RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

ANÁLISE DE EMOÇÕES NO SUBREDDIT DO BBB

Gabriel Giolo Meleti Nunes
Graduando em Engenharia de software – Uni-FACEF
gabrielgiolo25@gmail.com

Pedro Henrique Kruszynski Nascimento
Graduando em Engenharia de software – Uni-FACEF
pedrokruszynski@gmail.com

Carlos Alberto Lucas
Mestre em Educação – Uni-FACEF
projetos@profcarloslucas.com.br

Resumo
A cada ano acontecem no Brasil vários *reality shows* e o mais famoso dentre eles é o BBB (Big Brother Brasil) que ganhou destaque ainda maior através da internet e redes sociais, aumentando a interação entre os usuários que até então conversavam sobre o programa nos encontros casuais. Com essa mudança onde múltiplas plataformas difundem a informação, buscamos analisar mais sobre as emoções acerca da opinião expressa pelos comentários de uma comunidade do *reality show* no Reddit (o conjunto de fóruns conta com mais de 230 milhões de usuários ativos). O advento das plataformas difusas e facilitação ao acesso digital contribuiu fortemente para o surgimento de diversos dados desorganizados que podem conter informações relevantes para tomadas de decisão, com base nisso buscamos oferecer uma solução facilitadora de obtenção de informação no subreddit do BBB. Para isso utilizamos o nosso conhecimento em programação associado à linguagem Python e R efetuando assim todo o desenvolvimento do projeto. Na vigésima segunda edição do BBB em específico, como resultado da análise, tivemos majoritariamente a emoção tristeza que é relacionada à sentimentos negativos.

Palavras-chave: Reddit. BBB. Análise de emoções. PLN.

Abstract
Every year happens in Brazil several reality shows and the most famous among them is the BBB (Big Brother Brazil) which gained even greater prominence through the internet and social networks, increasing the interaction between users who until then talked about the program in casual encounters, with this change where multiple platforms disseminate information, we seek to analyze more about the emotions and the opinion expressed by the comments of a reality show community on Reddit (the set of forums has more than 230 million active users). The advent of fuzzy platforms and facilitation of digital access has contributed strongly to the emergence of several disorganized data that may contain information relevant to decision making, based on this we seek to offer a facilitating solution for obtaining information in the BBB subreddit. For this we use our knowledge in programming associated with Python and R language, thus carrying out the entire development of the project. In this specific edition of BBB, as a result of the analysis, we mostly had the emotion of sadness,

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

which is related to negative feelings.

Keywords: *Reddit. BBB. Análise de emoções. PLN.*

1 Introdução

Na vida cotidiana do brasileiro já virou algo comum acompanhar os *reality shows.* Todo ano acontecem edições novas de programas que já são costumeiros, onde o foco está nas pessoas envolvidas, seus participantes e os sentimentos que elas causam no público. É o público que decide a permanência dos envolvidos no programa gerando uma espécie de relação ou laço entre membros do programa e espectadores com poder de voto, aflorando o desejo de expressar suas opiniões e acompanhar a reverberação do que foi expresso nas comunidades *online* com demais pessoas de interesses próximos.

Dentre várias opções que os usuários da internet podem utilizar para se socializarem nas redes, uma delas é o Reddit, um conjunto de fóruns onde pessoas podem criar grupos de interação com tópicos variados e conversar sobre algum assunto. Assim podemos ter um tópico de um nicho onde pessoas interessadas em participar possam criar um tema dentro do fórum que fica público para os seguidores e interessados no fórum, produzindo assim a interação entre os usuários da plataforma. O problema identificado foi quanto a ausência de uma solução que analisasse as emoções dos comentários referentes ao subreddit do BBB 22

Foi efetuada uma análise de dados no fórum de um dos *reality shows* mais famosos do Brasil. O BBB (Big Brother Brasil) que possui uma comunidade própria na plataforma para telespectadores e interessados conversarem, neste ambiente efetuamos uma pesquisa envolvendo os prováveis sentimentos que esses usuários expressavam no contexto de suas conversações, sobre os participantes e momentos marcantes do programa.

Para efetuar essa análise o projeto foi dividido em algumas etapas. A primeira delas é referente à mineração de dados, em que foi utilizada a linguagem Python para consumir a api pública do Reddit, registrar no banco NoSql e assim gerar um arquivo csv necessário para dar seguimento às demais partes, onde a próxima consiste na análise de dados. Essa etapa envolve a limpeza, alteração e demonstração dos dados.

Para executar a etapa de análise de dados utilizamos a linguagem R, que é uma linguagem que foi criada exclusivamente para manipular, analisar e visualizar dados, possui uma ótima base de usuários que reflete em uma comunidade ativa que pode auxiliar no desenvolvimento, manutenção e implementação de funcionalidades do software e, por isso, foi escolhida para efetuar essa etapa.

Através da linguagem R, por ser open *source*, a biblioteca syuzhet foi utilizada, sendo possível aplicar funções para efetuar a análise de sentimentos. A principal função utilizada é a get_nrc_sentiment, capaz de analisar um dataset de strings, retornando a presença dos seguintes sentimentos pré-definidos: anger (raiva), anticipation (antecipação), disgust (desgosto), fear (medo), joy (alegria), sadness (tristeza), surprise (surpresa), trust (confiança). Esses sentimentos são divididos em dois tipos de classificação: negative (negativo) e positive (positivo).

Após essas etapas, gráficos foram gerados para, enfim, efetuar uma análise organizada e mais assertiva ao estruturar os dados em plotagem virtual de

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

barras com coloração distinta entre cada emoção, deixando clara e amigável a informação ao usuário que solicita-la.

2 Referencial Teórico

Nesta seção são mostrados os referenciais utilizados como base para confecção deste trabalho de conclusão de curso.

2.1 Reddit

As redes sociais fazem parte do dia a dia da maioria dos cidadãos do mundo que possuem um dispositivo móvel, atuando como facilitador da comunicação entre os indivíduos, possibilitando o compartilhamento e manutenção de uma comunidade à distância, fazendo com que diariamente sejam geradas milhões de informações sobre variados assuntos (SAMPAIO; 2021)

Criado em Junho de 2005 em Medford, Massachusetts o Reddit consiste em um conjunto de fóruns denominados subreddits, esses por sua vez são divididos entre temas, reunindo indivíduos que compartilham gostos sobre um mesmo conteúdo, onde os usuários da plataforma são capazes de criar, compartilhar e comentar sobre o assunto do respectivo subreddit, atualmente contando com mais de 430 milhões de usuários (BELING; 2022).

O Reddit já existe há muito tempo e possui uma massa de usuários ativos diariamente sobre diversos temas, gerando conteúdo em abundância dentro dos subreddits.

Uma conta na plataforma pode ser criada por qualquer pessoa e ela é livre para expor seus posts na comunidade, algo que torna o Reddit uma fonte de dados que se destaca por simplesmente abdicar de filtragem de conteúdo já que todos podem criar e comentar o que desejarem (contudo os usuários não estão livres de sofrerem consequências relacionadas a suas atividades *online*, visto que há maneiras de denunciar conteúdos), transformando a plataforma em algo muito mais livre quando comparado com concorrentes similares, colaborando na veracidade dos dados ali contidos, tornando-os mais próximos da realidade.

