PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

Análise de Sentimento dos usuários do Twitter em relação à plataforma de streaming Globoplay

Miguel José Gonçalves da Silva

TRABALHO DE CONCLUSÃO DE CURSO

CENTRO DE CIÊNCIAS SOCIAIS - CCS

DEPARTAMENTO DE ADMINISTRAÇÃO

Graduação em Administração de Empresas

Rio de Janeiro, Junho de 2022.

Miguel José Gonçalves da Silva

Análise de Sentimento dos usuários do Twitter em relação à plataforma de streaming Globoplay

Trabalho de Conclusão de Curso

Trabalho de Conclusão de Curso, apresentado ao programa de graduação em Administração da PUC-Rio como requisito parcial para a obtenção do título de graduação em Administração.

Orientador(a): Evandro da Silveira Goulart

Rio de Janeiro Junho de 2022.

Agradecimentos

Agradeço todos os familiares presente em minha vida, sobretudo meus pais, tios e primos.

Agradeço a todos os professores que tive em minha vida pelos ensinamentos.

Agradeço aos meus colegas de faculdade pelas amizades durante essa trajetória, especialmente aos amigos do Cafil.

Também agradeço, a Vice-Reitoria Comunitária por me acompanhar e me dar todo suporte necessário.

Resumo

Silva, Miguel. Análise de Sentimento dos usuários do Twitter em relação à plataforma de streaming Globoplay. Rio de Janeiro, 2022. 68p. Trabalho de Conclusão de Curso – Departamento de Administração. Pontifícia Universidade Católica do Rio de Janeiro.

Em um contexto que as empresas estão cada vez mais tomando decisões de negócio baseada em dados, fica mais importante a profissionalização e especialização nas áreas relacionadas a Ciência de Dados. O presente trabalho tem como objetivo apresentar uma abordagem prática onde coletamos e analisamos dados da rede social Twitter, dos usuários que interagiram com a marca Globoplay, a fim de realizar análises quantitativas e análises de sentimentos desses tweets, utilizando técnicas de Machine Learning e Processamento de Linguagem Natural.

Palavras-chave

Análise de Sentimentos; Twitter; Globoplay; Emoticons; Dashboard; Python; Mineração de Texto; Machine Learning; Processamento de Linguagem Natural

Abstract

Silva, Miguel. Sentiment Analysis of Twitter users in relation to the Globoplay streaming platform. Rio de Janeiro, 2022. 68p. Term paper. – Administration Department. Pontifical Catholic University of Rio de Janeiro.

In a context where companies are increasingly making business decisions based on data, professionalization and specialization in areas related to Data Science becomes more important. The present work aims to present a practical approach where we collect and analyze data from the social network Twitter, from users who interacted with a streaming platform Globoplay, to perform quantitative analysis and sentiment analysis of these tweets, Machine Learning and Natural Language Processing techniques were used.

Palavras-chave

Sentiment Analysis; Twitter; Globoplay; Emoticons; Dashboard; Python; Text Mining; Machine Learning; Natural Language Processing

Sumário

Lista de Figuras

Lista de Gráficos

Lista de Tabelas

Lista de abreviaturas e siglas

API	Application Programming Interface
NLP	Natural Language Processing
BBB	Big Brother Brasil
BBB22	Big Brother Brasil edição 2022
DM	Data Mining
Pos-Tagging	Part-of-speech Tagging
ML	Machine Learning
NB	Naive-Bayes
CSV	Comma-separated values
RTs	Retweets

1 – Introdução

As redes sociais têm se tornado cada vez mais presentes no dia a dia da população. A sua ampla utilização traz oportunidades para empresas se comunicarem e estabelecerem uma base de relacionamento com seus clientes, principalmente por meio de postagens nessas plataformas e campanhas publicitárias. Esse aumento trouxe uma popularidade das análises de dados das redes sociais nas últimas décadas e uma oportunidade para empresas explorarem esse novo petróleo. É um desafio para as empresas e gestores acompanharem todas as interações e identificarem se os usuários que se relacionam com a marca estão interagindo de uma forma mais positiva ou negativa, principalmente para marcas com muito engajamento e popularidade, por conta do grande volume de dados.

Nesse contexto, surgem profissionais que analisam os dados dessas interações, como por exemplo: Analistas de Marketing Digital, Analistas de Dados e Cientistas de Dados. Esses profissionais comumente têm a função de realizar rotineiros relatórios gerenciais ou estudos mais profundos sobre o comportamento dos usuários. Esses relatórios e estudos têm diversas finalidades essenciais para os gestores das marcas, como por exemplo: acompanhar a volumetria das interações dos usuários com a marca ou explorar comportamentos que os usuários têm com a marca.

Este estudo visa, através de conceitos de Mineração de Textos, coletar dados relacionados a uma marca na rede social Twitter, e analisar as interações orgânicas dos usuários. O estudo também utilizará técnicas de Processamento de Linguagem Natural (PLN) e Machine Learning (ML), para analisarmos o sentimento dessas interações. Ao final do processo podemos visar um material rico de insights para os gestores da marca, com análises quantitativas dos dados coletados.

1.1 – Objetivos

Esse estudo tem como objetivo final, analisar os tweets dos usuários que interagiram com o Globoplay.

1.2 – Objetivos intermediários

Além disso, serão considerados os seguintes objetivos intermediários:

- Realizar análises quantitativas dos dados coletados no Twitter utilizando princípios de Ciência de Dados, Mineração de Dados, Processamento de Linguagem Natural (NLP).
- Identificar quais eventos foram mais relevantes para o engajamento da marca.
- Desenvolver um classificador sentimentos com técnicas de aprendizado de máquina para descobrir se os tweets foram mais positivos ou negativos.
- Elaboração de um dashboard com os dados coletados.

1.3– Delimitação do estudo

Esse trabalho abordará unicamente as interações da rede social Twitter. Os dados coletados contemplarão o período entre os dias 23/04/2022 até o dia 30/04/2022, das interações da marca Globoplay. Vale ressaltar que a coleta de dados pode não contemplar o total do universo de interações, por limitações da API gratuita do Twitter.

1.4– Justificativa e relevância do estudo

Tendo em vista o crescimento e a facilidade ao acesso à diversas tecnologias e especializações profissionais na área de Ciência de Dados, Análise de Dados e Marketing Digital, esse trabalho visa uma contribuição apresentando uma abordagem prática na temática de análise de dados em redes sociais de uma marca. Esse trabalho também visa uma contribuição

no desenvolvimento de um classificador de sentimentos em textos curtos da língua portuguesa do Brasil.

1.5 – Contexto da empresa

Esse estudo utilizará os tweets sobre o Globoplay para ser nosso objeto de pesquisa. O Globoplay, foi fundado pelo Grupo Globo em 2015, é uma plataforma de streaming de vídeos e áudios sob demanda. O motivo pela escolha desta marca é sua grande popularidade e engajamento. Em maio de 2022, o perfil do Globoplay no Twitter possuía cerca de 624 mil seguidores.

Outro motivo da escolha do Gloploplay para análise de tweets, é o contexto da reta final do Big Brother Brasil 22 e Carnaval 2022. Esses eventos historicamente impulsionam o engajamento dos usuários com os produtos da Globo, sobretudo o Globoplay, portanto uma boa ocasião para fazer uma grande coleta de dados. Analisando os tweets da conta do Globoplay, o “@globoplay”, vimos que a conta faz muitas postagens sobre os novos conteúdos da plataforma, ou sobre conteúdos das marcas do Grupo Globo. Outra característica relevante da conta do Globloplay, é a utilização de uma linguagem mais informal e ‘descolada’, mas com foco em sempre engajar o público com a marca. Alguns exemplos de postagens que ilustram isso:

Figura 1: Exemplo de tweet feito pelo perfil do Globoplay

Fonte: Perfil do Globoplay no Twitter (2022)

Figura 2: Exemplos de tweets feito pelo perfil do Globoplay

Fonte: Perfil do Globoplay no Twitter (2022)

Podemos associar essa prática a diversos conceitos de marketing digital apresentados por Torres (2009). Veremos mais à frente no referencial teórico.

2 – Referencial Teórico

Este capítulo apresenta os embasamentos teóricos para entendimento e compreensão deste trabalho.

2.1 – Marketing Digital

Peçanha (2018), diz que o marketing digital é a divulgação de produtos ou marcas via mídias digitais. É uma das formas no qual a empresa tem comunicação com o consumidor de uma forma mais simplificada, personalizada e direta. O Marketing Digital é uma nova forma de divulgação, que cresce conforme a tecnologia avança, para não ficar atrasada ela sempre tem que estar em sintonia com ela. O motivo do de sua ampla utilização pelas empresas está na facilidade de entendimento e impacto do público por conta dos recursos audiovisuais. (Oliveira e Lucena, 2012). Segundo Kotler (2017) o marketing digital não pretende substituir o marketing tradicional. Ambos devem coexistir, com papéis permutáveis, ao longo do caminho do consumidor. Torres (2009) cita que o marketing digital deve ser composto por 7 ações estratégicas:

- Marketing de conteúdo: Planejar, criar e publicar conteúdo em seu site para torná-lo mais visível na internet e mais atraente para o consumidor. O marketing de conteúdo é uma abordagem que envolve criar, selecionar, distribuir e ampliar conteúdo que seja interessante, relevante e útil para um público claramente definido com o objetivo de gerar conversas sobre esse conteúdo (Kotler, 2017). Seu objetivo é atrair sutilmente seu público de forma a ganhar sua confiança e conquistar seguidores fiéis para sua marca.

- Marketing nas Mídias Sociais: Ações de marketing nas mídias sociais para obter um relacionamento entre a empresa e o consumidor.

- Marketing Viral: Uso do efeito viral para transmitir uma mensagem de marketing.

- E-mail marketing: Adaptação do marketing direto, onde a empresa visa o contato direto com o consumidor através de e-mail eletrônico.

- Publicidade online: Início através de banners publicados em sites, com os avanços tecnológicos estes ganharam interação, animação, som, vídeo e outros recursos.

- Pesquisa online: Pesquisa é a base da atividade de marketing e a internet permite pesquisas mais elaboradas e baratas do que as pesquisas convencionais.

- Monitoramento: Uma vantagem do marketing digital é que seus resultados podem ser medidos. O monitoramento é a ação estratégica que integra os resultados de todas as outras ações estratégicas, táticas e operacionais, permitindo verificar os resultados, agindo para corrigir falhas ou melhorias nas ações.

2.2 – Redes Sociais

De acordo com Recuero (2011), as redes sociais são definidas como um grupo de pessoas, compreendido como uma estrutura interligada. As redes possuem um papel importante como meio de propagação de informação, ideias e influências (KEMPE; KLEINBERG; TARDOS, 2005). A interação entre pessoas presentes nas redes sociais gera grande interesse, principalmente pelas possibilidades ligadas à internet.

As redes sociais online nos dias de hoje é uma representação das redes sociais em um ambiente computadorizado. As redes sociais online herdaram da internet as características de encurtar distâncias entre as pessoas e estar presente em qualquer lugar a qualquer momento. Esse

encurtamento traz a possibilidade para pessoas que estão distantes fisicamente umas das outras, consigam se comunicar com facilidade. Outra característica relevante é que as redes sociais possuem diversas possibilidades de acesso, seja através do celular, computador de mesa, notebook ou tablet. Em uma rede social online, os usuários publicam em seus perfis suas informações pessoais, preferências e características pessoais, e com essas informações os algoritmos das redes sociais estimulam os usuários a se relacionarem, através da similaridade entre perfis.

2.2.1 – Twitter

O Twitter foi criado em 2006 por Jack Dorsey, Evan Williams e Biz Stone, nos Estados Unidos da América (EUA). É uma rede social que oferece aos usuários a possibilidade de enviar atualizações pessoais em forma de texto (até 280 caracteres, anteriormente eram 140), vídeo, foto ou GIFs. E visualizar essas atualizações de sua lista de contatos. Uma das principais funcionalidades do Twitter são os “Trending Topics” que é um painel dos assuntos mais falados do momento no Mundo ou no seu país. Outra funcionalidade marcante é a hashtags, o famoso #. Quando um usuário faz uma postagem, ele pode colocar um # acompanhado de uma palavra ou um conjunto de palavras, por exemplo: #feriado ou #feriadocomafamilia. Quando um tweet possui uma hashtag, ele está sendo categorizado. Ao acessar a plataforma, é possível ver as hashtags mais populares no momento, ou o usuário pode pesquisar por uma hashtag com algum de seu interesse, por exemplo: #yorkshire. Os usuários também podem compartilhar tweets de outros usuários, essa ação é o retweet (RTs).

O Twitter permite que seus usuários criem um perfil público, para interagirem com outras pessoas. A construção desse perfil ocorre a partir das trocas sociais dentro de uma rede constituída por pessoas que se relacionam mediante interações, proporcionando assim o sentimento de comunidade. Isso acontece porque os usuários do Twitter selecionam a quem desejam seguir. Portanto, caso a qualquer momento, por qualquer

motivo, os usuários podem se desvincular de seus “seguidores”, ou daqueles que estes “seguem”, podendo também optar por se desligar da rede a qualquer momento.

O Twitter tornou-se uma ótima mídia social para a relação de consumidor-marca e surgiu como canal onde os consumidores recorrem quando desejam expressar frustração por algo em particular, ou querem interagir de alguma forma com a marca. Está cada vez mais comum as marcas dedicarem suas contas do Twitter ao atendimento ao cliente, respondendo e engajando os consumidores que utilizam a plataforma para esclarecimento de dúvidas e preocupações. Diversos estudos já foram realizados a fim de entender o comportamento do usuário na plataforma, como por exemplo, Silva (2016) em seu estudo concluiu que os alguns usuários do Twitter utilizam a plataforma para fins hedônicos, porém justificam seu uso como sendo para fins utilitários.

