Aplicativo interativo de mineração de textos para análises de experiências turísticas

Author1[a,1,*], Author2[b], Author3[b,2]

[a] *Department Street City State Zip*
[b] *Department Street City State Zip*

Abstract

Investigating the experience economy, observing the levels of interaction experienced by tourists in destinations, can be an important tool for the best use of and sustainable development of tourist regions. This article presents an approach for evaluating the domains of experience of tourist attractions based on comments posted on the Tripadvisor platform. The interpretation was performed using netnography, through text mining techniques, natural language processing and text classification with machine learning. The attractions were evaluated and classified considering the reported experiences. The approach presented here contributes to the identification of gaps regarding public and private initiatives that make it possible to meet the expectations of tourists, making the most competitive and viable tourist spots in terms of sustainable development. The methodological conception of the research can constitute an important tool for the study of tourism and enable theoretical deepening and improvements in the offer of ecotourism products.

Keywords: Experience Economics, Text Mining, Natural Language Processing, Machine Learning, Textual Analysis Tool

1. Resumo

Investigar a economia da experiência, observando os níveis de interação vivenciados pelos turistas, pode ser uma importante ferramenta para o melhor aproveitamento e desenvolvimento sustentável de regiões turísticas. Este artigo apresenta uma abordagem para a avaliação dos domínios da experiência de pontos turísticos a partir dos comentários postados na plataforma *Tripadvisor*. Foi construído um aplicativo interativo e amigável para auxiliar na interpretação e classificação das experiências relatadas pelos turistas, empregando-se a netnografia, através de técnicas de mineração de textos, processamento de linguagem natural e classificação de textos com aprendizado de máquina. A abordagem aqui apresentada contribui para a identificação de lacunas no que tange iniciativas públicas e privadas que possibilitem o atendimento das expectativas dos turistas, tornando os pontos turísticos mais competitivos e viáveis na perspectiva do desenvolvimento sustentável. A concepção metodológica da pesquisa pode se constituir em uma importante ferramenta para o estudo do turismo e possibilitar aprofundamentos teóricos e melhorias da oferta de produtos do ecoturismo.

palavras-chave: Economia da Experiência, Mineração de Textos, Processamento de Linguagem Natural, Aprendizado de Máquina, Ferramentas de Análise textual

2. Introdução

O turismo constitui-se em um complexo processo de decisão realizado por diferentes motivações como, por exemplo, a hospedagem, a alimentação, o lazer, a informação turística, o entretenimento, dentre outras

*Corresponding author
Email addresses: a@example.com (Author1), b@example.com (Author2), c@example.com (Author3)
[1]This is the first author footnote.
[2]Another author footnote.

variáveis (M. C. Beni 2019). Os serviços turísticos não podem ser entendidos como um produto estático, pois eles necessitam de evolução para o crescimento do turismo no Brasil e no mundo (Coelho and Ribeiro 2007). Tal avanço também perpassa pelo consumidor, visto que ele busca, muito além de produtos e serviços turísticos, novas experiências, alterando gostos e preferências referentes à demanda anterior (Mário Carlos Beni 2004).

Segundo (Mário Carlos Beni 2004) deve-se buscar a harmonia entre o que o destino turístico pode oferecer e as experiências turísticas que o visitante busca ao viajar. Há uma mudança significativa nos gostos e preferências dos turistas, que anteriormente buscavam produtos e serviços, e atualmente a procura perpassa, também, pela ambição de experiências novas (Mário Carlos Beni 2004). Essas experiências, entendidas como uma avaliação que o sujeito faz de forma subjetiva quando há submissão às experimentações turísticas (afetivas, cognitivas e comportamentais) se iniciam com a preparação para a imersão, se alongam durante ela, e estendem-se até a completude da experiência, deixando bastante evidente a amplitude de significados gerados pelas experiências (Tung and Ritchie 2011). Dessa forma, o resultado torna-se extenso, passando pela fidelização do consumidor, a perpetuação do sentido experienciado em sua memória e até a recomendação para outros potenciais consumidores (Coelho and Ribeiro 2007; Alencar et al. 2019; LoBuono et al. 2016).

Diante dessa perspectiva, Pine and Gilmore (1998) fazem uma diferenciação quanto aos serviços e as experiências: respectivamente, de um lado tem-se um conjunto de atividades intangíveis, e do outro, eventos ou experiências memoráveis. Estas experiências memoráveis são planejadas para engajar o turista ao processo, e não somente entretê-lo. Assim, quatro domínios de experiência foram propostos, a partir de dois eixos, chamados de estágios da estruturação de uma experiência (Figura 1).

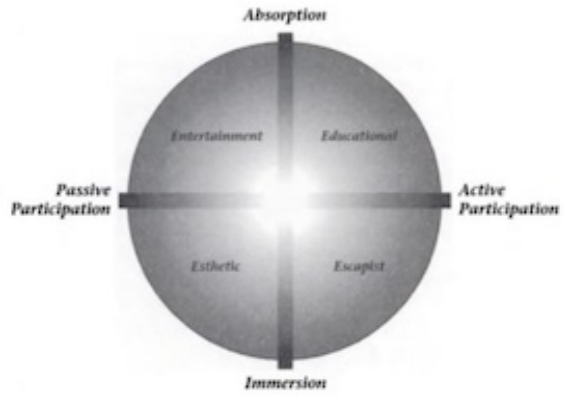

Figure 1: Categorias utilizadas para a classificação da experiência Fonte: Pine e Gilmore (1999)

O primeiro dos eixos, proposto horizontalmente, refere-se à participação do indivíduo, podendo ser classificada como passiva (passive participation) ou ativa (active participation). Já o eixo vertical representa o aspecto ambiental que interliga indivíduo e experiência: de um lado está a imersão (immersion) e do outro a absorção (absorption). O cruzamento desses dois eixos cria as quatro dimensões ou domínios de experiência: entretenimento, aprendizagem, estética e evasão/escapismo (Pine and Gilmore 1998).

O primeiro quadrante envolve a dimensão caracterizada como **entretenimento**. É quando o turista busca experiências que sejam divertidas e emocionantes, como parques de diversão, shows, esportes radicais, vida noturna ou atividades em grupo. Trata-se de uma dimensão mais passiva, em que o indivíduo responde aos elementos que lhe são apresentados levando-o a expressar sinais de satisfação, riso e relaxamento. Dessa forma, a fim de desenvolver um serviço turístico que contemple esta dimensão, deve-se torná-lo mais divertido e admirado (Horodyski, Fernandes, and Gândara 2015; Alencar et al. 2019; Corrêa and Gosling 2023). Em vista disso pode-se resumir que se trata de oferecer opções pessoais de lazer no destino escolhido.

A segunda dimensão consiste na **aprendizagem**. Envolve uma participação ativa do indivíduo com a atividade turística. É quando o turista busca uma experiência que o faça aprender algo novo, como uma

cultura diferente, um idioma, uma habilidade ou conhecimento histórico. O aprender demanda que o sujeito interaja e se envolva com o objeto de apreensão. Nesse sentido, ao se preparar um serviço que contemple a economia da experiência, deve-se definir quais informações pretende-se que o turista absorva, como também, quais habilidades pretende-se que o mesmo pratique durante o consumo, implicando em vivenciar tanto a perspectiva sensorial quanto intelectual LoBuono et al. (2016). Aqui depreende-se que contemple a obtenção de contato com aspectos do ambiente, da cultura e da história dos habitantes do local/região visitado.

A terceira dimensão corresponde à **estética/contemplação**. É quando o turista busca experiências que envolvam a apreciação da beleza, seja ela natural ou criada pelo homem. Isso pode incluir visitas a museus, galerias de arte, parques naturais ou cidades históricas. Essa dimensão abrange elementos que atraem o indivíduo por razões visuais, fazendo com que tome a decisão de adentrar em um local e ali permanecer. Cabe ressaltar que a escolha em incluir a "contemplação" no conceito original Pine and Gilmore (1998) parte do estudo de Andrukiu and Gândara (2015) que considera a dimensão da contemplação como um fator característico da "estética". Ao se propor a oferta de um serviço, a fim de que proporcione ao turista vivenciar esta dimensão, deve-se criar um ambiente convidativo, interessante e confortável, para que ele se sinta impelido a ficar ali. Pode-se inferir que esta dimensão se caracteriza com a qualidade aparente dos atrativos visitados, que despertam no indivíduo a conduta de admirar o ambiente.

A quarta dimensão refere-se a **evasão/escapismo**. É quando o turista busca uma experiência que o ajude a escapar da realidade e do estresse do dia a dia. Isso pode incluir viagens para lugares isolados e tranquilos, retiros espirituais, spas, praias ou montanhas. Diz respeito à capacidade de fazer com que o turista fique imerso nas atividades que lhe são propostas. Ao desenvolver produtos turísticos, deve-se criar condições que possibilitem ao indivíduo vivenciar situações que lhe demandem uma participação ativa, bem como, despertem nele (o domínio dos seus sentidos), uma completa imersão, suscitando a manifestação de sentimentos e emoções. Pode-se dizer que esta dimensão implica em sensações de desprendimento pessoal. Já Aroeira, Dantas, and Gosling (2016), classifica experiências dessa natureza como equivalentes ao pensamento, à absorção e imaginação do cliente, como por exemplo, o "desligar", "se conectar" com o lugar.

É importante lembrar que esses domínios não são mutuamente exclusivos e que muitas vezes uma única experiência pode incluir elementos de mais de um domínio. Além disso, a maneira como os turistas experimentam cada um desses domínios pode variar de acordo com suas personalidades, interesses e expectativas. Dessa forma, para que a experiência seja memorável, deve-se proporcionar ao turista vivenciar as quatro dimensões. Essa perspectiva analítica oferece ao produtor de serviço turístico, um diagnóstico que permite compreender o atendimento da expectativa do cliente, nestas quatro dimensões, exibindo opções e diretrizes que possam proporcionar uma melhor experiência ao mesmo.

