PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Cléber Rodrigo de Souza

PERCEPÇÃO AMBIENTAL ESTRANGEIRA SOBRE CONSERVAÇÃO AMBIENTAL NA AMAZÔNIA: UMA ABORDAGEM UTILIZANDO MINERAÇÃO DE TEXTO DE REDES SOCIAIS

Belo Horizonte

2022

Cléber Rodrigo de Souza

PERCEPÇÃO AMBIENTAL ESTRANGEIRA SOBRE CONSERVAÇÃO AMBIENTAL NA AMAZÔNIA: UMA ABORDAGEM UTILIZANDO MINERAÇÃO DE TEXTO DE REDES SOCIAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2022

SUMÁRIO

1. Introdução

1.1. Contextualização

A floresta tropical amazônica consiste em um dos locais mais ecologicamente importantes do planeta, em função dos seus serviços ecossistêmicos prestados e da sua enorme biodiversidade. Esta região abriga mais de 15 mil espécies de árvores (25 % do total global) (CARDOSO et al., 2017), além de ter um papel essencial na dinâmica hidrológica e biogeoquímica no continente sul-americano (BARLOW et al., 2018), principalmente no que se refere ao seu papel como estoque e sumidouro de carbono emitido por fontes antrópicas (PAN et al., 2013).

No entanto, desde a aceleração da sua ocupação na segunda metade do século XX, a região tem passado por um processo avançado de redução de cobertura vegetal, que segundo o MapBiomas, foi de 44.5 Milhões de ha entre 1985 e 2020 (MAPBIOMAS – SOUZA et al., 2020), uma área que representa 9 vezes o estado do Rio de Janeiro, por exemplo. Tal redução está associada à ocupação de áreas e expansão de atividades agrícolas na região, que muitas vezes é realizada de maneira não planejada e com danos expressivos aos ecossistemas e à sua biodiversidade.

Em função da sua importância global para o funcionamento do planeta, muita atenção tem sido dada a região nos últimos anos, seja em campanhas publicitárias, músicas, filmes e outras iniciativas políticas e sociais que tem como objetivo mobilizar a sociedade e os governos para a conservação da região. Exemplos são projetos de restauração florestal na região como o Bonn Challenge (VERDONE; SEIDL, 2017), projetos de pagamento por serviços ambientais (Floresta+ - https://www.gov.br/mma/pt-br/assuntos/servicosambientais/florestamais) e ainda o envolvimento de figuras artísticas importantes em causas relacionadas a conservação da região.

Além da importância como abrigo de grande quantidade de florestas, a região amazônica e os estados adjacentes estão associados a um grande volume de atividades econômicas relacionadas principalmente a exportação de produtos da agricultura e pecuária. O estado do Mato Grosso, por exemplo, é o maior produtor brasileiro de soja, milho, algodão e rebanho bovino, que juntamente a outras culturas fazem com que cerca de

17.6 % do valor bruto da produção agropecuária do país em 2021 seja oriundo do estado de acordo com Dados do Observatório de Desenvolvimento da Secretaria de Estado de Desenvolvimento Econômico (SEDEC - http://www.sedec.mt.gov.br/observatorio-desenvolvimento). Outros estados da região também apresentam grande volume de produção agropecuária, tendo participação importante na balança econômica do país, especialmente através da quantidade de exportações realizadas (SEDEC, 2022).

Dentro do contexto de exportação dos produtos agropecuários produzidos na região, tem crescido cada vez mais a preocupação dos mercados consumidores estrangeiros em relação a origem dos produtos oriundos da Amazônia (RAJÃO et al., 2020). Assim, um consumidor médio na Europa, Ásia ou na América do Norte, por ex, busca por garantias de que a produção daquele produto não está associada a práticas incompatíveis com a conservação ambiental dos ecossistemas da Amazônia. Como resposta do mercado a essa exigência, tem-se medidas como a atribuição de selos de certificação de produção, criação de cadeias de custódia para rastreio de etapas de produção e o desenvolvimento das chamadas abordagens ESG (Environamental, Social and Governance – Ambientais, sociais e de governança) para as organizações (CANAVARI; CODERONI, 2019; LIM et al., 2022). Com isso, o próprio mercado consumidor tem levado os produtores agrícolas a aderirem a determinadas práticas que permitam com que seus produtos adentrem determinados mercados que são mais exigentes, mas que estão dispostos a pagar mais caro pela certeza de estar não contribuindo com a destruição ambiental (CANAVARI; CODERONI, 2019; RAJÃO et al., 2020; LIM et al., 2022).

No entanto, desde 2019 tem sido registrada uma elevação da ocorrência de incêndios e de desmatamento na região (GIBBENS, 2019; BARLOW et al., 2020). Somente em 2020, SILVA-JUNIOR et al. (2020) registraram o desmatamento de 11.088 km², o que representa um aumento de 47 % e 9,5 % em relação ao observado em 2018 e 2019, respectivamente, e a maior taxa de desmatamento da década. Tal tendência foi amplamente noticiada, tendo motivado uma ampla pressão internacional de países sobre o governo brasileiro, em busca de garantias da implementação de medidas para uma proteção mais efetiva da floresta (BARLOW et al., 2020).

Dentre deste cenário, a atenção para a origem dos produtos brasileiros se intensificou nos centros de comercialização ao redor do mundo, influenciando inclusive a comercialização de produtos de outras regiões do país (BARLOW et al., 2020; RAJÃO et al., 2020). No entanto, é um desafio para a comunidade científica e órgãos governamentais avaliar como os consumidores percebem tais eventos, tendo em vista o impacto que esta mudança de percepção pode influenciar a aceitação dos produtos e a importância do desenvolvimento de formas de certificação (CANAVARI; CODERONI, 2019).

A avaliação da chamada "percepção ambiental", que consiste basicamente no entendimento de uma pessoa ou grupo em relação a um componente do ambiente, é uma forma de se buscar avaliar esta relação entre o público consumidor e a região amazônica, tendo como objetivo os produtos que são originados dali (OZDEMIR, 2010; BENNET, 2019). As formas mais comuns de avaliar a percepção ambiental de um grupo incluem aplicação de questionários, realização de entrevistas e audiências públicas, entre outas formas (OZDEMIR, 2010; SILVEIRA-JUNIOR et al., 2021). No contexto atual de digitalização e grande importância de redes sociais, este trabalho traz uma nova abordagem utilizando informações provenientes de postagens relacionadas ao tema feitas em uma rede social utilizando exclusivamente a língua inglesa. Com isso, tem-se uma forma rápida, escalável e de grande alcance geográfico e capacidade analítica.

1.2. O problema proposto

O problema aqui proposto envolve a análise de um grande dataset de publicações de redes sociais feitas em inglês associados à região amazônica, tendo como objetivo explorações a percepção ambiental estrangeira sobre a região. A seguir serão dadas mais informações sobre as formas de análise, os dados coletados e os objetivos, utilizando para isso a ferramenta dos 5-Ws:

- **(Why?) Por que avaliar a percepção ambiental estrangeira sobre a região Amazônica?**

Avaliar a percepção ambiental estrangeira sobre uma região tão ambientalmente crucial como a Amazonia é importante por diversos aspectos: **primeiro**, esta percepção está diretamente associada a percepção estrangeira sobre o Brasil, o que pode influenciar o consumo de produtos comerciais, agrícolas e midiáticos; **segundo,** fornece um bom indicativo de quão permeável são as redes sociais, em que as percepções podem corresponder ou não aos dados obtidos por órgãos de pesquisa, sobre queimadas, por ex; **terceiro,** entender como este público externo percebe os acontecimentos locais pode ajudar no planejamento de estratégias de comunicação que possam buscar potencializar a sua participação em iniciativas de conservação da região.