No Brasil e no mundo, há várias empresas que acompanham as atividades nas redes sociais, como o Twitter por exemplo. Segundo a Statista (Empresa alemã especializada em dados de mercado) em março de 2022, 19,05 milhões de brasileiros acessaram a rede social, quarta maior base do mundo. Isso mostra a grande relevância que essa rede social ainda possui no Brasil.

Em relação a dados sobre marcas no Twitter, um estudo feito por Salgado (2021) para OpinionBox (Empresa que oferece soluções de pesquisa de mercado online), mostra que 74% dos usuários seguem uma empresa ou marca. Além disso, segundo a OpinionBox, “38% já compraram algum produto ou contrataram algum serviço que conheceram por meio da rede social”. Outro dado sobre esse estudo é que 51% já solicitaram contato com empresas para tirar dúvidas ou fazer reclamações.

2.2.1.1 – Twitter API

A sigla API (Application Programming Interface), pode ser compreendida como uma interface de programação de aplicação. Ou seja, APIs são “tradutores” com a função de conectar sistemas, softwares e

aplicativos. A API do Twitter permite que diversos aplicativos se conectem a ele para os mais variados fins.

2.3 – Data-Driven

Data-Driven é um adjetivo que qualifica processos orientados por dados, ou seja, embasados na coleta e análise de informações. No mundo dos negócios, significa colocar os dados no centro da tomada de decisão e do planejamento estratégico, buscando fontes confiáveis ao invés de gerir a empresa por intuição.

Em uma entrevista ao site programaria.org (site sobre mulheres e tecnologia), em julho de 2021, Luana Hohmann, na época Gestora de Big Data e IA da Globo, citou seu posicionamento sobre o tema: “Quando falamos em uma cultura data-driven, nosso foco é promover mentalidade e práticas analíticas para toda a organização. Utilizamos dados para validar hipóteses, entender o impacto do desenvolvimento nos produtos, verificar a qualidade e consistência de dados, prever eventos (tal como as “rajadas” de acesso ao Globoplay no período de BBB), identificar melhorias nas soluções de machine learning, entre outros“ Hohmann (2021).

Isso mostra como é importante o amadurecimento do conhecimento das empresas em relação aos dados, tanto dados internos, quanto dados externos sobre a empresa como é o exemplo de dados do Twitter sobre a Globo.

2.4 – Dados Como Ativo Estratégico

Nas últimas décadas, vimos altos investimentos em infraestrutura de negócios que têm melhorado a capacidade de coletar dados em toda a empresa. A tendência é que todos os aspectos dos negócios sejam instrumentados para tal operação. A informação está amplamente disponível de forma pública, como tendências de mercado, notícias industriais e os movimentos dos concorrentes ou redes sociais. Essa ampla disponibilidade de dados levou ao aumento do interesse em métodos para extrair informações úteis e conhecimento a partir de dados.

Agora, com essa grande disponibilidade de dados, as empresas

estão focadas em explorá-los para obter vantagem competitiva. Os dados, e a capacidade de extrair conhecimento relevante a partir deles, devem ser considerados importantes ativos estratégicos. Muitas empresas consideram a análise de dados como pertencentes, principalmente, à obtenção de valores a partir de alguns dados existentes. E, portanto, existem evidências convincentes de que a tomada de decisões orientada em dados e tecnologias de big data melhoram substancialmente o desempenho nos negócios. (PROVOST, FOSTER; FAWCETT, TOM. 2016).

2.5 – Mineração de Dados

A Mineração de Dados, ou Data Mining (DM), é o processo de explorar dados relevantes em uma grande quantidade de dados à procura de padrões consistentes, ou informações relevantes para o negócio.

Em um artigo, a Microsoft define o processo de Data Mining em 6 etapas:

- Definindo o problema: Inclui a análise dos requisitos de negócio, a definição do escopo do problema, a definição das métricas usadas para avaliar o modelo e a definição de objetivos específicos para o projeto de mineração de dados.

- Preparando os dados: Consolidação e limpeza dos dados identificados na etapa anterior. A limpeza de dados não envolve apenas a remoção de dados incorretos ou interpolação de valores ausentes, mas também a localização de correlações ocultas nos dados, a identificação de fontes de dados mais precisas e a determinação de quais colunas são mais apropriadas para a análise. Antes de ir para as próximas etapas, é necessário identificar esses problemas e determinar como solucioná-los.

- Explorando os dados: É preciso o total entendimento dos dados para tomar decisões apropriadas ao criar os modelos de mineração. Diversos cálculos matemáticos e estatísticos

podem ser realizados para avaliar a representatividade dos dados.

- Criação de modelos: A exploração da etapa anterior auxiliará na definição e criação de modelos (Ver posteriormente na sessão de Inteligência Artificial). Aqui serão escolhidas as técnicas mais adequadas para modelagem, com base em algoritmos de mineração. O processamento de um modelo é geralmente chamado de treinamento. Treinamento refere-se ao processo de aplicação de um algoritmo matemático com a finalidade de extrair padrões. Os padrões que você localiza no processo de treinamento dependem da seleção de dados de treinamento.

- Avaliando e explorando o modelo: Aqui exploramos os modelos de mineração criados e testamos a eficiência deles. A avaliação vai checar se o modelo elaborado condiz com as expectativas da organização e do que foi definido anteriormente na fase inicial do processo. O resultado desta avaliação pode ser aceitável ou pode resultar na necessidade de revisão das fases anteriores, a fim de redefinir alguns passos.

- Implantação e atualização dos modelos: Implantar os modelos que tiveram o melhor desempenho em um ambiente de produção. Faz parte dessa etapa a apresentação dos resultados obtidos e possíveis alternativas de ação no processo de descoberta de conhecimento aplicado na organização.

O processo de mineração de dados se assemelha ao de Mineração de Textos, a diferença está na natureza dos dados em questão, o DM normalmente lida com dados estruturados enquanto a Mineração de Textos é aplicada sobre dados de texto, ou seja, dados não estruturados.

2.6 – Mineração de Textos

O texto está por toda parte. Em sistemas de informação temos muitos exemplos: E-mails, registros médicos, avaliação de loja de aplicativos, comentários em redes sociais, registros de reclamação do consumidor. Explorar essa vasta quantidade de dados requer a sua conversão em uma forma adequada. A Mineração de Textos aplica as mesmas funções analíticas da Mineração de Dados (GOMES, 2013), porém para dados textuais. O texto costuma ser chamado de dados “não estruturados”, ou seja, não pode ser utilizado por computadores para extração de conhecimento, uma vez que os mesmos a tratam apenas como uma sequência de caracteres. Portanto, é necessária a aplicação de diferentes métodos e algoritmos para dar estrutura aos dados textuais, assim facilitando a extração de conhecimento dos respectivos dados.

Jiakang Chang, et al (2018) propõe uma sequência de 5 etapas para extrair informações de forma eficiente de um documento, sentença ou texto.

Figura 3: Passos da mineração de texto

Fonte: Jiakang Chang, et al (2018)

Começamos com o processo de coleta, que consiste em reunir dados de uma ou várias fontes como websites, e-mails, comentários em fóruns, arquivos, etc. (CHANG, et al, 2018). A fase de pré-processamento é formada por três etapas principais: Limpeza do texto: São removidos os caracteres ou trechos de texto indesejados. Tokenização: em que o texto é dividido em sentenças ou entidades, dependendo do objetivo da aplicação. Seleção de Atributos faz a análise quantitativa do texto, podendo obter a frequência de uma determinada palavra no texto, por exemplo (CHANG, et al, 2018). A fase de armazenamento cria índices na base de dados e outros elementos que possibilitem o acesso mais rápido aos dados dentro de um banco de dados (CHANG, et al, 2018). Na fase de mineração e análise, ocorre a utilização de algoritmos e técnicas de exploração de dados

(Machine Learning e PLN, por exemplo. Falaremos desses termos posteriormente) para possivelmente revelar alguma informação de um texto. Por fim, temos a fase de apresentação, que busca visualizar os dados para então avaliá-los e posteriormente chegar em alguma conclusão. (CHANG, et al, 2018).

2.7 – Técnicas de Mineração de Texto

A seguir veremos as principais técnicas de mineração de texto

2.7.1 – Tokenização

Tokenização é o processo de tokenizar ou dividir um texto em uma lista de tokens. É uma das primeiras etapas na PNL.
Por exemplo: Eu gosto de dançar.
O algoritmo de Tokenização irá pegar cada palavra dessa frase e colocar em uma lista da seguinte forma: [‘Eu’, ‘gosto’, ‘de’, 'dançar','.']

2.7.2 – Bag of Words

A técnica Bag of Words é utilizada para converter os textos em vetores para facilitar o estudo da frequência de todas as palavras distintas presentes em um texto. Basicamente essa técnica cria uma lista que contém todas as palavras únicas que estão nos textos e utilizamos ela no NLP para poder identificar as palavras mais recorrentes.

Um exemplo:
Frase1: Eu gosto de comer doces.
Frase2: Futebol na praia, futebol na TV.
Frase3: Comer na praia sentado.

Aqui a lista de palavras únicas das três frases:

eu
gosto
de
comer
doces

futebol
na
TV
praia
sentado

Agora pegamos nossa lista de palavras únicas e olhamos para cada palavra das nossas 3 frases e contamos as ocorrências:

Tabela 1: Exemplo Bag of Words

	eu	gosto	de	comer	doces	futebol	na	TV	praia	sentado
Frase1	1	1	1	1	1	0	0	0	0	0
Frase2	0	0	0	0	0	2	2	1	1	0
Frase3	0	0	0	1	0	0	1	0	1	1
Soma ocorrencias	1	1	1	2	1	2	3	1	2	1

Fonte: Elaboração pelo autor

Por fim, listamos as palavras mais frequentes: na: 3, comer:2, futebol:2, praia 2, eu:1, gosto:1, de:1, doces:1, TV:1, sentado:1

2.7.3 – Stopwords

Para auxiliar na etapa de pré-processamento de dados textuais, uma técnica que normalmente é utilizada é a remoção de stopwords. Stopwords são palavras muito comuns que aparecem no texto e carregam pouco significado; servem apenas com uma função sintática, mas não indicam importância ao assunto. Alguns exemplos de stopwords: as, e, os, de, para, com.

2.7.4 – Pos-Tagging

O Pos-Tagging (part-of-speech tagging) é uma rotulação gramatical de uma palavra. Quando começamos o estudo da língua portuguesa aprendemos a identificar palavras como: Substantivos, adjetivos, preposições, artigos etc. Vejamos um exemplo:

O Miguel adora ver TV.

Agora o pos-tagging da frase:
O - artigo, Miguel - substantivo, adora - verbo, ver - verbo,
TV - substantivo, . - pontuação

Fazendo essa rotulação das palavras, podemos fazer filtros em frases.

2.7.5 – Stemming

A técnica de Stemming visa reduzir as formas flexionadas das palavras, como por exemplo palavras no plural ou no gerúndio. Porém essa redução pode tirar o significado original da palavra. Vamos ver um exemplo de Stemming com as palavras “Amigo” e “Amigas”:

Amigo->Amig
Amigas->Amig

Como podemos perceber “Amig” é uma palavra que não existe no dicionário da língua portuguesa. Entra então o conceito de Lemmatization.

2.7.6 – Lemmatization

Na lemmatization, também queremos reduzir a palavra à sua raiz, retirando todas as inflexões e chegando ao lemma. Entretanto, essa redução sempre resultará em uma palavra que realmente existe na gramática. No mesmo exemplo das palavras “Amigo” e “Amigas”:

Amigo->Amigo
Amigas->Amigo

A vantagem da lemmatization é que ela garante que vai gerar palavras gramaticalmente corretas e com maior precisão já que leva a classe gramatical em consideração.

2.7.7 – Expressões Regulares na limpeza de texto

As Expressões Regulares no contexto de manipulação de textos, podem ser utilizadas para selecionar caracteres de uma frase. Veremos alguns exemplos úteis na limpeza dos textos de tweets:

-Remoção de Acentos
-Remoção de Pontuação
-Remoção de Links
-Remoção de Hashtags
-Remoção de caracteres repetidos seguidos

No anexo III podemos conferir os códigos em Python desses processos.

2.8 – Inteligência Artificial

Inteligência Artificial (IA) pode ser definida como um sistema computacional que pode simular, usando matemática e lógica, o raciocínio que os humanos empregam para aprender com novas informações, aprendendo com exemplos e experiências, podendo reconhecer objetos, compreender e responder, tomar decisões e resolver os mais diversos problemas.

Podemos considerar quatro subconjuntos principais de IA.

1) Raciocínio de Máquina ou Machine Reasoning

2) Processamento de Linguagem Natural (PLN)

3) Planejamento Automatizado ou Planning

4) Aprendizado de Máquina ou Machine Learning (ML)

2.8.1 – Processamento de Linguagem Natural

Processamento de Linguagem Natural (PNL) ou Natural Language Processing (NLP) é a capacidade de treinar modelos computacionais para entender tanto o texto escrito quanto a fala humana. Portanto, as técnicas de PNL são utilizadas para capturar o significado de textos não

estruturados de documentos ou da comunicação do usuário, seja ela escrita ou em áudio. Portanto, a PNL é a principal forma de os sistemas interpretarem o texto e a linguagem falada.

2.8.2 – Análise de Sentimentos

Classificada como uma vertente da Inteligência Artificial, a Análise de Sentimentos (AS) é o estudo computacional de opiniões, sentimentos e emoções expressos em texto. É um assunto de grande interesse e evolução na pesquisa acadêmica e está em constante desenvolvimento por possuir muitas aplicações práticas. Sobretudo no entendimento da opinião das pessoas sobre um determinado assunto, produto, marca etc. analisando grandes quantidades de informações. Usualmente se refere a um problema de classificação em que o foco é encontrar e classificar a polaridade das sentenças (ou textos) em negativo, positivo ou neutro (JOSE; CHOORALIL, 2016).