2.1. Apresentação da abordagem realizada

Muitas pessoas e empresas utilizam a internet diariamente para expressar suas opiniões, aumentando a quantidade de dados textuais existente, como é o caso da plataforma *TripAdvisor*. Esses dados podem conter informações valiosas, que muitas vezes, podem ser obtidas com rapidez e baixo custo financeiro, através de técnicas de mineração de texto e processamento de linguagem natural (PLN), as quais envolvem o uso de tecnologias computacionais para processar e analisar dados em linguagem natural (Tereza Medeiros 2021).

A mineração de textos envolve a extração de informações úteis e conhecimento a partir de grandes volumes de texto não estruturado, como documentos, artigos, posts de redes sociais, entre outros. O objetivo é transformar esse texto em dados estruturados, tornando-o mais fácil de ser analisado. As principais etapas da mineração de textos incluem: pré-processamento (limpeza, tokenização, stemming), análise de sentimentos, extração de informações (entidades, relações), classificação e clusterização (Sharma and Rana 2020; Yadav, Chahande, and Wandre 2018).

Já o processamento da linguagem natural é uma área da inteligência artificial que envolve a compreensão da linguagem humana e a geração de respostas adequadas em linguagem natural, normalmente envolvendo algoritmos de aprendizagem de máquina, que podem ser supervisionados, quando os dados já classificados efetivamente definem as categorias e são usados como "dados de treinamento" para construir modelos que possam ser usados para classificar novos dados (Tan 1999); ou não supervisionados, quando não existem categorias pré-definidas. Uma das técnicas PNL não supervisionada é a análise de tópicos, que permite

identificar os principais temas e conceitos presentes em um conjunto de dados textuais, como avaliações de usuários e comentários em redes sociais, permitindo identificar padrões e tendências nos comportamentos dos turistas, bem como as percepções e opiniões dos viajantes sobre destinos, atrações e serviços turísticos (Silge and Robinson 2017).

Uma questão central na mineração de textos e no processamento de linguagem natural é como quantificar do que se trata um documento. Uma medida de quão importante uma palavra pode ser é sua frequência de termo (`tf`), ou seja, a frequência com que uma palavra ocorre em um documento. Outra abordagem é observar a frequência inversa de documento (`idf`) de um termo, que diminui o peso das palavras comumente usadas e aumenta o peso das palavras que não são muito usadas em uma coleção de documentos. Isso pode ser combinado com a frequência do termo para calcular o `tf-idf` de um termo (as duas quantidades multiplicadas juntas), ou seja, a frequência de um termo ajustada para quão raramente ele é usado (Fay 2018).

Zipf (1949) mostrou que uma das características das linguagens humanas, populações das cidades e muitos outros fenômenos humanos e naturais, seguem uma distribuição similar, a qual denominou de "Princípio do Menor Esforço". Esta lei de Zipf define que, tomando um determinado texto, o produto $k_t \log(f_t)$ é aproximadamente constante, em que f_t é o número de vezes que o termo $_t$ ocorre no texto e k_t é a posição deste termo em uma relação de todos os termos daquele texto, ordenados pela frequência de ocorrência (Bakulina 2008).

Por outro lado, Luhn sugeriu, em 1958, que a frequência de ocorrência das palavras em um texto pode fornecer uma medida útil sobre a expressividade das mesmas, pois o "autor normalmente repete determinadas palavras ao desenvolver ou variar seus argumentos e ao elaborar diferentes enfoques sobre do que se trata o assunto". As palavras com maior frequência de ocorrência deveriam ser consideradas pouco expressivas porque este conjunto de palavras é composto normalmente por artigos, preposições e conjunções. Também as palavras que muito raramente ocorrem deveriam ser consideradas pouco expressivas justamente em razão da baixa frequência. Restam como expressivas as palavras com maior frequência de ocorrência intermediária (Bezerra and Guimarães 2014).

Dessa forma, medidas de frequência de termos são valiosas nas análises textuais. Esses termos são definidos como *tokens*, ou seja, uma unidade significativa de texto, que podem ser palavras, uma sequência delas, frases, parágrafos, entre outros. Portanto, a tokenização é uma etapa importante do processamento de linguagem natural e é usada em muitas aplicações, como a classificação de textos. Os n-gramas, por sua vez, são conjuntos de n tokens adjacentes em um texto (Silge and Robinson 2017). A escolha do melhor método de tokenização e do tamanho dos n-gramas depende do objetivo do problema. O tamanho ideal dos n-gramas depende do tamanho do texto e da complexidade do problema. N-gramas maiores podem ajudar a capturar contextos mais amplos, mas podem levar a um aumento na dimensionalidade e na esparsidade dos dados, o que pode prejudicar o desempenho de modelos. O tamanho ideal dos n-gramas é geralmente determinado empiricamente, testando-se vários valores em um conjunto de dados de validação (Hvitfeldt and Silge 2021).

Nesse contexto, esta pesquisa tem como objetivo disponibilizar uma ferramenta informatizada que possa ser utilizada para a análise dos níveis de experiência e de satisfação de turistas em ambientes naturais, empregando-se técnicas de mineraçao de textos e processamento de linguagem natural. A partir das análises geradas, será possível definir quais os n-gramas mais adequados para serem rotulados com as categorias entretenimento, aprendizagem, estética e escapismo, possibilitando a utilização de algoritmos de aprendizado supervisionado para a classificação dos textos. Assim, modelos de *machine learning* poderão ser treinados para classificar os comentários em cada uma das categorias, através de algoritmos como regressão logística, árvores de decisão e redes neurais. Estudos de casos de pontos turísticos da plataforma *Tripadvisor* serão apresentados neste artigo, de forma ilustrativa para evidenciar e validar a ferramenta desenvolvida. Portanto, pretende-se com o trabalho, gerar um instrumento que possa ser generalizado e utilizado para qualquer ponto turístico.

3. Material e Métodos

Foi construído um aplicativo usando o pacote *Shiny* do ambiente R de computação estatística, com a proposta de fornecer informações suficientes para contribuir na análise de experiências turísticas. Foi utilizado

o *software* estatístico R, que incorpora as características de software livre, e o *Rstudio*, que é o ambiente de desenvolvimento integrado (IDE) para a linguagem R. O aplicativo foi desenvolvido utilizando o *Shiny*, framework que possibilita a criação de aplicações interativas na web, não sendo necessário ao programador ter conhecimento de HTML, CSS ou Javascript. Ele possui um conjunto de funções destinadas a promover a interface com o usuário (*ui*), onde são coletadas informações fornecidas pelo mesmo, e enviadas para o *server*, cujo papel é processar as informações coletadas e retorná-las para ui, que dará o *feedback* ao usuário (Xie 2016; RStudio 2018; Johnson 2020).

Para realizar a análise da experiência turística de acordo com os quatro domínios da experiência, foram selecionados dois pontos turísticos da Plataforma *TripAdvisor*: Jalapão e Superagui. A aquisição dos dados consistiu em obter os comentários dos turistas da plataforma *TripAdvisor*, utilizando técnicas de captura de dados (*web scraping*) (Munzert et al. 2014) para a extração automatizada dos dados, através do pacote *rvest* (Wickham 2019) do *software* R. Também foi utilizada a função *gsub* para retirar as marcações do *html* e obter o texto sem formatação.

Após a importação dos dados, e antes de realizar as análises, foi realizado um **pré-processamento** do mesmos. Este envolveu a transformação das palavras em letras minúsculas, retirada de pontuação e caracteres especiais (p.ex. "!", "@", "#","$") e eliminação de espaços em branco. Existem algumas funções disponíveis para remoção da pontuação, porém algumas substituem a pontuação por espaços e outras não. A função `removePunctuation` do pacote *tm* é muito utilizada, mas deve ser utilizada com cuidado. O ideal é que sua utilização seja somente em arquivos no formato *corpus* em que as palavras já estão separadas. Atenção também deve ser dada aos caracteres especiais, por exemplo o @, pois estes também podem ou não ser substituídos por espaço, alterando o resultado das análises (Hocking 2019).

Em geral as expressões regulares são uma ferramenta poderosa e flexível para trabalhar com texto e foram utilizadas nesse estudo. Elas são amplamente utilizadas em programação e em sistemas de busca e análise de dados, e podem ajudar a automatizar tarefas que envolvem manipulação de texto. As expressões regulares permitem que você especifique um conjunto de caracteres ou um padrão de caracteres que correspondem a uma determinada sequência de texto, para realizar operações de substituição e validação de texto (RStudio 2017).

Como parte do processo de limpeza ou pré-processamento do texto está a remoção de caracteres repetidos. A função `gsub("(\\w)\\1+", "\\1", data, perl = TRUE)` utiliza expressões regulares para manipulação de strings. A expressão regular `(\\w)\\1+` busca por qualquer caractere alfanumérico `(\\w)` seguido por uma ou mais ocorrências do mesmo caractere `(\\1+)`. O primeiro argumento `\\1` é uma referência à captura da primeira parte da expressão, que é o caractere encontrado inicialmente. O segundo argumento `\\1` da função `gsub` é o caractere que será utilizado para substituir a ocorrência de caracteres repetidos. Como resultado, essa função produzirá uma string em que todos os caracteres repetidos consecutivos são reduzidos a uma única ocorrência. O argumento `perl = TRUE` diz ao R para usar a sintaxe de expressão regular Perl compatível com PCRE (Perl Compatible Regular Expressions). Isso permite que a expressão regular faça uso de recursos avançados, como *lookbehind* e *lookahead*, que não estão disponíveis na sintaxe padrão do R (Silge and Robinson 2017).

