

Análise de sentimentos dos usuários brasileiros do Twitter com relação ao COVID-19

Nathália T. Guimarães¹, Ana Régia de M. Neves²

¹Eixo de Informação e Comunicação – Instituto Federal de Brasília (IFB)
Brasília DF Brasil

naeeguimaraes@gmail.com, ana.neves@ifb.edu.br²

Abstract. *In February 2020, the first case of contamination by the Covid-19 virus occurred in Brazil. Based on this, the country needed to take measures to contain the increase in the positive diagnosis of infected. At the same time, the Brazilian used social media as a vehicle for expressing his thoughts in the face of the pandemic and the events that involved it. Twitter was the most efficient platform used for this purpose. In this context, this article aims to make a sentiment analysis of tweets published about the Covid-19 epidemic during the period from February 4 to March 7, 2021. For this, the analysis will be based on the six basic sentiments of every human being, which were identified in Paul Ekman's thesis of the theory of feelings. The Naive Bayes machine learning technique was used to develop two models, one for the Paul Ekman classification and the other for the Polarity classification. A total of 1,651 manually classified tweets were used for training and testing the classifiers. The results showed an F1-score of up to 0.69 for negative tweets in the Polarity classification and 0.67 for tweets classified as angry in the Paul Ekman classification. In both models, the F1-score obtained better results for the classes in which the number of samples was greater. Based on the sentiment analysis carried out, it was possible to conclude that the predominant feeling of Brazilians regarding Covid-19 and its possible symptoms in the period from September 4 to March 7, 2021 was anger.*

Resumo. *Em fevereiro de 2020 ocorreu o primeiro caso de contaminação pelo vírus da Covid-19 no Brasil. Com base nisso, o país precisou tomar medidas para a contenção do aumento no diagnóstico positivo de infectados. Ao mesmo tempo, o brasileiro usou as redes sociais como veículo de manifestação do pensamento frente à pandemia e aos acontecimentos que a envolviam. O Twitter foi a mais eficiente plataforma utilizada com esse objetivo. Nesse contexto, este artigo tem como objetivo fazer uma análise de sentimento dos tweets publicados sobre a epidemia da Covid-19 durante o período de 4 de fevereiro a 7 de março de 2021. Para isso, a análise terá como base avaliativa os seis sentimentos básicos de todo ser humano, que foram identificados na tese da teoria dos sentimentos, de Paul Ekman. A técnica Naive Bayes de aprendizado de máquina foi utilizada para o desenvolvimento de dois modelos, um para a classificação de Paul Ekman e outro para a de Polaridade. Um total de 1.651 mil tweets classificados manualmente foram utilizados para o treino e teste dos classificadores. Os resultados demonstraram F1-score de até 0.69 para os tweets negativos da classificação de Polaridade e de 0.67 para os tweets classificados como raiva na classificação de Paul Ekman. Em ambos os modelos, o F1-score obteve melhor resultado para as classes em que o número de amostras era maior. Com*

base na análise de sentimentos realizada, foi possível concluir que o sentimento predominante dos brasileiros com relação à Covid-19 e seus possíveis sintomas no período de 4 de setembro a 7 de março de 2021 foi o de raiva.

1. Introdução

Após a identificação do primeiro caso de infectado pelo vírus da Covid-19 no Brasil, em 26 de fevereiro de 2020, as autoridades de saúde brasileiras iniciaram as medidas para contenção do alastramento da contaminação. Dentre elas, a principal foi a quarentena, a qual todos os estados adotaram durante os primeiros meses da epidemia.

Uma das consequências do isolamento durante a quarentena foi o aumento do uso da internet e suas mídias¹. Isso ocorreu porque as restrições adotadas para a contenção do alastramento da Covid-19 fizeram com que muitas atividades que antes eram realizadas presencialmente passassem a ser feitas de forma remota. Consequentemente, houve maior adesão do uso das redes sociais como meio de interação e manifestação do pensamento entre as pessoas. Nesse sentido, o *Twitter* é um micro blog utilizado para o compartilhamento imediato das reações de seus usuários frente às notícias que acontecem no mundo. Devido a isso, ele se tornou a principal ferramenta de análise do comportamento humano durante a pandemia.