No ambiente das redes sociais, a Análise de Sentimentos é, comumente, utilizada para verificar a polaridade de opiniões e pensamentos expressos pelos usuários, por exemplo, a partir de um tweet, queremos atribuir o sentimento positivo, negativo ou neutro.

2.8.3 – Machine Learning

O Machine Learning (ML) é um modelo preditivo que consiste na construção de sistemas que aprendem com a partir de uma base. É capaz de analisar uma grande quantidade de dados, além de usar uma variedade de algoritmos para encontrar padrões em uma base de dados. Com base nesses padrões, consegue fazer determinações ou predições.

O Machine Learning ou Aprendizado de Máquina, está dividida em quatro categorias:

1) Aprendizagem Supervisionada (Supervised Learning)

2) Aprendizagem Não Supervisionada (Unsupervised Learning)

3) Aprendizagem por Reforço (Reinforcement Learning)

4) Aprendizagem Profunda (Deep Learning)

Aprendizagem Supervisionada utiliza um conjunto de dados estruturados e previamente processados para descobrir padrões e fazer classificações ou regressões. Essa aprendizagem se destina a encontrar padrões em dados que podem ser aplicados em um processo analítico. É dito Aprendizado “Supervisionado”, pois os dados de treinamento do modelo são conhecidos, isto é, para um determinado conjunto de dados de entrada (input) há um conjunto de dados de saída (output) bem definidos.

Aprendizagem Não Supervisionada é mais adequada quando o problema requer uma enorme quantidade de dados não estruturados, isto é, que não são rotulados.

Aprendizagem por Reforço é um modelo de aprendizagem comportamental. O algoritmo recebe feedback da análise dos dados para que o usuário seja guiado para o melhor resultado. Esse tipo de aprendizagem aprende através de tentativa e erro.

Deep Learning incorpora a ideia das Redes Neurais Artificiais em camadas sucessivas para aprender com os dados de maneira iterativa. Nesse aprendizado, o algoritmo busca emular como o cérebro humano funciona permitindo lidar com vários tipos de dados e lidar com abstrações ou problemas mal definidos, Hurwitz e Kirsch (2018). Uma rede neural artificial possui uma primeira camada chamada de entrada (input layer), uma ou mais camadas intermediárias chamadas de ocultas (hidden layer) e uma última camada de saída (output layer). Cada valor de entrada e de saída está associado a um neurônio. O número de neurônios das camadas ocultas varia de acordo com a configuração escolhido no modelo. Durante esse processo iterativo, os dados de entrada são repassados e modificados tanto na camada oculta como na camada de saída, alterando-se os pesos que são aplicados em cada neurônio utilizando-se funções de ativação.

Além disso, Deep Learning pode gerar aprendizado considerando

uma combinação de algoritmos de Aprendizagem Supervisionada com algoritmos de Aprendizagem Não Supervisionada.

No contexto de PLN, o ML nos auxiliará a fazer predições de sentimentos nos tweets. No presente trabalho, por conta da natureza dos dados de texto já estruturados previamente, a melhor aplicação do ML para esse estudo é o Aprendizado Supervisionado. Para isto, utilizaremos um dos algoritmos de classificação mais famosos, o Naive Bayes.

2.8.3.1 – Naive Bayes

O classificador Naive Bayes é um algoritmo que se baseia nas descobertas de Thomas Bayes. O termo naive vem de ingênuo, por assumir que os termos de uma instância são condicionalmente independentes entre si, não possuindo influência um sobre o outro. Uma vantagem desse algoritmo é seu desempenho consideravelmente bom com classes múltiplas, principalmente em estudos de PLN.

2.9 – Métricas de performance

Ao construir modelos de Machine Learning, uma etapa muito importante é a validação do modelo. Não adianta construir um modelo sem a validação de sua performance. Atualmente existem diversas métricas de validação, elas apresentam o desempenho do modelo treinado em dados desconhecidos. Analisar essas métricas é tão importante quanto a preparação dos dados e o ajuste do modelo. A seguir vamos conhecer algumas métricas utilizadas no estudo.

2.9.1 – Acurácia

A acurácia é a métrica mais simples, ela representa o número de previsões corretas do modelo. Ele dá uma visão geral do quanto o modelo está identificando as classes corretamente. Por exemplo, temos essas 3 frases:

-Eu gosto de balas
-Odeio futebol
-Essa comida é maravilhosa

Jogando essas frases em um modelo de ML para classificar sentimentos, vamos supor que ele nos dê os seguintes resultados:

-Eu gosto de balas - Negativo
-Odeio futebol - Negativo
-Essa comida é maravilhosa - Positivo

Nesse caso o modelo acertou 2 dos 3 sentimentos, ou seja, o modelo teve uma acurácia de 66,66%.

2.9.2 – Validação Cruzada

O Validação cruzada (cross-validation) é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Na etapa de avaliação de um modelo de ML separamos nosso conjunto de dados em treino e teste. Essa separação é feita de forma aleatória. E por conta dessa aleatoriedade na separação dos dados, a acurácia do modelo pode ser impactada. Existem vários métodos de Cross-Validation, neste trabalho foi utilizado o K-fold.

2.10 – Dashboard

Few (2006), define Dashboard como uma apresentação visual das informações mais importantes para atingir um ou mais objetivos de negócio. Esses dados são representados em uma única tela, para serem consumidos de forma rápida e prática. Os dashboards são considerados sistemas de apoio à decisão, uma vez que sua interface gráfica traz indicadores de performance de um negócio, possibilitando a tarefa de gestão e servindo de apoio à tomada de decisão.

Portanto, na prática, fica mais fácil acompanhar métricas e indicadores, o que permite melhorar as decisões feitas dentro da empresa.

2.11 – Trabalhos Relacionados

Na literatura podemos encontrar diversos estudos utilizando o Mineração de Textos no Twitter como objeto de estudo. Thoring (2011), fez uma análise quantitativa dos dados baseado em: frequência, horário e dia

da semana dos tweets. GONÇALVES (2013), realizou uma análise mais voltada nos emoticons contidos nos textos dos tweets. Silva Filho et al. (2020) fez um estudo de caso múltiplo com abordagens qualitativa e quantitativa de análises, analisando tweets de 3 marcas. Com a utilização de análise de sentimentos e ranking de termos mais utilizados. Estudo similar de FILHO (2014), onde ele realizou uma coleta de tweets relacionada ao evento da Copa do Mundo de Futebol de 2014 e realizou uma análise de sentimento dos tweets coletados.

3 – Metodologia

O princípio da metodologia está no conceito do processo de Mineração de Textos proposto por Jiakang Chang, et al (2018), juntamente com o conceito apresentado por PROVOST (2016): "A partir de uma grande massa de dados, a tecnologia da informação pode ser usada para encontrar atributos descritivos informativos de entidades de interesse".

3.1 – Preparação do ambiente para desenvolvimento do sistema

O passo zero do projeto foi a configuração do computador pessoal do autor e a obtenção do token de acesso a API do Twitter. Para a obtenção do token é necessário criar uma conta de desenvolvedor no site de Desenvolvedor do Twitter. O projeto foi feito utilizando a linguagem de programação Python. O passo seguinte, foi o desenvolvimento de um script em Python para se conectar com a API do Twitter. Para acessar a API foi utilizada a biblioteca Tweepy. Após a validação da conexão com a API, seguimos para a coleta dos dados.

3.2 – Coleta de Dados

A coleta dos tweets foi a primeira parte da execução deste trabalho. O script recebe como parâmetro o nome a ser pesquisado (no nosso caso a palavra “globoplay”), e retorna os tweets que contêm esse nome. Este processo de coleta foi iniciado no dia 23/04/2022 as 00:00:00 e encerrado no dia 30/04/2022 as 23:59:59. Lembrando que os períodos foram escolhidos para contemplar o período da etapa final do Big Brother Brasil 22 e o início dos desfiles do carnaval do Rio de Janeiro e São Paulo. Esses tweets foram armazenados em arquivos com valores separados por vírgula (CSV). Este arquivo gerado, será a principal base de dados para as próximas etapas, chamaremos esse arquivo de “datasetgplay”. O código fonte utilizado nesse processo está no ANEXO I.

3.3 – O Datasetgplay

A seguir temos os campos obtidos na coleta de dados, com sua respectiva explicação:

id_user - Número de identificação única do usuário
user - Nome do usuário na plataforma
number_of_followers - Número de Seguidores na data de extração
number_of_followings - Número de pessoas que esse usuário segue na data de extração
id tweet - Número de identificação única do tweet
tweet_created_at - Data do tweet
text - Texto do Tweet
mentions - Lista de usuários mencionados no tweet
retweeted - Se o tweet é um Retweet
location - Localização do Usuário

Para se ter uma ideia prática do dataset, no Anexo II podemos encontrar uma amostra do arquivo.

3.4 – Exploração e análise dos dados

O objetivo agora é, a partir do nosso dataset, explorar quantitativamente os dados. Para isso vamos analisar:

-Volumetria dos dados
-Evolutivos diários
-Rankings dos principais campos

3.5 – Análise dos Emoticons

Com base nos estudos de Gonçalves (2013), queremos realizar análises dos tweets com emoticons. Como por exemplo, ranking dos emoticons mais utilizados dos tweets que contêm a palavra Globoplay e volumetria do tweets com emoticons. Assim podemos analisar quais emoticons foram mais ligados a marca.

3.6 – Implementação Análise de Sentimentos

A ideia é construir um classificador de sentimento dos textos coletados, fazendo a atribuição dos sentimentos positivo, negativo ou neutro para cada tweet.

O início do processo foi uma classificação manual de alguns tweets para o algoritmo utilizar como validação e teste. Foram classificados 1317 tweets, sendo 548 como positivos, 229 como neutro e 540 como negativo. Essa etapa também tem grande importância para posteriormente medir o desempenho do algoritmo.

O próximo passo segundo o modelo proposto por Jiakang Chang, é a etapa de pré-processamento. Aqui trabalhamos na limpeza e normalização do texto do tweet, isso visa facilitar o algoritmo na predição do sentimento. Diversos processos de limpeza de url foram aplicados.

Agora temos um momento crítico do processo. Qual método utilizar? Qual algoritmo utilizar? Diversas variações foram testadas, abaixo segue uma tabela com as metodologias testadas com suas respectivas acurácias. Lembrando que as acurácias foram calculadas baseadas nos 1317 tweets classificados manualmente. No anexo IV estão os links com as respectivas fontes de cada modelo:

Tabela 2: Listagem dos modelos testados

Modelo	Metodologia Modelo	Acurácia
Modelo 1	Biblioteca do Dicionário Léxico Leia	28,66%
Modelo 2	Tradução dos tweets para o ingles e utilização do classificador da Biblioteca TextBlob	43,83%
Modelo 3	Tradução dos tweets para o ingles e utilização do classificador da Biblioteca Vader	20,77%
Modelo 4	Dicionário Léxico Reli-Lex com classificador de NaiveBayes	59,52%
Modelo 5	Dicionário Léxico Sentilex com classificador de Score	31,15%
Modelo 6	Dicionário Léxico Sentilex com classificador de NaiveBayes	46,52%
Modelo 7	Dicionário Léxico Optxlex com classificador de NaiveBayes	60,50%
Modelo 8 *	Utilização de outra base de tweets pré-classificada com classificador de NaiveBayes	25,89%
Modelo 9	Utilização de 70% da própria base pré-classificada com classificador de NaiveBayes	85,20%

* A acurácia utilizada foi a de validação cruzada.

Fonte: Elaborada pelo autor

Ao final dos testes foi escolhido a metodologia com maior acurácia, o Modelo 9. No Anexo V está o pseudo-código do modelo.

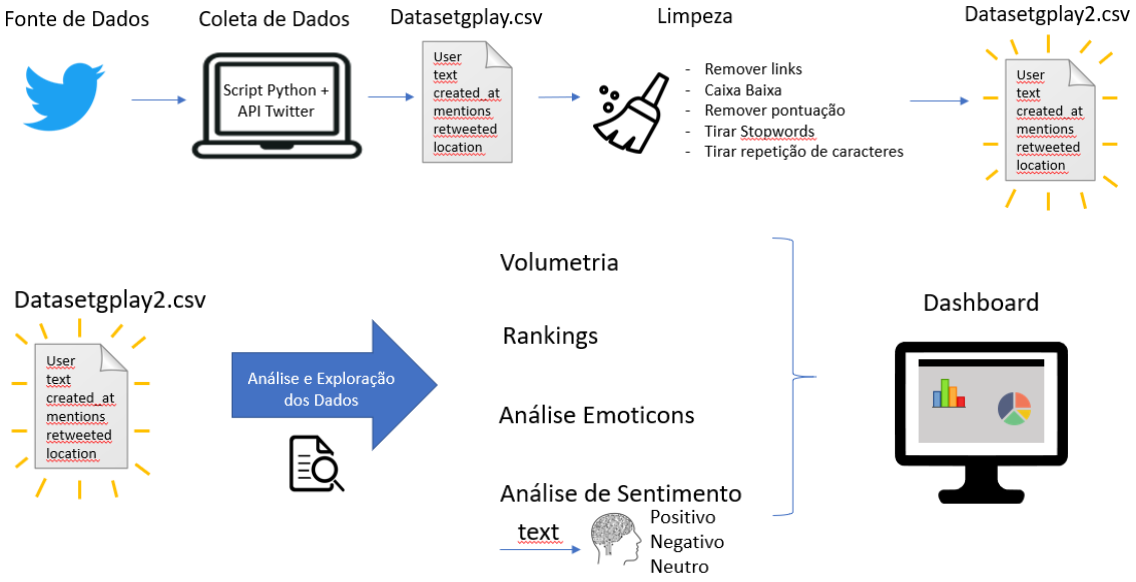
3.7 – Construção do Dashboard

Ao final do processo queremos juntar todas as informações apresentadas e colocar em um dashboard para a melhor apresentação dos dados. A ferramenta escolhida foi o Data Studio desenvolvido pela Google. O Data Studio é uma ferramenta gratuita online de visualização de dados. Com ele podemos transformar dados brutos em dashboards personalizáveis. Outra vantagem é alta compatibilidade com diferentes fontes de dados. Um software concorrente do DataStudio bastante presente no mercado, é o PowerBI desenvolvido pela Microsoft.

