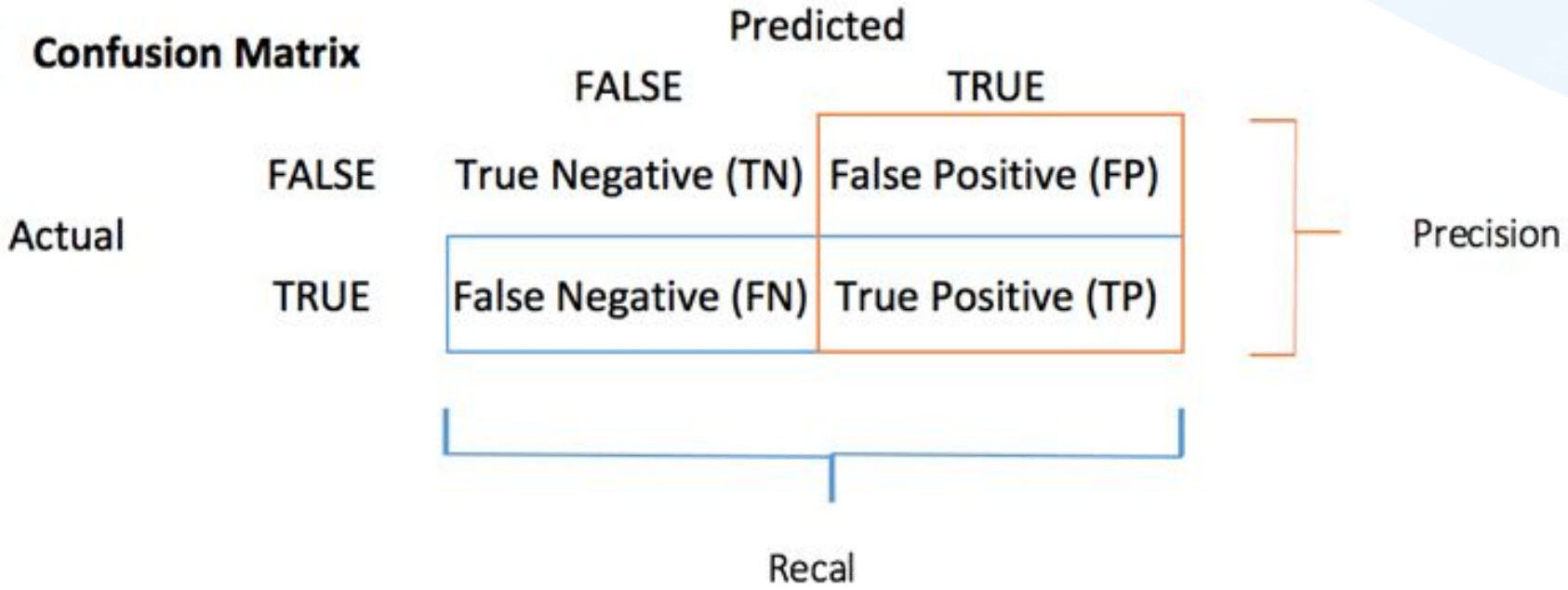
# MÉTRICAS PARA REGRESSÃO

- **Erro quadrático médio (RMSE)**
  - Média das diferenças entre as predições e observações reais ao quadrado
  - Pense na distância Euclidiana entre o vetor de valores corretos e o vetor de valores previstos mediados por 'n', onde n é o número de pontos.



# CLASSIFICAÇÃO

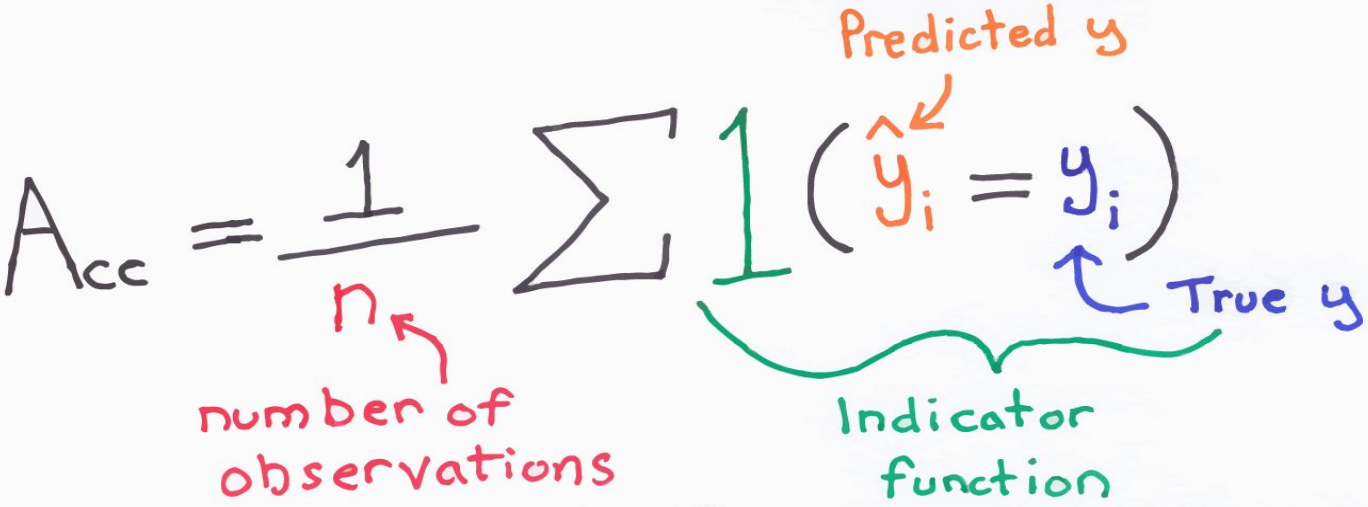
# MATRIZ DE CONFUSÃO



# MATRIZ DE CONFUSÃO

- **Verdadeiros Positivos (TP)**
  - Classificação correta da classe Positivo.
- **Verdadeiros Negativos (TN)**
  - Classificação correta da classe Negativo.
- **Falsos Positivos (Erro Tipo I) (FP)**
  - Erro em que o modelo previu a classe Positivo quando o valor real era classe Negativo.
- **Falsos Negativos (Erro Tipo II) (FN)**
  - Erro em que o modelo previu a classe Negativo quando o valor real era classe Positivo.

# ACCURACY




A common metric in classification. Fails when we have highly imbalanced classes. In those cases F1 is more appropriate.

ChrisAlbon

# PRECISION

Precision is the ability a classifier to not label a true negative observation as positive.