- **(Who?) Quais foram os dados coletados?**

Os dados analisados de percepção ambiental estrangeira sobre a região amazônica consistem em publicações não privadas realizadas na rede social Twitter que utilizaram a língua inglesa como idioma principal e que incluíram em seu conteúdo o termo "*Amazon rainforest*", que é a denominação mais comum em língua inglesa para se referir a região amazônica. Destas publicações foram coletadas o texto principal escrito pelo usuário, assim como os seus metadados associados (data de publicação, conta que publicou, se é um retweet ou não, entre outras coisas).

- **(What?) Quais os objetivos a serem alcançados com esta análise?**

Com esta análise, a ideia é identificar através de ferramentas de mineração de textos quais os principais padrões de percepção estrangeira sobre a região, identificando principais termos, qual o principal sentimento presente nas publicações (positivo/negativo), suas variações temporais e principais associações entre termos. Para isso, será analisado o conteúdo textual das publicações ao longo do tempo, que será particionado em termos (*tokenization*) para em seguida ser analisado quanto aos objetivos propostos.

- **(Where?): Quais são os aspectos geográficos relacionados à análise?**

Os dados coletados para análise estão associados ao domínio fitogeográfico da Amazônia, ou simplesmente “floresta Amazônica”, que se distribui por nove países na América do Sul, porém tendo 60 % de sua extensão localiza no Brasil, principalmente na região Norte do país. A região corresponde à maior área de floresta tropical úmida do planeta, sendo reconhecia como uma das principais áreas do mundo para conservação da biodiversidade e serviços ecossistêmicos.

- **(When?): Qual o período está sendo analisado?**

O período de análise consiste em 18 semanas, compreendidas entre os dias 31 de maior e 04 de outubro de 2021. Este período coincide com uma transição de estações climáticas na região, que historicamente está associada eventos de distúrbios antrópicos.

1.3. Objetivos

Este trabalho tem como objetivo principal avaliar a percepção ambiental estrangeira sobre a região amazônica, buscando identificar principais associações, sentimentos e tendências temporais de variações de importância de termos. Especificamente, buscou-se responder os seguintes questionamentos:

i) Quais os principais termos presentes em publicações estrangeiras em redes sociais relacionadas à região amazônica e como a sua importância variou ao longo da série temporal avaliada?

ii) A composição de termos nas publicações estrangeiras sobre a região está associada mais fortemente a sentimentos positivos ou negativos? Existe uma variação destas proporções?

iii) Quais são as principais regras de associação entre termos identificadas em publicações estrangeiras sobre a região?

2. Coleta de Dados

Os dados de publicações sobre a Amazonia foram obtidos na rede social Twitter, utilizando para isso o pacote *twitteR* versão 1.1.9 (GENTRY et al., 2016) no software R v. 4.2.0 (R CORE TEAM, 2022), que se conecta a API de desenvolvimento disponibilizada pela plataforma (https://developer.twitter.com/en) para obter as informações. Contudo, antes de realizar a coleta é necessário que seja criado um projeto na plataforma mediante aprovação da equipe do Twitter, a partir do qual são fornecidas chaves de acesso aos dados e um código de identificação do projeto. O projeto que originou esse trabalho está cadastrado com o APP ID 19715654.

A obtenção das publicações ("tweets") foi feita utilizando a função *searchTwitter*, que realiza uma busca no Twitter baseada em uma série de parâmetros definidos dentro da função e de acordo com suas limitações. Dentre estes parâmetros estão a definição do assunto de interesse, do idioma de publicação, do intervalo temporal de busca, da inclusão ou não de retweets e do número máximo de tweets a retornar. Já as limitações principais compreendem o limite máximo de 18000 tweets por busca e o intervalo temporal máximo de 8 dias anteriores a execução da função.

Assim, neste trabalho, a busca por foi publicações realizadas no idioma inglês nos 8 dias anteriores a cada execução do código que contivessem o termo "Amazon rainforest", que é a forma mais comum se referir a Floresta Amazônica em língua inglesa. A busca também incluiu retweets, considerando que a replicação de um mesmo conteúdo conta é um indicativo da sua importância e está associada a percepção ambiental. Em função da limitação temporal, a coleta foi realizada ao longo de 18 semanas, definindo-sempre o número máximo de tweets de retorno. O elevado número de semanas foi adotado para que a amostra de dados contivesse um intervalo temporal que compreendesse a transição de estações climáticas observada na região. Os intervalos de cada semana de coleta estão apresentados na Tabela 1.

Tabela 1: Data inicial e final do intervalo de coleta de cada semana de coleta de dados realizada.

Semana	Data inicial	Data final
1	31/05/2021	07/06/2021
2	07/06/2021	14/06/2021
3	14/06/2021	21/06/2021
4	21/06/2021	28/06/2021
5	28/06/2021	05/07/2021
6	05/07/2021	12/07/2021
7	12/07/2021	19/07/2021
8	19/07/2021	26/07/2021
9	26/07/2021	02/08/2021
10	02/08/2021	09/08/2021
11	09/08/2021	16/08/2021
12	16/08/2021	23/08/2021
13	23/08/2021	30/08/2021
14	30/08/2021	06/09/2021
15	06/09/2021	13/09/2021
16	13/09/2021	20/09/2021
17	20/09/2021	27/09/2021
18	27/09/2021	04/10/2021

O objeto de dados gerado pela execução do código em cada semana tem 16 colunas, sendo que algumas são nulas ou não trazem informações relevantes para este trabalho. Assim, foi feito uma seleção de colunas principais, assim como adicionadas outras colunas identificadoras que pudessem facilitar o trabalho com os dados. Vale ressaltar que neste momento já foi realizada a remoção de publicações duplicadas que possam ter entrado em mais de uma coleta, usando para isso um código identificador existente. Após a remoção, os dados de todas as semanas foram unidos em um só *dataset* de 63.773 linhas que tem as colunas apresentada as seguir:

Tabela 2: Lista de colunas existentes no conjunto de dados, juntamente a sua descrição e tipo.

Variável	**Descrição**	**Tipo**
week	Semana de coleta dos dados	Numérica
initial_date	Data de início do intervalo	Data
final_date	Data de término do intervalo	Data
id_tweet	Identificador único do tweet	Numérica
date_time	Data e horário de publicação	Data
day	Dia de publicação	Data
hour	Horário de publicação	Data
text	Conteúdo textual do tweet	String
user	Login do usuário	String
retweetCount	Contagem de retweets	Numérica
is_retweet	Identificador se é aquela publicação é ou não retweet	Boleana

3. Processamento/Tratamento de Dados

O processamento dos dados coletados se iniciou com a seleção e união de dados das várias semanas e remoção de tweets duplicados, que teve como resultado um dataset de 574051 linhas, em que cada uma delas corresponde a uma publicação feita na rede social. Na sequência, iniciou-se a etapa de limpeza de dados, na qual foram realizadas uma série de etapas de padronização e remoção:

i) Remoção de hashtags;

ii) Remoção de palavras com @ (menções);

iii) Remoção de “https://” e “http://”;

iv) Remoção de caracteres gráficos como emotions;

v) Remoção de pontuação.

vi) Remoção de caracteres de controle;

vii) Remoção de números;

viii) Substituição de “_” por espaço;

ix) Remoção espaços excessivos e desnecessários;

x) Remoção de quebras de linha.

Na sequência foi realizada a etapa de Tokenização (“tokenization”), que consiste na quebra do texto dos tweets em palavras isoladas que vão ocupar as linhas. Assim, um mesmo tweet que antes ocupava uma linha é particionado em várias linhas relativas a cada uma das suas palavras. Esta atividade foi realizada utilizando a função *“unnest_tokens”* do pacote *tidytext* v. 0.3.3. Com isso, o *dataset* passou a ter um total de 966.325 linhas. No entanto, parte destas linhas consistem em palavras que não tem significado de valor para este trabalho, tais como preposições, conjunções e artigos, entre outros conjuntos de palavras que são normalmente chamadas de “Stop words”. Estas palavras estão presentes em todas as línguas em alta frequência e não tem um significado específico, podendo assim prejudicar a análise. Para lidar com isso, foi feita assim a remoção destas palavras no conjunto de dados utilizando como base o dicionário de stop words “Lektek”

(http://www.lextek.com/manuals/onix/stopwords1.html) presente no pacote *tidytext.* Além destas palavras, foram removidas as palavras que fazem parte do termo de busca ("*amazon*" e "*rainforest*"), considerando sua esperada importância, e o termo "rt", que está presente em retweets. Após a remoção destas palavras, o conjunto de dados foi reduzido a 386.602 linhas.