3.8 – Resumo do processo Metodológico

Na figura a seguir vemos visualmente como foi realizado conceitualmente o processo metodológico do projeto.

Figura 4: Resumo do processo Metodológico



Fonte: Elaborado pelo autor

4 – Análise e Discussão dos Resultados

4.1 – Volumetria

Durante o período dos dias 23/04/2022 até o dia 30/04/2022 temos as seguintes informações:

- No total, 66.566 tweets foram coletados.
- Desse total, 54% (35.728) deles foram retweets.
- 58% dos tweets totais tiveram o preenchimento da informação de localização.
- 84% dos tweets totais possuem alguma menção.
- A média de menções dos tweets que possuem ao menos uma menção é de 1,7 Menções/Tweet.
- Do total dos tweets, foram identificados 38.492 usuários únicos.
- Portanto, temos em média 1,73 Tweets/Usuário.
- Ao considerarmos os tweets únicos, temos o total de 33.583 tweets, ou seja, 50,3% dos tweets totais são únicos.

Abaixo temos o sumário das principais informações sobre a volumetria dos dados coletados.

Tabela 3: Sumários das principais métricas coletadas

Resumo	
Nome Métrica	Total
Total de Tweets	66.566
Total de Retweets	35.728
Tweets Únicos	33.583
Usuários Únicos	38.492
Tweet/Usuário	1,73
Tweets com Geolocalização	38.901
Tweets com Menções	55.705
Média de Menções/Tweet	1,7

Fonte: Elaborado pelo autor

Para um melhor entendimento da formação da volumetria, vamos categorizar os tweets coletados em temas. Por exemplo, um tweet com esse texto: “Adorei o BBB no Globoplay”. Esse tweet será categorizado como BBB. A regra de categorização são palavras específicas no texto do tweet. A seguir temos os eventos identificados com sua respectiva regra de atribuição:

Tabela 4: Temas com sua respectiva explicação e regras

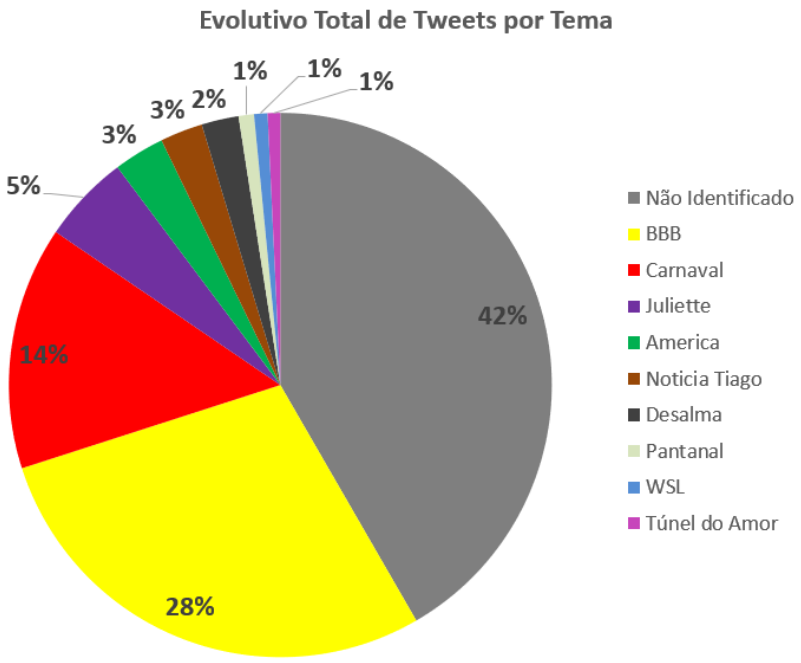
Temas	Explicação	Regras
BBB22	Reality Show	bbb, big brother, big boss, jade, arthur, eslo, dia 101, jadre, paredao, tiago abravanel, pa , dg , lais
Canaval22	Evento comemorativo	carnaval, sapucai, enredo, mocidade, mangueira, salgueiro, portela, desfile, vai-vai, gavioes, viradouro, estacio de sa, grande rio, comissao de frente, apuracao, beija flor, rosas de ouro, tuiuti, sao clemente, vila isabel, samba
Juliette	Influenciadora Digital, campeã do BBB21	juliette
América	Novela da TV Globo	america
Desalma	Série produzida pelo Globoplay	desalma
Pantanal	Novela da TV Globo	pantanal
Túnel do Amor	Reality Show	tunel do amor
WSL	Campeonato de Surfe	wsl
Notícia Tiago Leifert	Notícia sobre a volta de Tiago para a Globo, para narrar os jogos da copa do mundo no globoplay	tiago leifert
Não Identificado	Tweets sem nenhuma relação com os temas relacionados	-

Fonte: Elaborado pelo autor

Com as atribuições realizadas, a seguir podemos identificar os temas mais relevantes no período. Identificamos a grande relevância do BBB22 e do Carnaval 2022. 42% dos tweets não receberam a classificação de tema, pois foram tweets sem relação com os principais temas/eventos mapeados. Há uma limitação nessa classificação, pois a atribuição dos temas é feita observando o texto do tweet, porém existem tweets que são associados a

um tema de forma implícita, mas no texto do tweet não tem nada relacionado ao tema, portanto para esses casos o tema recebido será o “Não Identificado”.

Gráfico 1: Total de Tweets por tema

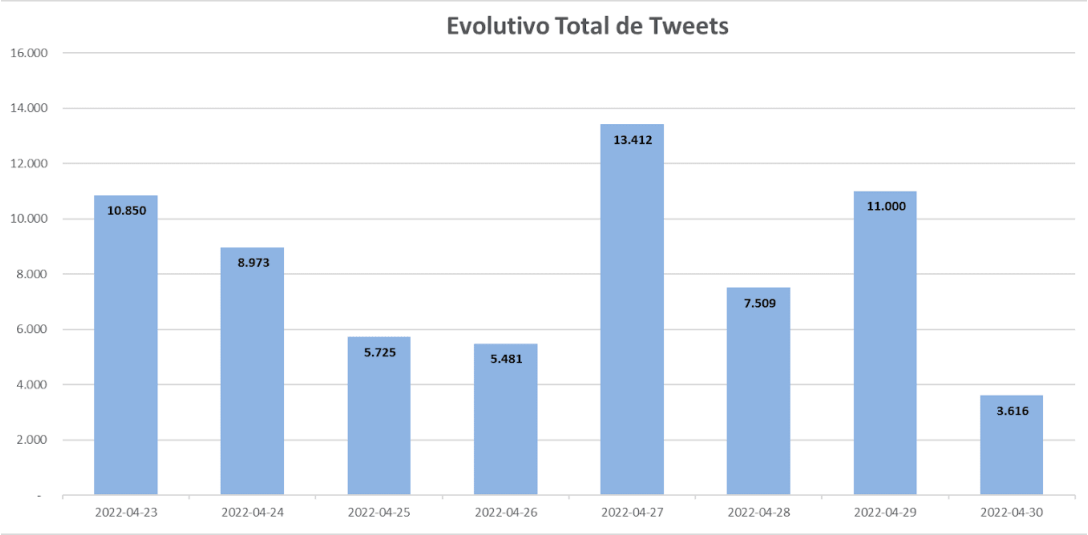


Fonte: Elaborado pelo autor

4.2 – Evolutivos

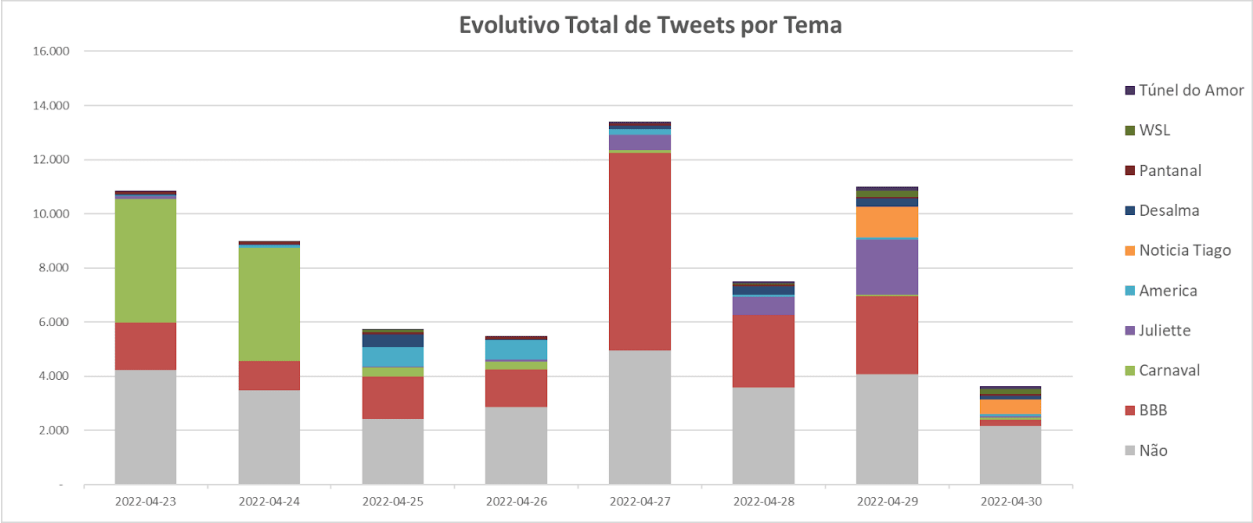
Uma abordagem comum em análise de dados é a investigação dos números ao longo do tempo. Desta forma podemos entender os picos de consumo e correlacionar com os eventos ocorridos. Abaixo temos os evolutivos diários do total acompanhado pela quebra por tema, entre os dias 23 de abril de 2022 e 30 de abril de 2022.

Gráfico 2: Evolutivo diário do total de tweets



Fonte: Elaborado pelo autor

Gráfico 3: Evolutivo diário do total de tweets quebrado por tema



Fonte: Elaborado pelo autor

No evolutivo, vemos que grande parte do volume de tweets está concentrada nos dias 23,24,27,29. No gráfico 3 podemos associar esses picos principalmente aos eventos da final do BBB22 e os desfiles do carnaval de 2022. Nos dias 23 e 24 vemos uma grande representatividade do carnaval. Já nos dias 27 e 29, vemos uma maior representatividade de tweets sobre o BBB22, impulsionada pela data da final do reality show e a exibição do programa especial BBB101. Outros

eventos relevantes que podemos citar, foi o lançamento da segunda temporada da série ‘Desalma’ (dia 18/04/2022), a adição da novela ‘América’ (lançamento no dia 25/04) no catálogo da plataforma, tweets sobre o anúncio (dia 29/04/2022) da volta de Tiago Leifert a Globo, para narrar os jogos da Copa do Mundo de Futebol de 2022 no Globoplay, e uma ação publicitária da Juliette (realizada no dia 28/04/2022).

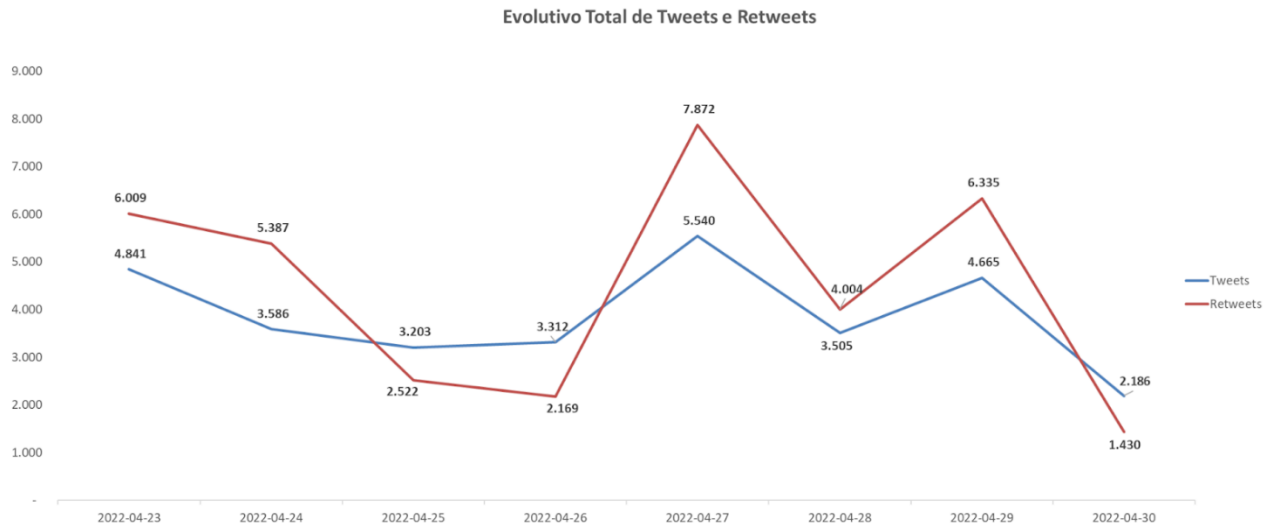
Analisando os dados hora a hora dos tweets do gráfico 4, podemos concluir que grande parte dos tweets foram realizados em horários mais noturnos, justamente no horário do programa do BBB22 e do horário dos desfiles das escolas de samba de 2022. Ou seja, os usuários se manifestam na rede social praticamente em tempo real em relação aos acontecimentos externos.