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

ChrisAlbon

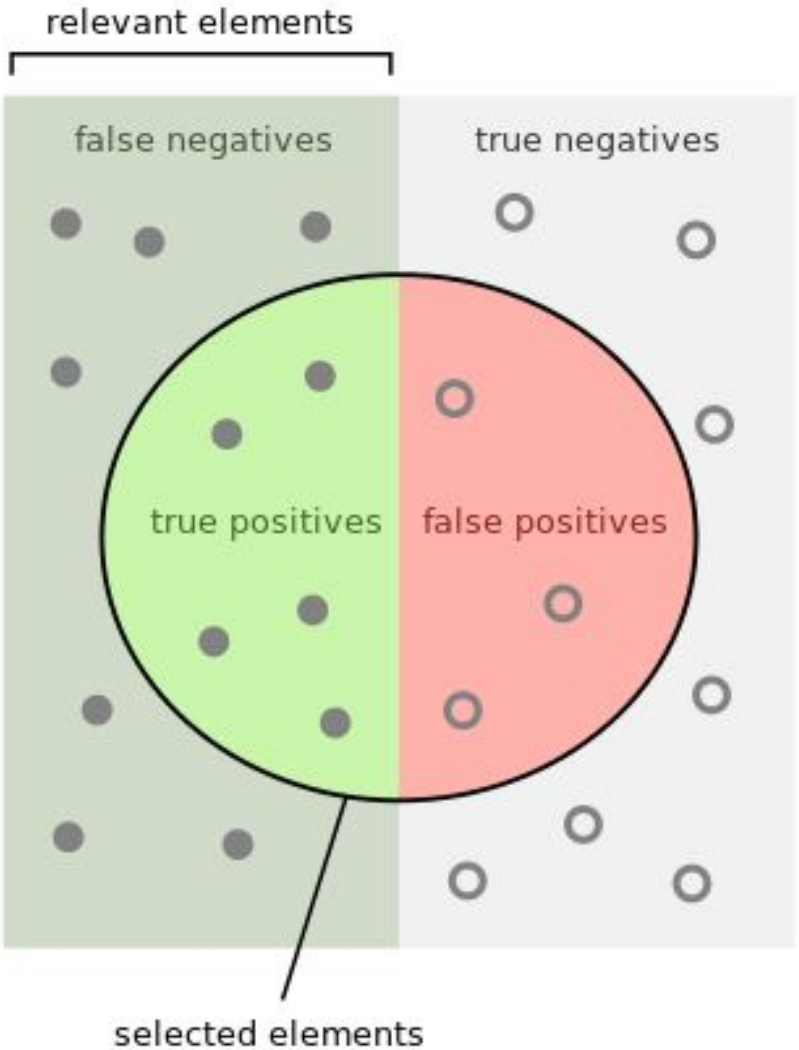
# RECALL

"Recall is about the real positives"

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is the ability of the classifier to find positive examples. If we wanted to be certain to find all positive examples, we could maximize recall.

Chris Albon




How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

# F1 SCORE

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

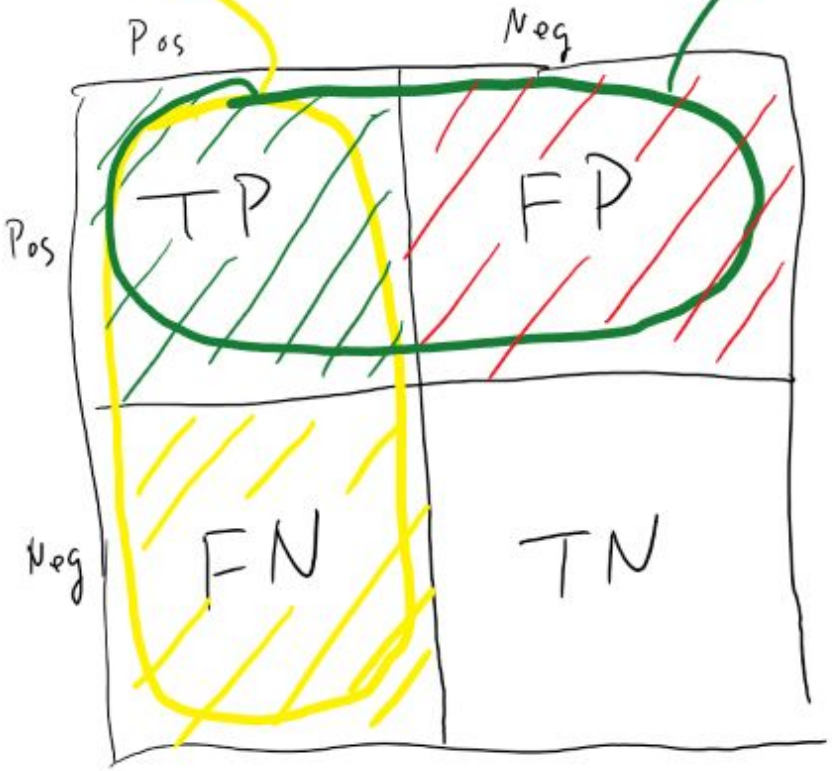
F1 score is the harmonic mean of precision and recall. Values range from 0 (bad) to 1 (good).

Chris Albon

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$



# 3. CONSTRUINDO UM DETECTOR DE FAKE NEWS ACERCA DO COVID-19.

# INTRODUÇÃO

A popularização das **redes sociais** emponderou o cidadão, trazendo consequências positivas e negativas.

Uma das consequências negativas trazidas pelo mal uso das redes sociais é a difusão de notícias falsas ou tendenciosas.

A atual **pandemia por COVID-19** tornou-se um dos **temas mais utilizados para produção de notícias** que geram desinformação.

# OBJETIVOS

## Objetivo geral

- Criar um detector de fake news relacionados à COVID-19.

## Objetivos Específicos

- Formar uma base de dados de fake-news de COVID-19;
- Gerar modelos capazes de classificar notícias acerca do COVID-19 extraídas de redes sociais como falsas ou verdadeiras;
- Comparar o desempenho dos modelos criados na busca do THE BEST.

# PROPOSIÇÃO DE VALOR

- Ajudar no combate da propagação de fake news;
- Ter uma aceitável taxa de erro na classificação correta das notícias verdadeiras ou falsas relacionadas ao COVID-19;
- Ajudar a população quanto ao esclarecimento sobre uma notícia acerca do COVID-19 e sua veracidade, visto ser um assunto novo e ainda com bastante incertezas.

# FONTE DE DADOS

- Dados utilizados de notícias falsas brasileiras sobre o COVID-19, dispostos no Chequeado;
- Notícias verdadeiras obtidas através de um *web crawler* dos links das notícias utilizadas para comprovar que a notícia é falsa no Chequeado (Aos Fatos, Piauí Folha, Estadão Verifica, Agência Lupa);
- Fato Ou Fake do G1*.

# INTRODUÇÃO

| AFIRMAÇÃO | CLASSIFICAÇÃO |
|---|---|
| Medicamento ivermectina cura a Covid-19. | Falso |
| É falso que número de mortes por Covid-19 caiu após Moro anunciar investigação | Verdadeiro |

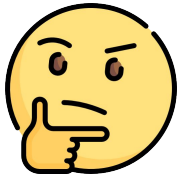
**Tabela 1.** Amostras do dataset.

# FUNDAMENTAÇÃO TEÓRICA

## FAKE NEWS

*"Quaisquer **notícias e informações falsas que são compartilhadas como se fossem reais e verdadeiras**, divulgadas em contextos virtuais, especialmente em redes sociais ou em aplicativos."*

**Mas como representar notícias em modelos de aprendizado de máquina?**



# FUNDAMENTAÇÃO TEÓRICA

## BAG OF WORDS (BOW)

- Representação simplificada e esparsa
- Iremos ter a bolsa de palavras (vocabulário) como colunas do dado
- 1 representa as palavras que ocorrem no texto. 0 pras demais colunas

## TF-IDF

- Term Frequency - Inverse Document Frequency.
- Indica a importância de uma palavra em um documento.
- Enquanto TF está relacionada à frequência do termo, IDF busca balancear a frequência de termos mais comuns/frequentes que outros.

# FUNDAMENTAÇÃO TEÓRICA

## OUT OF VOCABULARY

- Consiste nas **palavras presentes no *dataset* que não estão presentes no vocabulário da *word embedding*,** logo elas não possuem representação vetorial.

## EDIT DISTANCE

- Maneira de **quantificar a diferença entre duas palavras,** contando o **número mínimo de operações necessárias para transformar uma palavra na outra**.