Por fim, foi realizado ainda um cruzamento adicional das palavras obtidas com um dicionário de sentimentos de palavras ("*bing*") através da função *get_sentiments* do pacote *tidytext,* de forma a obter para cada uma delas a identificação da tendência de sentimento do seu significado. Esta identificação busca indicar se a palavra tende a estar associada mais a sentimentos positivos ou negativos. No entanto, é importante ressaltar que nem todas as palavras estão presentes no dicionário de sentimentos, não tendo assim preenchimento total desta coluna no conjunto de dados. No caso destes dados, a amostra com preenchimento de informações de sentimentos tem 44132 linhas, representando assim um preenchimento de pouco mais de 11 %.

4. Análise e Exploração dos Dados

Com os dados de publicações já processados e tratados, as primeiras atividades de análise e exploração de dados foram realizadas para responder o primeiro questionamento levantado: "*Quais os principais termos presentes em publicações estrangeiras em redes sociais relacionadas à região amazônica e como a sua importância variou ao longo da série temporal avaliada?".* Para isso, foram levantadas quais as 20 palavras mais citadas em publicações ao longo de todo o período de monitoramento, quantificando ainda a sua representatividade dentro do conjunto total de citações de palavras identificado. Dentro desse grupo, foi avaliado ainda como este número de citações variou para as 5 palavras mais importantes. A definição do número de palavras selecionadas foi feita considerando o grande número de palavras disponível e a dificuldade de apresentação resultados para todas elas. No conjunto de dados como um todo também foram levantadas medidas descritivas da importância das palavras e de sua variação ao longo do período monitorado.

Com os dados de sentimentos obtidos na etapa de processamento de dados, na sequência foram realizadas atividades para responder o segundo questionamento: *"A composição de termos nas publicações estrangeiras sobre a região está associada mais fortemente a sentimentos positivos ou negativos? Existe uma variação destas proporções?"* Para isso, o primeiro passo foi quantificar a representatividade de termos de cada sentimento (positivo/negativo) em cada publicação, obtendo assim um % de positivos e % de negativos em cada uma delas. Estes dados foram então utilizados para quantificar a média de composição dos tweets no geral e avaliar se existe uma diferença significativa entre a representatividade de cada sentimento nas publicações. Para avaliar tal diferença, foi realizado uma análise utilizando modelos lineares em que o percentual de cada sentimento foi utilizado como variável resposta e o sentimento (positivo/negativo) como variável explanatória. Foi utilizada a distribuição de resíduos *gaussian*, considerando que os dados apresentaram normalidade dos resíduos e homogeneidade da variância, além de adoção de nível de significância de 0.05. Além disso, para estes dados de composição de sentimentos foram obtidas medidas descritivas de cada categoria. Também foi obtido a média de cada sentimento ao longo do período de avaliação, permitindo assim a análise das tendências temporais.

5. Criação de Modelos de Machine Learning

Com base nos dados de ocorrência das palavras nas publicações (eventos) foi avaliado os padrões de associação entre palavras usando análise de regras de associação (ARA; AGRAWAL et al., 1993), que é uma ferramenta de mineração de dados que identifica associações entre observações categóricas em conjuntos de dados extensos. O ARA utiliza os dados para propor uma rega no molde: “se a palavra X, então a palavra Y”, de forma que a partir da sua ocorrência conjunta seja possível obter regras de associação entre elas, quantificando ainda métricas de ocorrência e importância.

As regras obtidas utilizando esta técnica tem duas partes principais: o lado esquerdo da regra (LER) e o lado direito da regra (LDR). O LER é a parte de referência da categoria (no caso, a palavra) que tem uma determinada frequência de ocorrência no conjunto de dados, a qual é chamada de “Suporte” (Support). Assim, um suporte de 0.5 indica que a palavra presente no LER ocorre em metade dps eventos (publicações). O LDR consiste em outro nível da categoria que que está relacionado ao LER em uma determinada frequência, chamada de “Confiância” (Confidence), de forma que os valores de confiança são sempre relativos aos dados de suporte. Por exemplo, em uma regra que tem um suporte de 0.5 e uma confiança de 1, a palavra presente do LDR ocorre junto a palavra do LER em todos os seus registros de ocorrência, ou seja, em metade dos eventos/publicações. Depois das regras obtidas, ainda pode ser obtida uma métrica adicional, o valor p de Fisher, obtido por teste de significância. O valor p de Fisher corresponde à significância da regra obtida através do teste exato de contingência de Fisher para pequenas amostras, avaliando se a regra é mais frequente do que o esperado por acaso (HAHSLER, 2006). O processo completo de obtenção de regras pela aplicação de restrições é resumido na figura a seguir.

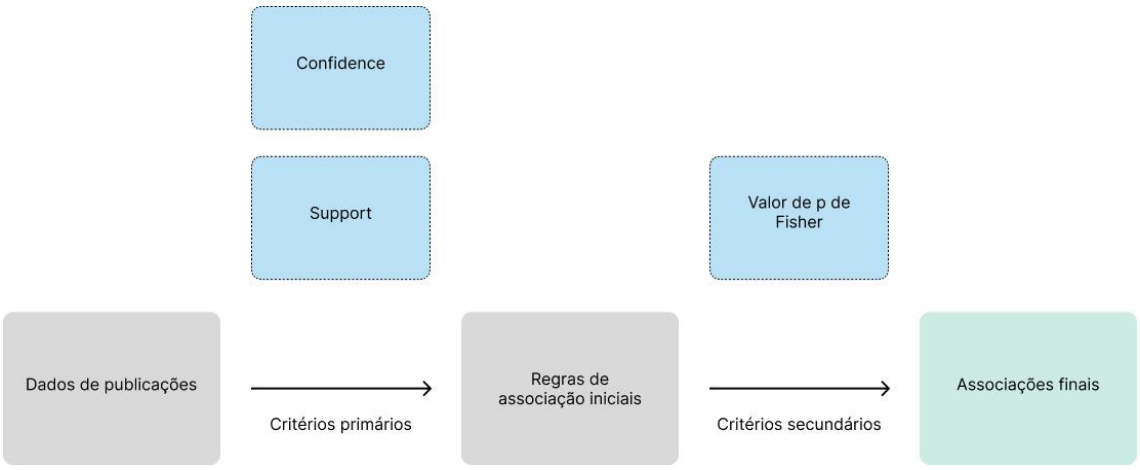


Figura 1: Referencial teórico do processo de obtenção de regras de associação entre palavras utilizando análise de regras de associação (ARA). O esquema mostra que a partir do conjunto de ocorrências de palavras nas publicações, duas limitações são inicialmente propostas (suporte e confiança) para a obtenção das regras iniciais. Este conjunto de regras iniciais passa então por um segundo processo de seleção através do p-valor de Fisher, para finalmente obter as regras finais.

