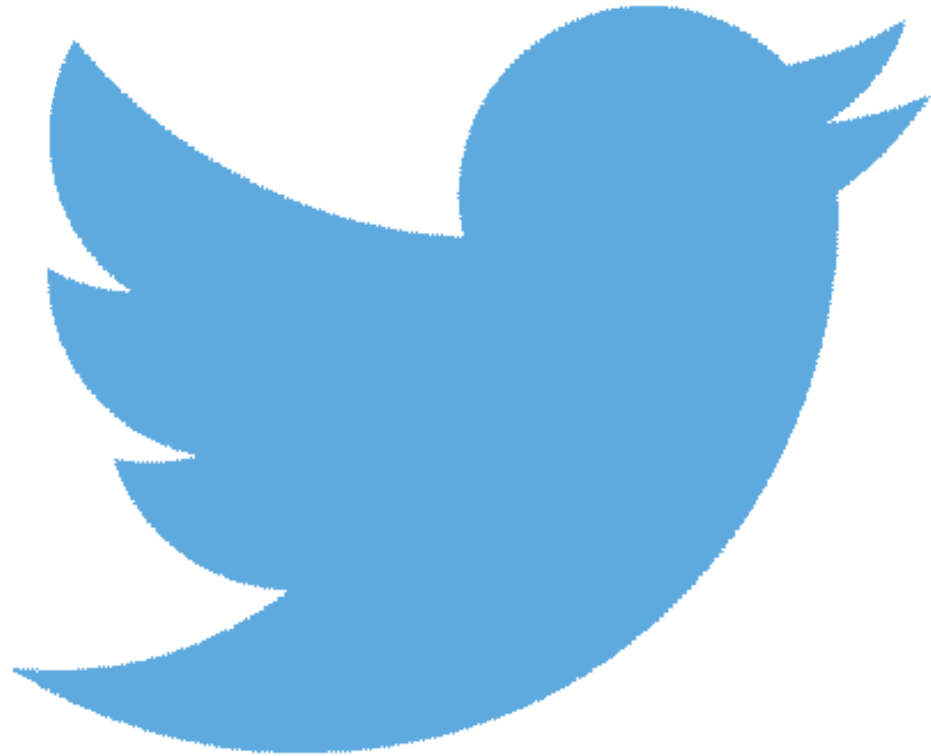


# SOLUÇÃO DO CASE: ANÁLISE DE TWEETS

PEDRO ALMEIDA

21.01.2024



# AGENDA

1. PROBLEMA DE NEGÓCIO E OBJETIVO
2. DEFINIÇÃO TÉCNICA
3. PLANEJAMENTO DA SOLUÇÃO
4. ANÁLISE EXPLORATÓRIA
5. MODELAGEM E AVALIAÇÃO DE RESULTADOS
6. DEPLOY/IMPLANTAÇÃO
7. CONCLUSÃO






# 1. PROBLEMA DE NEGÓCIO E OBJETIVO

- O Twitter tornou-se um canal de comunicação importante em tempos de emergência.
- A onipresença dos smartphones permite que as pessoas anunciem uma emergência que estão observando em tempo real.
- Devido a isso, mais **agências** estão **interessadas** em **monitorar o Twitter** de forma programática (ou seja, organizações de auxílio em desastres e agências de notícias). Uma delas é a **Agência de Apoio a Catástrofes (AAC)**.
- Como estagiário na Agência de Apoio a Catástrofes (AAC), que desenvolve soluções de análise de dados, o meu **objetivo** é analisar um conjunto de tweets para determinar quais estão relacionados a desastres reais e quais não estão.
- A partir da identificação de tweets relacionados a desastres em tempo real, a agência pode receber alertas antecipados e prover uma resposta/auxílio mais rápido. Além disso, é possível mapear a extensão e o impacto de um desastre e engajar-se diretamente com a comunidade afetada.

## 2. DEFINIÇÃO TÉCNICA

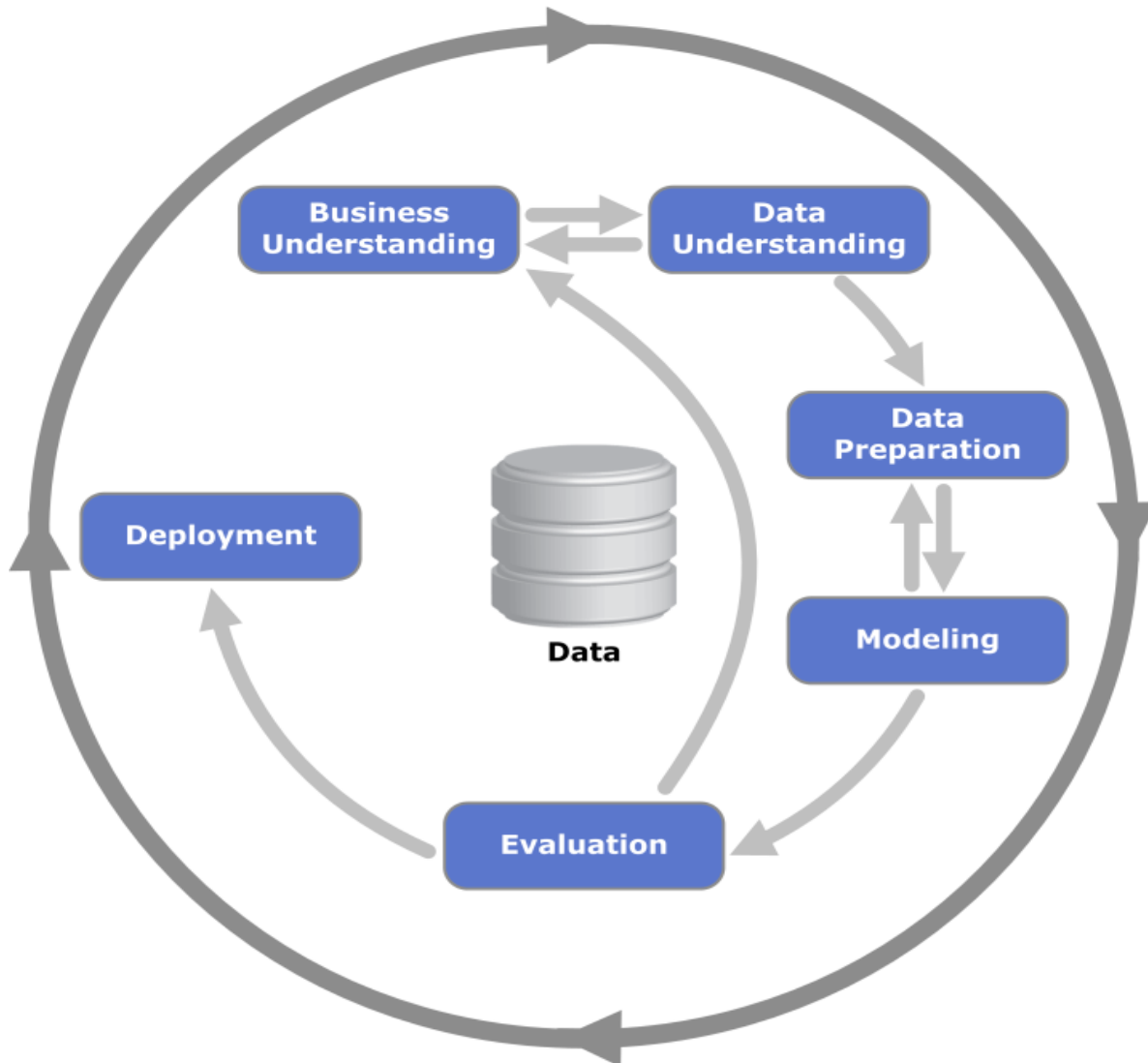
- Para resolver o problema de negócio, e conseguir identificar tweets relacionados a desastres e tweets não relacionados, coletei **dados** de milhares de **tweets** já rotulados com ambas as classes. Então, decidi aplicar a ciência de dados, mais especificamente, utilizando técnicas de processamento de linguagem natural (**nlp**), dividindo minha solução em duas tarefas que serão abordadas no próximo documento.
- Os dados coletados possuem as seguintes **variáveis**:
  - **id**: Identificador único do tweet.
  - **keyword**: Palavra-chave associada ao tweet.
  - **location**: Localização de onde o tweet foi postado.
  - **text**: Texto do tweet.
  - **target**: Categoria do tweet, 0 (tweets não associados a catástrofes) ou 1 (tweets relacionados a desastres)0.