# FUNDAMENTAÇÃO TEÓRICA - MODELOS

**REGRESSÃO LOGÍSTICA (*)**

- Classificador linear

**K-NN (*)**

- Modelo não-paramétrico

**ANÁLISE DISCRIMINANTE GAUSSIANO (*)**

- Modelo que não possui hiperparâmetros

**SVM**

- Encontra ótimo global

**ÁRVORE DE DECISÃO**

- Abordagem heurística gulosa

**RANDOM FOREST**

- Ensemble Bagging de Árvores de Decisão

**XGBOOST**

- Amplamente utilizado em competições do Kaggle, Ensemble

**Encode-LSTM-Dense**

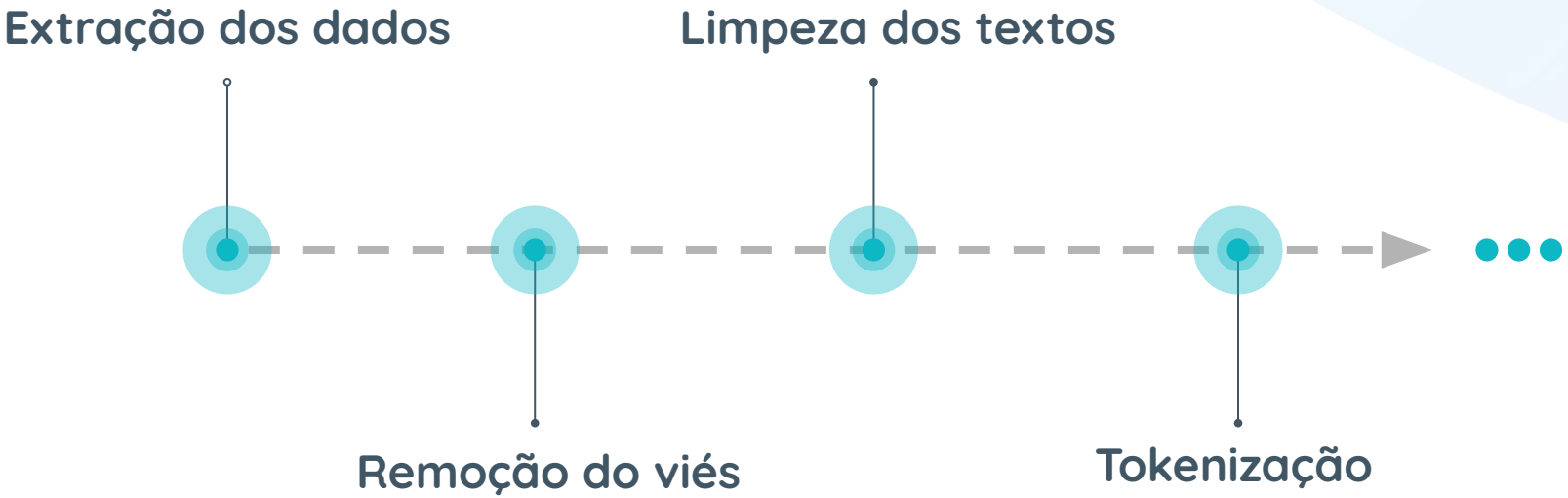
- Exemplo de técnica de deep learning

# FUNDAMENTAÇÃO TEÓRICA - MÉTRICAS

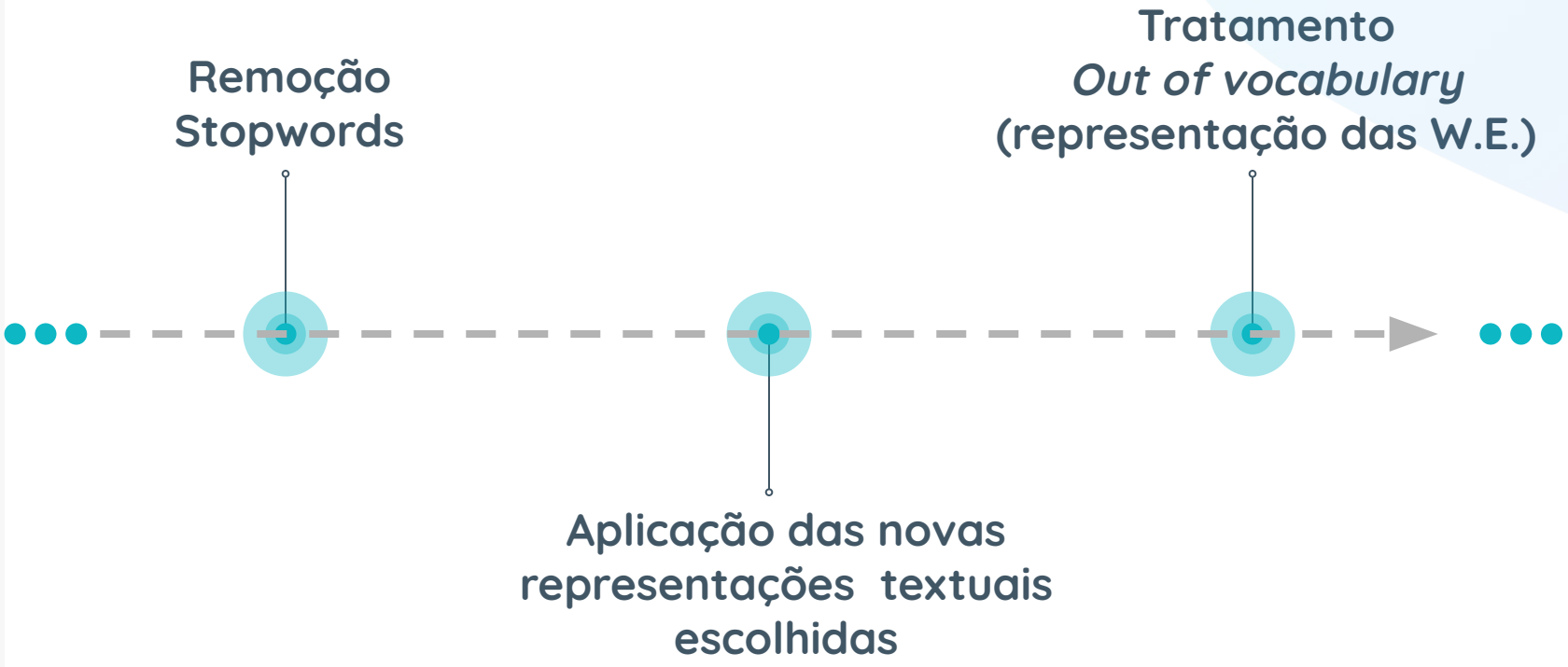
- Acurácia
- Precision
- Recall
- F1-Score
- Curva ROC
- Matriz de Confusão