Com base nos dados de ocorrência (presença/ausência) das palavras nas publicações, foi utilizado o algoritmo de machine learning *apriori* (BORGELT; KRUSE, 2002) do pacote *arules* (HAHSLER et al., 2020) no software R v. 4.2.0 (R CORE TEAM, 2022) para obter todas as regras de associação entre pares de palavras que atendessem aos seguintes critérios: suporte igual ou superior a 0.01875117 (1.875117 %), que corresponde à ocorrência da espécie no LER em pelo menos 1000 publicações; e confiança de 0,8 para a regra. Esses critérios indicam que as palavras em LDR devem estar relacionadas com as palavras do LER em pelo menos 80% de suas ocorrências. Das regras obtidas, foram selecionadas aquelas com valor de p de Fisher significativo (<0,05), para então serem obtidas as regras finais com associação com alto nível de confiabilidade. A figura abaixo traz o código de execução da análise.

```
######### Obtenção de regras

rules<- apriori(trans, parameter = list(supp = 0.01875117, conf = 0.8, target="rules", minlen=2,maxlen=2))
rules
summary(rules)


########## Obtenção de métricas das regras


measures<-interestMeasure(rules,
                          measure=c("fishersExactTest"),
                          transactions = trans)

############ Adição das métricas às regras


rules@quality$fisher_p<-measures$fishersExactTest


## Visualização final das regras

rules
```

Figura 2: Recorte do código de execução da obtenção das regras de associação utilizando a linguagem R.

6. Interpretação dos Resultados

A ocorrência das publicações virou ao longo da série temporal avaliada, oscilando entre momentos de maior e menor ocorrência. Os maiores valores foram observados no mês de julho de 2021, quando foram observados em dias seguidos os três maiores valores diários de 5372 publicações em 15/07, 2720 publicações em 16/07 e 1006 publicações em 17/07 que podem ser observados na Figura 2. Em contrapartida, os menores valores diários foram de 98 publicações em 14/06, 132 publicações em 25/09 e 151 publicações em 05/07.

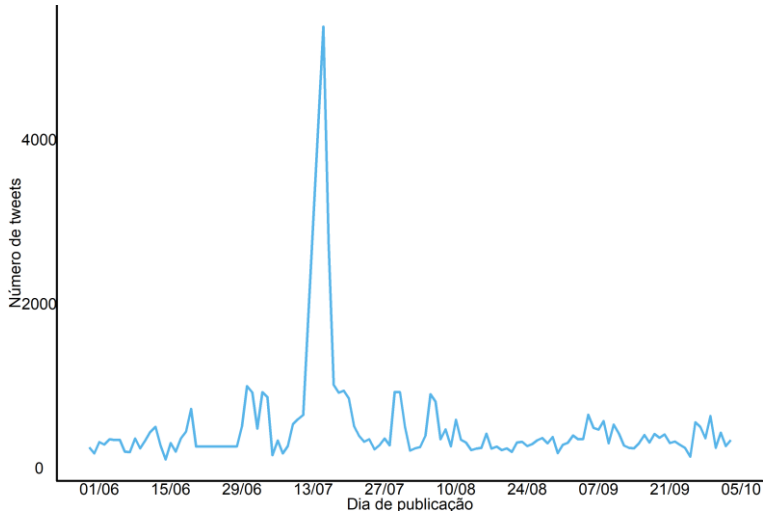


Figura 2: Número diários de publicações em redes sociais contendo o termo “Amazon Rainforest”.

Considerando o período todo de avaliação, as 20 principais palavras em número de citações nas publicações realizadas envolvendo o termo “Amazon Rainforest” foram, na sequência: carbono (carbono), “fire” (fogo), ”people” (pessoas), ”indigenous” (indígenas), ”deforestation” (desflorestamento/desmatamento), ”brazil” (Brasil), ”brazilian” (brasileiro), ”emitting” (emitindo), ”absorbs” (absorvem), ”climate” (clima), ”forest” (Floresta), ”world” (mundo), ”Bolsonaro” (sobrenome do presidente brasileiro), ”dying” (morrendo), ”destruction” (destruição), ”absorb” (absorve), ”ice” (gelo), ”dioxide” (dióxido), ”president” (presidente), ”protect” (protege/proteger). Assim, as principais palavras estão associadas principais a assuntos relacionados a mudanças climáticas (”carbon”, ”dioxide”, ”climate”), conservação (”protect”) ou destruição da floresta (”fire”, ”deforestation”, ”dying”, ”destruction”, etc.). Os

valores de número de citações e sua importância percentual no conjunto de dados geral podem ser observados na figura 3.

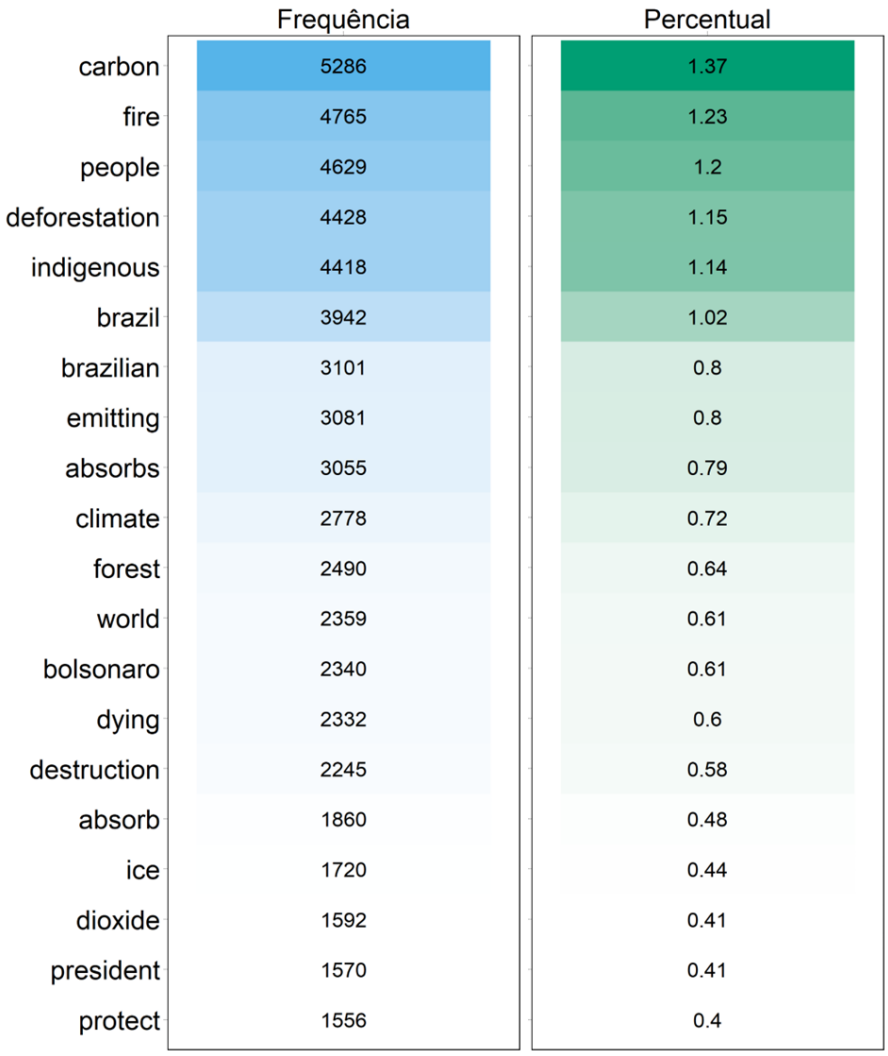


Figura 3: Heatmap apresentando o número de citações (frequência) e importância percentual das 20 palavras mais frequentes em publicações em redes sociais contendo o termo “Amazon Rainforest” ao longo de todo o período temporal avaliado.

Considerando as 5 palavras mais frequentes ao longo do período de avaliação (“Carbon”, “deforestation”, “fire”, “indigenous” e “people”), pôde ser observado a existência de picos de maior frequência concentrados entre o final de junho e o final de julho. As palavras “”people” e “indigenous” tiveram um picos de frequência no fim de junho e início de julho, enquanto “carbono” e as demais palavras tiveram seus picos no meio do mês de julho. Em

especial, os maiores pico da palavra “carbon” foram observados nesse período, especialmente nos dias 15 e 16 de julho, com 1781 e 1062 citações da palavra, respectivamente.