- 
- **1. Análise exploratória de dados:** Foi realizada uma análise para desvendar padrões e **insights** ocultos nos dados acerca de tweets relacionados a desastres e tweets não associados.
  - **2. Modelagem preditiva:** Foi construído um modelo de machine learning, utilizando técnicas de processamento de linguagem natural para **prever** acuradamente a **probabilidade** de um tweet estar relacionado a uma catástrofe ou desastre.

Nesse sentido, as principais **tecnologias** e ferramentas utilizadas foram:

- Python (Pandas, Numpy, Matplotlib, Seaborn, Flask, Optuna, NLTK, Spacy, TextBlob, Scikit-Learn, Ambientes virtuais).
- Jupyter Notebook em VSCode (ambiente de desenvolvimento).
- Git e Github (versionamento de código).
- Algoritmos de machine learning para classificação.
- Estatística.

# 3. PLANEJAMENTO DA SOLUÇÃO



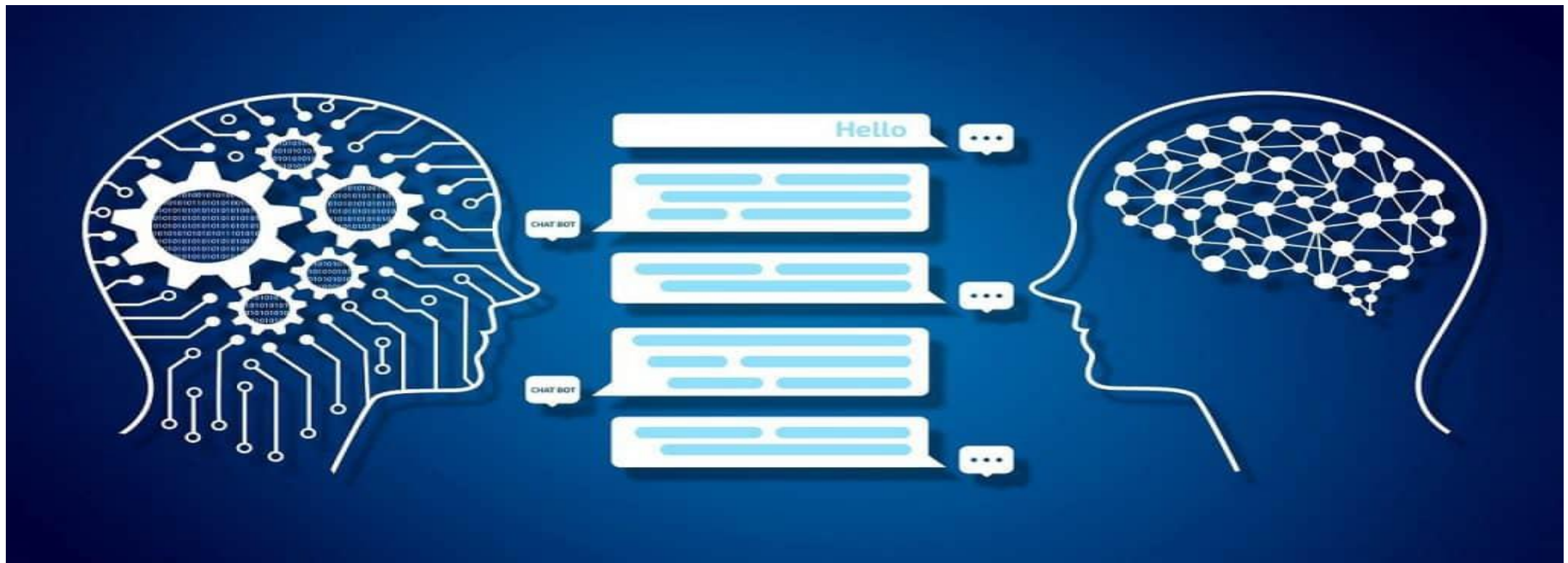
## Framework CRISP-DM:

1. Entendimento do negócio.
2. Entendimento dos dados.
3. Preparação dos dados.
4. Modelagem.
5. Avaliação.
6. Deploy ou implantação.



# \*UMA BREVE EXPLICAÇÃO SOBRE NLP





- Daqui em diante, tanto a análise exploratória de dados quanto a modelagem, para atingir os objetivos propostos e solucionar o problema de negócio, serão realizados utilizando técnicas de processamento de linguagem natural (NLP).
- NLP, ou Processamento de Linguagem Natural, refere-se à capacidade de os computadores compreenderem e interpretar a linguagem humana.
- Em resumo, NLP no contexto de prever tweets de desastres significa utilizar tecnologias que permitem que os computadores entendam e classifiquem automaticamente se um tweet está associado a um evento catastrófico ou não, com base na linguagem utilizada no tweet.



# 4. ANÁLISE EXPLORATÓRIA DE DADOS

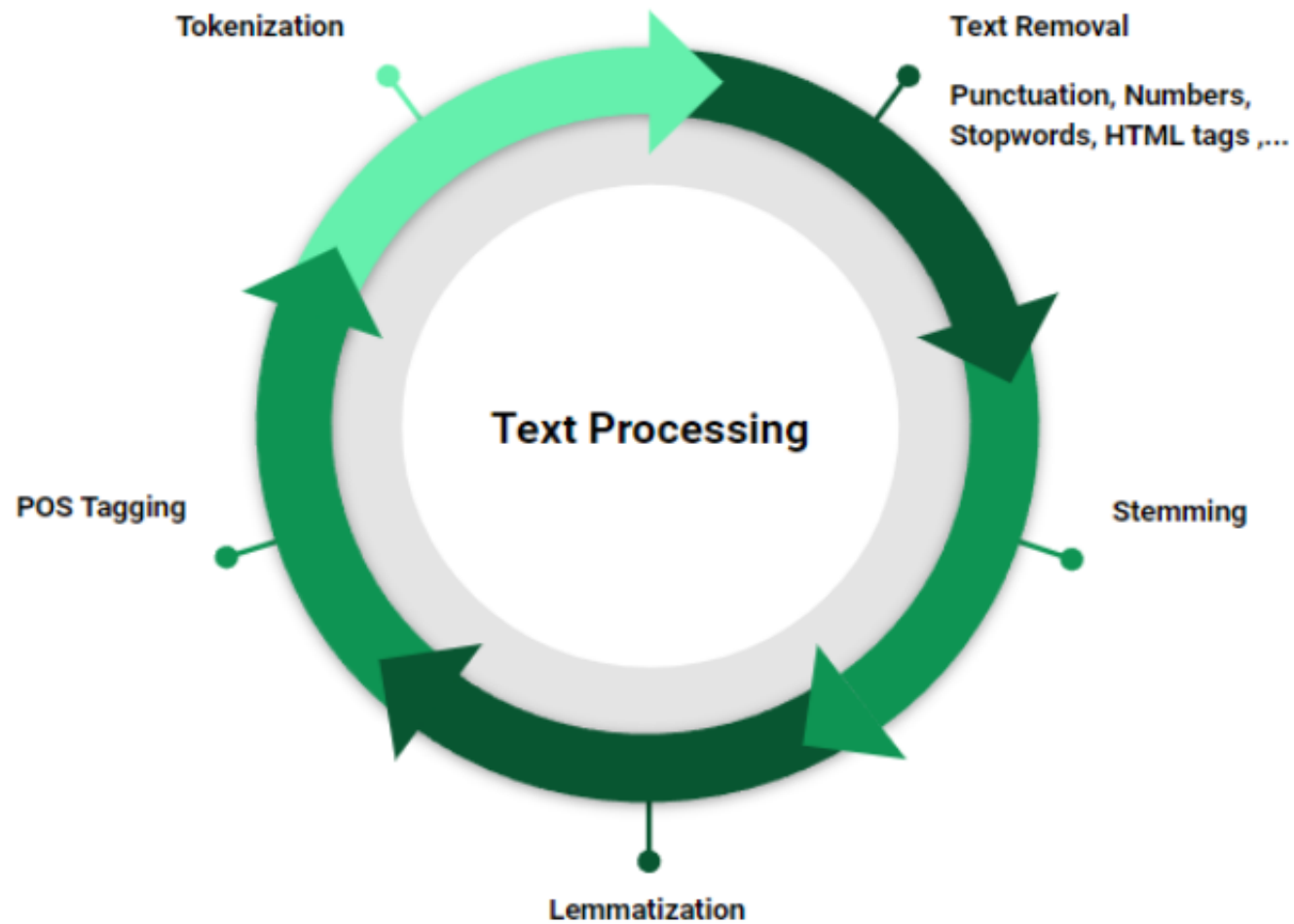
Antes de aplicar qualquer algoritmo de machine learning, foi realizada a análise dos dados, com o **objetivo** de obter **insights** e padrões acerca de tweets relacionados a desastres e tweets não relacionados. Esta etapa engloba os processos 1 e 2 do framework citado no slide anterior.