# 3.
# METODOLOGIA

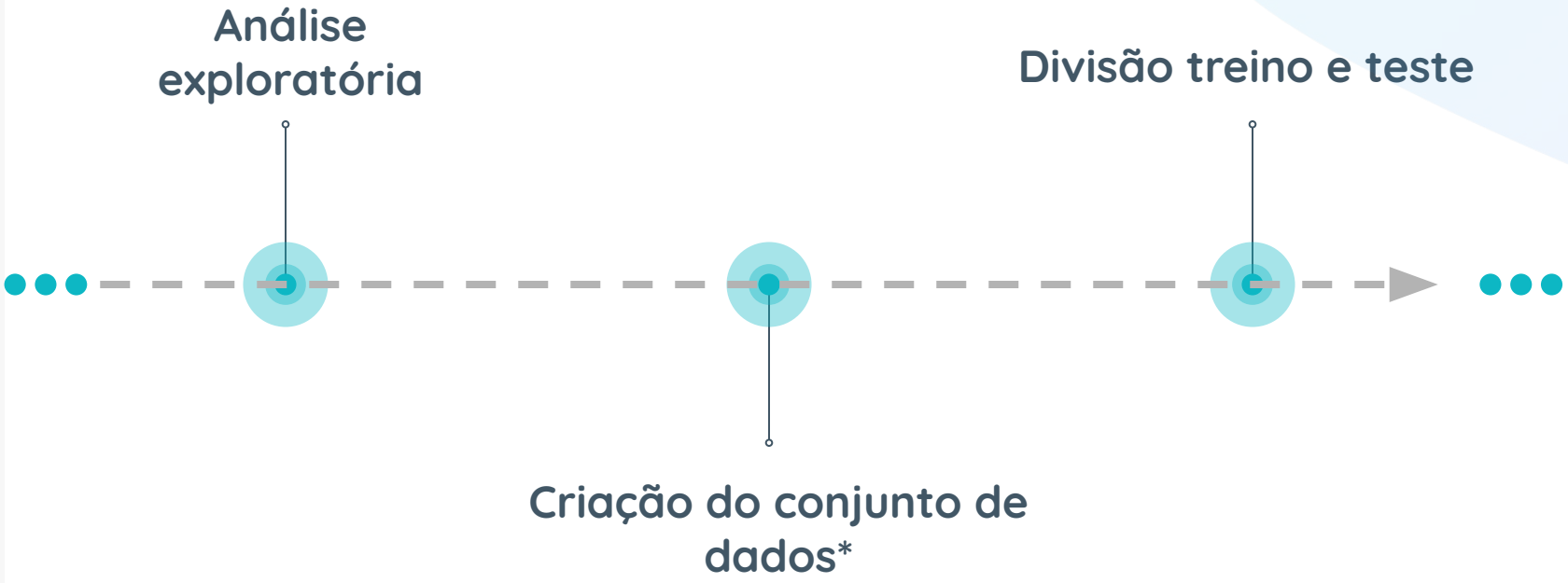
# METODOLOGIA



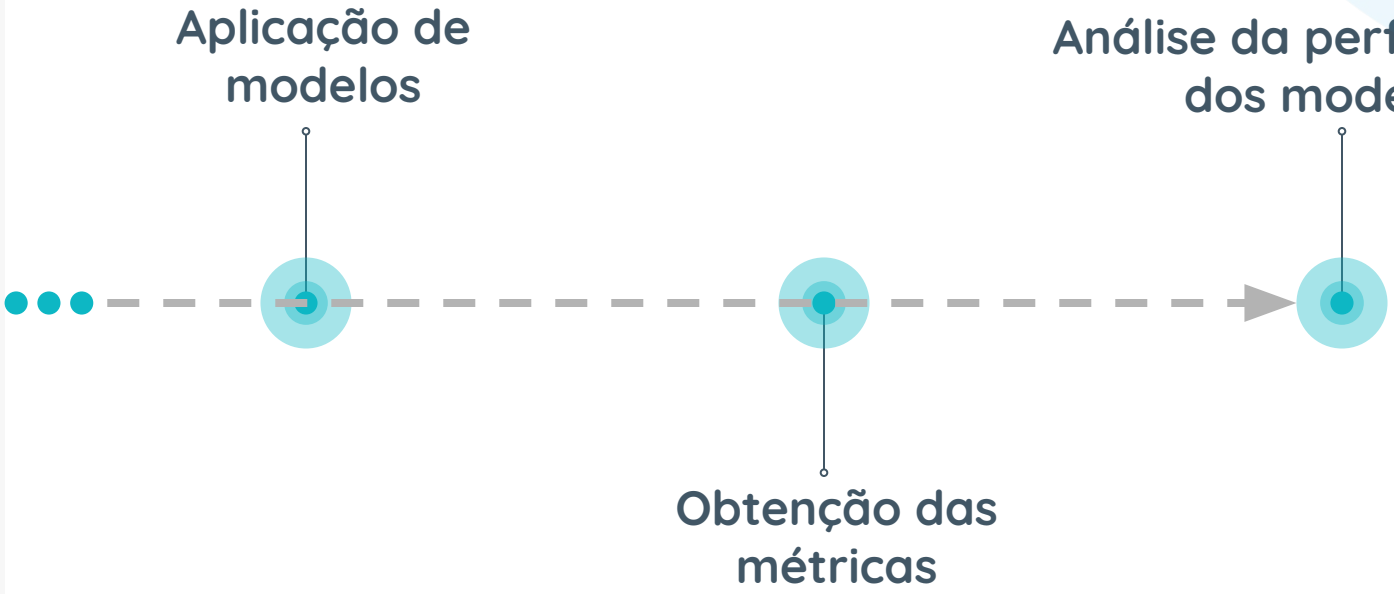
# METODOLOGIA



# METODOLOGIA



# METODOLOGIA



# 5. RESULTADOS

# EXPERIMENTOS - DATASET

- **Colunas utilizadas:** Text e classification
- **1.753 padrões**
  - 808 *fake news*
  - 945 *true news*
- **Tamanho do vocabulário do dataset**: 3.698
  - True: 2.328
  - False: 2.663
- **80% Treino e 20% Teste**

# EXPERIMENTOS - DATASET

- **OOVs:** 32 palavras
  - **Observação**: palavras chaves do nosso contexto são mapeadas para palavras que não tem muita conexão.

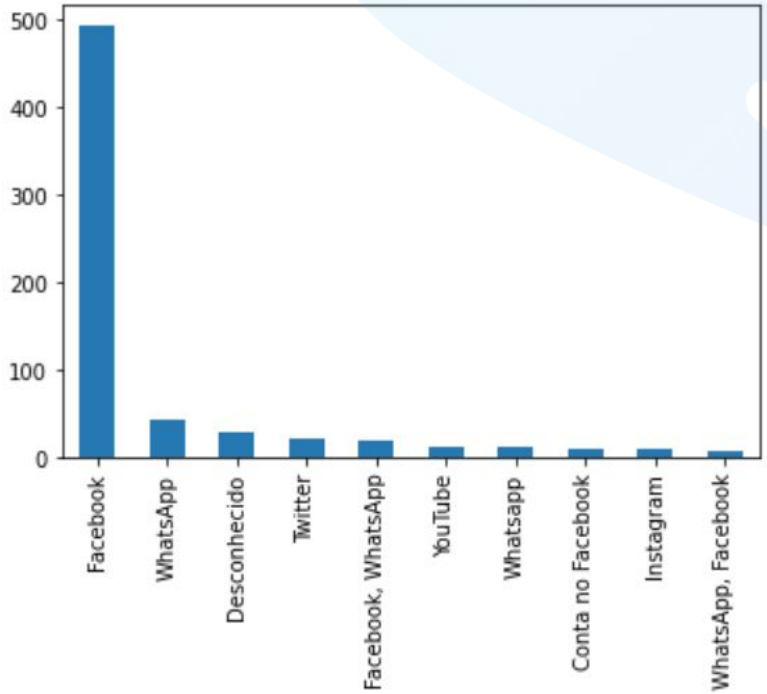
```
{'adhanom': 'phanom',
 'alquingel': 'alquil',
 'arpen': 'aspen',
 'autocontaminação': 'autocontemplação',
 'autodiagnosticar': 'autodiagnóstico',
 'bolsonarista': 'bolsonaro',
 'bolsonaristas': 'bolsistas',
 'certillamado': 'certificado',
 'chadox': 'chador',
 'coronavac': 'coroava',
 'covid': 'covil',
 'cracolância': 'cracolândia',
 'espancadoem': 'espancado',
```

```
'friston': 'frisson',
'ghebreyesus': 'hebreus',
'incint': 'incine',
'invermectina': 'ivermectina',
'kathlenn': 'kathleen',
'ncov': 'nov',
'ozonioterapia': 'ozonoterapia',
'paracovid': 'paranoid',
'pisoetada': 'pilotada',
'radarbox': 'radar',
'reinfectadas': 'reinjectadas',
'reinfectam': 'reinfectar',
'remdesivir': 'redefinir',
'ruilan': 'ruslan',
'sinovac': 'inovar',
'subnotificam': 'subnotificado',
'supernotificam': 'supernotificação',
'tedros': 'cedros',
'zelenko': 'elenko'}
```

# EXPERIMENTOS - Fonte de Dados

**Frequência das fake news por rede social**

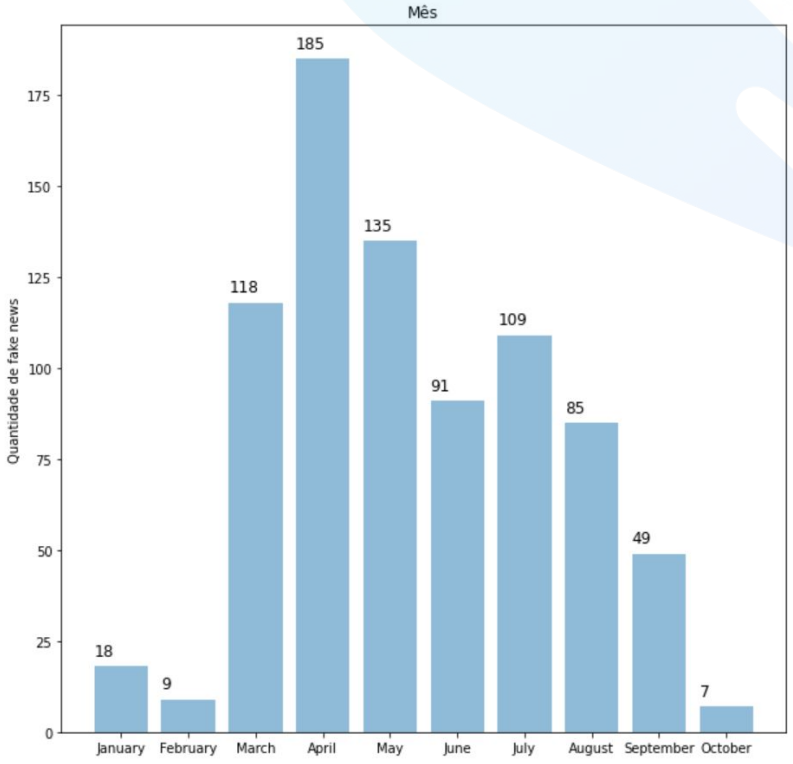
- Há repetição de redes sociais
- Maiores veículos de propagação de fake news
  - Facebook
  - Whatsapp.