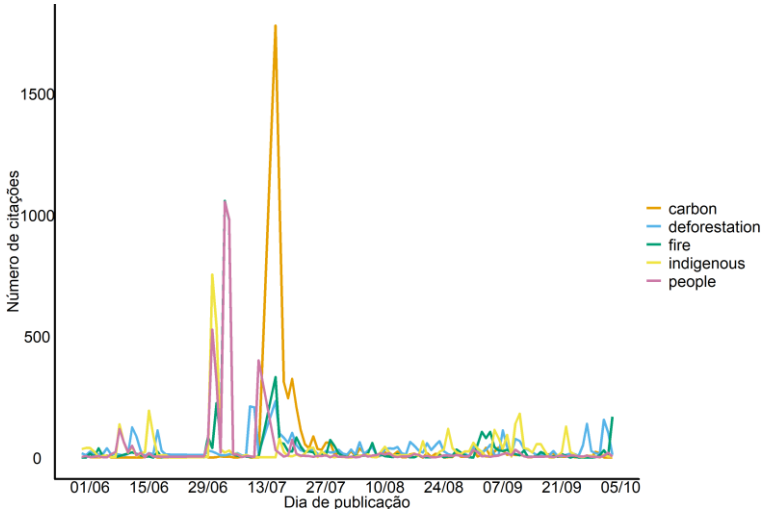


Figura 4: Variação diária do número de citações das 5 palavras mais frequentes em publicações em redes sociais contendo o termo “Amazon Rainforest” ao longo de todo o período temporal avaliado.

De acordo com a análise de sentimentos realizada em cima das publicações envolvendo o termo “Amazon rainforest”, encontrou-se que ao longo do período avaliado as publicações têm em média 68.47 % de termos com conotação negativa, contra 31.53 % de termos considerados positivos (Figura 5). O mínimo e o máximo observado foi de 0 % e 100 % para algumas publicações. Ou seja, são publicações em que todos os termos são negativos ou positivos, de acordo com o dicionário de sentimentos utilizado. A análise utilizando modelos lineares generalizados indicou que tais médias são significativamente diferentes ao nível de significância de 0.05. Assim, no geral as publicações envolvendo o termo “Amazon rainforest” foram mais negativas que positivas.

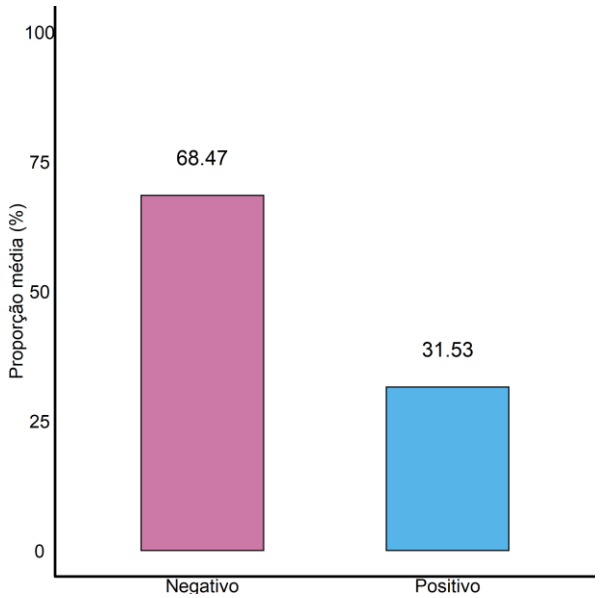


Figura 5: Proporção média de termos de cada sentimento nas publicações em redes sociais contendo o termo “Amazon Rainforest” ao longo de todo o período temporal avaliado.

Considerando a série temporal avaliada, a proporção média diária de termos de cada sentimento nas publicações foi mais negativa que positiva ao longo de toda a série temporal, com picos de termos negativos especialmente no mês de julho, como pode ser observado na figura 6. O menor valor médio diário negativo observado foi de 44.5 no dia 15 de setembro, enquanto o maior foi de 76.13 no dia 23 de julho. Nestes mesmos dias, o maior valor positivo foi de 55.5 % e o menor de 23.86 %. Assim, além das publicações serem compostas por mais termos negativos no geral, este padrão também foi observado ao longo de toda a série temporal.

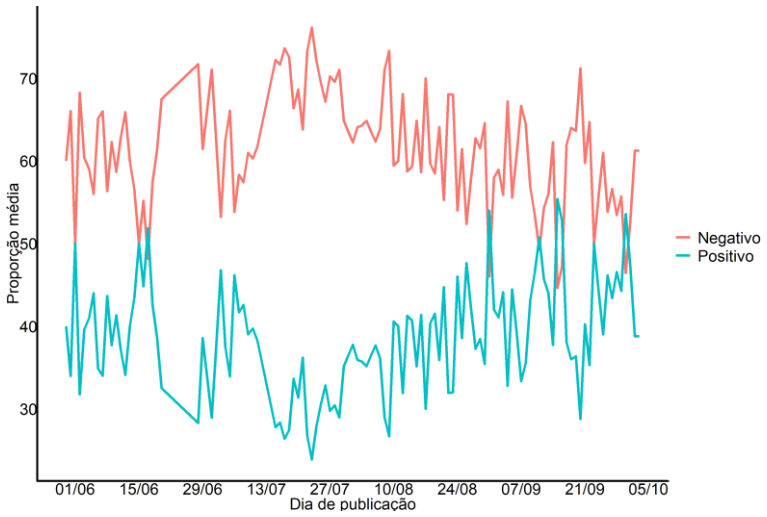


Figura 6: Proporção média diária de termos cada sentimento (%) em publicações em redes sociais contendo o termo “Amazon Rainforest” ao longo de todo o período temporal avaliado.

A análise de regras de associação (ARA) utilizando o algoritmo *apriori* e os critérios estabelecidos para identificação de associações consistentes resultou em 75 regras entre pares de termos, compostas por 30 palavras que se relacionam entre si. As associações e interações entre as 64 regras consistentes obtidas estão apresentadas na figura 7 e na tabela 3. Parte destas palavras estão presentes nas 20 mais importantes apresentadas anteriormente (como “carbono” e “fire”), enquanto outras não foram citadas anteriormente (como “ice” e “heatwaves”). Dentro do contexto avaliado de percepção ambiental, chama a atenção dentre as regras obtidas as associações consistentes entre “unnatural” e “heatwaves”, “dying” e “fire”; “emitts” e “absorbs”; “peoples” e “indigenous”; “carbono” e “dioxide” e “absorb”. Todas estas associações indicam que a ocorrência de uma palavra na publicação está fortemente relacionada a ocorrência da outra, em uma relação forte de coexistência. Também podem ser observadas outras associações consistentes de acordo com os critérios considerados, mas que não são relevantes dentro do contexto avaliado de percepção ambiental sobre a região amazônica.