A fim de alcançar os resultados esperados, defini, antes de tudo, as **perguntas a serem respondidas**:

1. Qual é a distribuição da variável dependente - tweets relacionados e não relacionados a desastres?
2. Existe alguma relação entre o tamanho do texto e a probabilidade de um tweet estar relacionado a um desastre?
3. Quais são as palavras-chave mais frequentes associadas aos tweets de desastres? E aos tweets não relacionados a desastres? E as menos frequentes?
4. Quais são as palavras mais frequentes associadas aos tweets de desastres? E aos tweets não relacionados a desastres? Existe alguma diferença considerando os textos limpos e os textos da forma original?
5. Quais tendências, representadas por hashtags, são mais frequentes entre tweets relacionados a desastres? E entre os tweets não relacionados a desastres?
6. Qual é a emoção dominante nos tweets relacionados a desastres e não relacionados a desastres?



## 4.1 PIPELINE DE ANÁLISE



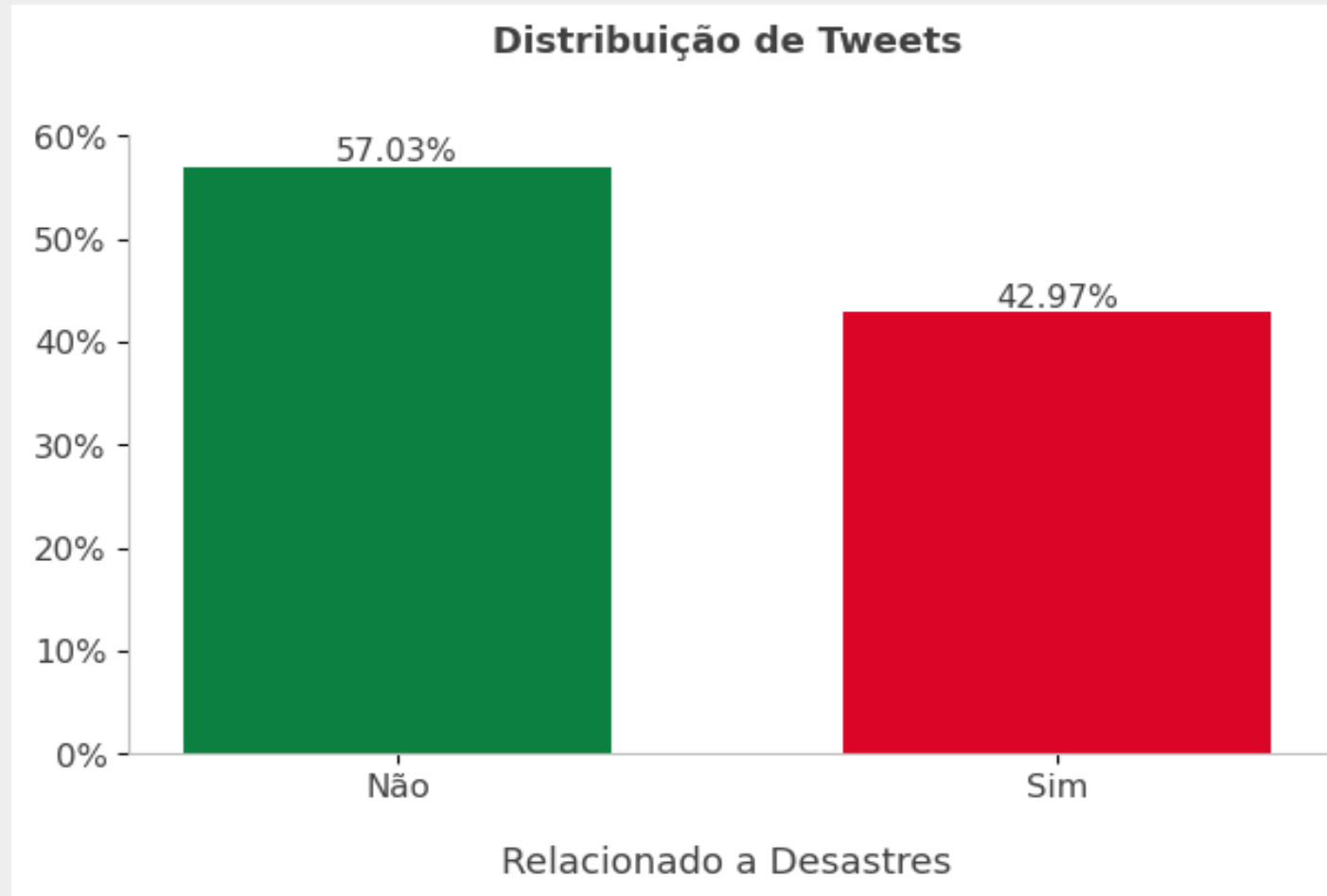
- Inicialmente, realizamos uma rápida visualização e entendimento da base de tweets. Incluem-se neste passo atividades como acesso a **informações gerais** (número de linhas e de colunas e tipos de dados das variáveis), identificação de valores **nulos** e identificação de valores **duplicados**. De início, foi possível perceber que os tweets necessitam de **limpeza** (alguns incluem nomes de usuários, por exemplo) e que não é necessário tratar valores nulos, pois 'location' apresenta altíssima cardinalidade e categorias iguais podem estar presentes nas mais variadas formas, sendo descartável para a análise.
- Para responder as perguntas propostas, as colunas 'keyword' e 'text' supriram a nossa necessidade. Portanto, foi necessário apenas realizar uma **limpeza** na variável contendo os **tweets** ('text') aplicando técnicas de **NLP**, considerando que 'keyword' já estava em um formato adequado.
- Para essa **limpeza**, as seguintes **atividades** foram realizadas:
  - Remoção de links, tags html, @s de usuários, pontuações, caracteres especiais, números e stopwords (palavras sem natureza informativa, como 'is').
  - Padronização da formatação de todos os termos para lowercase.
  - Lematização utilizando part of speech tags para reduzir as palavras a suas raízes ou lemas garantindo que o resultado será um termo existente na língua inglesa. Exemplo: "burning" -> "burn".
  - Tudo isso foi efetuado através de expressões regulares, que buscam padrões de texto dentro dos tweets. Por exemplo, `@[\w]*` identifica qualquer sequência que comece com "@" seguida por palavras.

- Os **objetivos** dessas transformações foram:
  - Reduzir a dimensionalidade na etapa de modelagem, filtrar apenas os termos relevantes do ponto de vista semântico e de contexto nos tweets, padronizar ocorrências de palavras de mesmo sentido, possibilitando um melhor reconhecimento de padrões e agregações tanto na obtenção de insights quanto na construção do modelo.
  - Considere o **exemplo**:
  - **"I'm a runner and I saw my friend @joe123 running!!!"**
  - Para a máquina e para a nossa análise, é muito mais interessante que a frase seja transformada para **"runner saw friend run"**. Note como @s de usuários, pontuações, stopwords como "a" e palavras derivadas que poderiam ser reduzidas a sua raiz não apresentam contribuição semântica/contextual e é possível simplificar a frase removendo-as. Note como tudo torna-se mais padronizado em lowercase. Este é o nosso objetivo para uma análise e modelagem de qualidade!
  - Uma observação interessante é que não removi hashtags. Afinal, elas indicam tendências no Twitter e fornecem insights valiosos para a nossa análise!