Porém, essa função não considera caracteres especiais e elimina dígrafos, o que não é uma boa alternativa para o português. Por exemplo, **carro** e **caro** possuem significados diferentes. Uma das alternativas para corrigir isso é tratar **r** e **s** separadamente. Portanto, outra alternativa é utilizar a função `gsub("([^rs])(?=\\1+)|(rr)(?=r+)|(ss)(?=s+)` em que três expressões regulares entre parênteses que estão agrupadas por `\\` (ou lógico), sendo: `([^rs])(?=\\1+)` uma expressão correspondente a qualquer caractere que não seja "r" ou "s", desde que não seja seguido de um ou mais caracteres idênticos a ele mesmo. `(?=\\1+)` é um *lookahead* positivo que verifica se o caractere seguinte é igual ao primeiro caractere. `\\1+` se refere à primeira captura do grupo entre parênteses, que é o caractere correspondido na primeira parte da expressão. A expressão `(rr)(?=r+)` corresponde a duas ocorrências consecutivas do caractere "r", desde que sejam seguidas por um ou mais caracteres "r". O mesmo vale para expressão `(ss)(?=s+)` considerando o caractere "s". Por fim, a função `gsub("\\b\\w{1,2}\\b\\s*", "", dados)` remove todas as palavras com menos de três caracteres.

A remoção de *stopwords*, que são palavras que ocorrem muitas vezes, mas não fornecem nenhuma contribuição na identificação do conteúdo do texto, também foi realizada (Sarica and Luo 2021). Por exemplo

advérbios, artigos, conjunções, preposições e pronomes. Nesse trabalho optou-se por utlizar um conjunto de listas de *stopwords* de diferentes fontes, de forma a contemplar o máximo de palavras possíveis. Foram utilizadas: 1) a lista padrão do pacote *tm*, com 203 palavras; 2) a lista de da RSLP (Redução de Sequências de Letras e Stemming para Português) que é uma lista mais completa e específica para o português do pacote *SnowballC*, com 560 palavras e 3) a lista disponível no repositório GitHub de Jodavid Ferreira, com 577 palavras. Essas listas foram unidas e retiradas as palavras duplicadas, resultando em 644 palavras. Além disso, a palavra "não" pode ser considerada como *stopword* dependendo do contexto. Nesse trabalho ela não foi incuída na lista de *stopwords*.

O *stemming* e o seu melhor representante é um processo que tem a função de diminuir a variação de uma mesma palavra nos documentos, chegando aos radicais e substituindo estes radicais pela palavra mais frequente originalmente. O processo *stemming* foi implementado usando a função `stemDocument` do pacote *tm* (Feinerer, Hornik, et al. 2014) e, para selecionar o seu melhor representante, foi desenvolvida a função *Representante*. Esta função realiza uma contagem das palavras considerando um dado radical e elege um representante para esse radical, substituindo-o pela palavra com a maior ocorrência.

Posteriormente ao pré-processamento, foi realizada a *geração de n-gramas*, que é um conjunto de *tokens* candidatos, que podem ser relevantes na análise. Para isso, foi utilizado o pacote *tidyverse*. Na composição do núcleo do pacote, existe um conjunto de outros pacotes que atendem as premissas da ciência de dados, como o *dplyr* (Wickham and Francois 2016), *tidyr* (Wickham 2016), *ggplot2* (Wickham 2009) e *broom* (Robinson 2017), que subsidiam as ações de importar, arrumar e visualizar dados (Fay 2018). Após definido o formato de texto organizado como sendo uma tabela com um *token* por linha, os dados foram manipulados sob a forma de *strings*, ou vetores de caracteres, *corpus*, que são objetos anotadas com metadados e detalhes adicionais e *matrizes documento-termo*, que é uma matriz esparsa que descreve uma coleção (ou seja, um *corpus*) de documentos com uma linha para cada documento e uma coluna para cada termo. As proximas análises foram realizadas para cada n-grama, sendo eles: unigramas ou palavras, bigramas, trigramas, tetragramas e pentagramas. A disponibilização de dados com esta formatação (n-gramas) possibilita ao analista interpretar os sentidos das comunicações, facilitando a categorização das palavras, aplicando-se os procedimentos de análise temática proposta por Bardin (2011).

Foi calculada a frequência absoluta (FA) de n-gramas ou termos, que é a contagem de quantas vezes um determinado token aparece em um documento ou em uma coleção de documentos e a frequência relativa (FR), que refere-se a quantidade de vezes que um termo aparece em uma coleção de documentos, mas sem considerar as repetições em um mesmo documento. Assim, FR resulta em uma listagem de termos e a porcentagem de documentos em que estes aparecem pelo menos uma vez. A necessidade de se obter FA e FR ocorreu, visto que elas se complementam. A FA mostra as frequências das palavras diante de outras palavras na coleção de documentos, enquanto FR mostra quais palavras são relevantes considerando todos os documentos. Ou seja, uma palavra pode ter uma frequência elevada segundo o índice de FA, todavia, pode-se verificar que, apesar disso, essa palavra não aparece em todos os documentos, o que é indicado pela FR. Trata-se, portanto, de operações para refinamento dos dados (Silge and Robinson 2017).

Foi realizada a clusterização ou análise hierárquica de agrupamentos, através da construção do dendrograma Euclidiano, com o auxílio do pacote *dendextend* (Galili 2014), em que palavras com FA semelhantes e com frequência nos mesmos documentos tendem a ter uma distância euclidiana menor entre si, proporcionando indícios de quais palavras são usadas no mesmo contexto. A utilidade deste tipo de apresentação de dados, consiste em oportunizar ao analista identificar conjuntos de termos que podem indicar aspectos relevantes do contexto, facilitando a comparação entre conjuntos de dados.

Outra abordagem é a análise de sentimentos, em que as palavras são classificadas como positivas ou negativas com base no dicionário de sentimentos e os resultados são visualizados em um gráfico de dispersão que mostra a frequência de palavras positivas e negativas. Existem vários dicionários para a língua portuguesa, como por exemplo: 1) *SentiLex-PT*: um dicionário de sentimentos para o português que contém cerca de 5.000 palavras classificadas em termos de polaridade (positiva, negativa e neutra) e intensidade (forte, média e fraca); 2) *OpLexicon*: um dicionário de opinião e sentimento que contém cerca de 30.000 palavras classificadas em termos de polaridade (positiva, negativa e neutra) e subjetividade; 3) *Bing*: um dicionário de sentimentos em inglês que pode ser utilizado também para a língua portuguesa. Ele contém cerca de 6.800 palavras classificadas em termos de polaridade (positiva ou negativa); 4) *LIWC (Linguistic Inquiry and Word Count)*:

um software que analisa o texto em várias dimensões linguísticas, incluindo a análise de sentimentos. Ele contém um dicionário de sentimentos para o português; 5) *Emotion-LexPT*: um dicionário de emoções para o português que contém cerca de 2.000 palavras classificadas em termos de emoções básicas (como alegria, tristeza, raiva, medo, nojo e surpresa) e emoções secundárias (como admiração, esperança, amor e confusão) (Freitas and Vieira 2016; Duarte 2012; Carosia, Coelho, and Silva 2020; Lopes et al. 2022; Oliveira and Campos Merschmann 2021; Pereira 2021). Nesse estudo foi utilizado o dicionário OpLexicon, através do pacote *lexiconPT* (Gonzaga 2022).

Por fim, a análise de tópicos também foi realizada, permitindo identificar tópicos latentes em um conjunto de documentos. Essa técnica busca agrupar as palavras em torno de temas comuns, identificando os tópicos mais relevantes presentes nos documentos. Foi utilizado o método LDA (*Latent Dirichlet Allocation*), que é um modelo probabilístico que assume que cada documento é composto por uma mistura de vários tópicos e que cada palavra dentro do documento é gerada a partir de um dos tópicos do documento. Em outras palavras, o LDA é uma técnica que busca descobrir os tópicos subjacentes em um conjunto de documentos, levando em conta a distribuição probabilística das palavras em cada tópico.

4. Resultados

Nesse trabalho, dois tipos de resultados foram gerados. O **aplicativo shiny**, que pode ser acessado aqui, e um **Bookdown** (ainda não publicado), que contém informações detalhadas de como o *app* foi construído, servindo também como documentação ou tutorial que explica as suas principais funcionalidades e usabilidade. Partindo do princípio de ciência aberta, as ferramentas desenvolvidas são baseadas em código aberto que permitem fluxos de trabalho reprodutíveis. A abordagem proposta e implementada se encontra disponível juntamente como os arquivos de dados no repositório do Git-Hub. Para fins de apresentação nesse artigo, não serão mostradas todas as figuras, principalmente dos tetragramas e pentagramas, pois os tamanhos dos outputs gerados não se adequam na renderização do PDF. Eles podem ser visualizados em *html* ou no próprio aplicativo. De modo geral, o aplicativo consistui-se em uma página inicial contendo uma barra de navegação de nível superior *navbarPage* e uma funcionalidade de ajuste de tamanho das figuras, com um conjunto de painéis separados *tabPanel*, referentes a sete abas temáticas: Importação dos dados; Frequência de N-Gramas; Vizualização das frequências; Wordclouds; Redes de Palavras; Análise de Agrupamentos; Análise de sentimentos e Análise de tópicos. O formato dos arquivos de texto a serem importados é livre, mas recomeda-se .csv; .txt ou .tsv. O número de colunas do arquivo também não impossibita a análise, já que o arquivo todo passará por um processo de limpeza e será tradado como caractere.