Gráfico 4: Total de tweets por dia por hora

Fonte: Elaborado pelo autor

Abaixo temos o evolutivo quebrado por tweets e retweets:

Gráfico 5: Total de Tweets e Retweets por dia

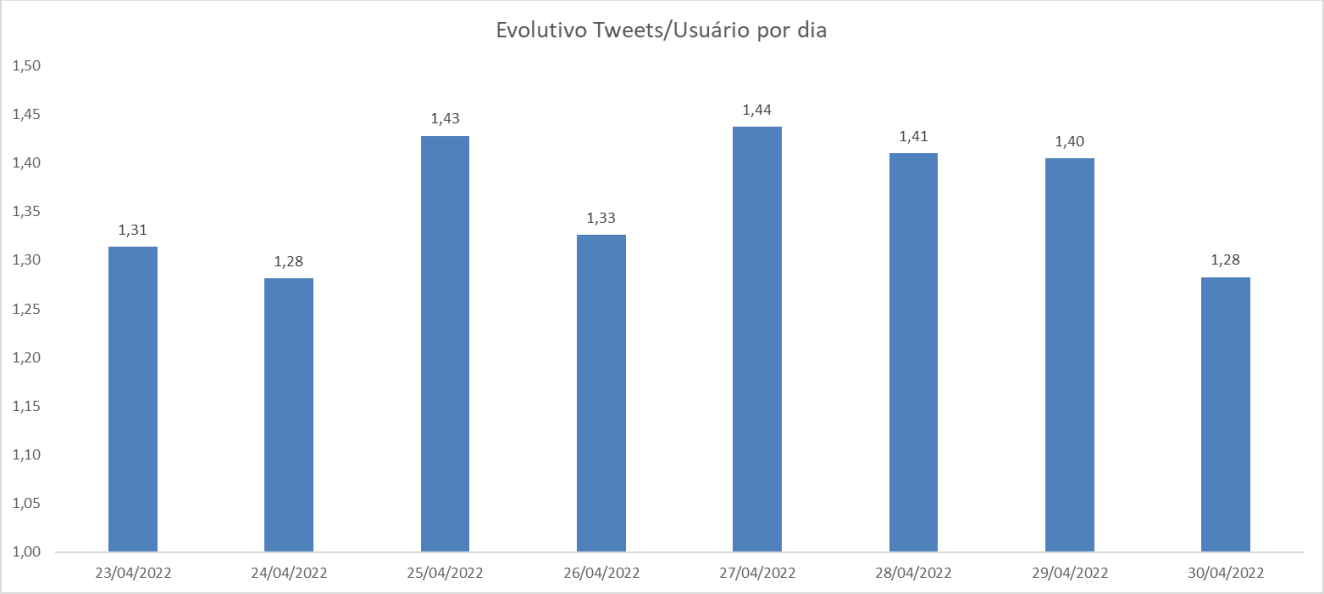


Fonte: Elaborado pelo autor

Vemos no gráfico que os picos de engajamento estão relacionados com o aumento na quantidade de RTs. Fica evidente o fator viral nas datas de picos.

O gráfico 6 mostra o evolutivo diário de tweets/usuário. Essa métrica é calculada pegando o total de tweets de um dia e dividindo pelo total de usuários únicos deste dia. A ideia é identificar se quando temos um aumento nos números de tweets é por conta da maior quantidade de tweets que um mesmo usuário realiza. Porém, verificando os dados no gráfico 6, não vemos altas correlações.

Gráfico 6: Tweets/Usuário por dia



Fonte: Elaborado pelo autor

4.3 – Rankings

Nessa sessão exploramos os principais campos dos tweets para identificar os destaques.

4.3.1 – Ranking de Tweets

Agora vamos observar os tweets mais populares do período de coleta. Para isso vamos somar todos os retweets que cada tweet teve.

Tabela 5: Top 10 Tweets mais populares

Rank	Tweet	User	Total Tweets+Retweets
1	Para mim, o maior influenciador dessa casa é o @globoplay . É filme, novela, série, documentário e mais um monte de benefício	juliette	1519
2	MANO, o Tiago Leifert vai narrar jogos da Copa do Mundo no Globoplay, no Sportv e no Globo Esporte! Tem que comentar co...	curiosidadesdb	840
3	simplesmente	Casimiro	803
4	Pabllo Vittar nos Estados Unidos e RuPaul no Brasil. #Globeleza	forumpandlr	793
5	P.A. e Jade	globoplay	780
6	A química de milhões! #BBB22 #BBBNoGloboplay	globoplay	719
7	Os ♡Jadré Lovers♡ vão à loucura! #PrêmioRedeBBB #BBB22 #BBBNoGloboplay	globoplay	697
8	Errada não tá!	globoplay	626
9	Pode gritar, Arthur: É CAMPEÃO! #BBB22 #FinalBBB22 #BBBNoGloboplay	globoplay	528
10	Nem parece que Sabrina Sato estava em SP há poucas horas desfilando na Gaviões da Fiel. Fôlego de sobra na Vila Isabel. Rainha...	ZAMENZA	517

Fonte: Elaborado pelo autor

O tweet da ação publicitária de Juliette foi o tweet mais popular do período, estando na frente com uma larga vantagem do segundo colocado. Uma outra observação está na quantidade relevante de tweets do perfil do Globoplay presentes no TOP 10. Aqui podemos ver na prática um bom exemplo do Globoplay no conceito de marketing de conteúdo.

4.3.2 – Ranking de Palavras

Utilizando o algoritmo de Bag of Words, juntamente com os algoritmos de limpeza de texto, iremos fazer uma contagem das palavras únicas relevantes, a fim de descobrir quais palavras aparecem com maior frequência. Um ponto muito importante aqui é que só iremos considerar os tweets únicos. Uma característica da escrita da língua portuguesa é a grande utilização de artigos, preposições e conjunções. Para esse exercício os artigos, preposições e conjunções não têm valor. Portanto, iremos aplicar o conceito de stopwords e POS-Tagging para retirar palavras que não agregam a análise. Na tabela abaixo temos o ranking das 40 palavras mais presentes nos tweets coletados.

Avaliando o resultado podemos identificar a presença de palavras ligadas aos eventos de BBB e Carnaval, palavras ligadas ao consumo de streaming e palavras ligadas aos programas presentes na plataforma. Um ponto positivo é que nenhuma palavra negativa está no top 40. A primeira palavra ruim está a partir do ranking 140 com a palavra ‘cancelar’, seguido pela palavra ‘merda’ na posição 149.

Tabela 6: Top 40 palavras mais presentes

Rank	Palavra	Ocorrencias		Rank	Palavra	Ocorrencias
1	globoplay	14393		21	tenho	647
2	ver	2500		22	conta	610
3	assistir	2211		23	vendo	604
4	globo	1765		24	amor	599
5	bbb	1099		25	assistindo	574
6	novela	1071		26	queria	574
7	tv	972		27	melhor	549
8	serie	963		28	alguem	538
9	gente	953		29	final	537
10	arthur	945		30	acho	522
11	dia	929		31	pantanal	513
12	america	906		32	deus	504
13	desfile	833		33	documentario	497
14	temporada	761		34	transmissao	497
15	desalma	761		35	bom	491
16	kkk	727		36	vem	491
17	vivo	708		37	pode	479
18	voce	687		38	mundo	467
19	quero	654		39	faz	458
20	sao	648		40	boa	455

Fonte: Elaborado pelo autor

Figura 5: Nuvem de palavras das palavras mais presentes

Fonte: Elaborado pelo autor

4.3.3 – Ranking de Palavras - Bigrama/Trigrama

Vimos o rank de palavras únicas, agora veremos os rankings de duas e três palavras seguidas, aqui não foi feita a filtragem de stopwords e POS-Tagging. Analisando os resultados vemos uma presença marcante de globoplay nos bigramas, porém na maioria dos casos sempre

acompanhado de artigos, o que não nos dá muita informação. Já no trigrama, podemos perceber a palavra globoplay juntamente com outras palavras ligadas aos conteúdos da plataforma.

Tabela 7 e 8: Bigrama/Trigrama mais presentes

Rank	word1	word2	Ocorrencias
1	no	globoplay	3983
2	na	globoplay	1651
3	globoplay	e	1261
4	o	globoplay	1052
5	do	globoplay	997
6	a	globoplay	872
7	e	o	830
8	da	globoplay	799
9	e	a	793
10	ao	vivo	672
11	com	o	636
12		a	624
13	que	a	590
14	globoplay	pra	582
15	o	que	547
16	pra	ver	542
17	que	o	536
18		eu	530
19	o	desfile	496
20	pelo	globoplay	488
21	a	globo	487
22	nao	tem	474
23	e	nao	465
24	pra	assistir	453
25	que	nao	430

Rank	word1	word2	word3	Ocorrencias
1	no	globoplay	e	358
2	desalma	no	globoplay	293
3	america	no	globoplay	221
4	tunel	do	amor	202
5	temporada	de	desalma	193
6	o	desfile	da	170
7	segunda	temporada	de	166
8	a	segunda	temporada	165
9	na	globoplay	e	162
10	tem	no	globoplay	143
11	canais	ao		141
12	copa	do	mundo	140
13	globoplay	pra	ver	127
14	globoplay	pra	assistir	118
15	tem	na	globoplay	111
16	pelo	amor	de	107
17	da	copa	do	107
18	vou	ter	que	106
19	disponivel	no	globoplay	103
20	jogos	da	copa	102
21	o	globoplay	e	99
22	no	globoplay	pra	99
23	conta	do	globoplay	99
24	escolas	de	samba	98
25	amor	de	deus	97

Fonte: Elaborado pelo autor

4.3.4 – Ranking de Hashtags

Foram identificadas 10.412 hashtags, abaixo temos as hashtags mais presentes. As principais hashtags seguem a tendência de acompanhar os principais eventos do período. O destaque fica com o BBB22 dominando as primeiras posições. A seguir as 25 hashtags mais populares acompanhada pela núvem de palavras dos 250 hashtags mais populares.

Tabela 9: Top25 Hashtags

Rank	Hashtag	Ocorrencias	% do Total
1	#bbb22	1374	13,2%
2	#globoplay	897	8,6%
3	#globeleza	546	5,2%
4	#redebbb	357	3,4%
5	#bbbnogloboplay	286	2,7%
6	#desalma	246	2,4%
7	#finalbbb22	176	1,7%
8	#carnaval2022	167	1,6%
9	#pantanal	166	1,6%
10	#bbb	159	1,5%
11	#wslbrasil	146	1,4%
12	#américa	144	1,4%
13	#globo	139	1,3%
14	#carnavalnogloboplay	135	1,3%
15	#carnaval	92	0,9%
16	#carnavalrj	89	0,9%
17	#túneldoamor	88	0,8%
18	#bigbrotherbrasil	78	0,7%
19	#americanogloboplay	71	0,7%
20	#batepapobbb	65	0,6%
21	#bbb101	65	0,6%
22	#desalma2	63	0,6%
23	#america	56	0,5%
24	#g1	53	0,5%
25	#tuneldoamor	50	0,5%

Fonte: Elaborado pelo autor

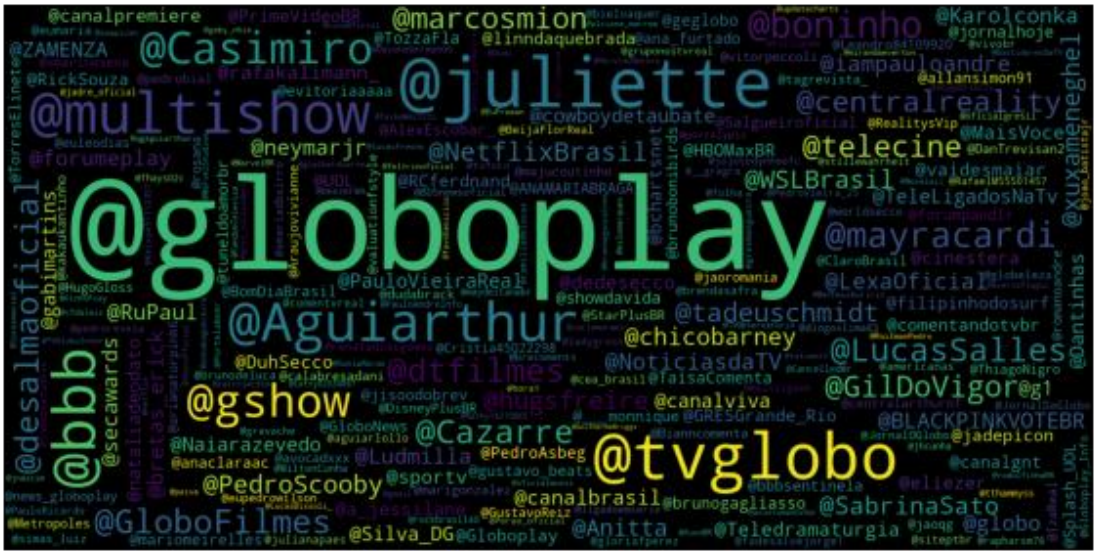
Figura 6: Nuvem de Palavras Hashtags

Fonte: Elaborado pelo autor

4.3.5 – Ranking de Menções

Ao total 42.989 menções foram realizadas. Abaixo temos o ranking e a nuvem de palavras com as menções mais presentes. O destaque fica na grande representatividade dos perfis de produto do Grupo Globo, como por exemplo bbb, tvglobo e multishow, e dos participantes do BBB22. Outro ponto interessante, foi a forte presença do perfil de Juliette, tendo mais menções que o perfil do Arthur Aguiar, campeão da edição do BBB22, mostrando grande força dos seguidores e admiradores da influenciadora.