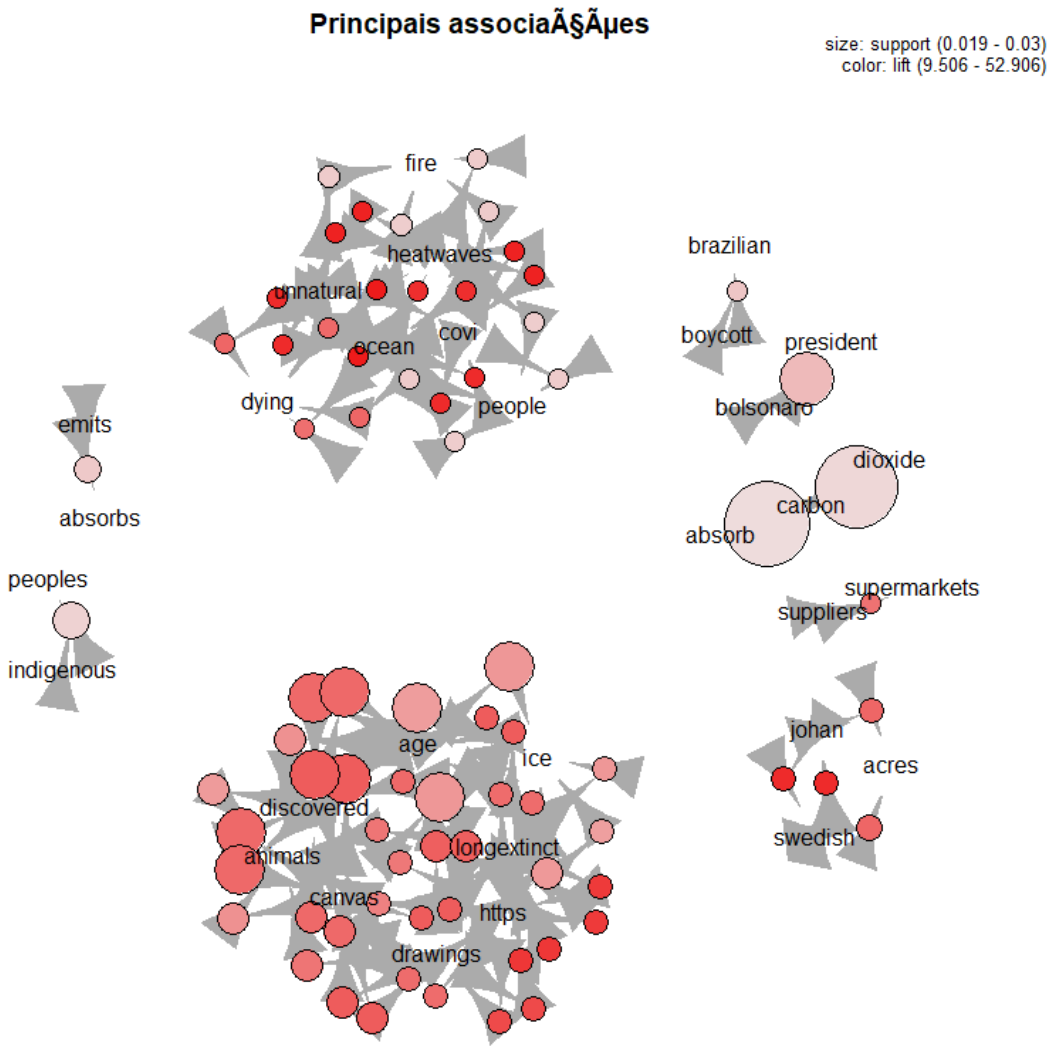


Figura 7: Representação das relações entre termos obtidas nas 64 regras de associação ide identificadas nos dados analisados de publicações em redes sociais contendo o termo “Amazon Rainforest” ao longo de todo o período temporal avaliado.

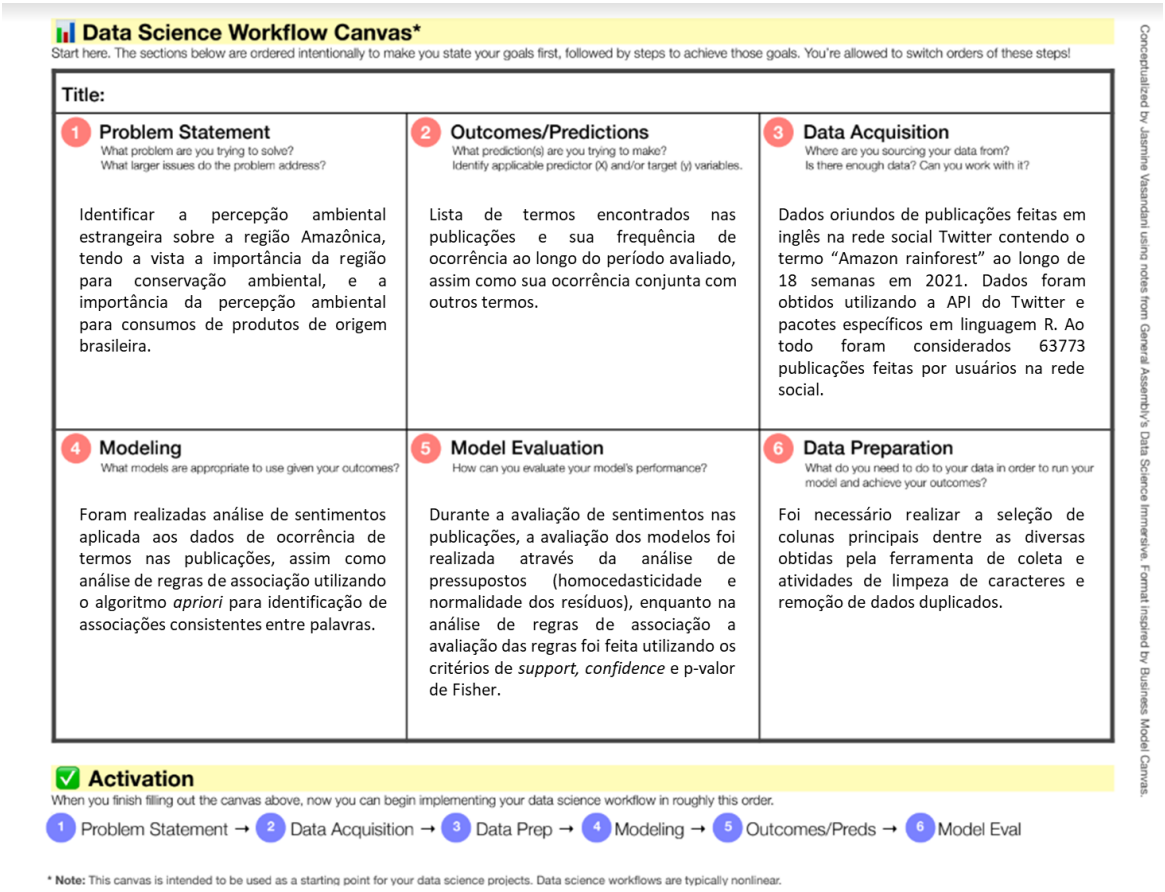
Tabela 3: Resultados obtidos para as 75 regras de associação entre termos identificadas nos dados analisados de publicações em redes sociais contendo o termo “Amazon Rainforest” ao longo de todo o período temporal avaliado, trazendo ainda medidas diagnósticas das associações.

Lado esquerdo da Regra	Lado direito da regra	Suporte	Confiança	P-valor de Fisher
covi	unnatural	0.019	0.999	0
unnatural	covi	0.019	0.994	0
covi	heatwaves	0.019	0.999	0
heatwaves	covi	0.019	0.980	0
covi	ocean	0.019	0.999	0
ocean	covi	0.019	0.952	0
covi	dying	0.019	1.000	0
covi	fire	0.019	0.999	0
covi	people	0.019	1.000	0
unnatural	heatwaves	0.019	0.994	0
heatwaves	unnatural	0.019	0.980	0
unnatural	ocean	0.019	0.994	0
ocean	unnatural	0.019	0.952	0
unnatural	dying	0.019	0.994	0
unnatural	fire	0.019	0.999	0
unnatural	people	0.019	0.994	0
heatwaves	ocean	0.019	0.980	0
ocean	heatwaves	0.019	0.952	0
heatwaves	dying	0.019	0.980	0
heatwaves	fire	0.019	0.980	0
heatwaves	people	0.019	0.980	0
ocean	dying	0.019	0.952	0
ocean	fire	0.019	0.962	0
ocean	people	0.019	0.952	0
emits	absorbs	0.020	0.871	0
longextinct	https	0.019	0.996	0
https	longextinct	0.019	0.933	0
longextinct	drawings	0.020	1.000	0
drawings	longextinct	0.020	0.936	0
longextinct	canvas	0.020	1.000	0
canvas	longextinct	0.020	0.814	0
longextinct	age	0.020	1.000	0
age	longextinct	0.020	0.813	0
longextinct	discovered	0.020	1.000	0
longextinct	animals	0.019	0.999	0
longextinct	ice	0.020	1.000	0
https	drawings	0.019	0.933	0
drawings	https	0.019	0.932	0