4.1. Frequência de n-gramas

Os dados obtidos da platafora *TripAdvisor* resultaram em 171 comentários para Superagui e 2104 para o Jalapão. As frequências de ocorrência dos 15 primeiros unigramas podem ser visualizados na Tabela 1. As palavras mais comuns de Superagui foram “praia” e “ilha” com 195 e 115 ocorrências respectivamente. Já para o Jalapão, “água” e “cachoeira” foram as mais frequentes, com 1165 e 984 ocorrências, respectivamente. É possivel observar que a palavra “não” apareceu entre as 3 palavras mais citadas nos dois atrativos turísticos. Portanto, convém observar quais são as palavras subsequentes da negação para compreender melhor o contexto.

Sendo assim, duas tabelas de bigramas foram geradas. Uma com todos os bigramas do *corpus* (Tabela 2) e outra com os bigramas que iniciam com a palavra “não” (Tabela 3). Pode-se perceber que os bigramas negativos de Superagui “não esqueça” e “não estrutura” sugerem algum tipo de alerta, enquanto que os do Jalapão “não afundar”, “não nadar” e “não permitido” correspondem a um tipo de relato, com certa expressão de frustação, o que já demonstra um tipo de indício da experiência turística do lado ativo.

Os bigramas mais comuns de Superagui foram praia deserta (63) e vale pena (27). Já para o Jalapão, vale pena (309) e água cristalina (148) foram os mais frequentes (Tabela 3).

Do mesmo modo que os n-gramas menores, as frequências absolutas e relativas podem ser geradas para qualquer n-grama (Tabela 4). Aí vai do especialista que está analisando os resultados decidir qual é o mais apropriado, dependendo de seu objetivo. No app *shiny*, é possível gerar e visualizar os resultados até os

Table 1: Frequência de unigramas

Superagui	FA	FR	Jalapão	FA	FR
praia	195	64.91	água	1165	40.31
ilha	115	35.67	não	1131	38.22
não	99	35.09	cachoeira	984	35.78
deserta	76	35.09	linda	660	27.81
trilha	69	29.82	jalapão	600	25.05
superagui	60	24.56	duna	540	17.18
mar	57	22.22	sol	525	21.37
barco	54	22.81	vale	484	21.85
água	48	20.47	visita	364	15.03
levar	45	22.22	pena	345	15.94
chegar	42	19.88	banho	338	14.36
natureza	39	22.81	ficar	336	14.17
pousada	38	19.30	areia	300	11.93
vale	38	20.47	chegar	280	11.93
linda	37	19.30	acesso	267	12.02

Table 2: Frequência de bigramas

Superagui	FA	FR	Jalapão	FA	FR
praia deserta	63	30.41	vale pena	309	14.41
vale pena	27	15.20	água cristalina	148	7.06
ilha superagui	19	8.19	pôr sol	127	5.63
levar água	14	8.19	cachoeira formiga	106	4.96
parque nacional	13	7.60	capim dourado	95	4.39
ilha peça	11	6.43	queda água	93	4.20
vila pescador	11	5.85	cachoeira linda	85	4.06
nacional superagui	10	5.85	água transparente	76	3.63
barra superagui	9	4.68	não deixe	75	3.53
alugar bike	8	4.68	cachoeira velha	74	3.29
chegar praia	8	4.68	nascer sol	68	3.10
contato natureza	8	4.68	espírito santo	67	3.10
passeio barco	8	4.09	tirar foto	65	2.96
praia linda	8	4.09	serra espírito	64	2.96
barco paranaguá	6	3.51	tomar banho	61	2.77

pentagramas. No caso das tabelas, o formato datatable do pacote DT (Yihui Xie 2018), permite visualizar tabelas interativas, que podem ser facilmente personalizadas com várias opções, como ordenação, pesquisa, paginação e seleção de linhas.

4.2. Nuvens de Palavras

Depois de calculadas as frequencias de termos, foram gerados gráficos úteis para visualizar os resultados, por exemplo as nuvens de palavras, que são representações visuais de um conjunto de palavras, em que as palavras mais frequentes são exibidas com maior destaque (Figura 2 e 3). Nuvens de palavras são uma forma de visualização de dados que ajuda a identificar rapidamente as palavras mais relevantes em um texto ou

Table 3: Frequência de bigramas de negação

Superagui	**FA**	**Jalapão**	**FA**
não esqueça	5	não deixe	75
não estrutura	4	não afundar	45
não encontrar	3	não nadar	45
não avisado	2	não permitido	35
não barco	2	não banho	26
não cara	2	não entrar	26
não deixe	2	não conseguir	16
não existe	2	não fria	16
não movimento	2	não gelada	16
não opção	2	não vontade	16
não venda	2	não tomar	15
não água	2	não esqueça	14
não aceitei	1	não fácil	14
não acho	1	não ficar	13
não ajudem	1	não não	13

Table 4: Frequência de tetragramas

Superagui	**FA**	**FR**	**Jalapão**	**FA**	**FR**
acampar aconselhável levar comida	2	1.17	erosão serra espírito santo	10	0.48
aconselhável levar comida preço	2	1.17	cachoeira linda água cristalina	9	0.43
afinando desaguar estreito largura	2	1.17	formada erosão serra espírito	8	0.38
alta temporada levar água	2	1.17	trilha serra espírito santo	8	0.38
alternativa busca sossego acampar	2	1.17	cachoeira linda água transparente	5	0.24
aproximadamente praia mar baixa	2	1.17	duna formada erosão serra	5	0.24
baixa vimo barranco vegetação	2	1.17	jalapão não deixe conhecer	5	0.24
barco indo volta diariamente	2	1.17	água não deixe afundar	5	0.24
barco voadeira chegar cal	2	1.17	artesanato capim dourado peça	4	0.19
barranco vegetação praia mar	2	1.17	comprar artesanato capim dourado	4	0.19
busca sossego acampar aconselhável	2	1.17	famosa artesanato capim dourado	4	0.19
cal embarcação carnaval recebe	2	0.58	linda vale pena conhecer	4	0.19
cara devido isolada alta	2	0.58	não permitido tomar banho	4	0.19
carnaval recebe visitante réveilon	2	0.58	rio água potável mundo	4	0.19
chegar cal embarcação carnaval	2	0.58	serra espírito santo fundo	4	0.19

conjunto de dados. Elas são úteis para descobrir rapidamente as palavras-chave ou tópicos mais frequentes em um texto, ajudando a resumir as informações de forma concisa e visualmente atraentes (Hvitfeldt and Silge 2021). No aplicativo *shiny*, existem 3 opções de ajuste das *wordclouds*. São elas: o tamanho das letras, o espaçamento entre elas e a quantidade de n-gramas postados. Dessa forma, consegue-se ajustar a escala da figura para que os n-gramas sejam mostrados sem cortes.

Figure 2: Wordcloud de unigramas

Figure 3: Wordcloud de bigramas

4.3. Frequencia td-idf

A ideia do `tf-idf` é encontrar as palavras importantes para o conteúdo de cada documento diminuindo o peso das palavras comumente usadas e aumentando o peso das palavras que não são muito usadas em uma coleção ou *corpus* de documentos, neste caso, o grupo de atrativos turísticos. O cálculo do `tf-idf` tenta encontrar as palavras que são importantes/comuns em um texto, mas não muito comuns.

A função `bind_tf_idf()` no pacote *tidytext* recebe um conjunto de dados de texto organizado como entrada com uma linha por *token* (termo), por documento. Uma coluna contém os termos, outra coluna contém os atrativos e a última coluna contém as contagens, ou quantas vezes cada atrativo contém cada termo. Normalmente, `idf` e, portanto, `tf-idf` são zero para essas palavras extremamente comuns. Estas são todas as palavras que aparecem em todos os dois atrativos, então o termo `idf` (que será então o logaritmo natural de 1 é zero. A frequência inversa do documento `tf-idf` é muito baixa (próxima de zero) para palavras que ocorrem em muitos dos documentos em uma coleção. É assim que essa abordagem diminui o peso das palavras comuns. A frequência de documento inversa será um número maior para palavras que ocorrem em menos documentos na coleção.