## 4.2 PRINCIPAIS INSIGHTS OBTIDOS

## Nuvem de Palavras para Tweets Relacionados a Desastres

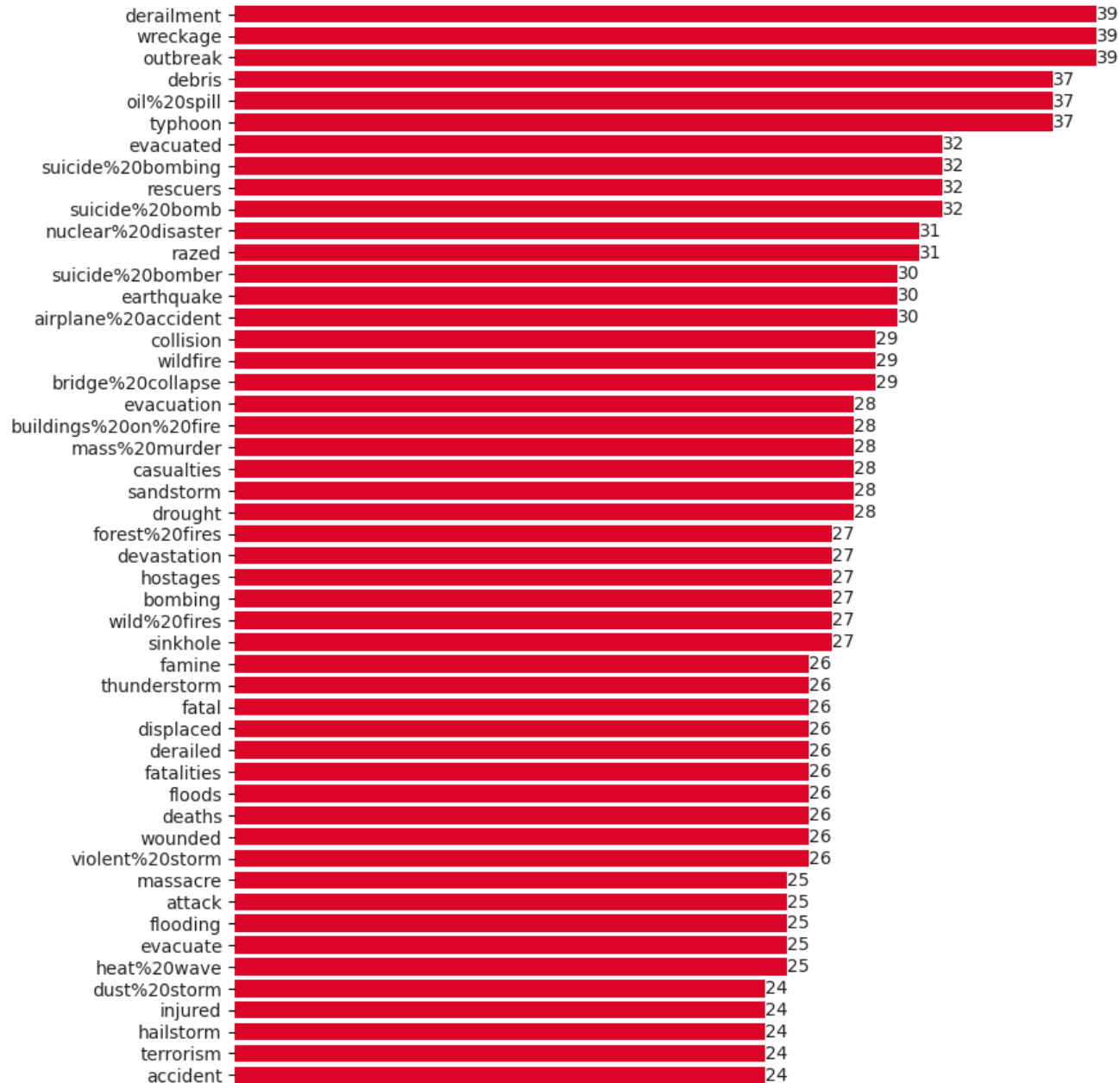




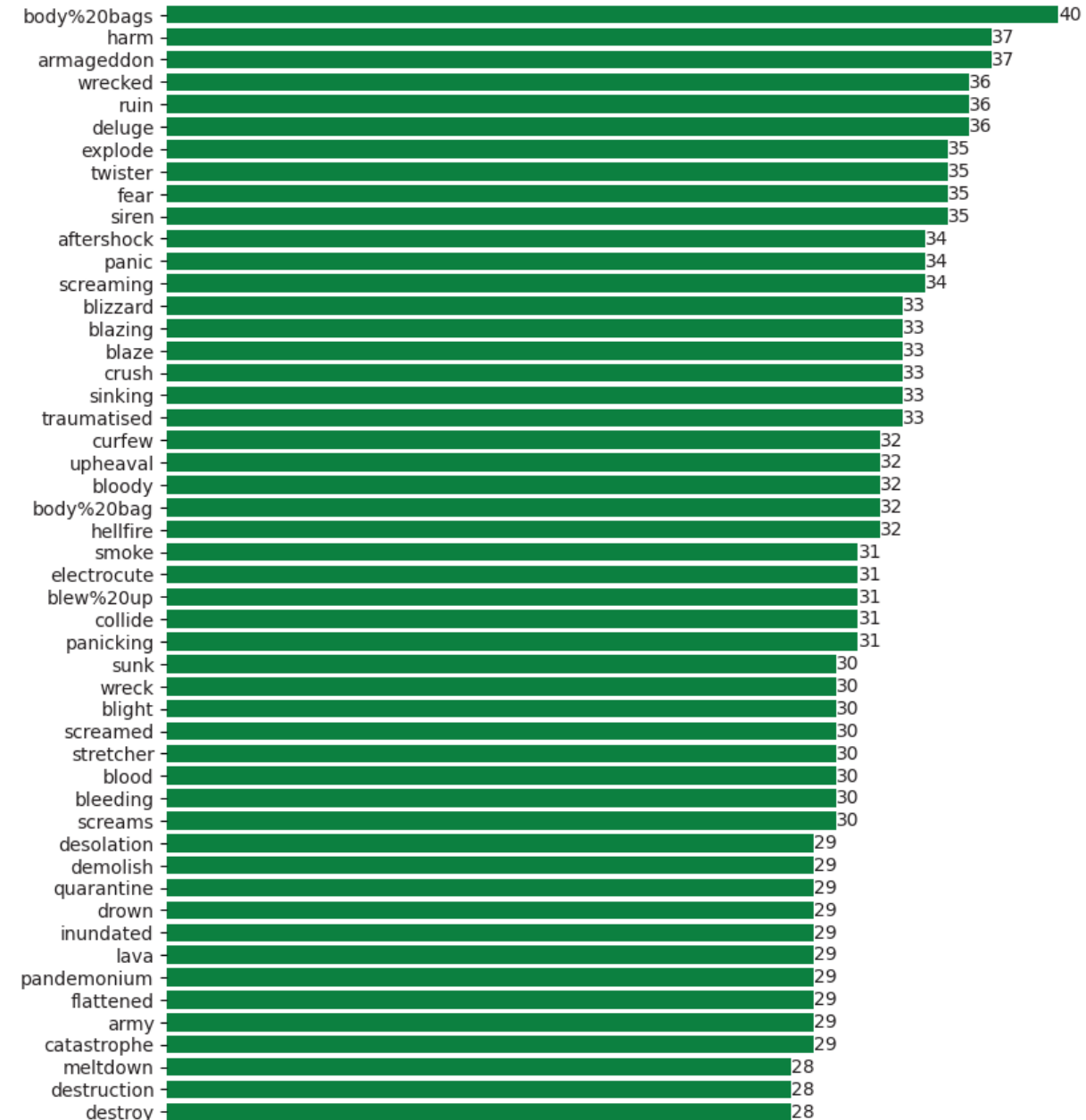
- Aproximadamente 43% dos tweets da base de dados são relacionados a desastres de fato.
- Isso indica que há um leve desbalanceamento da variável resposta. Para fins de rigorosidade, apesar de ser sutil, serão adotadas técnicas para lidar com isso na modelagem.



**Top 50 Palavras-chave em Tweets Relacionados a Desastres**



**Top 50 Palavras-chave em Tweets NÃO Relacionados a Desastres**



- **Tweets relacionados** a desastres têm uma ênfase clara em **eventos naturais, acidentes e atos de violência**.

- Exemplos de **palavras-chave**:

- **Desastres naturais**: "earthquake", "wildfire", "heat wave", "dust storm".
- **Acidentes**: "derailed", "accident".
- **Violência e terrorismo**: "suicide bomber", "bombing", "terrorism".
- **Saúde**: "fatalities", "wounded".
- **Desastres ambientais**: "oil spill", "forest fires".