Figura 7: Nuvem de Palavras Menções



Fonte: Elaborado pelo autor

Tabela 10: TOP 25 Menções

Rank	Perfil	Ocorrencias	% do Total
1	globoplay	13887	32,3%
2	juliette	1155	2,7%
3	tvglobo	972	2,3%
4	bbb	819	1,9%
5	multishow	769	1,8%
6	Aguiarthur	687	1,6%
7	gshow	461	1,1%
8	Casimiro	433	1,0%
9	boninho	304	0,7%
10	mayracardi	270	0,6%
11	GloboFilmes	251	0,6%
12	desalmaoficial	245	0,6%
13	dtfilmes	243	0,6%
14	Cazarre	241	0,6%
15	LucasSalles	241	0,6%
16	telecine	205	0,5%
17	GilDoVigor	204	0,5%
18	marcosmion	201	0,5%
19	centralreality	192	0,4%
20	tadeuschmidt	153	0,4%
21	NetflixBrasil	145	0,3%
22	WSLBrasil	135	0,3%
23	hugsfreire	133	0,3%
24	xuxameneghel	131	0,3%
25	PedroScooby	130	0,3%

Fonte: Elaborado pelo autor

4.4 – Análise de Emoticons

Dos 33.583 tweets únicos, 24,6% contêm pelo menos um emoticon. Ao todo, 17.261 emoticons foram coletados. Abaixo temos o ranking dos emoticons mais presentes. Podemos perceber um maior volume de emoticons positivos como coração, cara apaixonada e risada. Detalhe para o emoticon de pão, associado ao participante Arthur Aguiar do BBB22.

Além disso, apenas dois emoticons negativos apareceram no top 15. Isso pode nos levar a uma hipótese onde tweets negativos, para o nosso universo de dados, não costumam vir acompanhado de emoticons em seu texto.

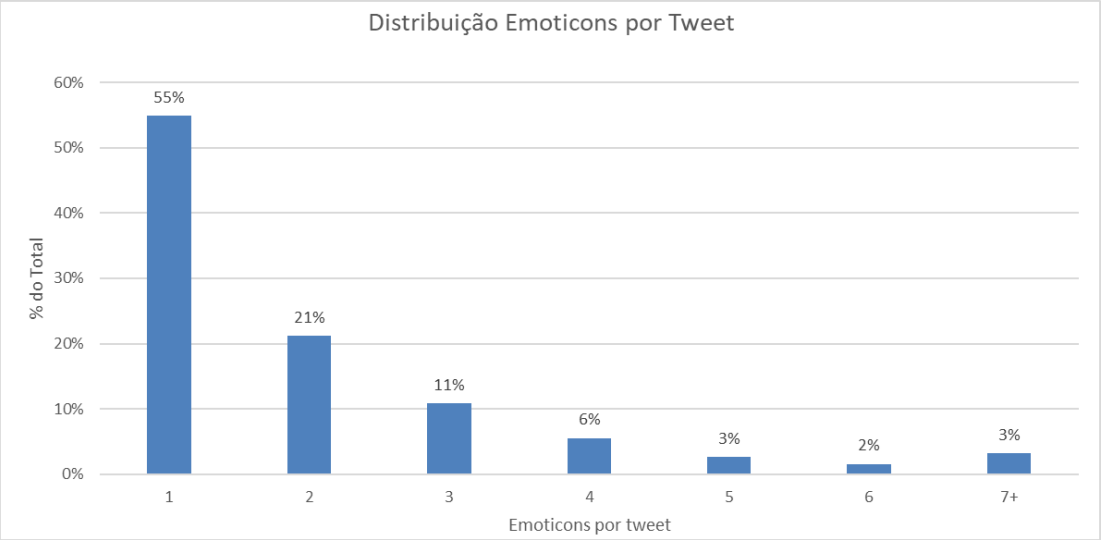
Tabela 11: TOP15 Emoticons mais utilizados

	Emoticon	total
1	😍	1296
2	❤️‍🔥	1172
3	😭	734
4	😂	687
5	🤣	526
6	👏	521
7	🤍	501
8	😻	477
9	😕	423
10	🤡	372
11	🤦‍♂️	354
12	🍞	338
13	✨	287
14	🗣️	271
15	🔥	269

Fonte: Elaborado pelo autor

Quando olhamos para a distribuição da quantidade de emoticons por tweets, temos o seguinte resultado (Tweets com pelo menos um emoticon):

Gráfico 7: Distribuição de Emoticons por Tweet



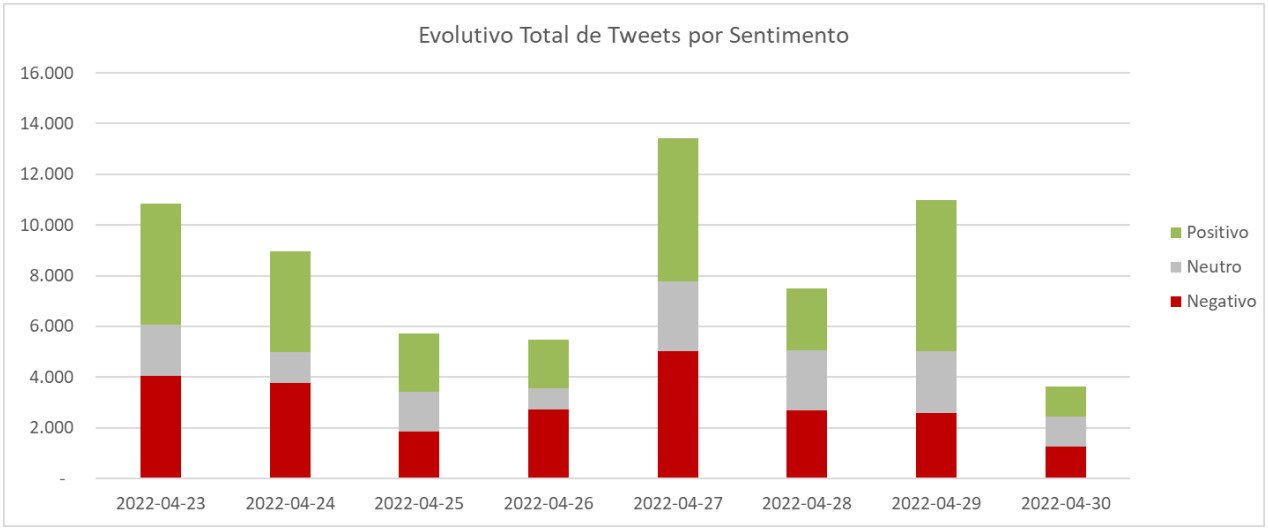
Fonte: Elaborado pelo autor

Ou seja, dos tweets com emoticons, um pouco mais da metade (55%) dos tweets únicos possuem apenas um emoticon.

4.5 – Análise de Sentimentos

Após a etapa de coleta e limpeza dos tweets, aplicando o Modelo 9 visto na metodologia, obtivemos os seguintes resultados. Dos 66.566 tweets coletados, 42% foram classificados como positivo, 36% foram classificados como negativo e 22% como neutro. Abaixo o evolutivo diário quebrado por sentimento.

Gráfico 8: Total de tweets por dia por sentimento



Fonte: Elaborado pelo autor

Aplicando o modelo para os tweets únicos temos o seguinte resultado. Dos 33.583 tweets únicos coletados, 34% foram classificados como positivo, 46% foram classificados como negativo e 20% como neutro. Temos uma diferença significativa comparando os valores proporcionais dos sentimentos positivos e negativos. Para entendermos o porquê disso, vamos olhar para a métrica - Soma Tweets + Retweets/Tweet Únicos, quebrado por sentimento. Essa métrica é basicamente a quantidade média que um tweet apareceu para cada um dos sentimentos.

Tabela 12: Tweets + Retweets/Tweet Únicos por sentimento

Sentimento	Valor
Positivo	2,54
Neutro	2,26
Negativo	1,56

Fonte: Elaborado pelo autor

Observando a tabela, podemos concluir que tweets positivos de uma forma geral tiveram uma repercussão maior que os negativos, por outro lado os tweets negativos foram mais variados.

A seguir alguns exemplos dos tweets de cada sentimento:

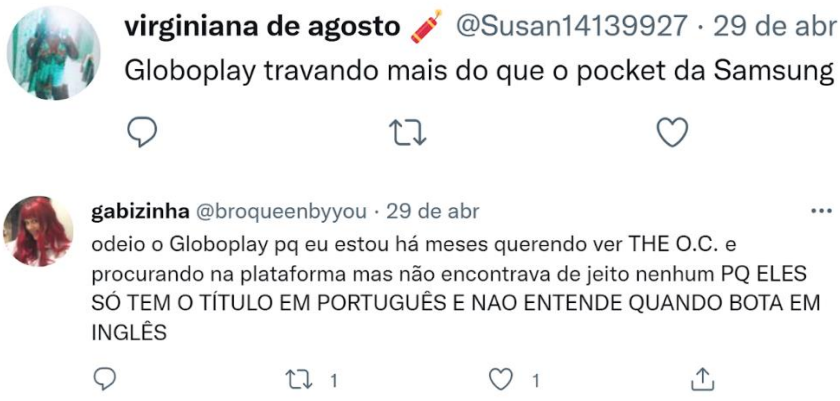
Tweets positivos:

Figura 8: Exemplos tweets positivos classificados pelo modelo

Fonte: Twitter

Tweets negativos:

Figura 9: Exemplos tweets negativos classificados pelo modelo



Fonte: Twitter

Tweets Neutros:

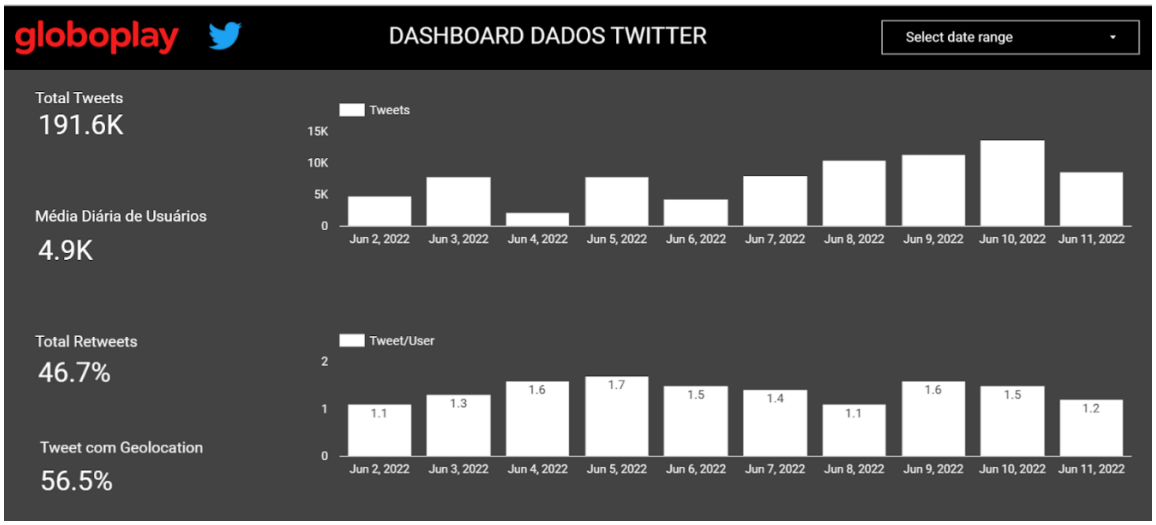
Figura 10: Exemplos tweets neutros classificados pelo modelo

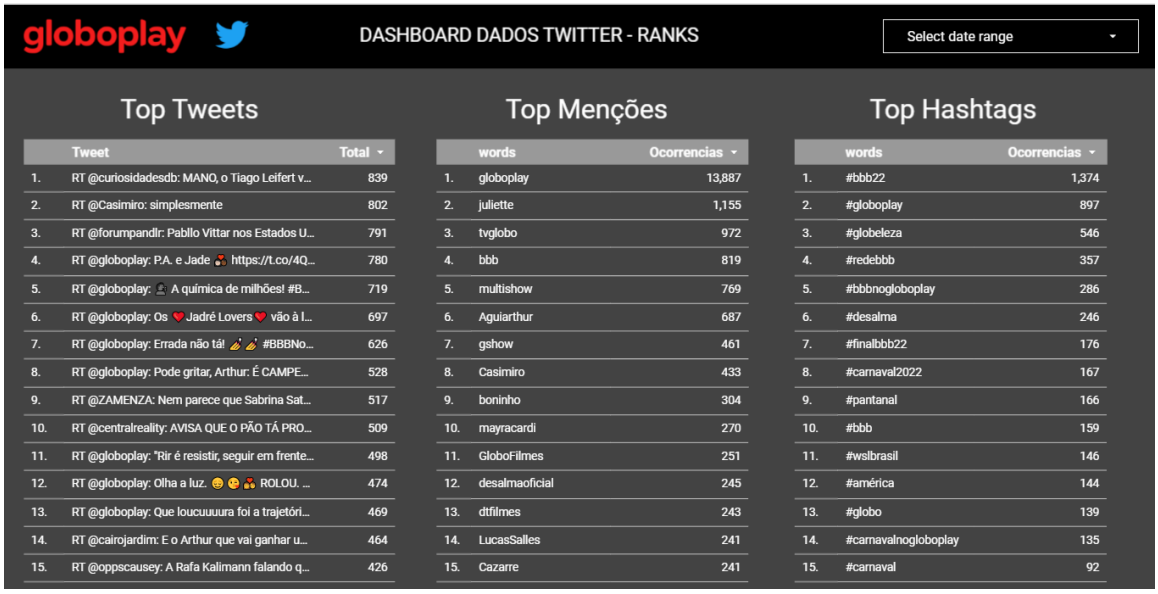
Fonte: Twitter

4.6 – Dashboard Executivo

Por fim, queremos colocar todos os valores encontrados em um dashboard, para que os gerentes, executivos ou outros stakeholders da marca possam acompanhar todos os dados apresentados no estudo. As imagens a seguir ilustram como seriam as telas do dashboard.