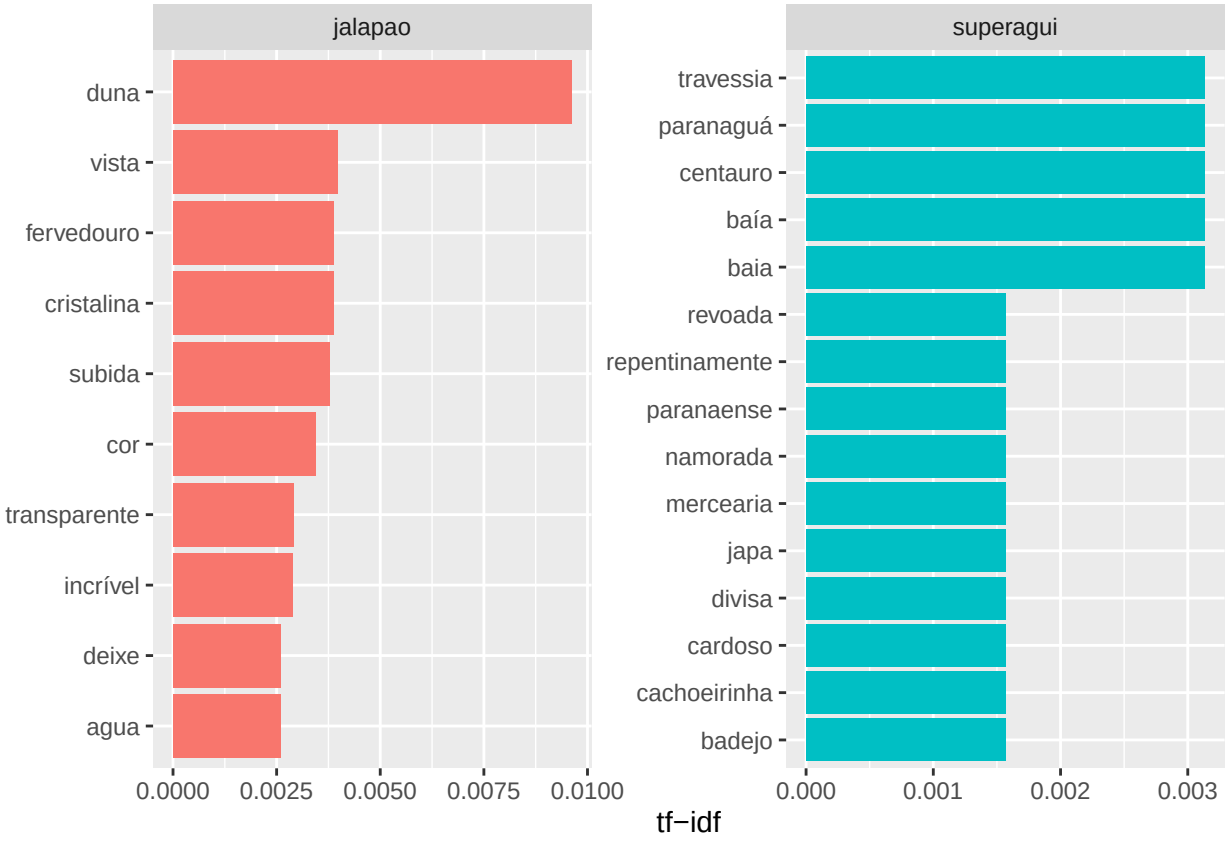

Figure 4: Frequência td-idf dos atrativos Jalapão e Superagui

A Figura 4 mostra a frequência `td-idf` entre os atrativos turísticos utilizados como caso de estudo.

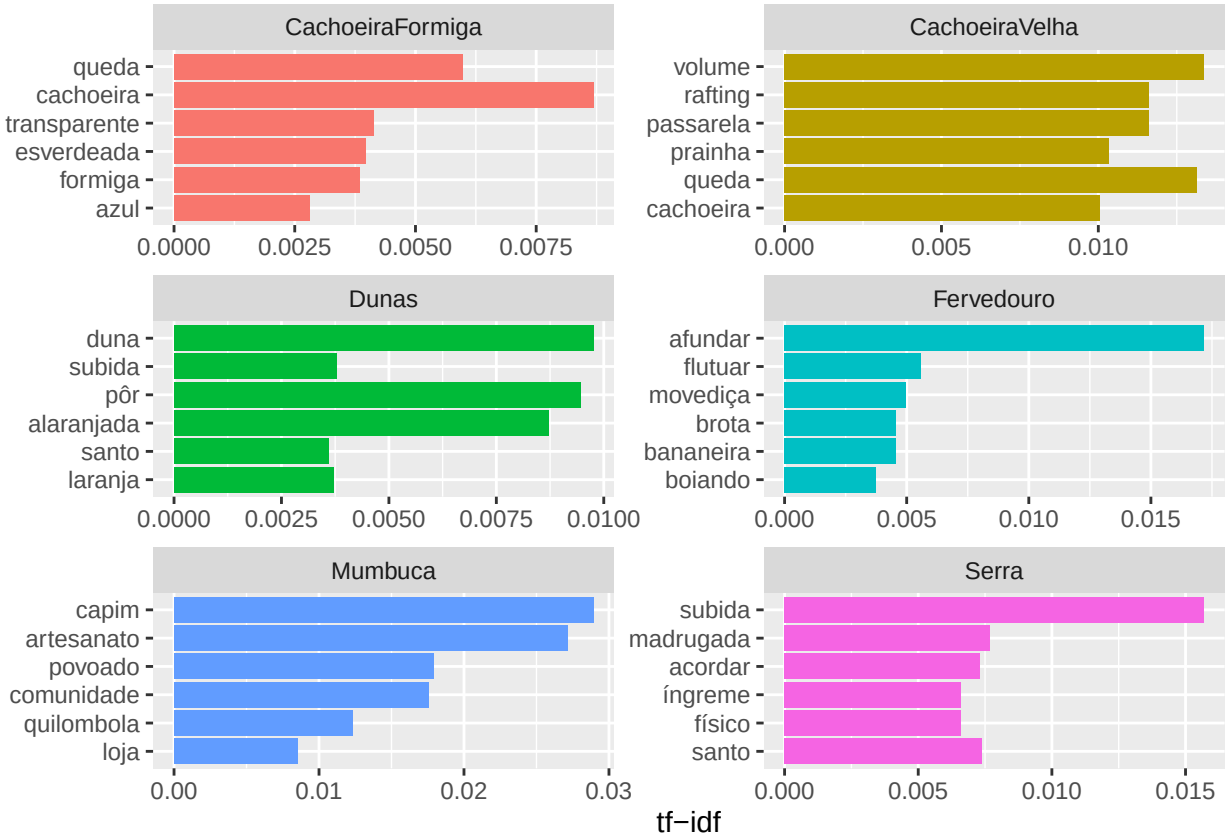

Figure 5: Frequência td-idf dos atrativos do Jalapão

A Figura 5, mostra outro exemplo, agora dividindo os atrativos existentes dentro do Jalapão.

4.4. Análise de sentimentos

A avaliação da experiência turística é um processo complexo, que envolve a análise de diversos fatores, como o atendimento ao cliente, a qualidade dos serviços oferecidos, o ambiente e a cultura local, entre outros. A análise de sentimentos pode ser uma ferramenta útil para avaliar a experiência turística, pois permite capturar e quantificar as emoções expressas pelos turistas em relação a diferentes aspectos da viagem. Com a análise de sentimentos, é possível identificar as emoções positivas e negativas associadas a cada experiência turística, e utilizar essas informações para melhorar a qualidade dos serviços oferecidos (Figuras 6 e 7).

Figure 6: Análise de sentimentos de Superagui

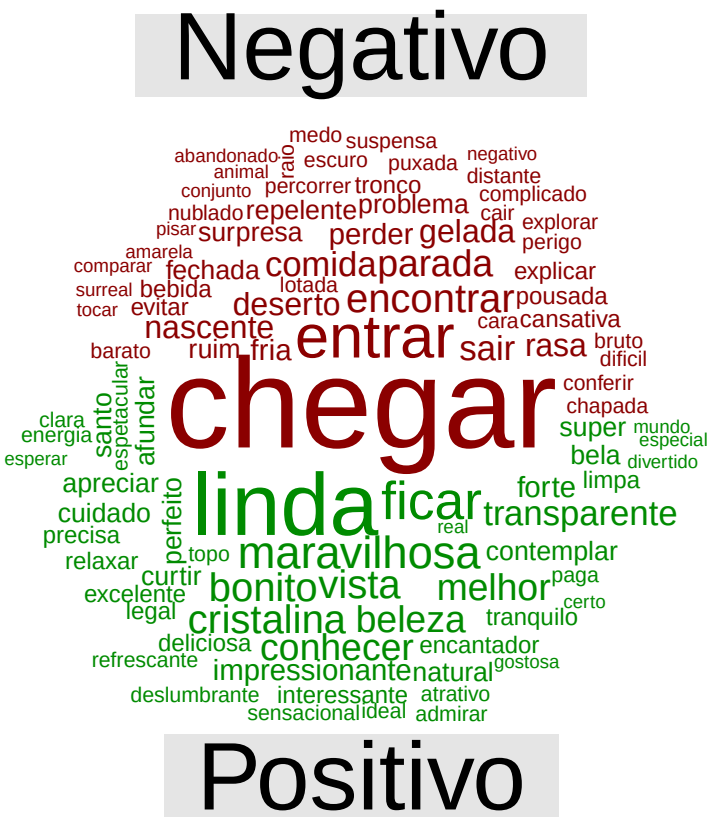

Figure 7: Análise de sentimentos do Jalapão

4.5. Clusterização

Em termos de clustering hierárquico, a esparsidade refere-se à presença de grupos de dados com poucas observações. Esses grupos são assim considerados porque têm poucos pontos de dados em relação aos outros grupos. Ao realizar uma análise de cluster, é importante levar em consideração a esparsidade e como ela pode afetar o resultado. Uma forma de lidar com ela é aplicar técnicas de agrupamento aglomerativo que considerem a distância entre grupos ponderando a densidade dos grupos. Alguns métodos, como o método "ward.D2" no R, tentam minimizar a variância total dos grupos e levando em consideração a esparsidade (Figuras 8 e 9).

No aplicativo *shiny*, existe a opção de ajuste de esparsidade de forma a ajudar na interpretação dos resultados da análise. O método "ward.D2" ainda não foi implementado no aplicativo, que, por enquanto, está usando apenas a distância euclidiana.