https	canvas	0.019	0.934	0
canvas	https	0.019	0.812	0
https	age	0.019	0.934	0
age	https	0.019	0.811	0
https	discovered	0.019	0.934	0
https	animals	0.019	0.933	0
https	ice	0.019	0.934	0
johan	swedish	0.019	0.995	0
swedish	johan	0.019	0.987	0
johan	acres	0.019	0.991	0
swedish	acres	0.020	0.992	0
drawings	canvas	0.021	0.991	0
canvas	drawings	0.021	0.862	0
drawings	age	0.021	0.991	0
age	drawings	0.021	0.861	0
drawings	discovered	0.021	0.996	0
discovered	drawings	0.021	0.821	0
drawings	animals	0.021	0.991	0
drawings	ice	0.021	0.991	0
peoples	indigenous	0.022	0.962	0
canvas	age	0.024	0.998	0
age	canvas	0.024	0.996	0
canvas	discovered	0.024	0.997	0
discovered	canvas	0.024	0.944	0
canvas	animals	0.021	0.861	0
canvas	ice	0.024	0.998	0
age	discovered	0.024	0.995	0
discovered	age	0.024	0.943	0
age	animals	0.021	0.859	0
age	ice	0.024	0.998	0
discovered	animals	0.021	0.813	0
discovered	ice	0.024	0.944	0
suppliers	supermarkets	0.019	0.995	0
boycott	brazilian	0.019	0.933	0
president	bolsonaro	0.025	0.840	0
dioxide	carbon	0.030	0.992	0
absorb	carbon	0.030	0.863	0

7. Apresentação dos Resultados

A apresentação dos resultados neste tópico seguirá o modelo de Canvas proposto por Jasmine Vasandani, que resume o fluxo de trabalho de um projeto de Ciência de Dados em etapas bem definidas, que serão detalhadas na sequência. O modelo com preenchimentos gerais sobre este trabalho está apresentado na figura abaixo.



Data Science Workflow Canvas*

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title:

1 Problem Statement
What problem are you trying to solve?
What larger issues do the problem address?

Identificar a percepção ambiental estrangeira sobre a região Amazônica, tendo a vista a importância da região para conservação ambiental, e a importância da percepção ambiental para consumos de produtos de origem brasileira.

2 Outcomes/Predictions
What prediction(s) are you trying to make?
Identify applicable predictor (X) and/or target (y) variables.

Lista de termos encontrados nas publicações e sua frequência de ocorrência ao longo do período avaliado, assim como sua ocorrência conjunta com outros termos.

3 Data Acquisition
Where are you sourcing your data from?
Is there enough data? Can you work with it?

Dados oriundos de publicações feitas em inglês na rede social Twitter contendo o termo "Amazon rainforest" ao longo de 18 semanas em 2021. Dados foram obtidos utilizando a API do Twitter e pacotes específicos em linguagem R. Ao todo foram considerados 63773 publicações feitas por usuários na rede social.

4 Modeling
What models are appropriate to use given your outcomes?

Foram realizadas análise de sentimentos aplicada aos dados de ocorrência de termos nas publicações, assim como análise de regras de associação utilizando o algoritmo *apriori* para identificação de associações consistentes entre palavras.

5 Model Evaluation
How can you evaluate your model's performance?

Durante a avaliação de sentimentos nas publicações, a avaliação dos modelos foi realizada através da análise de pressupostos (homocedasticidade e normalidade dos resíduos), enquanto na análise de regras de associação a avaliação das regras foi feita utilizando os critérios de *support, confidence* e p-valor de Fisher.

6 Data Preparation
What do you need to do to your data in order to run your model and achieve your outcomes?

Foi necessário realizar a seleção de colunas principais dentre as diversas obtidas pela ferramenta de coleta e atividades de limpeza de caracteres e remoção de dados duplicados.

Activation

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

* **Note:** This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.

Conceptualized by Jasmine Vasandani using notes from General Assembly's Data Science Immersive. Format inspired by Business Model Canvas.

Figura 8: Modelo de Canvas proposto por Jasmine Vasandani preenchido com informações referentes a este trabalho.

Etapa 1 – Definição do problema.

Que problema você está tentando resolver? E quais questões maiores esse problema aborda? Esta seção ajuda a abordar o "porquê" do seu projeto.

O problema a ser resolvido neste trabalho está associado a identificação da percepção ambiental estrangeira sobre a região Amazônica, considerando para isso publicações feitas em redes sociais. A percepção ambiental é um fator que está fortemente relacionado a percepção estrangeira sobre o Brasil, podendo assim influenciar o consumo de produtos oriundos do país, sejam eles comerciais, agrícolas e midiáticos. Além disso, avaliar este problema pode ajudar a entender melhor a forma de percepção das redes sociais a eventos externos, como por ex. a crise de incêndios na região Amazônica.

Com isso, pode-se entender melhor como o público estrangeiro percebe os acontecimentos na região e como tal percepção pode impactar a imagem do Brasil como um todo. Além disso, é importante para entender a permeabilidade das redes sociais e o quão bem o público conhece o cenário de conservação e degradação da região, que presta serviços ecossistêmicos extremamente relevantes, além de ser um importante da biodiversidade global.

Etapa 2 – Resultados/Predições

Quais predições você está tentando fazer? Quais resultados pretende obter?

Os resultados obtidos neste trabalho consistem na lista de termos presentes nas publicações, com suas respectivas frequências de ocorrência ao longo do período de avaliação, sentimentos (negativo/positivo) e associação com outras palavras, se houver. Com isso, é possível avaliar quais os principais termos relacionados ao tema em questão, quais são mais importantes e quais os seus principais sentidos associados.

Ou seja, quando se analisa uma publicação que envolve o termo "Amazon rainforest" nesta rede social, quais as palavras mais prováveis de estarem presentes? Qual o sentimento mais provável de ser predominante nesta publicação? Esta publicação tem mais chances de ter um caráter positivo ou negativo? Estes são os resultados de interesse deste trabalho, que

podem permitir entender melhor a percepção ambiental do público externo ao país em relação à região.

Etapa 3 – Aquisição de dados

Quais as fontes dos dados utilizados? Os dados são suficientes para o objetivo proposto?

Os dados utilizados são oriundos de publicações em língua inglês feitas na rede social Twitter que utilizaram o termo "*Amazon rainforest*", que é a denominação mais comum em para se referir a região da Amazônia em língua inglesa. Com base nestes dados, foi realizada uma partição dos termos existentes, obtendo-se sua frequência de ocorrência, seu sentimento associado e as associações entre palavras.

O conjunto total de dados consistiu de quase 64 mil publicações feitas ao longo de 18 semanas, entre os dias 31/05 e 04/10 de 2021, abrangendo inclusive a troca de estações climáticas na região. Tais dados foram obtidos através da API da plataforma para obtenção de dados, mediante registro de projeto e aprovação de justificativas. A conexão com a plataforma foi realizada utilizando pacotes específicos desenvolvidos em linguagem R.

Etapa 4 – Modelagem/análise de dados

Quais modelos/análises são apropriadas para se obter os resultados esperados?

Além da identificação dos principais termos e sua variação de importância ao longo do período avaliado, a análise passa também pelo cruzamento destes termos com um dicionário de sentimentos, que associa cada palavra ao seu sentimento mais comum (positivo ou negativo). Com isso, é possível avaliar os principais sentimentos envolvidos nas publicações de uma maneira geral e de que foram tal percepção variou no tempo. Ou seja, é possível avaliar se uma publicação em língua inglesa feita na rede social contendo o termo "Amazon rainforest" tende a ser mais positiva ou mais negativa.

Além disso, a análise passou pela identificação de regras de associação entre termos utilizando o algoritmo apriori, buscando avaliar associações consistentes entre os termos

presentes nas publicações. Ao utilizar esta ferramenta, pode ser avaliado se existem associações consistentes entre termos de cunho ambiental, analisando seus principais significados e contextos.

Etapa 5 – Avaliação de modelos/análises

Como a performance dos modelos/análises pode ser avaliada?