Esse é o principal motivo da comunidade possuir grande expressividade, fazendo com que qualquer análise de dados nessas fontes seja um forte grupo para pesquisas dos mais variados temas, auxiliando em possíveis e eventuais tomadas de decisão no âmbito de negócios.

O Reddit disponibiliza uma API (Interface de programação de aplicações) onde uma aplicação externa consome os dados de outra aplicação, possibilitando extrair todas as informações necessárias sem muitas dificuldades. A API ainda é totalmente documentada, além de permitir que qualquer pessoa consiga gerar um token de acesso para começar a utilizar, com essa documentação e o livre acesso, é possível gerar análises desses dados.

Dois pesquisadores Judy Hanwen Shen e Frank Rudzicz da universidade de Toronto fizeram uma pesquisa onde analisaram uma base de dados consumida do Reddit para detectar ansiedade em posts, em sua conclusão foi possível identificar com eficácia de 91% de precisão ao detectar ansiedade em posts do redditSHEN, (Judy Hanwen; RUDZICZ, Frank., 2017).

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Constata-se que ao efetuar a análise de dados do Reddit é plausível obter resultados satisfatórios em áreas extremamente específicas ou de nicho que podem ser analisados para produção de conhecimento.

2.2 Mineração de texto

A mineração de texto deriva da mineração de dados que extrai dados em formato texto em linguagem natural, estabelecendo padrões de relacionamentos entre os dados ao se basear nos temas abordados e frequência de uso de determinadas palavras. Esse processo exige cooperação de diversas áreas como por exemplo aprendizado de máquina, recuperação de informação e a própria mineração de dados usados (PEZZINI; 2017).

A quantidade de dados desestruturados vem se tornando cada vez maior, trazendo à tona a possibilidade e algumas vezes necessidade de tratar esse oceano digital adequadamente, por isso a ciência por trás da análise e mineração de dados tem evoluído, polindo e aprimorando suas técnicas. A mineração de dados consiste na aplicação de diferentes técnicas para encontrar conexões contidas entre os dados que permanecem ocultas a menos que uma análise correlacional mais profunda seja feita, assim permitindo elaborar estratégias baseadas nos resultados obtidos e de certa forma prever eventuais causalidades.Essa funcionalidade é possível através da união de três pilares essenciais que formam a mineração de dados, sendo eles estatística, aprendizado de máquina e inteligência artificial (SANTOS; 2017).

Retomando um dos pilares da mineração de dados temos o avanço da inteligência artificial, essa por si só responsável por uma verdadeira revolução na psicologia, linguística e outras áreas impactadas pelo fenômeno. A partir dessa progressão em anos de história a IA se separou em subáreas e entre essas divisões podemos destacar o PLN (Processamento de Linguagem Natural)

2.3 Inteligência artificial com PLN

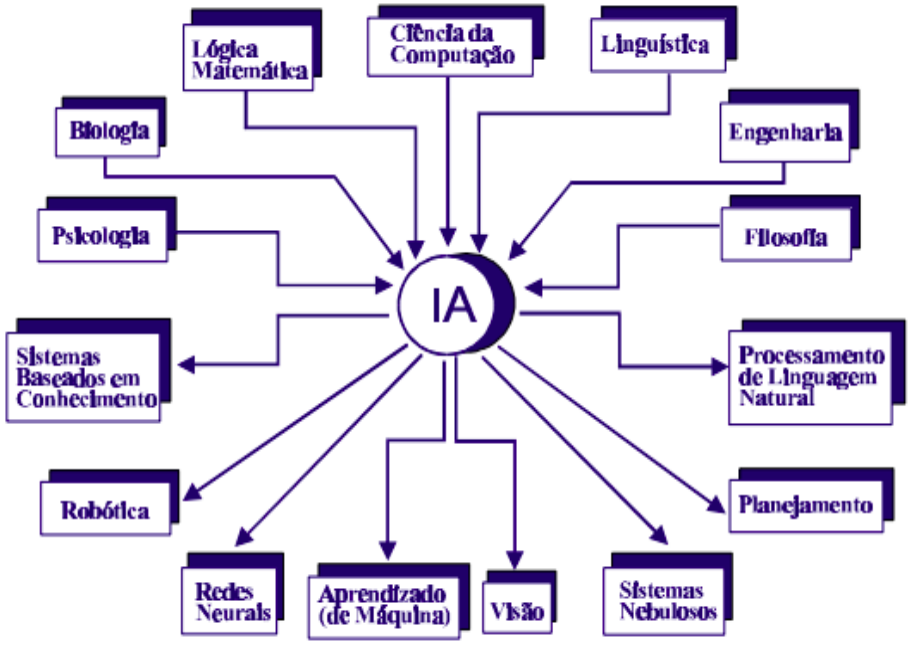
Para os pesquisadores relacionados à inteligência artificial, a mente do ser humano atua de maneira similar a um computador, sendo possível reproduzir aspectos da nossa inteligência por meio de programas, assim através da IA somos capazes de identificar objetos, falar e responder como um ser humano faria (TEIXEIRA; 2019).

A inteligência artificial é um ramo que possui diversas áreas de atuação, como psicologia, biologia, lógica matemática, linguística, engenharia, filosofia, entre outras, dentre elas vale destacar a utilização do processamento de linguagem natural ou PLN (GOMES; 2010).

O conjunto de ciências implementadas na inteligência artificial geram novos horizontes para a tecnologia, possibilitando o desenvolvimento da robótica, redes neurais, aprendizado de máquina além do Processamento de Linguagem Natural como representado na Figura 1.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 1: Campos de Estudo da Inteligência Artificial



Fonte:(MONARD, BARANAUKAS, 2000,p.2)

O Processamento de Linguagem Natural analisa a linguagem humana em suas diferentes expressões, para que programas de computadores sejam capazes de interpretá-las e assim realizarem tarefas levando em conta a subjetividade da linguagem; dentro dessa premissa a classificação de frases está presente na análise dos textos minerados a fim de diferenciar emoções básicas distintas e assim separá-las de acordo com o que foi identificado (SANTOS GUILHERME; 2017).

Essa questão pode ser explorada na confecção de representações gráficas dos textos estudados, assim como diferentes questões que impactam as pessoas inseridas naquele meio.

Os textos escritos e as falas humanas raramente são processadas corretamente, então analisar a linguagem humana é conseguir entender os conceitos e conseguir criar resultados. Para os humanos é fácil aprender uma nova linguagem, mas para computadores, implantar o PLN pode ser uma tarefa difícil (BARBOSA J; 2017).

Com isso vemos que tanto a fala quanto a escrita do ser humano é simples para um humano assimilar, todavia torna-se um desafio para o computador aprender adequadamente a ponto de utilizar e aplicar o PLN gerando significados, assim sendo deve ser implementada com cautela para alcançar os resultados adequados.