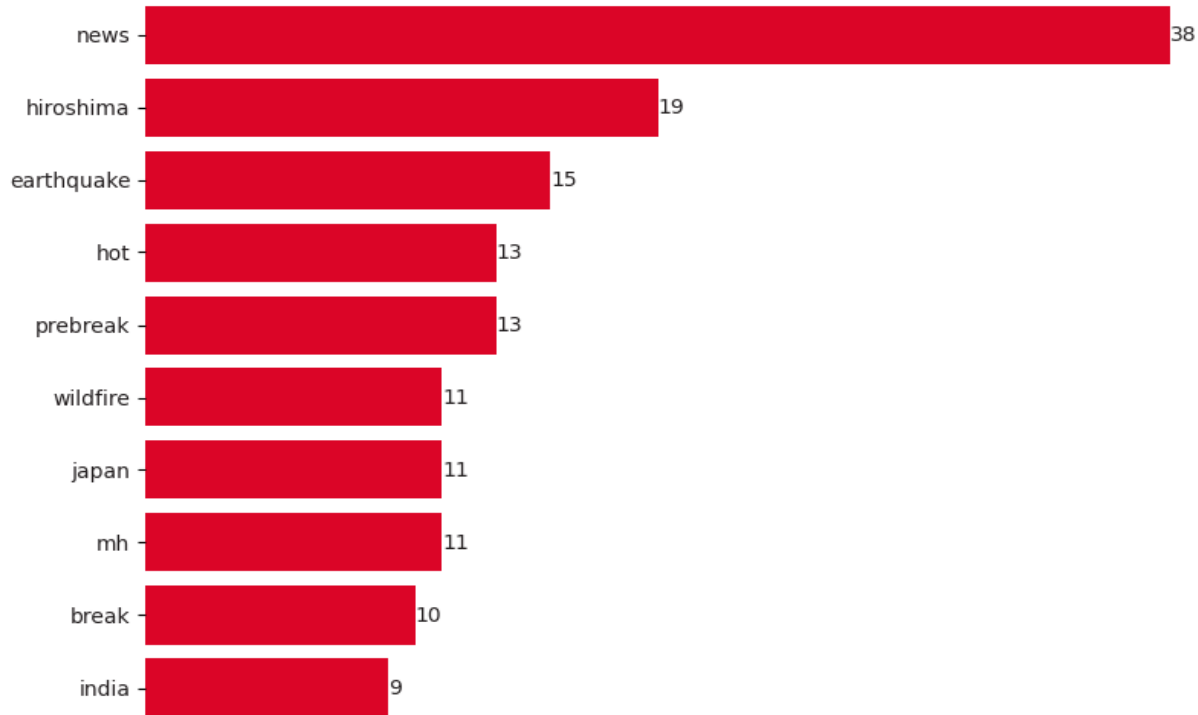
- **Tweets não relacionados** a desastres têm uma ênfase em **emoções intensas, cenários caóticos e elementos associados a consequências graves**. Isso faz sentido, pois, em se tratando de uma metáfora ou uma referência a algo que não é real, o **exagero** tende a prevalecer.

- Exemplos de **palavras-chave**:

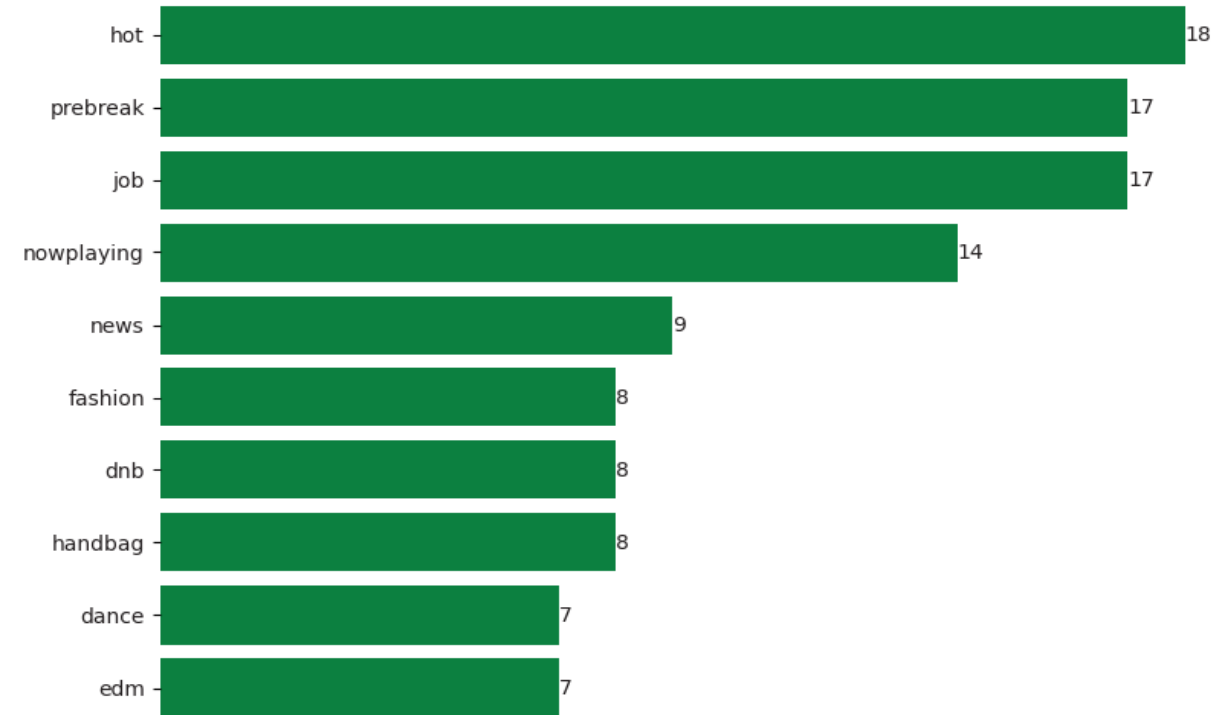
- **Destruição e caos**: "ruin", "explode", "destruction".
- **Emoções**: "panic", "fear".
- **Consequências graves**: "body bags", "bloody".
- **Referências a ambientes de emergência**: "army", "quarantine", "catastrophe".