# EXPERIMENTOS - Série temporal

## Quantidade de fake news ao longo dos meses

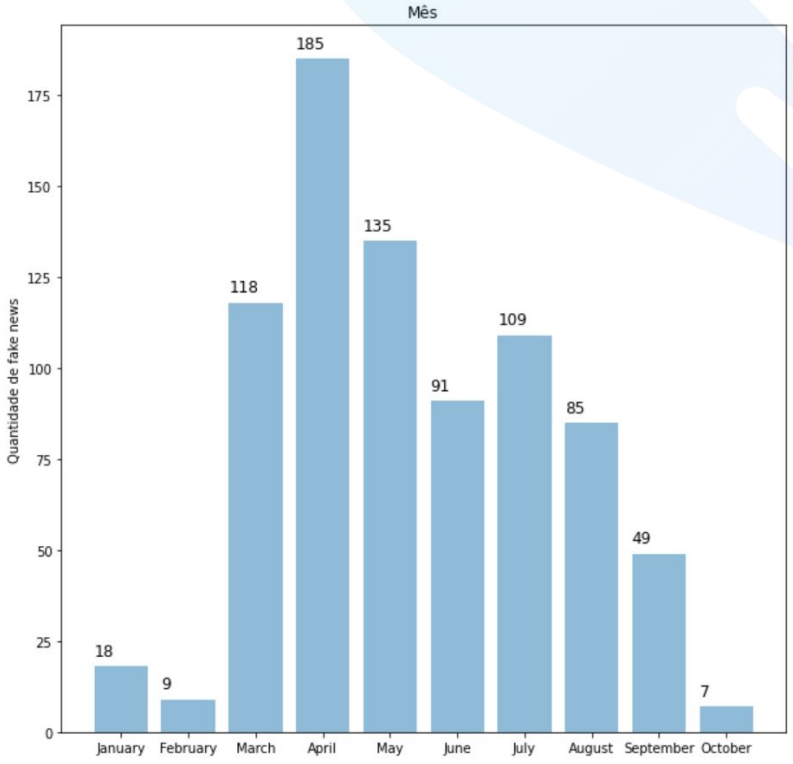
- Maior número de Fake News ocorreu em Abril
- Doença começou a se espalhar com maior velocidade no território Brasileiro.



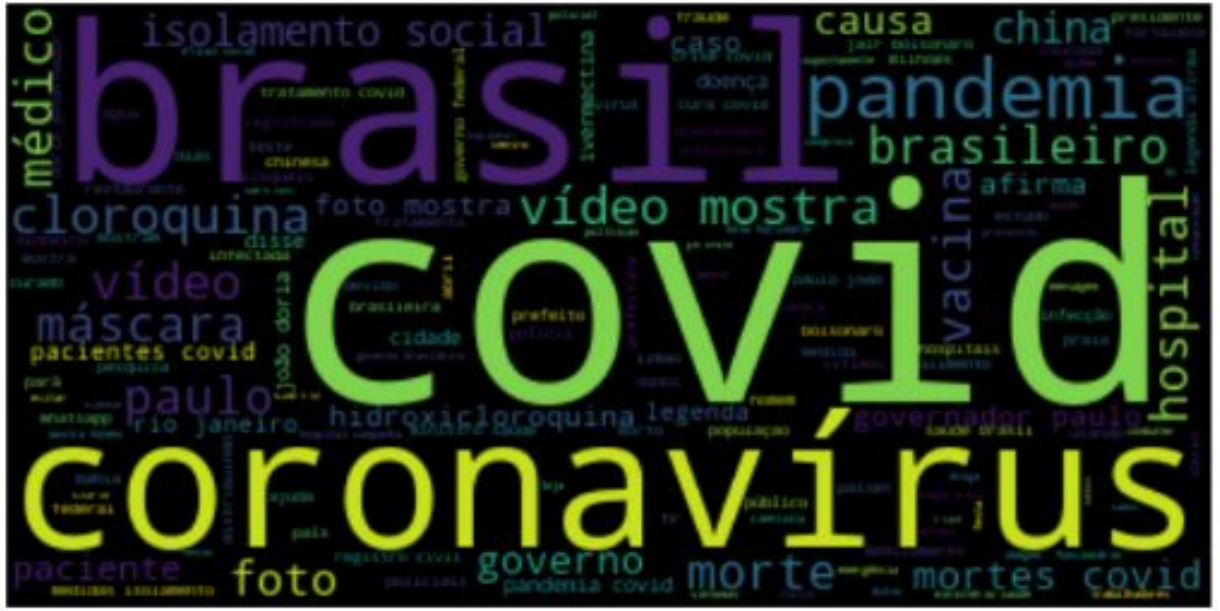
# EXPERIMENTOS - Série temporal

## Quantidade de fake news ao longo dos meses

- De acordo com o G1, em 28 de abril, o Brasil possuía **73.235 casos** do novo coronavírus (Sars-CoV-2), com **5.083 mortes**.
- Começaram a **surgir os boatos de combate do Coronavírus via Cloroquina**, além de remédios caseiros.



# EXPERIMENTOS - Nuvem de palavras




**Nuvem de palavras nas notícias falsas.**

# EXPERIMENTOS - Nuvem de palavras


**Nuvem de palavras nas notícias verdadeiras.**

# EXPERIMENTOS

- **Regressão Logística**
  - Valores testados
    - **alpha**: [0.01, 0.001, 0.0001, 0.00001, 5, 3, 10]
- **k-NN**
  - Valores testados
    - **k:** [3, 5, 7, 9, 11]
    - **distâncias:** ['euclidean', 'manhattan']
- **Árvore de Decisão**
  - Valores testados
    - **Máx. de nós:** [1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60]
    - **Critério de pureza**: ['gini', 'entropy']
- **Random Forest**
  - Valores testados
    - **Quantidade de árvores**: [100, 110, 120, 130, 140, 150]

# RESULTADOS - MELHORES REPRESENTAÇÕES POR ALGORITMO

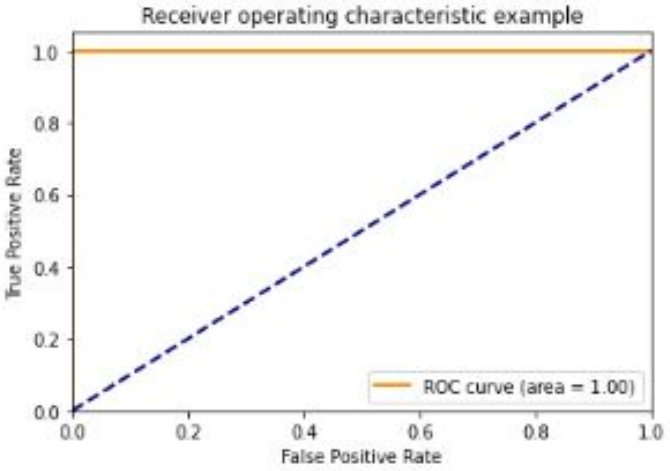
| MODELOS | PRECISION | RECALL | F1-SCORE | ACCURACY | ROC |
|---|---|---|---|---|---|
| **XGBoost BOW e TF-IDF*** | **1** | **1** | **1** | **1** | **1** |
| **SVM BOW E TF-IDF*** | **1** | **1** | **1** | **1** | **1** |
| Regressão Logística BOW | 0.7560 | 0.7549 | 0.7539 | 0.7549 | 0.7521 |
| LSTM FASTTEXT | 0.7496 | 0.7492 | 0.7493 | 0.7492 | 0.7492 |
| Random Forest TF-IDF | 0.7407 | 0.7407 | 0.7402 | 0.7407 | 0.7388 |
| Árvore de Decisão TF-IDF | 0.7120 | 0.7122 | 0.7121 | 0.7122 | 0.7111 |
| Análise Discriminante Gaussiano Word2Vec | 0.7132 | 0.7122 | 0.7106 | 0.7122 | 0.7089 |
| k-NN FastText | 0.6831 | 0.6809 | 0.6775 | 0.6638 | 0.6550 |