O processamento de linguagem natural não faz somente uma tratativa de uma sequência de caracteres, pois também considera sua estrutura hierárquica, podendo ser aproveitados em vários âmbitos de negócio como por exemplo, análise de sentimentos (Barbosa J; 2017).

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Por isso, a utilização do PLN se faz necessária para extrair e verificar os sentimentos de textos, levando em consideração uma base de dados como de uma comunidade, é possível supor os prováveis sentimentos dos usuários de uma rede social, resultando no surgimento de *insights* de negócio.

2.4 BBB (Big Brother Brasil)

O Big Brother Brasil já é um fenômeno no Brasil há algum tempo, o *reality show* televisionado pela rede Globo movimenta as redes sociais fazendo com que diferentes pessoas interajam ao redor de um ponto em comum. O ambiente cotidiano artificial gerado pelo programa que mescla ficção, realidade e jogo traz à tona questões relacionadas a meritocracia e autenticidade do indivíduo, fomentando discussões de ampla repercussão dentro e fora do grupo ligado ao entretenimento televisivo (ROBERTO; 2010).

Os sentimentos exprimidos pelos envolvidos em suas mensagens são compartilhados nos mais variados ambientes, alimentando uma onda que cresce gradativamente conforme acumula opiniões similares capazes de impactar no *reality*, mais comumente difundindo o programa.

Nas comunidades existentes no Reddit sobre BBB, podemos generalizar os usuários em duas categorias, aqueles que são ativos na comunidade, que se comunicam e discutem sobre os confinados do momento, gerando engajamento através de diversas postagens. A segunda categoria é referente aos interessados mais passivos os quais buscam apenas consumir as informações sem interação direta com os demais presentes nas discussões e assim o conjunto desses grupos colaboram para manter seu entretenimento favorito vivo nesses espaços bem como em seus cotidianos (CAMPANELLA; 2010).

A análise foi restringida para a versão de 2022 do programa a fim de obter resultados mais específicos e expressivos do BBB em sua apresentação mais recente, refletindo portanto em uma resposta emocional de um público recente que acompanha o programa.

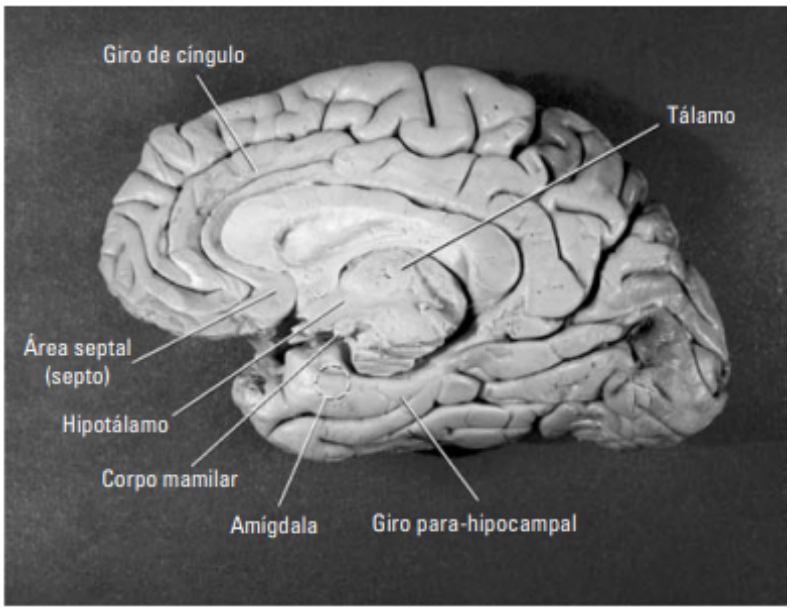
2.5 Análise de sentimentos

Com o avanço das neurociências a criação de hipóteses explicando sobre emoções a partir de estudos do sistema límbico têm aumentado, trazendo uma visão ampliada sobre a “natureza” das emoções, tema esse que intriga a religião, arte, filosofia e ciência desde os tempos antigos. Posteriormente a idéia de um único sistema das emoções foi severamente rebatida de acordo com evidências que áreas tidas como pertencentes a esse conjunto nem sempre participavam do processo, além disso a identificação de outros circuitos colabora ainda mais para a existência de um conjunto de sistemas de emoções, culminando em uma expansão que amplia os estudos do sistema límbico como destacado na Figura 2 (COLOMBO; 2007).

A figura 2 apresenta o sistema límbico sem a adição dos demais componentes citados que participam em algum grau nos sistemas das emoções, essa nova descoberta de integração proporciona melhor compreensão das respostas fisiológicas do organismo.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 2: Componentes do SL por Pepez



Fonte:(COLOMBO; 2007)

3 Análise do Problema e Implementação

A base selecionada passou por um processo de limpeza com o objetivo de estruturar e facilitar a análise de sentimentos, assumimos portanto o papel de Cientistas de Dados nesse processo e separamos através do python o que realmente seria relevante para a solução.

Os tópicos apresentados na sequência apresentarão as etapas em detalhes da análise de sentimentos a partir dos comentários presentes no subreddit do BBB 22 no espaço da plataforma do Reddit.

3.1 Processo de Data Mining

Os dados de comentários foram adquiridos no Reddit, onde estavam contidos no subreddit relacionado ao Big Brother Brasil edição 2022, através do consumo da lib praw capaz de coletar todos os dados do conjunto de fóruns daquele subreddit no período selecionado.

Essa atividade foi realizada pela linguagem Python por sua simplicidade, alta performance ao trabalhar com grandes contingentes de dados e comunidade ativa que auxilia nos programas desenvolvidos.

Através do MongoDB Client um ODB (object document mapper) que atua fazendo um mapeamento entre o mongoDB e a aplicação, foi selecionado devido a possibilidade de trabalhar eficientemente com grandes volumes de dados em Open Source, atingindo assim uma base de dados coesa.

O banco NoSQL foi criado utilizando a tecnologia Docker, que possibilita a conteinerização de imagens criadas por outros indivíduos na comunidade, com isso há a conexão com a aplicação Python sem que o computador com código em

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

execução necessite das dependências do MongoDB.

Figura 3: Classe principal em Python

```python
import praw
import pandas as pd
import database
from dotenv import dotenv_values

config = dotenv_values(".env")


You, 1 second ago | 1 author (You)
class RedditApi:

    def __init__(self):...

    def begin_the_mess(self):
        topics = []

        for submission in self.subreddit.top(time_filter="year"):...

        for submission in self.subreddit.search(query="DISCUSSÃO DIÁRIA - BBB22", time_filter="year"):
            topic_id = reddit_api.database.insert_topic({...
            submission.comments.replace_more(limit=0)
            for top_level_comment in submission.comments.list():...

        topics_frame = pd.DataFrame(topics, columns=["id", "title", "created", "comment"])
        topics_frame.to_csv("{name_csv}.csv".format(
            name_csv=config['NAME_CSV']
        ))


if __name__ == "__main__":
    reddit_api = RedditApi()
    reddit_api.begin_the_mess()

```

Fonte: os autores

Na figura 3, com os dados provenientes da biblioteca Praw o Python filtra apenas os dados relevantes e adiciona em uma tupla que será utilizada gerando o csv, por esse motivo foi utilizada a biblioteca Pandas.