Top 10 Hashtags em Tweets Relacionados a Desastres

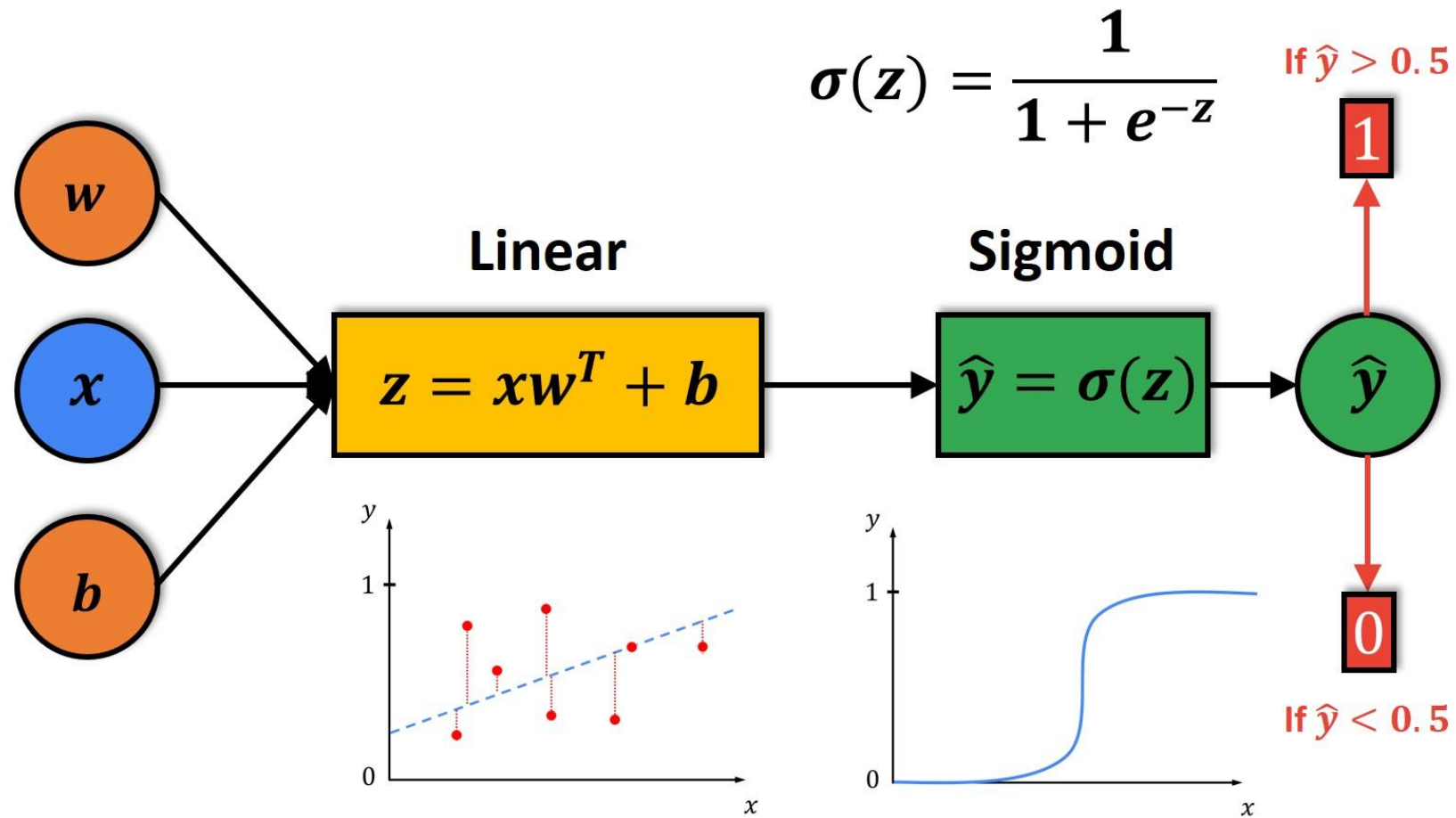


Top 10 Hashtags em Tweets NÃO Relacionados a Desastres



- É possível perceber que, de longe, a **hashtag** mais **comum** para **tweets relacionados** a desastres é a **"#news"**. Isso nos indica que a maioria desses tweets corresponde a **canais de notícia**, relatando os eventos em questão. Em seguida, **"earthquake"** e **"wildfire"**, ratificam a maior presença de palavras-chave relacionadas aos **desastres** em si. Neste caso, desastres ambientais. Outrossim, a presença de palavras como **"hiroshima"**, **"japan"** e **"india"** sugere uma ênfase em eventos específicos relacionados a desastres naturais em áreas geográficas específicas.
- Já nos **tweets que não estão relacionados** com desastres há um conteúdo diversificado nas hashtags. Isso pode ser visto em palavras como **"hot"**, **"job"**, **"fashion"**, **"dnb"** e **"dance"**. Dito isso, é notório um foco em **entretenimento, música e cultura**, indicando que esses tweets podem ser mais voltados para o lazer e interesses pessoais. Finalmente, a presença de **"nowplaying"** e **"fashion"** sugere uma possível associação com atualizações diárias sobre música e moda.

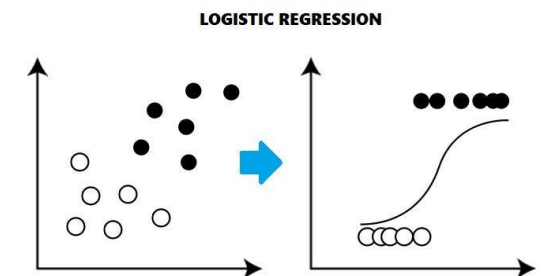
## 5. MODELAGEM E AVALIAÇÃO DE RESULTADOS



\*ALERTA - CONTEÚDO MAIS TÉCNICO

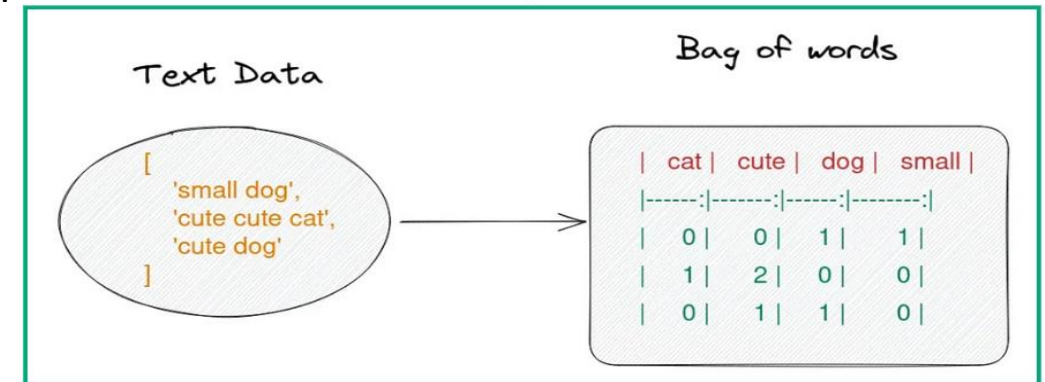


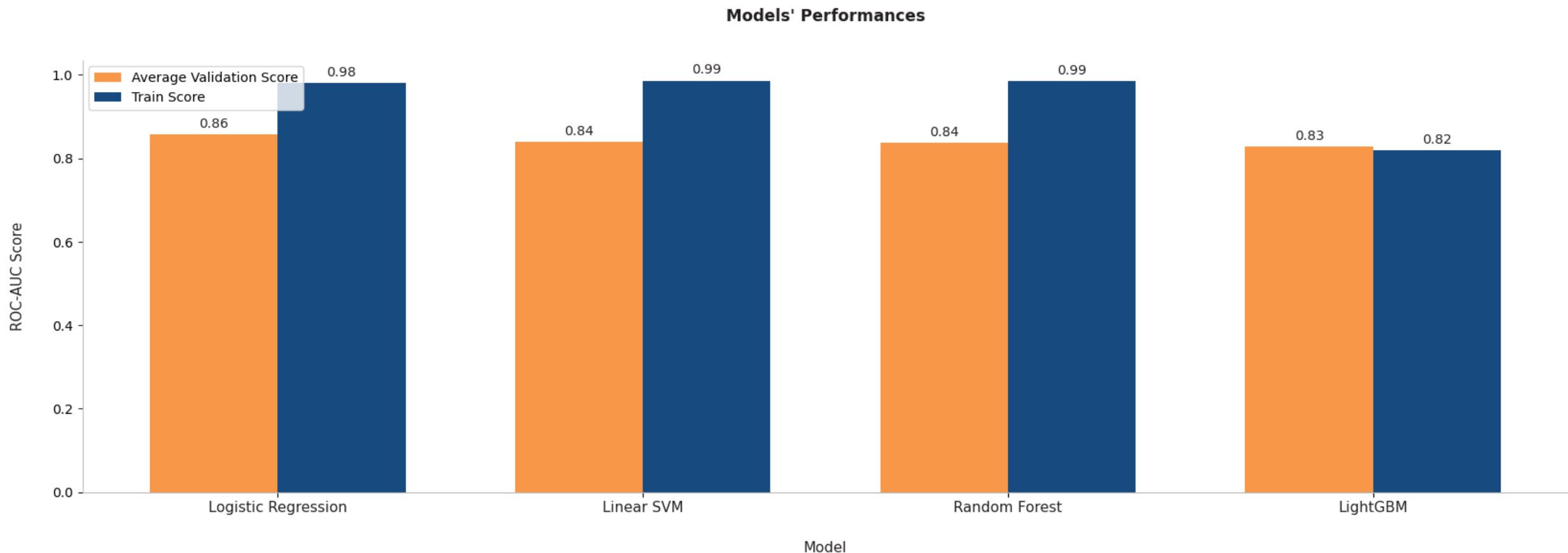
- Para solucionar o problema de negócio, isto é, permitir que a agência AAC identifique em tempo real tweets relacionados a desastres e não relacionados, iremos treinar um **modelo de machine learning** com os dados limpos já obtidos (como mencionado anteriormente).
- O **objetivo** da modelagem é construir um modelo de machine learning capaz de **prever** acuradamente a **PROBABILIDADE de um tweet estar de fato relacionado a uma catástrofe**.
- A escolha da abordagem probabilística leva em consideração a **informação de confiança**. Isso nos permitirá **avaliar quão provável** é que um determinado tweet represente uma catástrofe real. Tal abordagem facilita a **gestão de risco** por parte da Agência de Apoio a Catástrofes (AAC), possibilitando a **priorização** da atenção para tweets com maior chance de associação. Além disso, reduz a propagação de notícias falsas, uma vez que rótulos binários poderiam classificar erroneamente vários tweets, especialmente ao utilizar pontos de corte mais baixos para a classificação (balanceamento do trade-off precision-recall).
- Nesse sentido, **métricas** como ROC-AUC, PR-AUC e Brier Score são **priorizadas**. Entretanto, olhamos para diversas outras. Por exemplo, mesmo adotando o critério probabilístico, é interessante obter um bom recall, uma vez que é melhor que alcancemos o maior número de tweets relacionados a desastres de fato possível.
- O **pipeline de modelagem** consiste nos seguintes passos e será abordado em detalhes nas próximas transparências:
  1. Split dos dados em treino e teste, estratificado.
  2. Pré-processamento de dados (vetorização).
  3. Comparação de diversos modelos através de validação cruzada k-fold estratificada.
  4. Tunagem de hiperparâmetros do modelo com `class_weight` e otimização bayesiana.
  5. Avaliação final no conjunto de testes.
  6. Deploy.