# RESULTADOS - PIORES REPRESENTAÇÕES POR ALGORITMO

| MODELOS | PRECISION | RECALL | F1-SCORE | ACCURACY | ROC |
|---|---|---|---|---|---|
| **XGBoost Word2Vec** | **0.7238** | **0.7236** | **0.7227** | **0.7236** | **0.7211** |
| SVM Word2Vec | 0.7211 | 0.7179 | 0.7151 | 0.7179 | 0.7135 |
| Árvore de Decisão Word2Vec | 0.6391 | 0.6353 | 0.6351 | 0.6353 | 0.6372 |
| Random Forest Word2Vec | 0.6231 | 0.6210 | 0.6212 | 0.6210 | 0.62198 |
| Regressão Logística FastText | 0.6158 | 0.5982 | 0.5688 | 0.59829 | 0.5858 |
| Análise Discriminante Gaussiano TF-IDF | 0.5802 | 0.5811 | 0.5801 | 0.5811 | 0.5786 |
| k-NN BOW | 0.5140 | 0.5099 | 0.5087 | 0.5042 | 0.5127 |
| **LSTM WORD2VEC (*)** | 0.4660 | 0.4615 | 0.4367 | 0.4615 | 0.4717 |

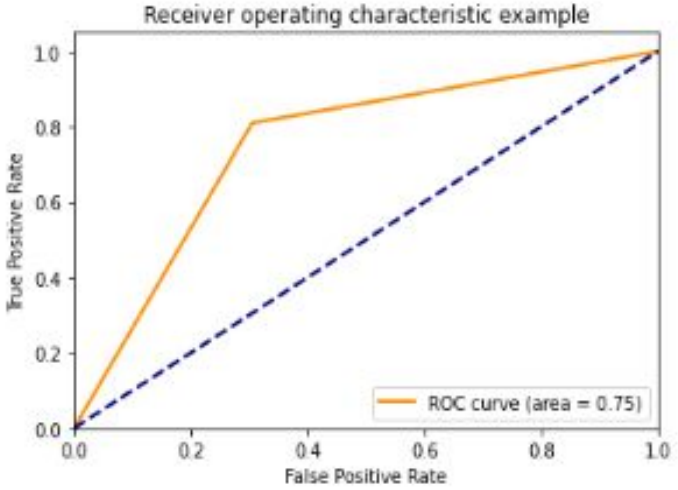
# DISCUSSÃO

- **THE BEST (?)**: SVM & XGBOOST com TF-IDF e BOW
  - Fecharam no 100%
  - Suspeita de overfitting

# DISCUSSÃO

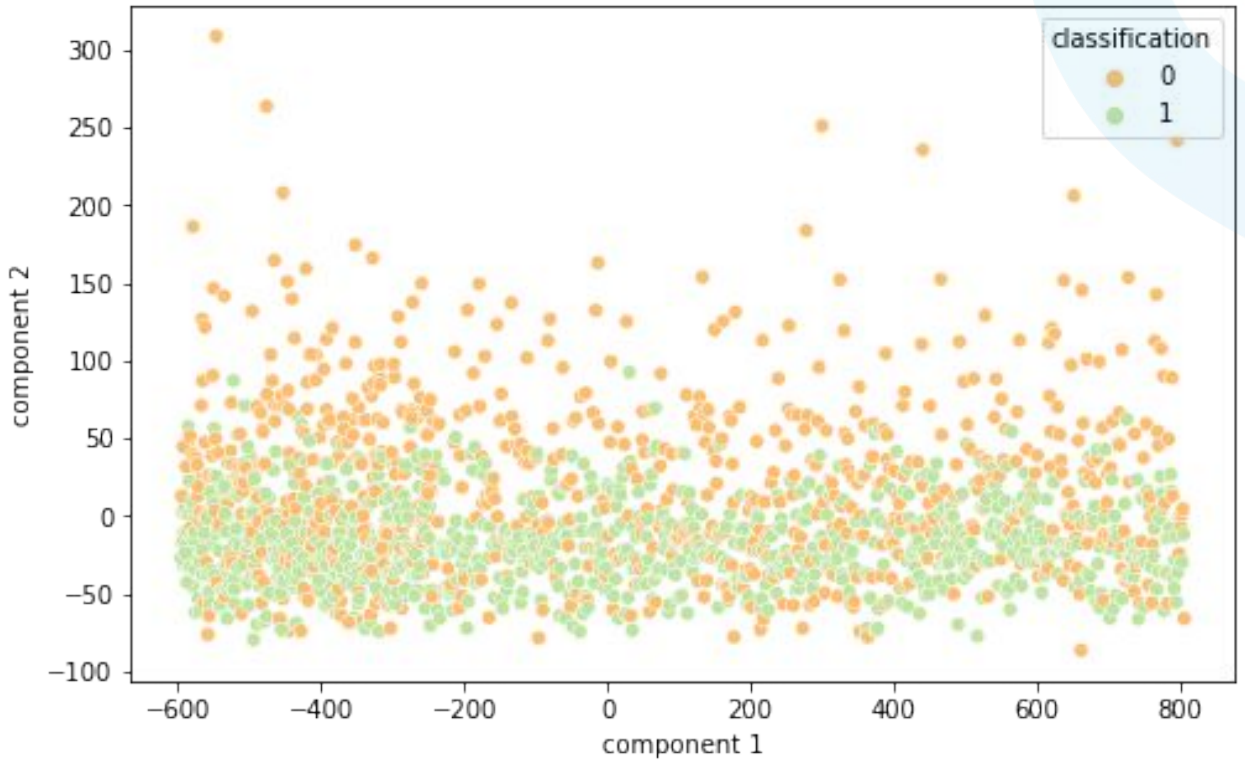
- **THE BEST**: Regressão Logística BOW ~75.49