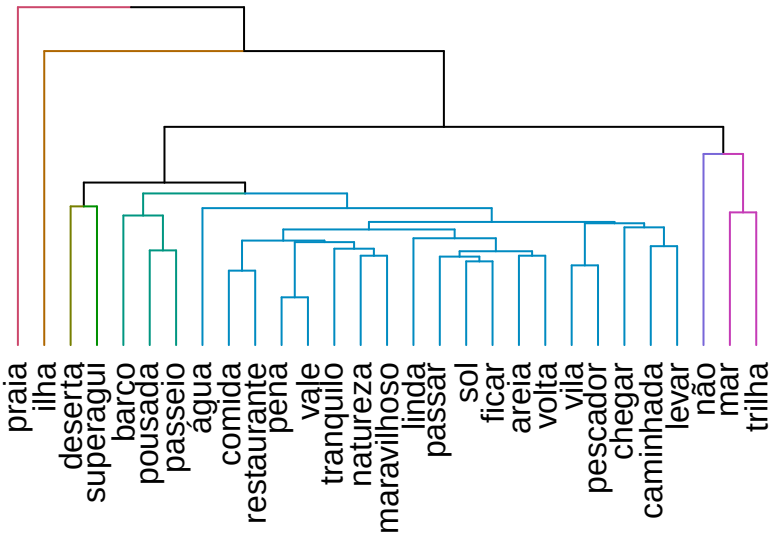

Figure 8: Dendrograma de Superagui

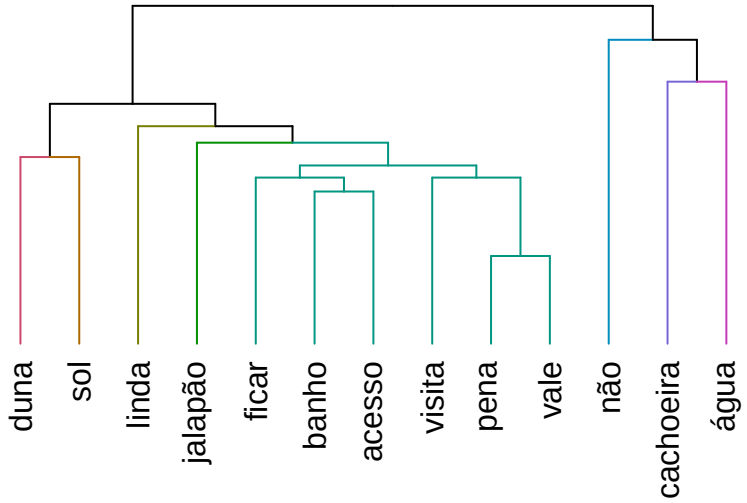

Figure 9: Dendrograma do Jalapão

4.6. Rede de Palavras

Foi realizada a distribuição de palavras sucessoras, ou redes de palavras (Figura 10 e 11), também conhecidas como redes léxicas ou grafos de palavras, que são representações gráficas que mostram as relações entre as palavras em um determinado *corpus* de texto. O resultado é a construção de conjuntos de palavras que fornecem maior sentido que termos isolados. Cada palavra é representada por um nó na rede, e as conexões entre elas são representadas por arestas que indicam a frequência e a força das relações entre as palavras. A utilidade das redes de palavras está na capacidade de mostrar as associações semânticas entre as palavras em um corpus, ajudando a identificar tópicos ou temas importantes, palavras-chave e padrões linguísticos (Fay 2018).

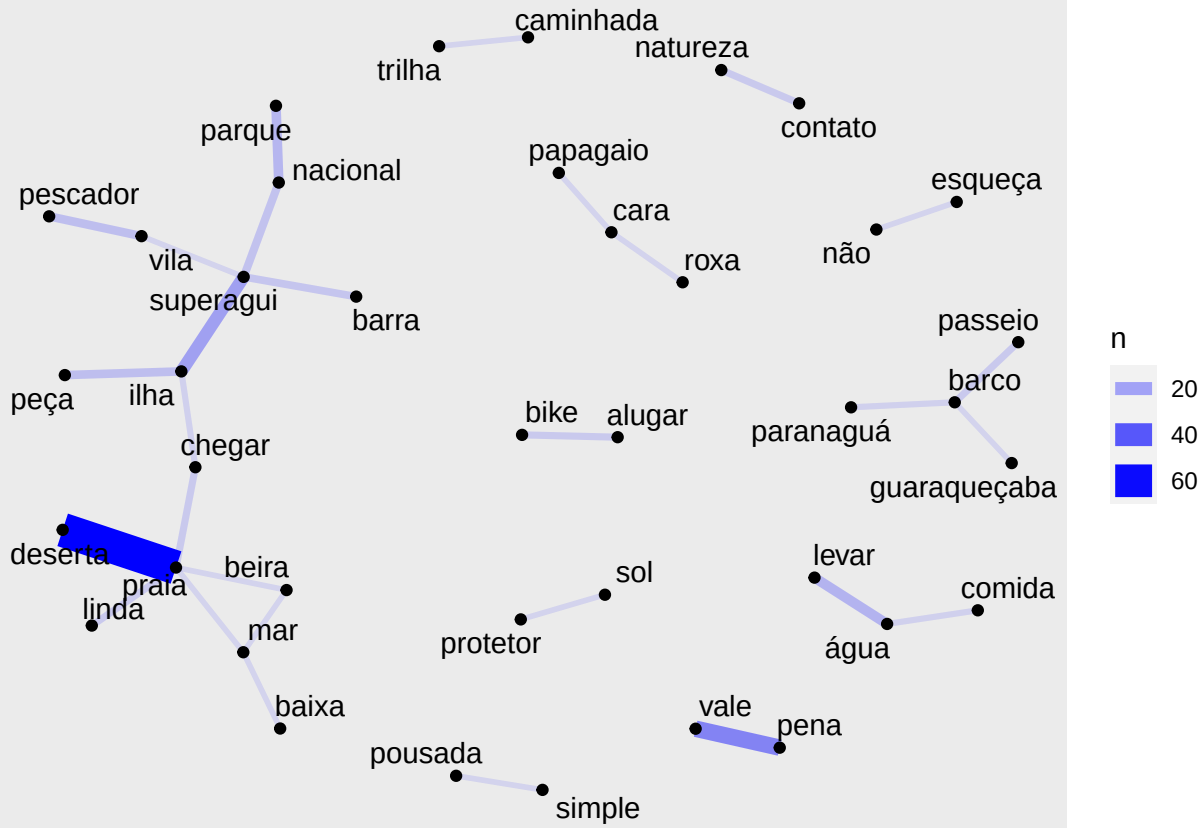

Figure 10: Redes de bigramas de Superagui

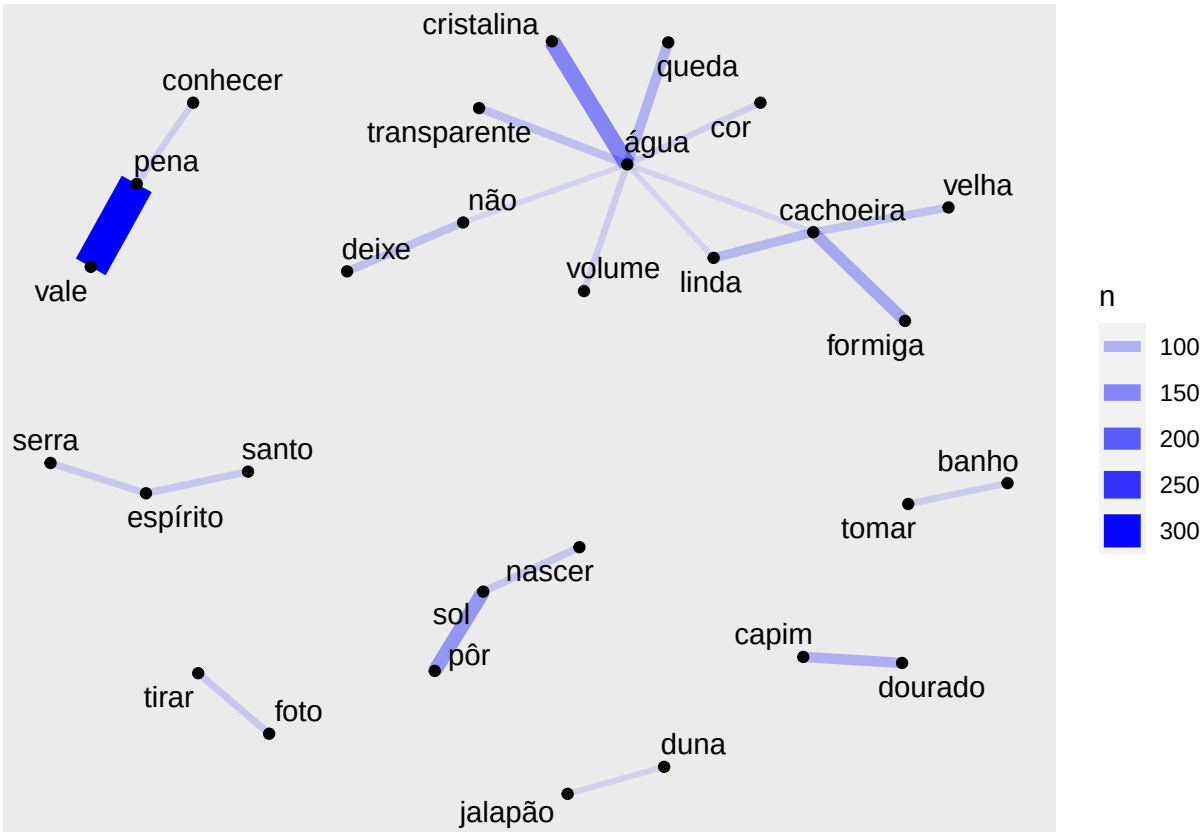

Figure 11: Redes de bigramas do Jalapão

4.7. Análise de tópicos

A LDA é um dos algoritmos mais comuns para modelagem de tópicos. Sem mergulhar na matemática por trás do modelo, podemos entendê-lo como sendo guiado por dois princípios. O primeiro é que cada documento é uma mistura de tópicos. Imaginamos que cada documento pode conter palavras de vários tópicos em proporções particulares. Por exemplo, em um modelo de dois tópicos, poderíamos dizer "Documento 1 é 90% do tópico A e 10% do tópico B, enquanto o Documento 2 é 30% do tópico A e 70% do tópico B". O segundo prinçipio é que cada tópico é uma mistura de palavras, onde cada tópico possui palavras relacionadas a ele, porém, as palavras podem ser compartilhadas entre os tópicos. A LDA é um método matemático para estimar ambos ao mesmo tempo: encontrar a mistura de palavras que está associada a cada tópico, ao mesmo tempo em que determina a mistura de tópicos que descreve cada documento. A vantagem da modelagem de tópicos em oposição aos métodos de "agrupamento rígido" é que os tópicos usados em linguagem natural podem ter alguma sobreposição em termos de palavras (Fay 2018).