1. **Split dos dados em treino e teste, estratificado:** Primeiro, dividi os dados limpos em conjuntos de treino e teste. Os dados de teste serão isolados e só serão utilizados para avaliação do modelo final, pois estes são dados que o modelo nunca deve ver, a fim de simular o ambiente de produção e obter uma mensuração de performance confiável. A estratificação tem por objetivo reproduzir a proporção da variável resposta nas diferentes amostras, isto é, que ambos os conjuntos contenham 43% de tweets relacionados a desastres, representando fielmente a distribuição real.
2. **Vetorização/Pré-processamento dos dados:** Aqui é aplicada a vetorização, que consistem em representar texto (os tweets) como vetores numéricos a fim de treinar algoritmos de machine learning. Estes algoritmos efetuam cálculos matemáticos, portanto, não lidam com texto. Três abordagens comuns são o Bag of Words, Word Vector e BERT. Escolhi aplicar o Bag of Words. Isso porque, apesar de essa técnica não levar em conta a semântica, a ordem e a estrutura gramatical das palavras (desvantagens), meus recursos computacionais são limitados, de forma que esse processo simples facilita o treinamento, validação e o deploy (vantagens). Além disso, considerando a limpeza efetiva que fiz, é esperado um desempenho bom, mesmo com o tratamento mais simples em questão. O **Bag of Words** consiste de 3 passos principais:
  1. **Tokenização:** Divide o tweet em palavras individuais (tokens), considerando limpezas como remoção de pontuações, stopwords, lowercase, entre outras já realizadas no nosso caso.
  2. **Construção do Vocabulário:** Cria um vocabulário único - Todas as palavras únicas presentes nos tweets.
  3. **Vetorização:** Representa cada tweet como um vetor, onde cada posição no vetor corresponde a uma palavra do vocabulário (coluna) e o valor na posição indica a contagem da palavra no tweet. Como resultado, temos uma matriz esparsa (guarda eficientemente os zeros).





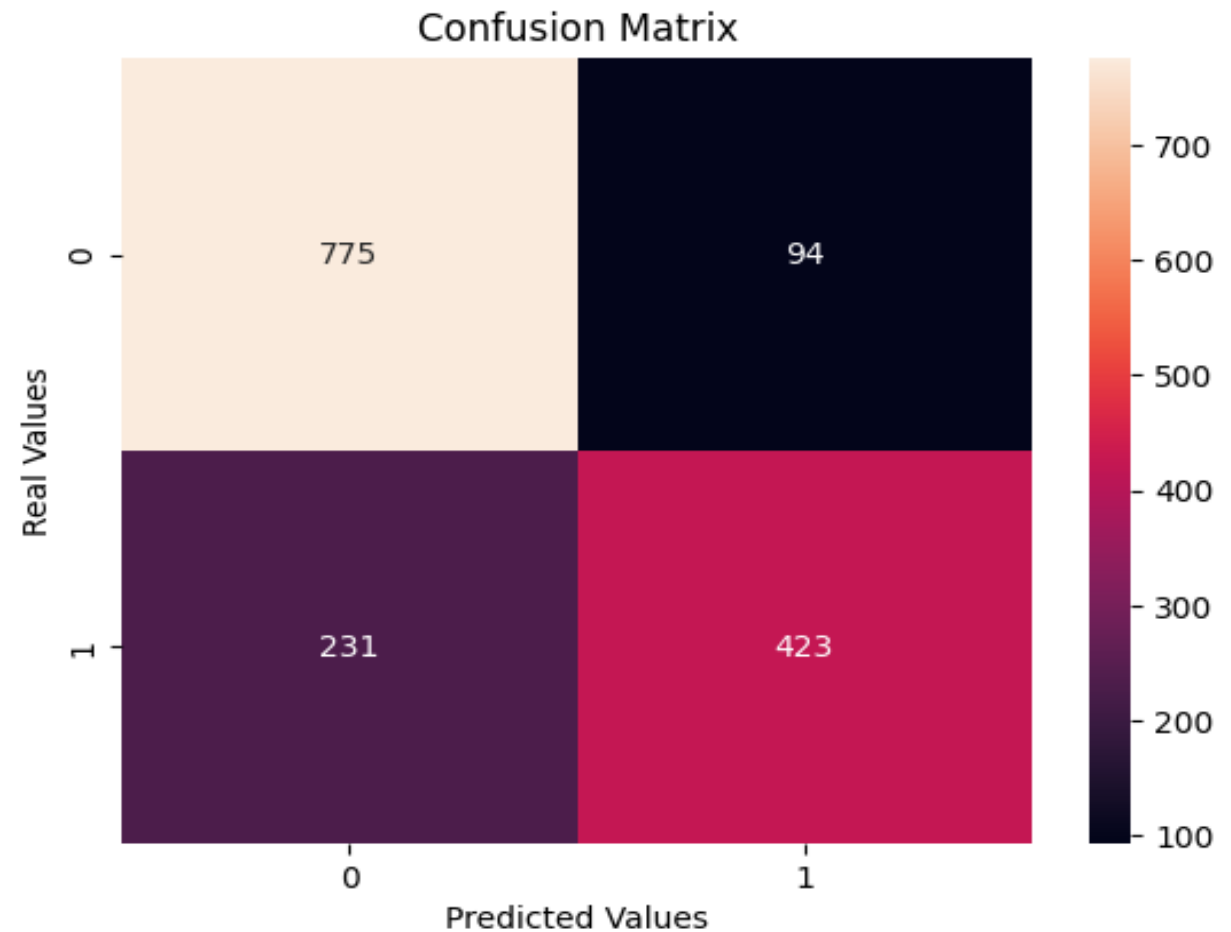
**3. Validação cruzada k-fold estratificada:** A fim de selecionar um modelo potencial para tunagem de hiperparâmetros e avaliação final, avaliei os quatro modelos acima com a validação cruzada k-fold. A validação cruzada k-fold permite obter uma mensuração de performance confiável de um modelo, dividindo o conjunto de treinamento em k conjuntos, e avaliando em cada um dos k-ésimos folds um estimador treinado nos outros k-1 restantes, agregando os scores com a média ao final. Novamente, a estratificação permite manter a proporção desbalanceada do target. Testei esses modelos pois, a Regressão Logística possui probabilidades que mais se aproximam das probabilidades reais calibradas, o Linear SVM costuma performar bem em classificações de texto, e o Random Forest e o LightGBM são ensembles, portanto, é esperado que performem melhor que os outros.

- O modelo de **Regressão Logística** apresentou o maior ROC-AUC score médio de validação, possui probabilidades mais próximas de probabilidades reais calibradas e um potencial para melhorias dado o overfit (através de técnicas de regularização l1 e l2, por exemplo). Portanto, ele foi escolhido.

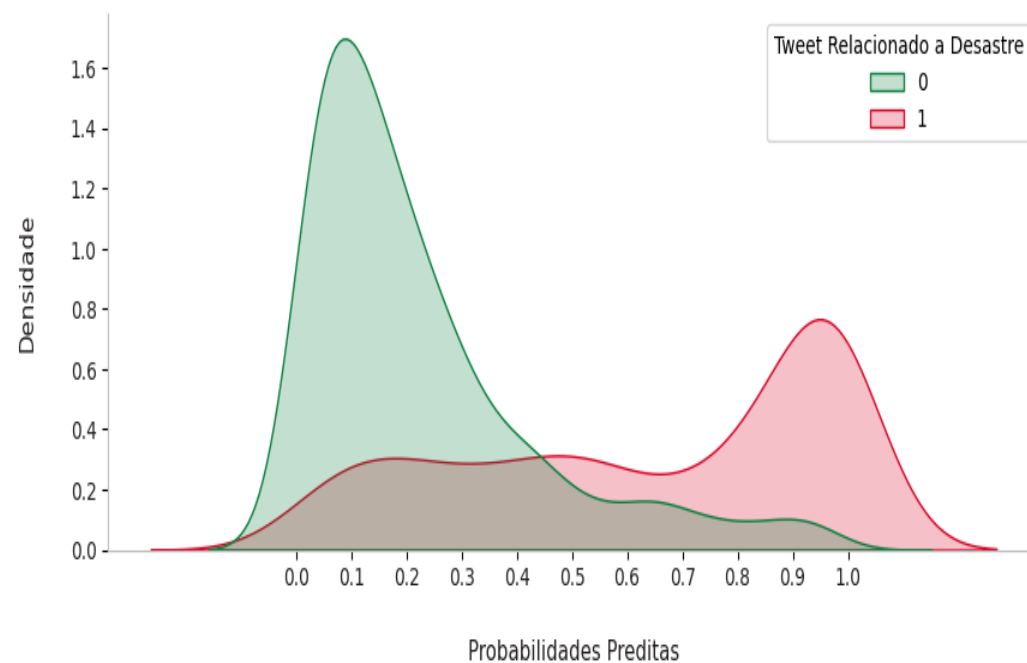
**4. Tunagem de hiperparâmetros:** A tunagem de hiperparâmetros foi aplicada através da otimização bayesiana, dado que ela explora inteligentemente o espaço de hiperparâmetros balanceando o trade-off exploration-exploitation.