# DISCUSSÃO

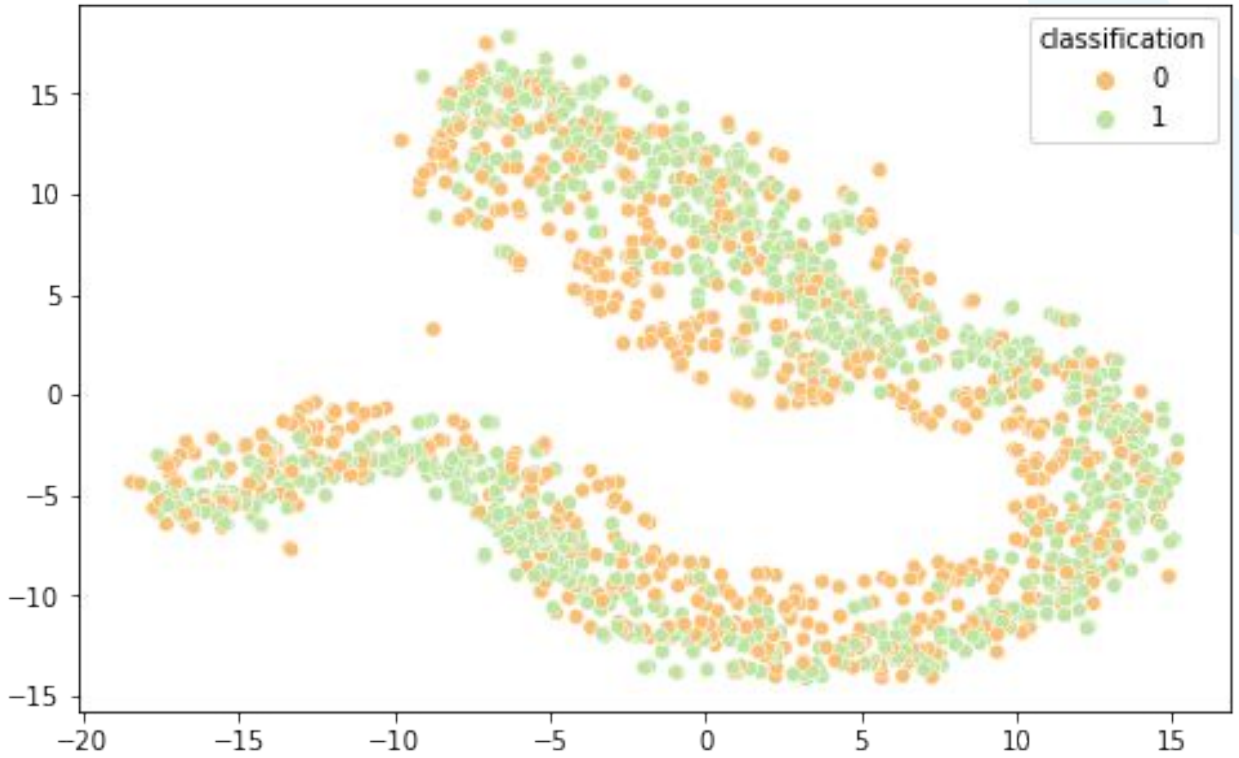
## Exemplos de classificações da Regressão Logística

- **True Positive (corretamente classificada)**
  - Texto que diz que vitamina C e limão combatem o coronavírus
- **True Negative (corretamente classificada)**
  - Notícia divulgada em 2015 pela TV italiana RAI comprova que o novo coronavírus foi criado em laboratório pelo governo chinês.
- **False Positive (erroneamente classificada)**
  - Vitamina C com zinco previne e trata a infecção por coronavírus
- **False Negative (erroneamente classificada)**
  - Que neurocientista britânico publicou estudo mostrando que 80% da população é imune ao novo coronavírus

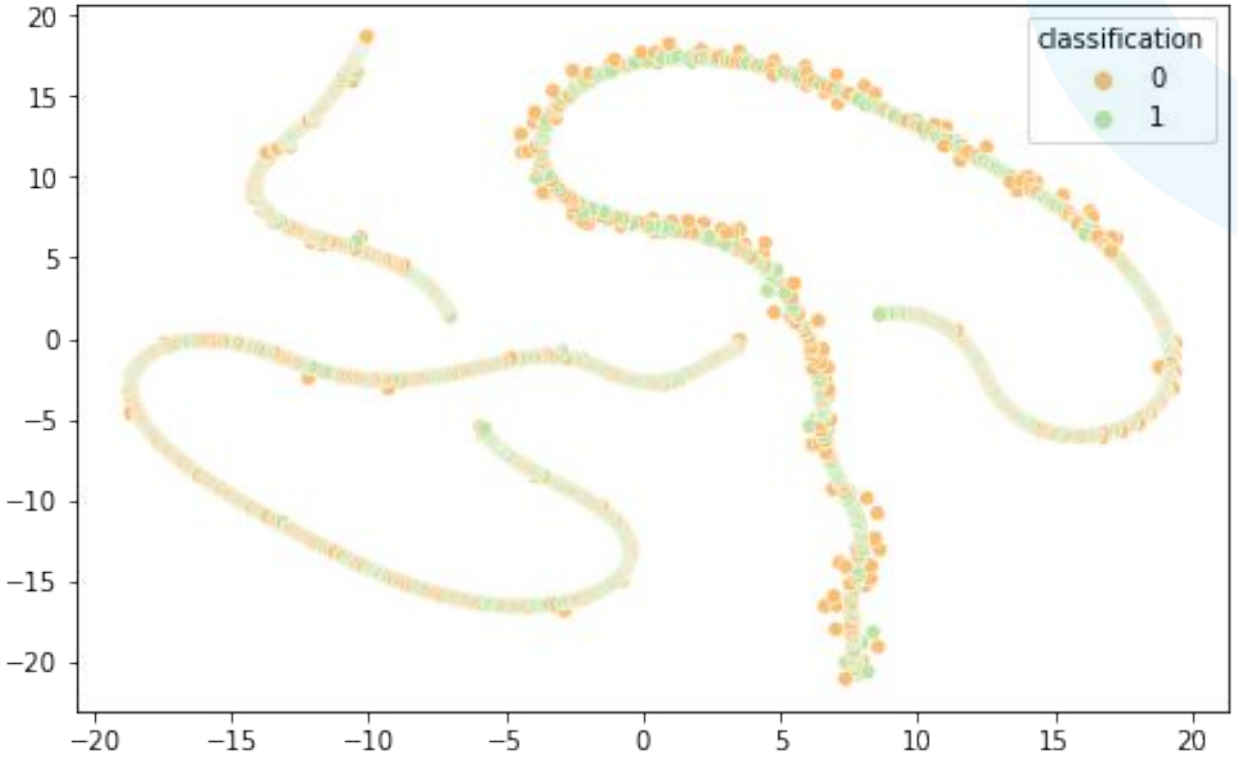



**FastText PCA**

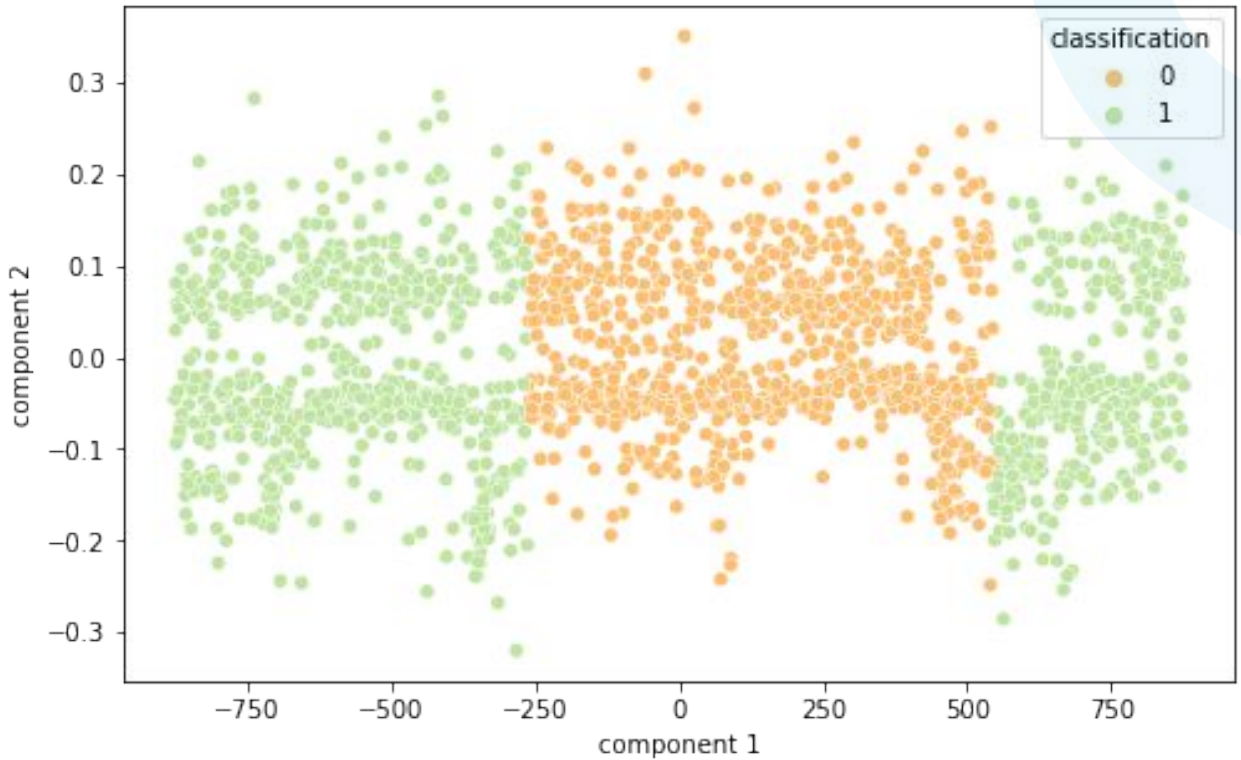
(Semelhante ao Word2Vec)