O código fonte será disponibilizado no GitHub, seu link estará localizado no tópico de referências do artigo.

3.2 Importação em R e pré processamento

Após executar o script em Python o arquivo gerado é um arquivo csv (Comma Separated Values) com todos os comentários no período do BBB 22, contendo seu cabeçalho descrito no Quadro 1.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Quadro 1: Headers do csv gerado

Index	Função	Exemplo
id	Id do post no reddit	sjk1yw
title	Título do tópico	Nós éramos felizes e sabíamos
created	Data em que foi criado o tópico em timestamp	1643897226
comment	Comentário escrito pelo usuário	Que dupla… Nunca será superada.

Fonte:(Autores, 2022)

A fim de executar todo o processamento de dados foi utilizada a linguagem R além da utilização de uma IDE (Integrated Development Environment) chamada RStudio, necessária para esse desenvolvimento.

O RStudio é um software open source capaz de ser utilizado para executar códigos na linguagem R e em Python, um programa escolhido principalmente por experiência prévia dos autores, além de ser muito utilizado e sua integração com todas as funções da linguagem R que facilitam a codificação.

Após a importação do csv gerado pelo Python é iniciado o tratamento dos comentários que não são relevantes para o projeto, por esse motivo ocorre a remoção dos comentários em branco e que por quaisquer motivos adversos não possuem um comentário no csv,o que é relativamente simples de ser realizado em R, como é apresentado na Figura 4.

Figura 4: Comando no RStudio para importar

Fonte: os autores

3.3 Gerando gráficos com ggplot2

Em sequência foi produzida uma função que irá somente gerar gráficos, ela fará mais sentido logo à frente no artigo, devido à grande quantidade de gráficos que será gerada resultando em inúmeros códigos repetidos,que serão criados pela utilização da lib criada para o R chamada ggplot2.

Ggplot2 consiste numa lib para visualizar dados em código aberto escrito para a linguagem R, onde os dados são providos assim como os detalhes de

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

configuração referentes aos gráficos gerados e assim concluindo essa etapa da criação dos gráficos.

Dentre todas as funções possíveis no ggplot2 barplot foi a escolhida para gerar todos os gráficos, através dela foi passado o principal parâmetro, sendo esse o dataset, utilizado para gerar o gráfico.

Como todos os gráficos gerados tem um padrão, foi criada uma função capaz de receber o dataset, nome do gráfico e caminho que será salvo no computador, facilitando para o desenvolvedor utilizar a mesma função para todos os gráficos que forem gerados, como demonstrado na Figura 5.

Figura 5: Função no RStudio para gerar os gráficos

```
40
41
42
43 createBar ← function(dataset, name, path) {
44   png(sprintf('%sSentiment Analysis: %s.png', path, name))
45   barplot(
46     colSums(dataset),
47     las = 2,
48     col = rainbow(10),
49     ylab = 'amount',
50     main = sprintf('Sentiment Analysis: %s', name)
51   )
52   dev.off()
53 }
54
55
56
```

Fonte: os autores

3.4 Biblioteca syuzhet

Para a análise de sentimentos ser efetuada, foi utilizada a biblioteca chamada syuzhet, responsável por identificar o contexto e os sentimentos das frases em linguagem humana natural.

3.5 Gráficos por eliminações

O BBB 22 possui um padrão que a cada semana um participante é eliminado, nesse caso toda terça-feira sendo mais exato. Para filtrar do conjunto de dados gerado pelo Python importado nas etapas anteriores criamos um dataset que possuía o dia da eliminação que usaremos para filtrar qual foi o participante eliminado e razão para tal.

O motivo da eliminação no *reality* foi armazenado pois existem diferentes causas que podem acarretar na eliminação do participante sendo a retirada deste da casa através da votação. Um exemplo comum de eliminação, entretanto, já ocorreu em edições anteriores de uma participante sair do programa por vontade própria.

O *data frame* que foi gerado possui o nome de daysOfEliminations, como o programa prossegue por um período prolongado, com diferentes participantes e motivos foi diminuída a quantidade de informações do *data frame* como consta na Figura 6, justamente para não poluir a página com informações desnecessárias.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 6: Data frame que armazena todas as

```
daysWithEliminations <- data.frame(
  dayOfElimination=c(
    '2022-01-25',
    '2022-02-01',
    '2022-02-07',
    '2022-02-08',
    '2022-02-15',
    '2022-02-22'
  ),
  eliminated=c(
    'Luciano',
    'Rodrigo',
    'Maria',
    'Naiara',
    'Bárbara',
    'Brunna'
  ),
  reason=c(
    'Eliminated',
    'Eliminated',
    'Out by aggression',
    'Eliminated',
    'Eliminated',
    'Eliminated',
    'Give up'
  )
)
```

Fonte: os autores

Com as datas de todas as eliminações, é possível filtrar do dataset do Reddit, para isso algumas funções disponíveis pelo pacote dplyr foram utilizadas. Nesse pacote há uma biblioteca utilizada para manipular dados, na Figura 7 temos como instalar e carregar a lib para dentro do R.

Figura 7: Instalação e importação da lib dplyr

```
install.packages("dplyr")
library(dplyr)
```

Fonte: os autores

Com a biblioteca instalada existe agora a opção de usar *pipes* quando o *script* trabalha com *data frames*. Os *pipes* são representados pelos caracteres de %>% e com eles é possível utilizar funções em sequência facilitando assim a leitura do código.

A variável redditEliminations é responsável por armazenar o *data frame* filtrado, para definir os tópicos que estão dentro do alcance das eliminações ocorridas na edição BBB.

O Reddit armazena datas em timestamp, por esse motivo é necessário fazer uma conversão, utilizando a função as.POSIXct, que auxilia a manipular objetos e formatá-los como datas e horas válidas.

O subreddit do BBB possui um padrão que durante todo o programa, após ocorrer sua exibição é criado um tópico para os usuários da plataforma Reddit interagirem, esses tópicos possuem nome padronizado, sua característica é iniciar sempre pela DISCUSSÃO DIÁRIA.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Sabendo disso, é possível filtrar somente os comentários que forem relacionados aos tópicos de discussão diária, para isso foi utilizado uma função nativa do R denominada str_detect. A str_detect recebe dois parâmetros, sendo o primeiro a *string* que será utilizada para testar e o segundo parâmetro por padrão a função necessita que seja passado uma regex (DATACAMP, 2022).

O próprio R possui uma função para trabalhar com regex. Regex é um meio para que seja possível identificar padrões em strings, muito utilizado para validar strings e saber se ela é aceita ou não. Nesta etapa do código é necessário que o *script* consiga validar se o título do tópico possui a string “DISCUSSÃO DIÁRIA”.