**5. Avaliação final no conjunto de testes:** Os resultados foram ótimos!

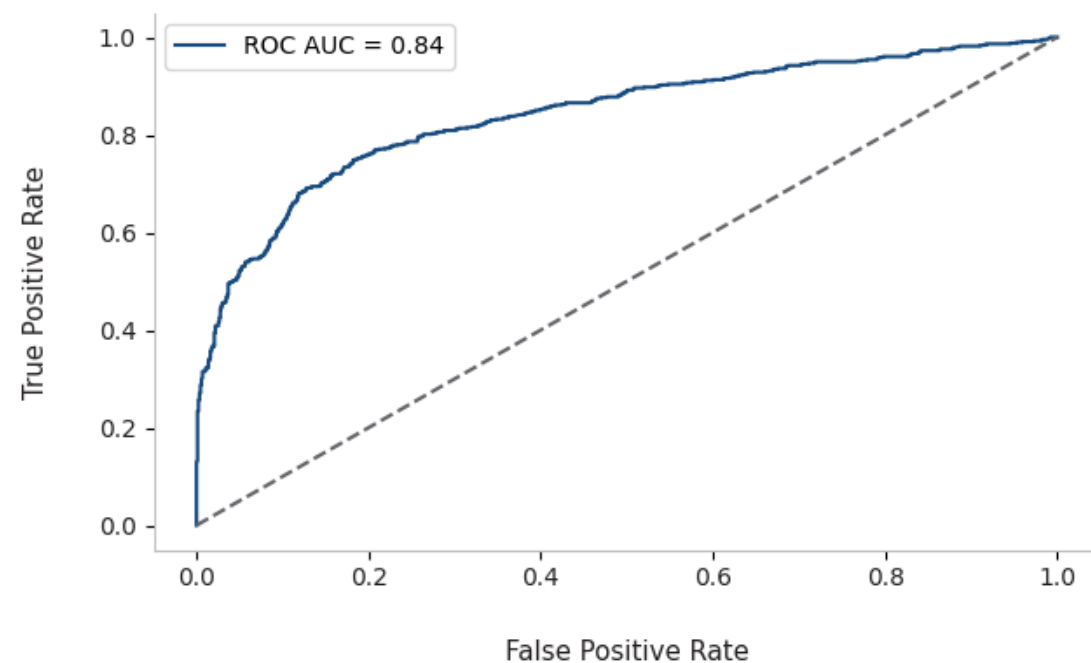
- **Recall (0,65):** O modelo identifica 65% dos tweets relacionados a desastres. Na prática, observando a matriz de confusão, o modelo foi capaz de prever corretamente 423 dos 654 tweets relacionados a desastres.
- **Precision (0,82):** De todos os tweets preditos como relacionados a desastres, 82% estavam relacionados de fato. Na prática, observando a matriz de confusão, dos 517 tweets preditos como relacionados a desastres, 423 deles realmente estavam associados.
- **ROC AUC (0,84):** Com um ROC AUC de 0,84, o modelo demonstra uma alta capacidade de diferenciar entre tweets relacionados a desastres e tweets não relacionados.



**Distribuição das Probabilidades Preditas entre Tweets Relacionados a Desastres e Não Relacionados**



**Receiver Operating Characteristic (ROC) Curve**



- Claramente há uma **separação** na **distribuição de probabilidade** de tweets relacionados a desastres e não relacionados, reforçando a qualidade do nosso modelo e o seu poder discriminante entre as duas classes.
- Nossos **scores** claramente **seguem uma ordenação**, o que é bom! É possível perceber que o percentual de tweets relacionados a desastres é muito maior para faixas de probabilidade mais altas. Isso sugere que esta Regressão Logística é confiável para prever a probabilidade de um tweet estar associado a uma catástrofe.
- Finalmente, comparando o ROC-AUC scores nas amostras de treino, teste e validação, ainda há **overfit** no conjunto de treinamento. **Entretanto**, os scores de teste e validação são semelhantes e indicam uma **excelente capacidade de generalização do modelo para novas instâncias**, com um **ROC AUC em torno de 0.85**. **Idealmente**, para **contornar este overfit**, **mais dados/tweets** deveriam ser incorporados ao conjunto de treinamento.



## 6. DEPLOY/IMPLANTAÇÃO

### Agência de Apoio a Catástrofes (AAC)

Análise de Tweets: Desastre ou Não?

**Bem vindo à Agência de Apoio a Catástrofes (AAC).**

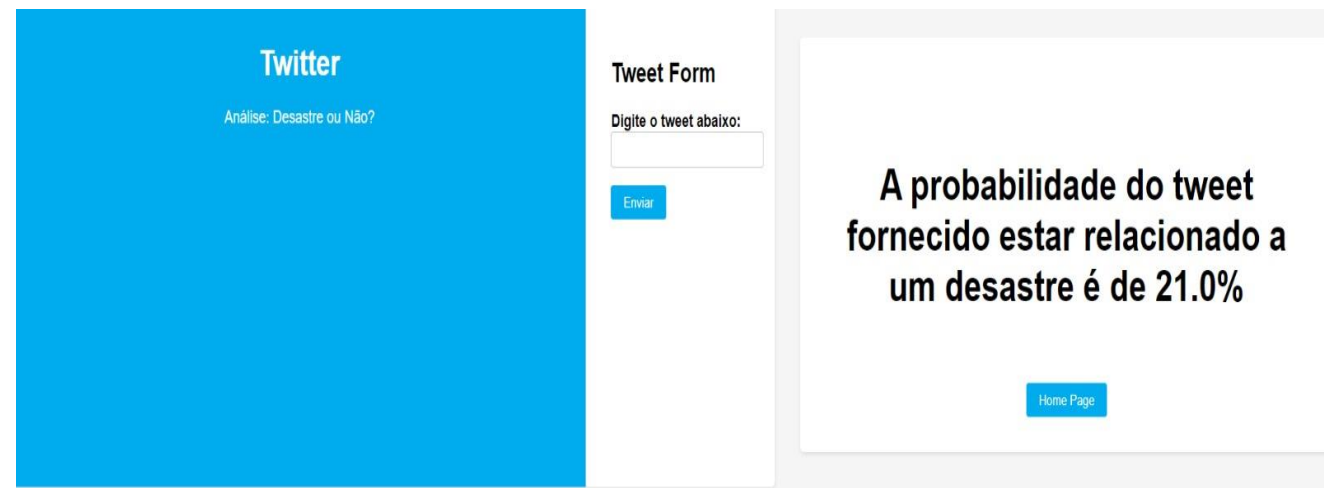
Preveja a probabilidade de um tweet estar de fato relacionado a um desastre ou catástrofe.

Prever



## 6. Deploy:

- O desenvolvimento do estudo nos notebooks foi convertido em scripts .py para produção.
- Esses scripts foram divididos em componentes de ingestão de dados, transformação de dados e treinamento de modelo, seguindo os mesmos passos do estudo.
- Uma vez desenvolvidos os componentes, foram implementados pipelines de treinamento e predição automatizados que os utilizam.
- O pipeline de treinamento executa esses componentes e obtém todos os artefatos do modelo de machine learning (modelo .pkl, preprocessor .pkl, dados de treino, teste e dados brutos), enquanto o pipeline de predição realiza as predições consumindo esses artefatos obtidos.
- Tudo isso foi implementado utilizando boas práticas como o uso de ambientes virtuais para isolamento de dependências, tratamento de exceções, logs, documentação, etc.
- Finalmente, foi desenvolvida uma API Flask integrando tudo que foi mencionado nos tópicos acima.
- O meu objetivo com isso foi seguir ao máximo um workflow real de um projeto de ciência de dados, construindo meu projeto inteiro como um pacote reproduzível.
- Entre os próximos passos, está o deploy em alguma cloud, como a aws, utilizando o serviço elasticbeanstalk.



# 7. CONCLUSÃO

- O nosso objetivo foi atingido, insights sobre os dois tipos de tweets foram desvendados e será possível prever acuradamente a probabilidade de um tweet estar relacionado a desastres com o modelo de Regressão Logística. O problema de negócio da Agência de Apoio a Catástrofes (AAC) está resolvido.
- Agora, a partir da identificação de tweets relacionados a desastres em tempo real, a agência poderá receber alertas antecipados e prover uma resposta/auxílio mais rápido. Além disso, será possível mapear a extensão e o impacto de um desastre e engajar-se diretamente com a comunidade afetada. O estagiário fez um bom trabalho e merece ser promovido :)!
- Link do repositório com o código completo: <https://github.com/allmeidaapedro/Twitter-Disaster-Analysis>

# OBRIGADO



<https://www.linkedin.com/in/pedro-henrique-almeida-oliveira-77b44b237/>



(61)99168-2702



[pedrooalmeida.net@gmail.com](mailto:pedrooalmeida.net@gmail.com)

<https://github.com/allmeidaapedro/Portfolio-Ciencia-de-Dados>