Nesse caso ilustrativo, utilizamos apenas um atrativo turístico para analisar os tópicos, porém a análise pode ser feita com mais de um atrativo ao mesmo tempo. Da mesma maneira que as outras análises, ela pode ser feita a partir de diferentes n-gramas, mas deve-se considerar a esparsidade com que longos n-gramas possuem.

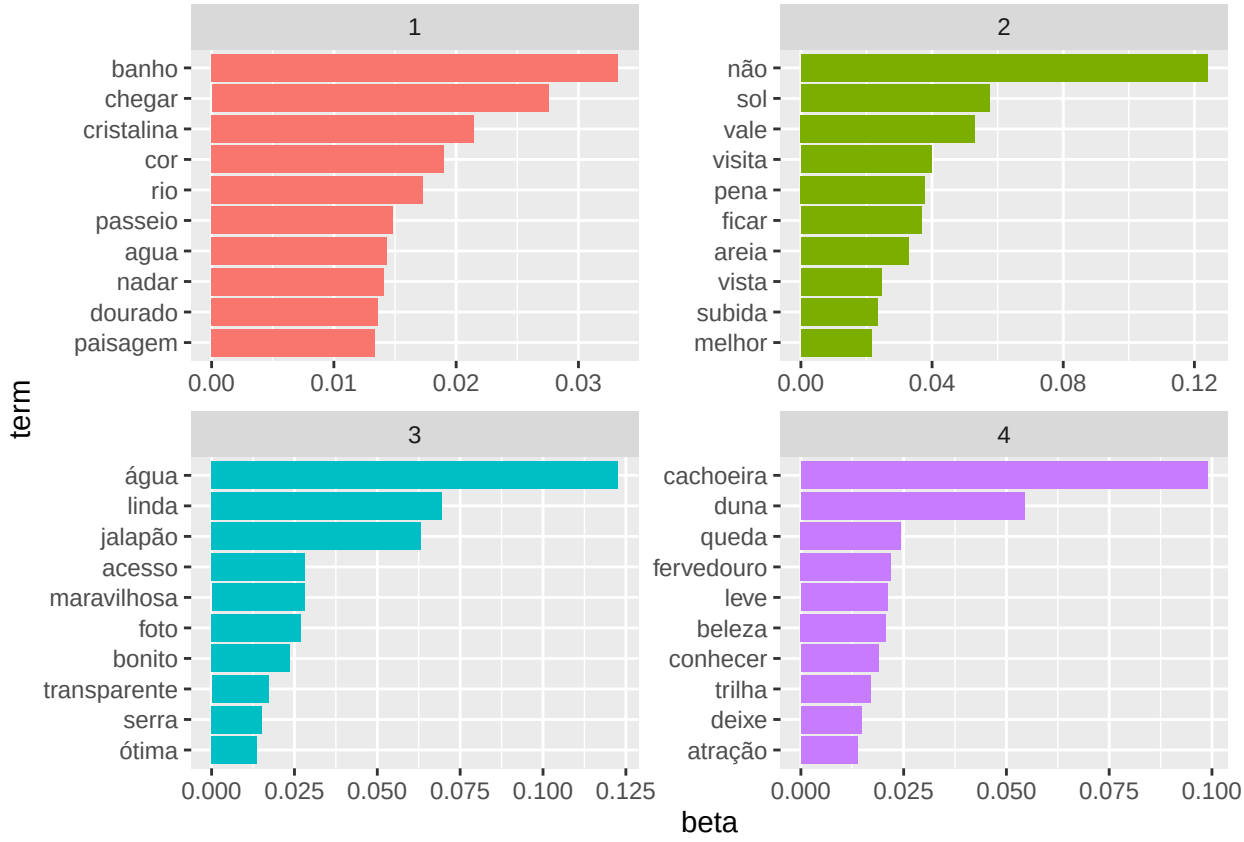

Figure 12: Análise de tópicos do Jalapão

A Figura 12 mostra uma análise de tópicos gerada para o Jalapão. O pacote de *tidytext* fornece um método para extrair as probabilidades por tópico por palavra, chamado β ("beta"), do modelo. Este transforma o modelo em um formato de um tópico por termo por linha. Para cada combinação, ele calcula a probabilidade daquele termo ser gerado daquele tópico. A função slice_max() do dplyr pode ser usada para encontrar os n termos mais comuns em cada tópico.

Como alternativa a primeira análise (Figura 13), pode-se considerar os termos que tiveram a maior diferença de β entre o tópico 1 e o tópico 2. Isso pode ser estimado com base na razão logarítmica dos dois (uma razão logarítmica é útil porque torna a diferença simétrica: $\beta 2$ sendo duas vezes maior leva a uma razão logarítmica de 1, enquanto $\beta 1$ sendo duas vezes maior resulta em -1). Para restringi-lo a um conjunto de palavras especialmente relevantes, pode-se filtrar por palavras relativamente comuns, como aquelas que têm um β superior a 1/1000 em pelo menos um tópico.

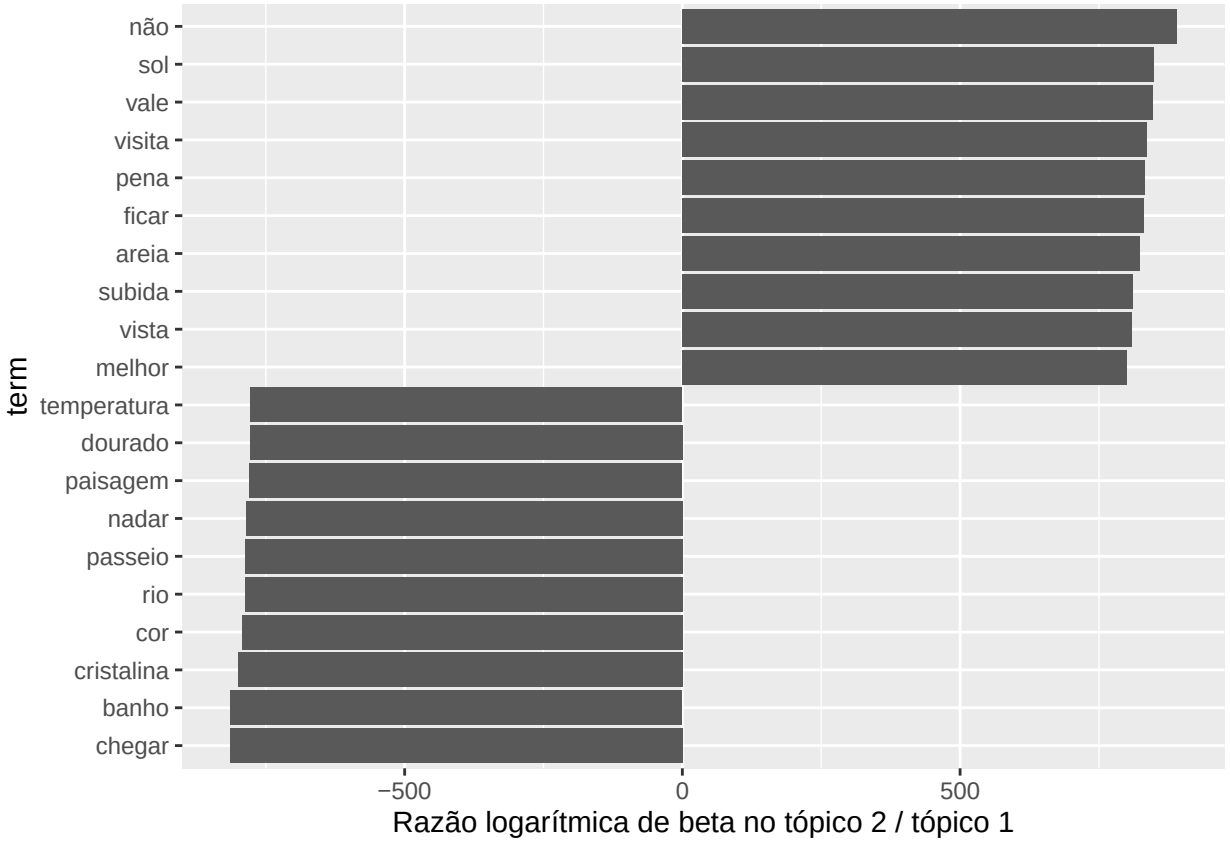

Figure 13: Análise de tópicos do Jalapão

5. Conclusão

Este artigo explorou as características e as potencialidades do pacote *Shiny* do software R para o desenvolvimento de um aplicativo *web* voltado à mineraçao de textos e análise de experiências turíticas. Através de um conjunto de passos que asseguram a captura e organização automática dos dados, pode-se realizar uma análise eficiente dos textos de comentários de turistas. Além da utilização de diversas funções estatísticas bastante conhecidas, foram desenvolvidas algumas funções específicas para o processamento de textos, as quais ainda não estavam disponíveis para o software R (R Core Team 2023). A proposta envolveu o desenvolvimento de um conjunto de instrumentos para o processamento de dados textuais e visualização destes na forma de tabelas, gráficos, dendrogramas, dentre outros. Isso pode ajudar empresas de turismo e hospitalidade a melhorar a experiência do usuário, identificando áreas de melhoria e desenvolvendo estratégias para atender às expectativas dos clientes.

Sabe-se que existem outros *softwares* que realizam este trabalho, todavia, o analista fica restrito as funcionalidades inerentes ao *software*. Nesse sentido, a proposta neste trabalho é fornecer uma ferramenta flexível e de baixo custo para o usuário, que permita a supressão ou inclusão de funcionalidades conforme as suas necessidades.