E por último é utilizado a função group_split da lib dplyr. Essa função foi selecionada para que seja possível separar todos os tópicos em frações diferentes, assim o *script* tem a sua disposição todos os comentários divididos por tópicos. Todo o código está representado na Figura 8.

Figura 8: Gerando data frame somente das

```
redditEliminations ← reddit_dataset %>%
  filter(
    as.Date(as.POSIXct(created, origin='1970-01-01', tz='UTC')) %in%
    as.Date(daysWithEliminations$dayOfElimination)
  ) %>%
  filter(
    str_detect(
      title, regex('DISCUSSÃO DIÁRIA', ignore_case = TRUE)
    )
  ) %>%
  group_split(created)
```

Fonte: os autores

3.6 Loop para gerar os gráficos

Com o conjunto de dados já filtrado e pronto o *script* utiliza um loop em cada tópico, através dele é chamada a função createBar já apresentada anteriormente, enquanto o caminho para salvar os gráficos permanecem declarado na variável pathByEliminations como consta na Figura 9.

Figura 9: Caminho para salvar os gráficos

```
pathByEliminations ← '~/Documents/TCC_project/DataAnalysis/Plots/ByEliminations/'
```

Fonte: os autores

Na ordem dos parâmetros, é passado como primeiro parâmetro o *score* gerado pela função get_nrc_sentiment da lib syuzhet, informando que o *array* de *strings* que será utilizado é a coluna de comentários do *data frame* e por último usando o portugues como linguagem.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Em seguida o título presente no gráfico aparecerá e o caminho que será salvo o gráfico, tudo isso englobado por um *loop for in* no conjunto de dados de eliminações. Na Figura 10 temos o código do *loop*.

Figura 10: Loop para gerar gráficos

```
for (i in redditEliminations) {
  createBar(
    get_nrc_sentiment(i$comment, lang = "portuguese"),
    sprintf(
      '%s - %s',
      i$title[1],
      as.Date(as.POSIXct(i$created[1], origin='1970-01-01', tz='UTC'))
    ),
    pathByEliminations
  )
}
```

Fonte: os autores

3.7 Gerando gráficos por participante

Para gerar os gráficos por participantes o processo foi dividido em três etapas, onde para efetuar a primeira há a filtragem pelos comentários que mencionam o participante e para isso é utilizado um filtro usando uma regex no conjunto de dados do reddit como demonstrado na figura 11.

Figura 11: Filtrando conjunto de dados do Reddit por

```
jadeComments = reddit_dataset %>% filter(grepl('jade|picon', comment))
arthurComments = reddit_dataset %>% filter(grepl('arthur|aguiar', comment))
tiagoComments = reddit_dataset %>% filter(grepl('tiago|abravanel', comment))
```

Fonte: os autores

Novamente para fins demonstrativos, durante o artigo será apresentado nas figuras somente uma parcela dos participantes para maior objetividade do artigo.

A segunda etapa utiliza a função get_nrc_sentiment já explicada anteriormente, a diferença é que é usado o dataset que foi filtrado gerados na Figura 11 para descobrir quais os prováveis sentimentos ligados àqueles participantes, como

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 12: Score por participante

```
analyzeJade ← get_nrc_sentiment(jadeComments$comment, lang = "portuguese")
analyzeArthur ← get_nrc_sentiment(arthurComments$comment, lang = "portuguese")
analyzeTiago ← get_nrc_sentiment(tiagoComments$comment, lang = "portuguese")
```

Fonte: os autores

demonstrado na Figura 12.Prosseguindo com a leitura do *script* e com o score de cada participante, o código chamará a função createBar apresentada anteriormente. Como o *script* estará gerando gráficos por participantes, o caminho para salvar os gráficos também é diferente como consta na Figura 13.

Figura 13: Caminho para salvar gráficos por participante

```
pathByBrother ← '~/Documents/TCC_project/DataAnalysis/Plots/ByBrother/'
```

Fonte: os autores

Por fim, o código chama a função completa com os três parâmetros necessários: o conjunto de score do participante, o título do gráficos e por último o caminho que será salvo o resultado, a Figura 14 demonstra como fica a finalização do *script*.

Figura 14: Gerando os gráficos por participante

```
createBar(analyzeJade, 'Jade', pathByBrother)
createBar(analyzeArthur, 'Arthur', pathByBrother)
createBar(analyzeTiago, 'Tiago', pathByBrother)
```

Fonte:(Autores, 2022)

4 Resultados de Análise e Projeto

O BBB 22 possui 22 participantes, por esse motivo foram designados quatro participantes como foco de análise e comparação, o primeiro eliminado chamado Luciano e mais três participantes chamados Douglas, Paulo e Arthur que disputaram a votação final do programa.

Dentre os comentários dos participantes em questão é possível notar a predominância de sentimentos negativos os quais são ligeiramente menores quanto ao participante Luciano e significativamente menores no caso do participante Paulo.

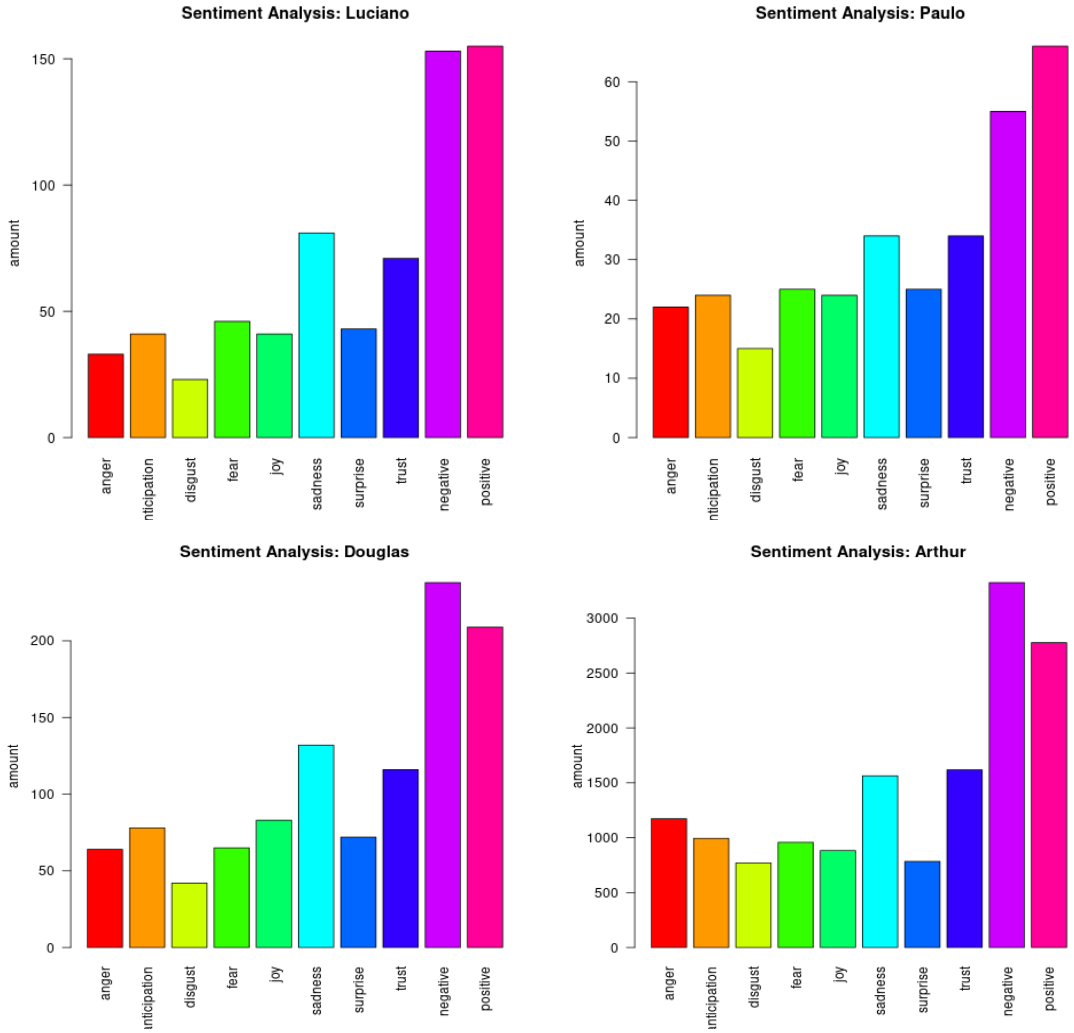
Esse suposto descontentamento deve levar em consideração que os