Na análise de sentimentos, a avaliação de sentimentos nas publicações, a avaliação dos modelos pode ser realizada através da análise de pressupostos (homocedasticidade e normalidade dos resíduos), que são critérios fundamentais para a execução da análise e obtenção de inferências confiáveis. Na análise de regras de associação, a avaliação da consistência das relações obtidas pode ser feita através do estabelecimento dos critérios de suporte, confiança e p-valor de Fisher, que selecionam relações consistentes entre os termos presentes nas publicações.

Etapa 6 – Preparação de dados

Quais são as atividades necessárias de serem realizadas nos dados para que a análise seja executada e os resultados esperados obtidos?

Os dados obtidos consistem a princípio do texto das publicações, além de metadados relacionados. Para tornar este texto possível de ser analisado é necessário realizar uma série de etapas como a chamada “tokenização”, que particiona o texto em termos isolados; remoção de “stop words”, que são palavras não úteis; remoção de publicações duplicadas; remoção de caracteres especiais como arrobas, endereços de e-mail e acentuação; e por fim o cruzamento com o dicionário de sentimentos, que vai atribuir para cada palavra um sentimento identificador, seja positivo ou negativo. Após estas etapas, os dados estão aptos a serem analisados como apresentado anteriormente.

Interpretação geral dos resultados

É importante destacar que a maior parte das palavras mais frequentes estão relacionadas ao contexto ambiental, tais como “carbon” e “emitting” que estão associadas a emissão deste elemento para a atmosfera e contribuição para o aquecimento global e mudanças climáticas; e “fire”, que faze referência aos incêndios que ocorrem na região e são responsáveis por impactos significativas nos ecossistemas amazônicos. Além disso, algumas palavras estão diretamente relacionadas a estes impactos, como “dying”, “deforestation” e “destruction”. Também existem termos relacionado a proteção da floresta (“protect”) e aos povos indígenas (“peoples” e “indigenous”) que ocupam as florestas da região e que são afetados pelo processo de destruição da floresta. Em conjunto, o resultado encontrado aponta um padrão de preocupação sobre o status de conservação ambiental das publicações envolvendo a região Amazônica.

Os picos observados no número de citações dos termos mais frequentes entre junho e julho de 2021 estão diretamente associados ao início da estação seca na região, na qual o número de incêndios aumenta substancialmente. De acordo com dados do Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE; 2020) o número de focos de incêndios identificados pelo sensor MODIS do satélite AQUA aumentou de 1166 em maio para 2305 focos em junho, 4977 em julho, 28060 focos em agosto e 16742 focos em setembro. Assim, o período de maior pico de citações dos termos marca o início do período de aumento substancial das queimadas na região, que é compatível com o aumento da publicidade internacional dada a estes eventos, que tem impactos ambientais na conservação dos ecossistemas da região.

Este resultado encontrado de prevalência de sentimento negativo nas publicações estrangeiras aponta para um cenário de percepção ambiental estrangeira predominantemente negativa sobre a região amazônica. Esta percepção é ainda potencializada em períodos de maior ocorrência de incêndios, nos quais assuntos relacionados ao tema são mais divulgados na mídia internacional, em função da importância dos ecossistemas da região em serviços ecossistêmicos essenciais para a população brasileira e global, assim como sua importância na conservação da biodiversidade. Considerando a importância dada ao tema ambiental pelos consumidores, tal percepção negativa pode influenciar a compra de produtos

brasileiros por consumidores estrangeiros, que ao vão associar a imagem do produto a destruição da floresta e ao não cumprimento de metas ambientais importantes.

As associações encontradas entre termos que se relacionam ao tema ambiental apontam para um cenário parecido ao das análises anteriores, de uma percepção ambiental com caráter negativa, relacionada a emissão de carbono, extinção de animais, ondas de calor não naturais e incêndios. Assim, estes resultados corroboram a percepção ambiental negativa estrangeira característica das publicações feitas sobre a região amazônica em redes sociais. Além disso, os resultados indicam o uso da técnica de análise de regras de associação como uma ferramenta de valor para a avaliação de percepção ambiental e identificação de associações consistentes entre termos de interesse.

8. Links

Link para o vídeo: https://youtu.be/k8OdlEN6Xsg

Link para o repositório:

https://github.com/crdesouza/TCC-PUC-Ciencia_de_dados_big_data.git

REFERÊNCIAS

AGRAWAL, Rakesh; IMIELIŃSKI, Tomasz; SWAMI, Arun. **Mining association rules between sets of items in large databases.** In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data. 1993. p. 207-216.

BARLOW, Jos et al. **The future of hyperdiverse tropical ecosystems**. Nature, v. 559, n. 7715, p. 517-526, 2018.

BARLOW, Jos et al. **Clarifying Amazonia's burning crisis**. Global Change Biology, v. 26, n. 2, p. 319-321, 2020.

BENNETT, Nathan James. **Using perceptions as evidence to improve conservation and environmental management**. Conservation Biology, v. 30, n. 3, p. 582-592, 2016.

BORGELT, Christian; KRUSE, Rudolf. **Induction of association rules: Apriori implementation**. In: Compstat. Physica, Heidelberg, 2002. p. 395-400.

CANAVARI, Maurizio; CODERONI, Silvia. **Green marketing strategies in the dairy sector: Consumer-stated preferences for carbon footprint labels**. Strategic Change, v. 28, n. 4, p. 233-240, 2019.

CARDOSO, Domingos et al. **Amazon plant diversity revealed by a taxonomically verified species list.** Proceedings of the National Academy of Sciences, v. 114, n. 40, p. 10695-10700, 2017.

GENTRY, Jeff et al. **Package 'twitteR'**. Cran. r-project, 2016.

GIBBENS, Sarah. **The Amazon is burning at record rates–and deforestation is to blame**. National Geographic, v. 21, 2019.

HAHSLER, Michael. **A model-based frequency constraint for mining associations from transaction data.** Data Mining and Knowledge Discovery, v. 13, n. 2, p. 137-166, 2006.

HAHSLER, Michael et al. **arules: Mining Association Rules and Frequent Itemsets.** R package version 1.6-6, 2020. https://CRAN.R-project.org/package=arules

INPE - Instituto Nacional de Pesquisas Espaciais, 2020. **Portal do Monitoramento de Queimadas e Incêndios Florestais.** Disponível em http://www.inpe.br/queimadas. Acesso em: 26/07/2022.

LIM, Weng Marc et al. **Environmental social governance (ESG) and total quality management (TQM): a multi-study meta-systematic review**. Total Quality Management & Business Excellence, p. 1-23, 2022.

OZDEMIR, Oguz. **The effects of nature-based environmental education on environmental perception and behavior of primary school students**. Journal of Education, v.27, 2010.

PAN, Yude et al. **The structure, distribution, and biomass of the world's forests**. Annual Review of Ecology, Evolution, and Systematics. 44 (1): 593-622., v. 44, n. 1, p. 593-622, 2013.

RAJÃO, Raoni et al. **The rotten apples of Brazil's agribusiness**. Science, v. 369, n. 6501, p. 246-248, 2020.

R Core Team, et al. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL http://www.R-project.org/

SILVA JUNIOR, Celso HL et al. **The Brazilian Amazon deforestation rate in 2020 is the greatest of the decade.** Nature Ecology & Evolution, v. 5, n. 2, p. 144-145, 2021.

SILVEIRA-JUNIOR, Wanderley Jorge et al. **Conservation conflicts and their drivers in different protected area management groups: a case study in Brazil**. Biodiversity and Conservation, v. 30, n. 14, p. 4297-4315, 2021.

SOUZA JR, Carlos M. et al. **Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine.** Remote Sensing, v. 12, n. 17, p. 2735, 2020.

VERDONE, Michael; SEIDL, Andrew. **Time, space, place, and the Bonn Challenge global forest restoration target.** Restoration ecology, v. 25, n. 6, p. 903-911, 2017.