A fim de testar o instrumento e validar os procedimentos, foram feitos dois estudos de casos, para os quais foi seguido o procedimento desenvolvido, possibilitando coletar os dados e organizá-los como pretendido. Com os testes preliminares, ficou demonstrado que a abordagem consegue capturar e organizar automaticamente os dados.

Algumas análises ainda estão em fase de implementação do app, como a análise de frequência *td-idf*, outros métodos de clusterização para dados esparsos (como o “ward.2”) e a análise de tópicos para um

conjunto de vários atrativos ao mesmo tempo, bem como a sua forma de visualização invertida. Além disso, alguns ajustes ainda precisam ser feitos, tanto na interface visual do usuário (front-end) como tamanho dos plots, quanto no tempo de processamento e otimização do código.

A continuidade da pesquisa envolve a escolha do n-grama mais representativo para a classificação manual destes por especialistas, de modo a criar um conjunto robusto de dados de treinamento para posterior aplicação de algoritmos de análise supervisionada, de forma a gerar classificações automáticas para novos conjuntos de dados; bem como a aplicação dessa metodologia juntamente com especialistas da área do turismo, a fim de desenvolver e verificar possibilidades analíticas em situações reais de tomada de decisão.

Referências

Alencar, Débora Gonçalves, Marina Lima dos Santos, Adriely Andrade e Souza, and José Manoel Gonçalves Gândara. 2019. "Produtos Turísticos Para Demandantes de Experiências Da Dimensão Entretenimento de Pine &Amp; Gilmore: Novas Características e Tendências Para o Paraná." *Turismo Visão e Ação* 21 (June): 46. https://doi.org/10.14210/rtva.v21n2.p46-67.

Andrukiu, A, and José M G Gândara. 2015. "As Emoções No Destino: Classificando Os Atrativos Turísticos de Antonina, Paraná (Brasil)." *Revista Hospitalidade* XII.

Aroeira, Tiago, Ana Carmem Dantas, and Marlusa De Sevilha Gosling. 2016. "Experiência Turística Memorável, Percepção Cognitiva, Reputação e Lealdade Ao Destino: Um Modelo Empírico." *Turismo - Visão e Ação* 18. https://doi.org/10.14210/rtva.v18n3.p584-610.

Bakulina, Marina P. 2008. "Application of the Zipf Law to Text Compression." *Journal of Applied and Industrial Mathematics* 2. https://doi.org/10.1134/S1990478908040042.

Bardin, Laurence. 2011. *Análise de Conteúdo*. Paperback; Edições 70.

Beni, M. C. 2019. *Análise Estrutural Do Turismo*. Editora Senac São Paulo. https://books.google.com.br/books?id=f9GCDwAAQBAJ.

Beni, Mário Carlos. 2004. "Turismo : Da Economia de Serviços à Economia Da Experiência." *Turismo Visão e Ação* 6.

Bezerra, Cicero Aparecido, and André José Ribeiro Guimarães. 2014. "Mineração de Texto Aplicada Às Publicações Científicas Sobre Gestão Do Conhecimento No Período de 2003 a 2012." *Perspectivas Em Ciência Da Informação* 19. https://doi.org/10.1590/1981-5344/1834.

Carosia, A. E. O., G. P. Coelho, and A. E. A. Silva. 2020. "Analyzing the Brazilian Financial Market Through Portuguese Sentiment Analysis in Social Media." *Applied Artificial Intelligence* 34. https://doi.org/10.1080/08839514.2019.1673037.

Coelho, André, and Leticia Ribeiro. 2007. "A Economia Da Experiência." *Observatório de Inovação Do Turismo* 2: 1–3. www.ebape.fgv.br/revistaoit.

Corrêa, Stela Cristina Hott, and Marlusa De Sevilha Gosling. 2023. "A Experiência Turística Inteligente Na Perspectiva Do Viajante." *Turismo: Visão e Ação* 25. https://doi.org/10.14210/rtva.v25n1.p72-93.

Duarte, Eduardo Santos. 2012. "Sentiment Analysis on Twitter for the Portuguese Language." *Lncs*.

Fay, Colin. 2018. "Text Mining with r : A Tidy Approach." *Journal of Statistical Software* 83. https://doi.org/10.18637/jss.v083.b01.

Feinerer, Ingo, Kurt Hornik, et al. 2014. "Text Mining Package." *R Reference Manual, R-Project.org*. https://doi.org/1.

Freitas, Larissa A. De, and Renata Vieira. 2016. "Exploring Resources for Sentiment Analysis in Portuguese Language." In *Proceedings - 2015 Brazilian Conference on Intelligent Systems, BRACIS 2015*. https://doi.org/10.1109/BRACIS.2015.52.

Galili, Tal. 2014. "Dendextend: Extending r's Dendrogram Functionality." *R Package Version 0.17* 5. https://doi.org/1.

Gonzaga, Sillas. 2022. "Lexicons for Portuguese Text Analysis." *The Journal of Open Source Software*, 5. https://cran.r-project.org/web/packages/lexiconPT/lexiconPT.pdf.

Hocking, Toby Dylan. 2019. "Comparing namedCapture with Other r Packages for Regular Expressions." *R Journal* 11. https://doi.org/10.32614/rj-2019-050.

Horodyski, Graziela Scalise, Diogo Lüders Fernandes, and José Manoel Gonçalves Gândara. 2015. "As Experiências Dos Turistas Em Estabelecimentos Comerciais de Souvenirs No Destino Curitiba-Brasil." *Investigaciones Turísticas*. https://doi.org/10.14198/inturi2015.10.08.

Hvitfeldt, Emil, and Julia Silge. 2021. *Supervised Machine Learning for Text Analysis in r. Supervised Machine Learning for Text Analysis in R*. https://doi.org/10.1201/9781003093459.

Johnson, Paul. 2020. "R Markdown: The Definitive Guide." *The American Statistician* 74. https://doi.org/10.1080/00031305.2020.1745577.

LoBuono, Raquel, Marlusa De Sevilha Gosling, Carlos Alberto Gonçalves, and Sandro Alves Medeiros. 2016. "Relações Entre Dimensões Da Experiência, Satisfação, Recomendação e Intenção de Retornar: A Percepção de Participantes de Evento Cultural Resumo." *PODIUM Sport, Leisure and Tourism Review* 5. https://doi.org/10.5585/podium.v5i2.158.

Lopes, Emerson, Larissa Freitas, Gabriel Gomes, Gerônimo Lemos, Luiz Hammes, and Ulisses Corrêa. 2022. "Exploring BERT for Aspect-Based Sentiment Analysis in Portuguese Language." In *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*. Vol. 35. https://doi.org/10.32473/flairs.v35i.130601.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons.

Oliveira, Douglas Nunes de, and Luiz Henrique de Campos Merschmann. 2021. "Joint Evaluation of Preprocessing Tasks with Classifiers for Sentiment Analysis in Brazilian Portuguese Language." *Multimedia Tools and Applications* 80. https://doi.org/10.1007/s11042-020-10323-8.

Pereira, Denilson Alves. 2021. "A Survey of Sentiment Analysis in the Portuguese Language." *Artificial Intelligence Review* 54. https://doi.org/10.1007/s10462-020-09870-1.

Pine, Joseph B, and James H Gilmore. 1998. "Welcome to the Experience Economy." *Harvard Business Review*, 97–105.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David. 2017. *broom: Convert Statistical Analysis Objects into Tidy Data Frames*. https://CRAN.R-project.org/package=broom.

RStudio. 2017. "Basic Regular Expressions in r." *Cheat Sheet*.

———. 2018. "Shiny - RStudio." *RStudio*.

Sarica, Serhad, and Jianxi Luo. 2021. "Stopwords in Technical Language Processing." *PLoS ONE* 16. https://doi.org/10.1371/journal.pone.0254937.

Sharma, Aman, and Rishi Rana. 2020. "Analysis and Visualization of Twitter Data Using r." In *PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing*. https://doi.org/10.1109/PDGC50313.2020.9315740.

Silge, Julia, and David Robinson. 2017. *Welcome to Text Mining with r. Development*.

Tan, Ah-Hwee. 1999. "Text Mining: The State of the Art and the Challenges." *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* 8. https://doi.org/10.1.1.38.7672.

Tereza Medeiros, José Mendes, Mariana Sousa. 2021. "A IMPORTÂNCIA DAS TECNOLOGIAS DE INFORMAÇÃO e COMUNICAÇÃO NO TURISMO SÉNIOR: UMA REVISÃO SISTEMÁTICA." *Turismo - Visão e Ação* 23. https://doi.org/10.14210/rtva.v23n3.p579-594.

Tung, Vincent Wing Sun, and J. R.Brent Ritchie. 2011. "Exploring the Essence of Memorable Tourism Experiences." *Annals of Tourism Research* 38. https://doi.org/10.1016/j.annals.2011.03.009.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

———. 2016. *tidyr: Easily Tidy Data with 'Spread()' and 'Gather()' Functions*. https://CRAN.R-project.org/package=tidyr.

———. 2019. "Rvest Package | r Documentation." *RDocumentation*.

Wickham, Hadley, and Romain Francois. 2016. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2016. *Bookdown: Authoring Books and Technical Documents with r Markdown. Bookdown: Authoring Books and Technical Documents with R Markdown*. https://doi.org/10.1201/9781315204963.

Yadav, Virendra, Srushti Chahande, and Ankita Wandre. 2018. "Review on Twitter Data Analysis Using r Language." *IJARCCE* 7. https://doi.org/10.17148/ijarcce.2018.71014.

Yihui Xie, Xianying Tan, Joe Cheng. 2018. "A Wrapper of the JavaScript Library DataTables." *The Journal of Open Source Software*, 5. https://rstudio.github.io/DT.