**FastText T-SNE**

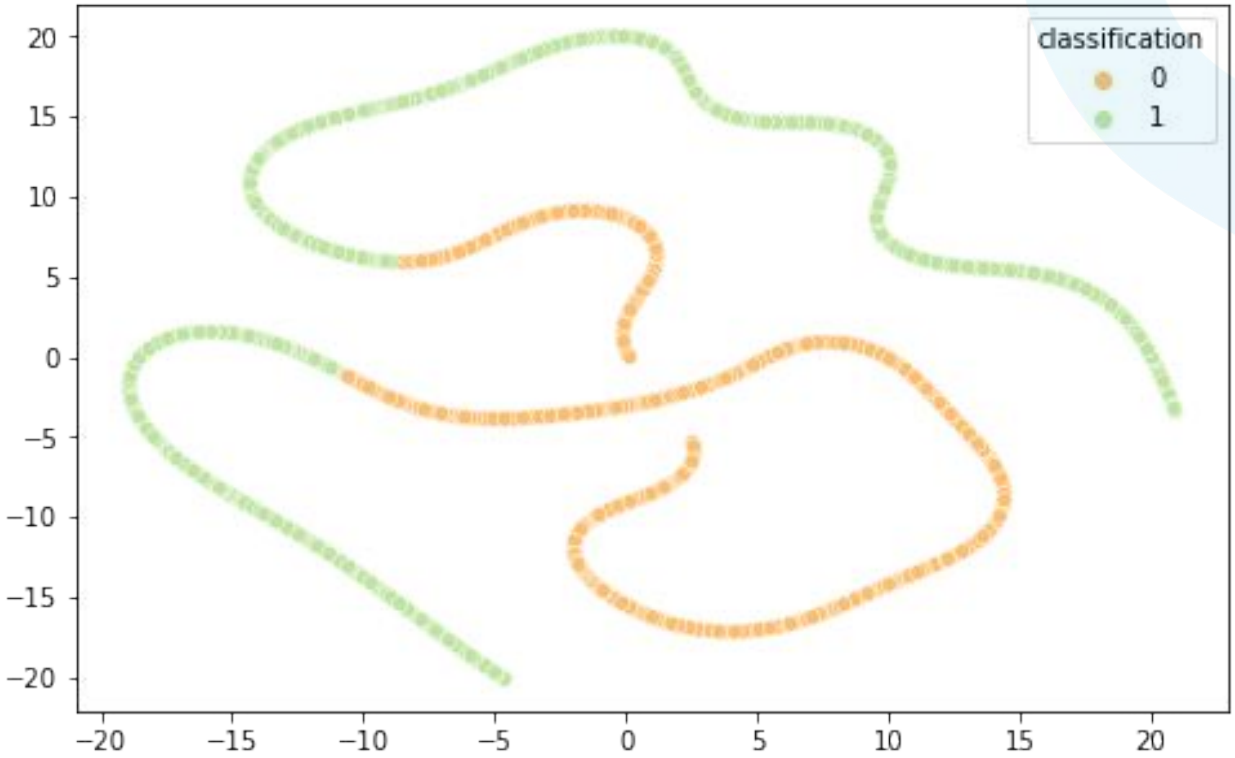



**Word2Vec T-SNE**

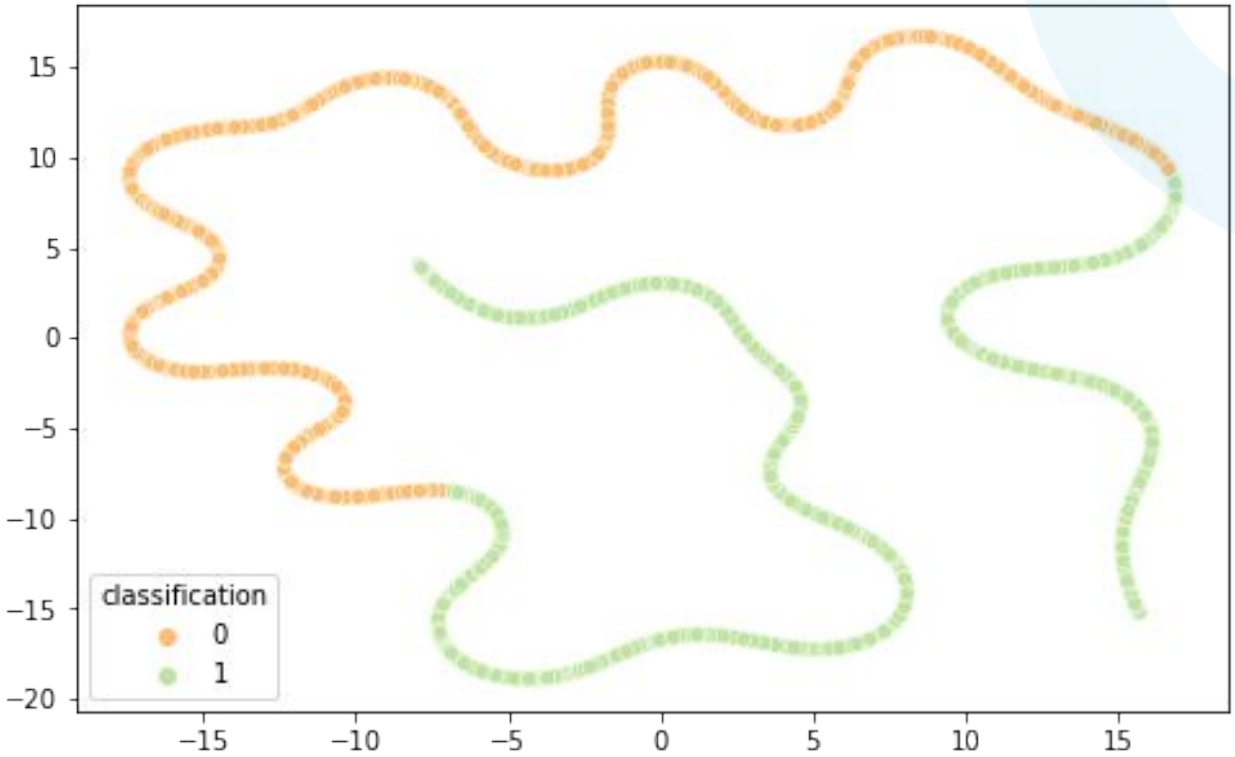



**BOW PCA**

(Semelhante ao TF-IDF)




**BOW T-SNE**




**TF-IDF T-SNE**

# 8.
# CONCLUSÃO

# CONCLUSÃO

- **Melhores modelos**: XGBoost(*) e SVM(*) e Regressão Logística.
- **Melhores Representações:** TF-IDF e BOW
- **Pior representação:** Word2Vec
- **Pior modelo**: LSTM (**) e kNN.
- Possibilidade de dados linearmente separáveis com BOW e TD-IDF.

# 10. BIBLIOGRAFIA

# REFERÊNCIAS


- Aldwairi, Monther, and Ali Alwahedi. "**Detecting fake news in social media networks.**" Procedia Computer Science 141 (2018): 215-222.
- Buntain, Cody, and Jennifer Golbeck. "A**utomatically identifying fake news in popular Twitter threads.**" In 2017 IEEE International Conference on Smart Cloud (SmartCloud), pp. 208-215. IEEE, 2017.
- Filho, Manuel; Oliveira, Andreza and Mattos, César. "**Detectando Fake News em Manchetes – Uma Comparação de Modelos de Aprendizagem de Máquina**" WTAG 2019.

# HANDS-ON!


LOONEY TUNES