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

comentários que tiveram as emoções analisadas podem se referir a um ou mais participantes em um mesmo contexto, não possuindo portanto uma filtragem tão precisa nesse aspecto a ponto de discernir essa divisão entre o que foi redigido.

Arthur foi o vencedor do *reality* e em sua representação gráfica obteve sentimentos negativos em maior quantidade quando comparado com Luciano (Primeiro Eliminado), Douglas e Paulo o que pode ser indicativo do descontentamento geral do público ante sua vitória. Os comentários do subreddit demonstram graficamente que os usuários ativos e expressivos na plataforma eram mais simpáticos a Paulo quando comparados aos resultados de Douglas e Arthur, considerando que o trio avançou para a final esse fator é, no mínimo, intrigante.

Figura 15: Gráficos por participante



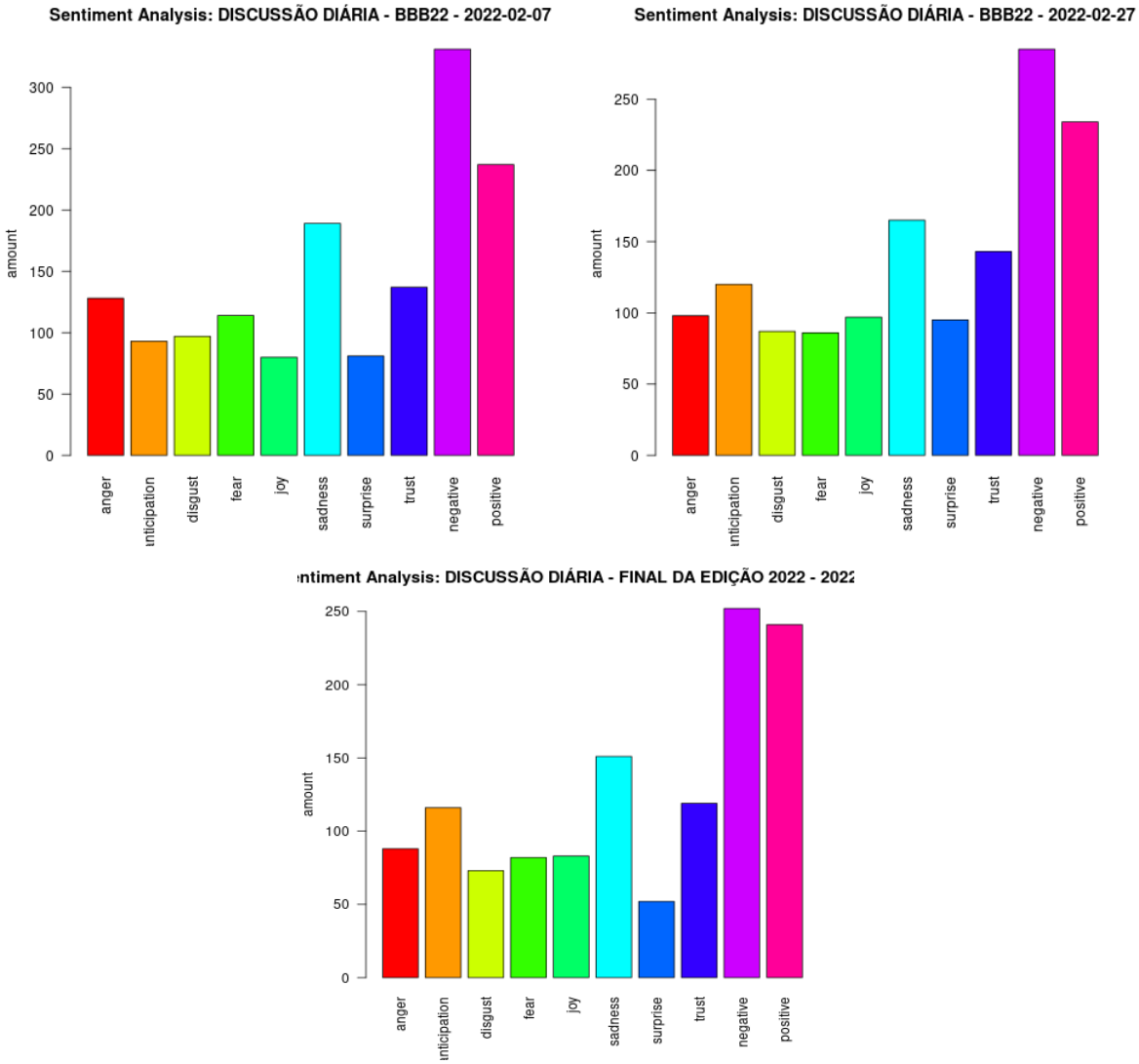
Fonte: os autores

Provavelmente dentro do grupo ativo do subreddit do Big Brother Brasil, Paulo era o finalista desejado para obter o prêmio, pois ao levar em conta o resultado que é apresentado em relação aos demais finalistas ele obteve mais comentários de teor emocional positivo mencionando seu nome.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Estes gráficos apresentados na Figura 16 foram selecionados para efetuar uma análise, compondo dias em que eventos de grande relevância ocorreram no BBB 22. O primeiro representa um episódio onde ocorreu agressão de uma participante com outra, resultando na eliminação da agressora.

Figura 16: Gráficos por eliminação



Fonte: os autores

O segundo foi um evento muito atípico do BBB que consistiu na desistência voluntária de um dos *brothers*. No *reality* é possível um participante cessar sua participação no programa através desse método que consiste em acionar o botão instalado na parede de um dos cômodos.