Figura 11: Telas Dashboard





Top Tweets

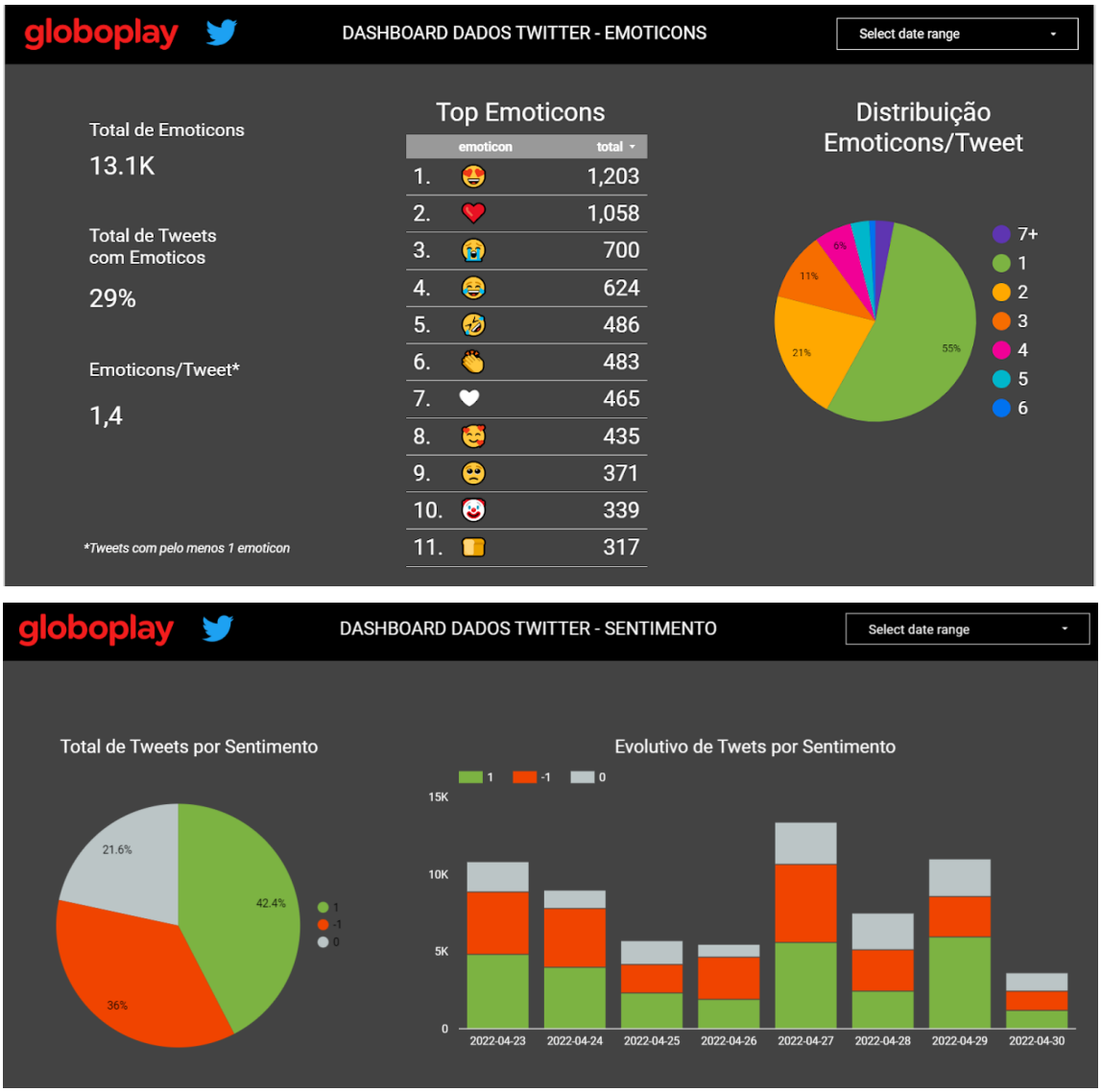
	Tweet	Total
1.	RT @curiosidadesdb: MANO, o Tiago Leifert v...	839
2.	RT @Casimiro: simplesmente	802
3.	RT @forumpandlr: Pabllo Vittar nos Estados U...	791
4.	RT @globoplay: P.A. e Jade https://t.co/4Q...	780
5.	RT @globoplay: A química de milhões! #B...	719
6.	RT @globoplay: Os Jadré Lovers vão à l...	697
7.	RT @globoplay: Errada não tá! #BBBNo...	626
8.	RT @globoplay: Pode gritar, Arthur: É CAMPE...	528
9.	RT @ZAMENZA: Nem parece que Sabrina Sat...	517
10.	RT @centralreality: AVISA QUE O PÃO TÁ PRO...	509
11.	RT @globoplay: "Rir é resistir, seguir em frente...	498
12.	RT @globoplay: Olha a luz. ROLOU. ...	474
13.	RT @globoplay: Que loucuuuura foi a trajetóri...	469
14.	RT @cairojardim: E o Arthur que vai ganhar u...	464
15.	RT @oppscausey: A Rafa Kalimann falando q...	426

Top Menções

	words	Ocorrencias
1.	globoplay	13,887
2.	juliette	1,155
3.	tvglobo	972
4.	bbb	819
5.	multishow	769
6.	Aguiarthur	687
7.	gshow	461
8.	Casimiro	433
9.	boninho	304
10.	mayracardi	270
11.	GloboFilmes	251
12.	desalmaoficial	245
13.	dtfilmes	243
14.	LucasSalles	241
15.	Cazarre	241

Top Hashtags

	words	Ocorrencias
1.	#bbb22	1,374
2.	#globoplay	897
3.	#globeleza	546
4.	#redebbb	357
5.	#bbbnogloboplay	286
6.	#desalma	246
7.	#finalbbb22	176
8.	#carnaval2022	167
9.	#pantanal	166
10.	#bbb	159
11.	#wslbrasil	146
12.	#américa	144
13.	#globo	139
14.	#carnavalnogloboplay	135
15.	#carnaval	92



Fonte: Elaborado pelo autor

5 – Conclusões

Visando o conhecimento das interações dos usuários que interagiram com uma determinada marca no Twitter, este trabalho utilizou um processo prático de Mineração de Texto. Através de um programa feito em Python, foi possível realizar a coleta dos tweets. Após a coleta, diversas análises quantitativas foram realizadas, adicionalmente, foi desenvolvido um classificador de sentimentos com técnicas de aprendizado de máquina, com a finalidade de termos mais riquezas nas análises. Por fim, todos os dados gerados foram colocados em um dashboard, no ponto de vista gerencial, o dashboard juntamente com seus dados, tem o intuito de ajudar os stakeholders interessados no monitoramento das diversas métricas apresentadas neste estudo. Além disso, o estudo pode ajudar os gestores na gestão de conhecimento do público-alvo da marca.

Analisando os resultados, foi possível entender como foi composta a volumetria dos tweets do Globoplay. Dentro do período, aproximadamente 66 mil tweets foram coletados. Foram identificados alguns eventos que tiveram grande contribuição para o amplo engajamento do público, como a final do BBB22 e os desfiles das escolas de samba do Rio de Janeiro e São Paulo de 2022. Tal engajamento está ligado aos benefícios oferecidos pela plataforma de streaming. No Globoplay, no caso do BBB22, é possível assistir diversos conteúdos do reality, como por exemplo, trechos do programa, íntegra do programa transmitido na TV Globo e acesso ao vivo às câmeras da casa onde o reality é realizado. Já para o Carnaval 2022, o Globoplay realizou uma transmissão ao vivo dos desfiles das escolas de samba do Rio de Janeiro e São Paulo. Outro ponto investigado foi o impacto significativo dos retweets para a formação dos picos de engajamento. Isso na prática é o comportamento viral sempre presente em redes sociais.

Explorando as diversas informações de cada tweet coletado, foi possível realizar diversos rankings a fim de encontrar outros destaques,

como por exemplo, o ranking dos tweets mais vistos, onde vimos que o Tweet da campanha publicitária de Juliette para o Globoplay foi o mais popular.

Por conta do grande volume de tweets, seria extremamente exaustivo entender para cada tweet, se a interação do usuário foi positiva ou negativa. Portanto, foi desenvolvido um classificador de sentimentos utilizando técnicas de aprendizado de máquina. Nos resultados obtidos, foi constatado uma proporção maior de tweets positivos, 42% contra 36% de tweets negativos. Porém, se considerarmos os tweets únicos, essa proporção muda para 34% de positivos e 46% de negativos, ou seja, os tweets positivo foram mais populares por conta dos RTs, já os negativos tiveram uma maior variedade de tweets únicos. Consolidando os resultados, podemos concluir que os tweets positivos foram mais predominantes. Podemos associar esse resultado com os dados de emoticons, onde vemos que a maioria dos emoticons identificados foram emoticons positivos. O que levantou uma hipótese de que os tweets negativos normalmente não veem acompanhado de emoticons negativos. Em relação ao conteúdo dos tweets, os tweets positivos foram elogios e exaltações aos participantes do BBB22 e aos conteúdos presentes na plataforma como séries, novelas, programas de TV e filmes. Já os tweets negativos foram marcados por reclamações sobre a performance da plataforma, como por exemplo travamentos, e críticas aos conteúdos disponíveis.

5.1 – Trabalhos futuros

- Implementar uma automatização do processo realizado, assim novos dados seriam coletados diariamente para o monitoramento via dashboard.
- Melhorar modelo de classificação de sentimento, explorando técnicas mais complexas como algoritmos de Deep Learing.
- Comparação de tweets de outros concorrentes como por exemplo a Netflix.
- Análise clusterizada dos usuários que interagiram com a marca.

- Procurar na literatura estudos sobre Netnografia. Com isso podemos ter uma análise qualitativa mais rica.

6 – Referências Bibliográficas

ARAÚJO, M.; PESSANHA, G; ALVES, R; SILVA FILHO, A. **Análise dos sentimentos dos usuários do Twitter em relação às marcas Havan, Madero e Giraffas: um estudo a partir dos posicionamentos dos seus proprietários na pandemia do covid-19.** 2021. Disponível em:< https://periodicos.ufpb.br/index.php/tematica/article/view/57701/32870>. Acesso em: 20 mai. 2022.

CHANG JIAKANG, Et al. **What is text mining, how does it work and why is it useful?** 2018. Disponível em: < https://bit.ly/2WWpaSr >. Acesso em 17 abr. 2022.

Dixon, S. **Countries with the most Twitter users 2022**. 2022. Disponível em: < https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>. Acesso em 17 abr. 2022.

FEW, S. **Information Dashboard Design: Effective Visual Communication of Data.** Sebastopol. O'Reilly, 2006

GOMES, Helder Joaquim Carvalheira. **Text Mining: análise de sentimentos na classificação de notícias**. Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on. Lisboa. 2013.

GONÇALVES, Pollyanna; BENEVENUTO, Fabrício; ALMEIDA, Virgílio. **O**

Que Tweets Contendo Emoticons Podem Revelar Sobre Sentimentos Coletivos?. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 2. , 2013, Maceió. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2013 . p. 128-139. ISSN 2595-6094. Disponível em:<https://homepages.dcc.ufmg.br/~fabricio/download/brasnam13.pdf>. Acesso em: 20 mai. 2022.

Hohmann, L. **Como a Globo promove a Cultura de Dados**. 2021. Disponível em: <https://www.programaria.org/como-globo-promove-cultura-dados/>. Acesso em 17 abr. 2022.

HURWITZ, J. and KIRSCH, D. **Machine Learning for dummies**, IBM Limited Edition. New York: Copyright Business Expert Press, 2018.

Jose, R. and Chooralil, V. S. (2016). **Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach.** In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), pages 64–67.

KEMPE, D.; KLEINBERG, J.; TARDOS, E**. Maximizing the Spread of Influence through a Social Network**. 2015. Disponível em: <https://theoryofcomputing.org/articles/v011a004/>. Acesso em 20 mai. 2022.

KNAFLIC, COLE NUSSBAUMER. **Storytelling com dados: Um guia sobre visualização de dados para profissionais de negócios.** Rio de Janeiro: Alta Books, 2019.

KOTLER, P; KARTAJAYA, H; SETIAWAN, I. Marketing 4.0: **Do Tracicional ao Digital.** Rio de Janeiro: Sextante, 2017.

LIDDY, E. **Natural Language Processing. Encyclopedia of Library and Information Science**. New York: Marcel Decker, Inc, 2001

OLIVEIRA, Renarth Bustamante de; LUCENA, Wellington Machado. **O uso da internet e das mídias digitais como ferramentas de estratégia de marketing**. V2, n.1. 2021. Disponível em: <http://revistaadmmade.estacio.br/index.php/destarte/article/viewFile/8742/47967092>. Acesso em 20 mai. 2022.

PAREDES, A. **Como são os usuários do Twitter e como se comportam com as marcas?.** 2020. Disponível em:<https://www.iebschool.com/pt-br/blog/social-media/redes-sociais/como-sao-os-usuarios-do-twitter-e-como-se-comportam-com-as-marcas/>. Acesso em: 20 mai. 2022.