O terceiro e último, foi selecionado por realizar a análise da edição final do Big Brother Brasil, onde ocorreu a vitória do participante Arthur e consequente derrota dos demais remanescentes no dia vinte e sete de abril de 2022.

Com base na análise dos três gráficos selecionados é possível constatar que o sentimento predominante foi negativo, mais especificamente ligado a tristeza, onde os usuários do Subreddit compartilharam seu descontentamento entre si e

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

citaram os pontos que julgam dignos de atenção, isso pode indicar que apesar da audiência aquele público ativo no fórum do Reddit estiveram desapontados, descontentes com o andamento do programa visto que as variações entre os gráficos não foram tão expressivas quanto era esperado já que comparamos dois momentos tidos como negativos com um que deveria ser supostamente positivo, pois o público é responsável por decidir quem sai vitorioso através da votação.

4.1 Evoluções futuras

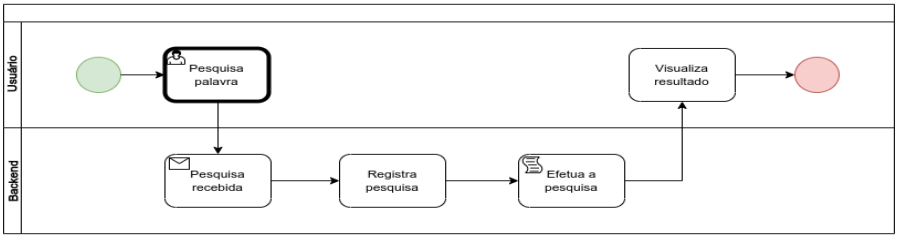
Nesta seção são comentadas as possíveis evoluções futuras para a continuação do projeto, contendo ideias que foram levantadas para acrescentar e melhorar o projeto e seus respectivos protótipos para uma melhor experiência ao usuário.

4.1.1. Implementar Front end

Um front end que foi prototipado para essa aplicação, consiste em uma interface web conectada com aplicação back end que faria a pesquisa dos comentários envolvendo uma palavra digitada pelo usuário na barra de pesquisa presente.

Seria uma SPA (Single Page Application), não havendo a necessidade de login, com a opção do usuário pesquisar uma palavra qualquer e; o back end receberia a requisição e gravaria um log, efetuando a pesquisa no banco e retornando para o front end transmitir a exibição dos comentários. Para demonstrar e auxiliar a visualização de todos os requisitos e prototipação referentes à implementação futura. A Figura 17 representa o BPMN do front end.

Figura 17: BPMN



Fonte: os autores

O documento de requisitos é um artefato de extrema importância para o desenvolvimento do front end da aplicação, esse documento auxilia na identificação das funcionalidades envolvidas no escopo da aplicação,através dele é possível identificar e visualizar os requisitos, como é ilustrado na Figura 18.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Figura 18: Documento de requisitos

ID: RF001	Nome Do Requisito: Pesquisar palavra
Descrição ➔	O sistema deverá consultar nos comentários a existência da palavra pesquisada
Categoria: Evidente	Prioridades: Essencial
Informações ➔	Palavra
Regra De Negócio ➔	O sistema deverá exigir uma palavra.

ID: RF002	Nome Do Requisito: Organizar comentários
Descrição ➔	O sistema organizará por paginação os comentários após a consulta
Categoria: Evidente	Prioridades: Desejável
Informações ➔	-
Regra De Negócio ➔	Obrigatoriamente o sistema terá encontrado comentários com a palavra pesquisada

ID: RNF001	Nome Do Requisito: Registrar pesquisa
Descrição ➔	O sistema produzirá um log de cada pesquisa feita pelo usuário
Categoria: Oculto	Prioridades: Importante
Informações ➔	Log
Regra De Negócio ➔	O sistema deverá armazenar o log no banco de dados

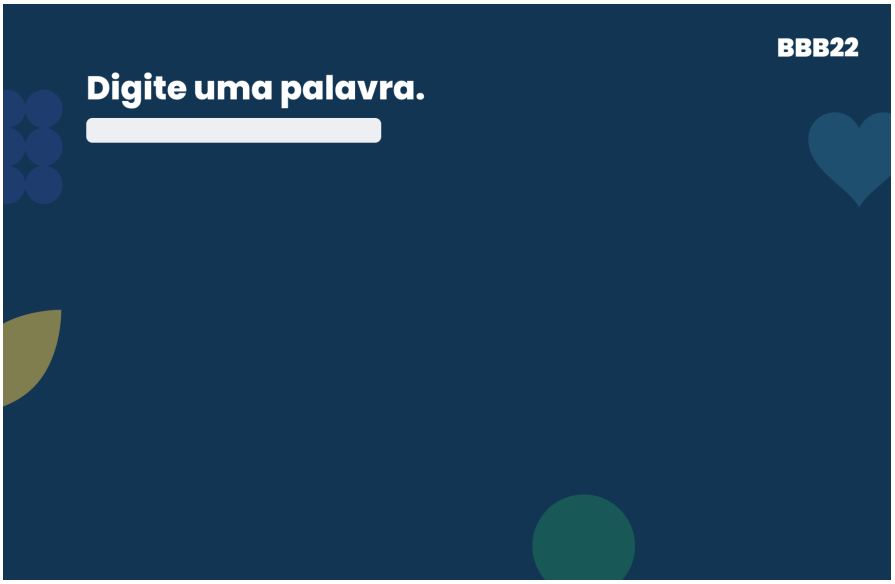
ID: RNF002	Nome Do Requisito: Ausência de login
Descrição ➔	Não será necessário um login para o usuário fazer uso da solução
Categoria: Oculto	Prioridades: Desejável
Informações ➔	-
Regra De Negócio ➔	-

Após identificação dos requisitos, é o momento para se efetuar a prototipação, utilizando uma plataforma especializada em protótipos de projetos e vetorização denominada Figma, selecionada por possuir uma interface intuitiva e comunidade atuante além de ser gratuito e sua utilização dispensar *download* já que é acessado por *browser*, sendo possível alterações serem feitas em equipe.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

A primeira tela do protótipo é intuitiva, tendo somente um campo de pesquisa para que o usuário saiba exatamente o que fazer a seguir, como indicado na Figura 19.

Figura 19: Primeira ação protótipo front end



Fonte: os autores

Após o usuário digitar a palavra o front end entrará em estado de loading, a tela adquire uma tonalidade cinza e no centro surge um ícone de carregamento como é possível identificar na Figura 20.

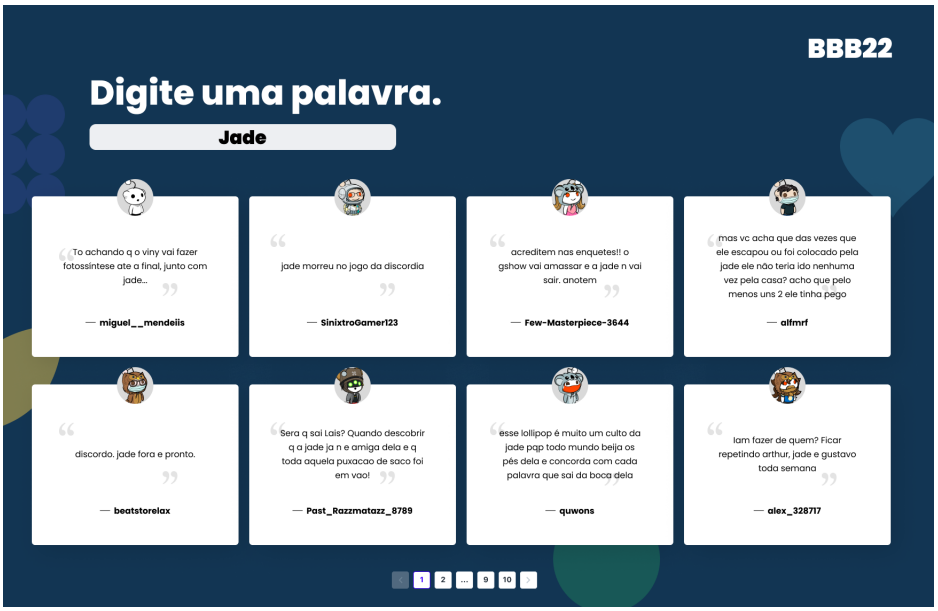
Figura 20: Loading protótipo front end

Fonte: os autores

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

Por fim os comentários são visíveis, carregados e disponibilizados para o usuário, quando a página terminar de carregar, caso haja mais de oito comentários com a palavra pesquisada, no final da página haverá uma paginação para que o usuário possa navegar entre os vários comentários organizadamente, como é visto na Figura 21.

Figura 21: comentários protótipo front end



Fonte: os autores

5 Conclusão

Certamente não era esperado um resultado tão negativo vindo da base observada, essa visão ligada ao subreddit conta apenas com usuários que comentaram no subreddit do BBB; considerando a popularidade do programa durante todos os anos de exibição, esse descontentamento por parte dos fãs pode apenas indicar o escopo de um grupo específico e não necessariamente apresentar toda a análise do público total.

Para fazer a mineração de dados utilizamos a linguagem Python e podemos dizer que ela se adequou bem para a implementação, foi alcançado um resultado muito satisfatório através de linhas de códigos bem simples e coesas. A fim de realizar a análise dos dados foi utilizado a linguagem R e também foi atingido um grau de resultados satisfatórios.

Analisando os resultados obtidos é possível notar que a maioria dos comentários tiveram como reação resultante emoções negativas o que não necessariamente está ligado a um fracasso por parte do programa, já que a ausência de uma filtragem mais detalhada quanto aos comentários em conjunto com os fatores de um grupo específico (usuários do Reddit frequentadores do subreddit do BBB e que são ativos nos comentários) e limitações da IA abrem margem para erros; entretanto esses indícios emocionais abrem portas para análises mais elaboradas do que é expressado pelas mensagens redigidas e a partir delas utilizar essas

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

informações em futuras tomadas de decisão.

Referências

ANDRADE, Vinicius Emanoel. oEmognizer: aplicação baseada em inteligência artificial para análise emocional de redes sociais. 2020.

BARBOSA, J. et al. Introdução ao processamento de linguagem natural usando python. III Escola Regional de Informatica do Piauí, v. 1, p. 336-360, 2017.

Bóbó, M., Campos, F., Stroele, V., David, J., & Braga, R. (2019, November). Identificação do perfil emocional do aluno através de análise de sentimento: Combatendo a evasão escolar. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 30, No. 1, p. 1431).

CAMPANELLA, Bruno. Perspectivas do Cotidiano: um estudo sobre os fãs do programa Big Brother Brasil. 2010. Tese de Doutorado. Doctorate thesis. Program for Post-Graduation in Communication and Culture through the Federal University of Rio de Janeiro.

CRAN.R. Introduction to the Syuzhet Package. In: JOCKERS, Matthew. Introduction. [S. l.], SD. Disponível em: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html. Acesso em: 19 set. 2022.

DATACAMP. In: As.POSIX*: Date-time Conversion Functions. [S. l.], SD. Disponível em: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/as.POSIX*. Acesso em: 18 set. 2022.

DATACAMP. In: Str_detect: Detect the presence or absence of a pattern in a string.. [S. l.], SD. Disponível em: https://www.rdocumentation.org/packages/stringr/versions/1.4.1/topics/str_detect/. Acesso em: 18 set. 2022.

DATACAMP. Modifiers: Control matching behaviour with modifier functions. [*S. l.*], sd. Disponível em: https://www.rdocumentation.org/packages/stringr/versions/1.4.1/topics/modifiers. Acesso em: 18 set. 2022.

DPLYR. A Grammar of Data Manipulation. In: Dplyr. 2.0.6. [*S. l.*], SD. Disponível em: https://dplyr.tidyverse.org/. Acesso em: 5 ago. 2022.

DPLYR. Split data frame by groups — group_split. In: Split data frame by groups. 1.0.10. [S. l.], SD. Disponível em: https://dplyr.tidyverse.org/reference/group_split.html. Acesso em: 18 set. 2022.

GGPLOT2. Create Elegant Data Visualisations Using the Grammar of Graphics. In: Ggplot2. 2.0.2. [*S. l.*], SD. Disponível em: https://ggplot2.tidyverse.org/. Acesso em: 5 ago. 2022.

RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 12 Nro. 2 2022
Uni-FACEF Centro Universitário Municipal de Franca

GOMES, D. dos S. Inteligência Artificial: conceitos e aplicações. Olhar Científico. v1, n. 2, p. 234-246, 2010.

MDN WEB DOCS. Expressões Regulares - JavaScript | MDN. In: Expressões Regulares. [S. l.], SD. Disponível em: https://developer.mozilla.org/pt-BR/docs/Web/JavaScript/Guide/Regular_Expressions/. Acesso em: 18 set. 2022.

PROGRAMMING HISTORIAN. Análise de sentimentos em R com 'syuzhet'. In: ISASI, Jennifer. Análise de sentimentos. [S. l.], SD. Disponível em: https://programminghistorian.org/pt/licoes/analise-sentimento-R-syuzhet#pacote-syuzhet. Acesso em: 24 set. 2022.

RSTUDIO. In: Publish your R and Python content with RStudio Connect. [*S. l.*], SD. Disponível em: https://www.rstudio.com/. Acesso em: 5 ago. 2022.

SAMPAIO, Amanda Gomes. Análise de sentimentos. 2021.

SHEN, Judy Hanwen; RUDZICZ, Frank. Detecting anxiety through reddit. In: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality. 2017. p. 58-65.

TEIXEIRA, João. O que é inteligência artificial. E-Galáxia, 2019.

WELLWO. Plutchik's Wheel: The 8 basic emotions | WellWo. In: ISASI, Jennifer. THE 8 BASIC EMOTIONS OF THE PLUTCHIK'S WHEEL. [S. l.], SD. Disponível em: https://wellwo.es/en/the-8-basic-emotions-of-the-plutchiks-wheel. Acesso em: 3 out. 2022.