PEÇANHA, Vitor. **O que é Marketing Digital?**. 2018. Disponível em: <https://marketingdeconteudo.com/marketing-digital/#01>. Acesso em: 26 jan. 2022.

PROVOST, FOSTER; FAWCETT, TOM. **Data Science para Negócios: O que Você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico de Dados.** Rio de Janeiro: Alta Books, 2016.

RECUERO, R. **Redes sociais na internet**, Ed Sulina, 2 edição. Porto Alegre.Co-edição CUBOCC, 2011

SALGADO, D. **Pesquisa sobre o Twitter no Brasil: entenda o comportamento dos usuários.** 2021. Disponível em: <https://blog.opinionbox.com/twitter-no-brasil/>. Acesso em: 20 mai. 2022.

SILVA, F. **TSCOOP - AVALIAÇÃO DE FERRAMENTAS PARA ANÁLISE**

DE SENTIMENTO: UM ESTUDO DE CASO NO TWITTER. 2019. Disponível em: <https://repositorio.utfpr.edu.br/jspui/bitstream/1/16012/1/PG_COCIC_2019_1_09.pdf >. Acesso em 20 mai. 2022.

SILVA, L; COSTA, E; GAMMARANO, I.; **FILHO, E. VALORES COMPORTAMENTAIS NA PREFERÊNCIA DE USO DA REDE SOCIAL TWITTER.** 2016. Disponível em: <https://periodicos.ufba.br/index.php/contemporaneaposcom/article/view/11252/11511>. Acesso em: 20 mai. 2022.

TERRA, E. **Ferramenta para Extração de Dados do Twitter para Mineração de Dados.** 2015. Disponível em: < https://repositorio.ucs.br/xmlui/bitstream/handle/11338/1124/TCC%20Edipo%20Deon%20Terra%20.pdf?sequence=1&isAllowed=y>. Acesso em 20 mai. 2022.

THORING, A. **Corporate Tweeting: Analysing the Use of Twitter as a Marketing Tool by UK Trade Publishers**. 2011. Pub Res Q, 27:141–158. Disponível em:< https://www.researchgate.net/publication/225552263_Corporate_Tweeting_Analysing_the_Use_of_Twitter_as_a_Marketing_Tool_by_UK_Trade_Publishers>. Acesso em 20 mai. 2022.

TORRES, C. **A Bíblia do Marketing digital. Tudo o que você gostaria de saber sobre marketing e publicidade na internet e não tinha a quem perguntar**.1 ed. São Paulo. Novatec, 2009.

7 – Anexo

7.1 – ANEXO I

```
#Imports + configuração do token
import tweepy
import csv
import numpy as np
import pandas as pd
import re

CONSUMER_KEY = 'codigo_pessoal'
CONSUMER_SECRET = 'codigo_pessoal'
ACCESS_TOKEN = 'codigo_pessoal'
ACCESS_TOKEN_SECRET = 'codigo_pessoal'
auth=tweepy.OAuth1UserHandler(CONSUMER_KEY,
CONSUMER_SECRET,ACCESS_TOKEN, ACCESS_TOKEN_SECRET)
api = tweepy.API(auth)

#header do arquivo final
TWEET_TABLE_HEADER = ['id_user', 'user', 'number_of_followers',
'number_of_followings', 'id tweet','tweet_created_at', 'text',
'mentions','retweeted','location']

keyword = "globoplay"
with open('%s_tweets.csv' %keyword, 'w+', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)
    csv_writer.writerow(TWEET_TABLE_HEADER)
    for tweet in tweepy.Cursor(api.search_tweets, keyword,
tweet_mode='extended', lang= 'pt', count=100).items():
        if re.match("^RT ", tweet.full_text ):
          csv_writer.writerow([tweet.user.id_str, tweet.user.screen_name,
tweet.user.followers_count, tweet.user.friends_count, tweet.id,
tweet.created_at, tweet.full_text, re.findall(r"@(\w+)", tweet.full_text),"yes",
tweet.user.location])
        else:
           csv_writer.writerow([tweet.user.id_str, tweet.user.screen_name,
tweet.user.followers_count, tweet.user.friends_count, tweet.id,
tweet.created_at, tweet.full_text, re.findall(r"@(\w+)", tweet.full_text),"no",
tweet.user.location])
```

7.2 – ANEXO II

id_user	user	number_of_followers	number_of_followings	id tweet	tweet_created_at	text	mentions	retweeted	location
1,4696E+18	thickxz	22	123	1,52057E+18	2022-05-01 00:54:02+00:00	@globoplay @marigonzalez HAHAHAHHAHA ELA É MUITO CARISMÁTICA E ESPONTÂNEA	['globoplay', 'marigonzalez']	no	
1232441281	AntGabsEv	5318	3917	1,52057E+18	2022-05-01 00:53:51+00:00	A fotografia, o elenco, o enredo, os efeitos, a ambientação de #Desalma da @globoplay são fant	['globoplay']	no	Rio de Janeiro, Brasil
63342281	QuelBPorto	358	341	1,52057E+18	2022-05-01 00:53:48+00:00	@globoplay Sempre bons momentos com Mari hahaha	['globoplay']	no	São Paulo, Brasil
55569724	globoplay	613608	259	1,52057E+18	2022-05-01 00:53:38+00:00	.@marigonzalez: Vou provar! 😎A produção: Vai não! ❌ #CarnavalNoGloboplay #Globeleza http	['marigonzalez']	no	
54409465	ERICALAMARA	309	98	1,52057E+18	2022-05-01 00:53:22+00:00	@MarisaGuimaes @globoplay De nada	['MarisaGuimaes', 'globoplay'	no	RIO DE JANEIRO, BRASIL
1354455728	bentojor	64	597	1,52057E+18	2022-05-01 00:53:13+00:00	RT @andretrig: A autorização p/este empreendimento é um dos maiores absurdos da história da	['andretrig']	yes	
95240960	danirenucci	326	1050	1,52057E+18	2022-05-01 00:52:58+00:00	RT @andretrig: A autorização p/este empreendimento é um dos maiores absurdos da história da	['andretrig']	yes	
8,3407E+17	wendeldesant	356	2775	1,52057E+18	2022-05-01 00:52:43+00:00	@kpoperflop Dá pra assistir pelo Globoplay.	['kpoperflop']	no	Duque de Caxias, Brasil
2688182382	anovelouca	962	668	1,52057E+18	2022-05-01 00:52:39+00:00	RT @barroswcb: Que grande palhaçada tornaram o desfile das campeãs ! Que vergonha essa @	['barroswcb', 'multishow']	yes	
1,472E+18	Adriell0151000	233	1076	1,52057E+18	2022-05-01 00:52:39+00:00	RT @globoplay: Globoplay é pra curtir muitas histórias originais. É pra ter uma diversidade de co	['globoplay']	yes	
1,4912E+18	commentsluni	299	94	1,52057E+18	2022-05-01 00:52:33+00:00	caraca não entrava/assistia nada na netflix/amazon/star/globoplay desde janeiro e quanta coisa	[]	no	bianchessi e afonso
3131487369	AlPossa7	288	1424	1,52057E+18	2022-05-01 00:52:25+00:00	RT @barroswcb: Que grande palhaçada tornaram o desfile das campeãs ! Que vergonha essa @	['barroswcb', 'multishow']	yes	Rio de Janeiro, RJ
1,469E+18	maesteves4	354	495	1,52057E+18	2022-05-01 00:52:06+00:00	RT @gruponostvreal: '8 Presidentes 1 Juramento - A História de Um Tempo Presente', de Carla C	['gruponostvreal']	yes	
3798083957	clebergeo	1089	2650	1,52057E+18	2022-05-01 00:51:37+00:00	Esqueçam os desfiles na Multishow. Tá uma merda. Assista os da semana passada disponíveis n	[]	no	

7.3 – ANEXO III

```
def clean_up_tweet_str(tweet_strip):
    cleaned_up1 = tweet_strip.replace(',',' ')
    cleaned_up1 = cleaned_up1.replace('\r\n',' ')
    cleaned_up1 = re.sub("^RT ","", cleaned_up1)
    cleaned_up1 = cleaned_up1.replace('.',' ')
    cleaned_up1 = cleaned_up1.replace('+','')
    cleaned_up1 = cleaned_up1.replace('|','')
    cleaned_up1 = cleaned_up1.replace('"','')
    cleaned_up1 = cleaned_up1.replace('\"','')
    cleaned_up1 = cleaned_up1.replace(':',' ')
    cleaned_up1 = cleaned_up1.replace(';',' ')
    cleaned_up1 = cleaned_up1.replace('!',' ')
    cleaned_up1 = cleaned_up1.replace('?',' ')
    cleaned_up1 = cleaned_up1.replace('', ' ')
    cleaned_up1 = cleaned_up1.replace('~', ' ')
    cleaned_up1 = cleaned_up1.replace('✿', ' ')
    cleaned_up1 = cleaned_up1.replace('❀', ' ')
    cleaned_up1 = cleaned_up1.replace('}', ' ')
    cleaned_up1 = cleaned_up1.replace('{', ' ')
    cleaned_up1 = cleaned_up1.replace('(?)',' ')
    cleaned_up1 = cleaned_up1.replace('□',' ')
    cleaned_up1 = cleaned_up1.replace('•','')
    cleaned_up1 = cleaned_up1.replace('é','e')
    cleaned_up1 = cleaned_up1.replace('ê','e')
    cleaned_up1 = cleaned_up1.replace('ú','u')
    cleaned_up1 = cleaned_up1.replace('ç','c')
    cleaned_up1 = cleaned_up1.replace('ã','a')
    cleaned_up1 = cleaned_up1.replace('à','a')
    cleaned_up1 = cleaned_up1.replace('á','a')
    cleaned_up1 = cleaned_up1.replace('â','a')
    cleaned_up1 = cleaned_up1.replace('í','i')
    cleaned_up1 = cleaned_up1.replace('iii','i')
    cleaned_up1 = cleaned_up1.replace('ó','o')
    cleaned_up1 = cleaned_up1.replace('ô','o')
    cleaned_up1 = cleaned_up1.replace('õ','o')
    cleaned_up1 = cleaned_up1.replace('ooo','o')
    cleaned_up1 = cleaned_up1.replace('ú','u')
    cleaned_up1 = cleaned_up1.replace('SKSKSKSKSKS', '')
    cleaned_up1 = re.sub("@[A-Za-z0-9_]+","", cleaned_up1)
    cleaned_up1 = re.sub("#[A-Za-z0-9_]+","", cleaned_up1)
    cleaned_up1 = cleaned_up1.replace('#','')
    cleaned_up1 = cleaned_up1.replace('"','')
    cleaned_up1 = cleaned_up1.replace("", "")
```

```
    cleaned_up1 = cleaned_up1.replace('...', "")
    cleaned_up1 = cleaned_up1.replace('…', "")
    cleaned_up1 = cleaned_up1.replace('/', " ")
    cleaned_up1 = cleaned_up1.replace("",'')
    cleaned_up1 = cleaned_up1.replace("",'')
    cleaned_up1 = cleaned_up1.replace("",'')
    cleaned_up1 = cleaned_up1.replace("",'')
    cleaned_up1 = cleaned_up1.replace('*','')
    cleaned_up1 = cleaned_up1.replace('&amp','')
    cleaned_up1 = cleaned_up1.replace('&gt','')
    cleaned_up1 = cleaned_up1.replace('--','')
    cleaned_up1 = re.sub('[()!?]', '', cleaned_up1)
    cleaned_up1 = re.sub('\[.*?\]','', cleaned_up1)
    cleaned_up1 = cleaned_up1.lower()
    cleaned_up1 = cleaned_up1.replace('  ',' ')

    return cleaned_up1


def add_space_between_emojies(text):
   EMOJI_PATTERN = EMOJI
   text = re.sub(EMOJI_PATTERN, r'', text)
   return text


def strip_links(tweet):
    tweet = re.sub(r'http\S+', "", tweet)
    return tweet


def reduce_sequence_word(word):
    pattern = re.compile(r'(.)\1*')
    return ''.join([match.group()[:3] if len(match.group()) > 3 else
match.group() for match in pattern.finditer(word)])
```

7.3 – ANEXO IV

- Modelo 1 - https://github.com/rafjaa/LeIA
- Modelo 2 - https://www.youtube.com/watch?v=gURY_e5KT3o
- Modelo 3 - https://medium.com/turing-talks/como-fazer-uma-n%C3%A1lise-de-sentimentos-com-vader-21bbe3f3e38d
- Modelo 4 - https://github.com/lucasfranklinsilva/Analise-de-Sentimentos/blob/master/Twitter%20Sentiment%20Analisys.ipynb
- Modelo 5 - https://carlosbonfim.com/pages/Modelagem_dos_topicos_e_analise_de_sentimentos.html
- Modelo 6 - Elaborado pelo autor baseado no modelo 4
- Modelo 7 - Elaborado pelo autor baseado no modelo 4
- Modelo 8 - https://www.kaggle.com/code/leandrodoze/sentiment-analysis-in-portuguese/notebook
- Modelo 9 - Elaborado pelo autor baseado no modelo 8

7.3 – ANEXO V

```
tweets = df['text'] #dataframe com o campo de texto

classes = df['sentimentmanual'] #campo de sentimento preenchido
manualmente

vectorizer =
CountVectorizer(stop_words=stopwords,strip_accents='unicode')

freq_tweets = vectorizer.fit(tweets)

predictionData = vectorizer.transform(tweets)

treino,teste, classe_treino , classe_teste = train_test_split(predictionData,
sentimento, test_size = 0.30, random_state = 42)


modelo = MultinomialNB(fit_prior=False, alpha=0.6)

modelo.fit(freq_tweets,classes)
```