A mineração de texto e a análise de sentimento são ferramentas que viabilizam identificar a variação comportamental nas mídias sociais. A mineração é o processo de transformar textos não estruturados em um formato estruturado para identificar padrões significativos e novos insights². Já a análise, é uma abordagem que usa Processamento de Linguagem Natural (PNL) para analisar as opiniões, sentimentos, avaliações, atitudes e emoções das pessoas a partir da linguagem escrita.

Alguns projetos de análise de sentimento utilizam como qualificadores de um texto a polarização de positivo e negativo para determinar o teor sentimental predominante em uma mensagem. Esse modelo limita a real percepção emotiva apresentada por seu autor. Por esse motivo, este trabalho analisou a metodologia de análise implementada em outros projetos com o intuito de fazer uma análise sentimental baseada em parâmetros qualificadores que permitissem determinar de forma mais clara e objetiva o sentimento trazido em uma publicação de texto (*tweet*) no *Twitter*.

Após a pesquisa, a Teoria da Universalidade das Emoções, de Paul Ekman, foi escolhida como base para a análise sentimental realizada neste trabalho, pois esta foi uma das mais utilizadas nos trabalhos investigados. Ekman dedicou uma vida estudando o comportamento e as emoções humanas, criou um método de análise e interpretação da expressão facial, publicou inúmeras produções, entre livros e artigos e é considerado um dos 100 psicólogos mais influentes do séc. XX³. Sua teoria agrupa as emoções humanas em seis categorias principais: raiva, nojo, medo, felicidade, tristeza e surpresa⁴.

¹<https://wearesocial.com/sg/blog/2021/07/digital-audiences-swell-but-there-may-be-trouble-ahead/>

²<https://ibm.com/cloud/learn/text-mining>

³<http://cicem.com.br/quem-e-paul-ekman/>

⁴Shiota, Michelle N. (2016). "Teoria das emoções básicas de Ekman". Em Miller, Harold L. (ed.). A enciclopédia Sage de teoria em psicologia . Thousand Oaks, CA: Sage Publications . pp. 248–250 . doi : 10.4135 / 9781483346274.n85 . ISBN 9781452256719.

Nesse contexto, este artigo tem como objetivo realizar uma análise de sentimentos dos *tweets* publicados pelos brasileiros sobre a Covid-19 durante o período de 4 de fevereiro a 7 de março de 2021 baseada nas seis emoções universais de Paul Ekman.

O artigo está organizado da seguinte forma: a Seção 2 descreve a metodologia utilizada na revisão sistemática da literatura; a Seção 3 apresenta os materiais e os métodos utilizados na preparação da base de dados e na construção dos modelos de classificação da análise de sentimento; a Seção 4 apresenta e discute os resultados encontrados; e, por último, a Seção 5 traz as considerações finais e os trabalhos futuros.

2. Revisão Sistemática da Literatura

Este artigo é fundamentado com base em uma revisão sistemática da literatura e direcionado pela seguinte questão: "Quais foram os principais sentimentos dos brasileiros no *Twitter* com relação à Covid-19 durante a pandemia?".

As pesquisas foram baseadas no título e no resumo dos trabalhos e ocorreram no período de outubro a novembro de 2021. Os critérios de inclusão e exclusão são descritos na Tabela 1. As fontes onde as buscas foram realizadas e o conjunto de strings utilizadas são apresentados na Tabela 2.

Tabela 1. Critérios definidos para a revisão sistemática

Critérios de Inclusão	Critérios de Exclusão
Análise de sentimento no <i>Twitter</i>	Análise de sentimento fora do <i>Twitter</i>
Análise de sentimento para o idioma português no <i>Twitter</i>	Análise de sentimento para idiomas que não sejam o português no <i>Twitter</i>
Análise de sentimento no <i>Twitter</i> durante a pandemia da Covid-19	Análise de sentimento aplicada fora do <i>Twitter</i> durante o período da pandemia da Covid-19
Análise de sentimento fazendo uso dos sentimentos da teoria de Paul Ekman	Análise de sentimento sem o uso dos sentimentos da teoria de Paul Ekman

Tabela 2. Fontes e *Strings* de busca utilizadas

Base de dados	Palavra-chave	Resultados
IEEE XPLORE	(Covid19 OR Covid-19) AND (Twitter) AND (Sentimental Analysis OR Opinion Mining) AND (Paul Ekman) - Open Access	78
Google Scholar	(coronavirus)(covid19)(covid-19) (twitter) (sentimental analysis) (Paul Ekman) - Open Access - Artigos de Revisão	111
Google Scholar	(twitter) (sentimental analysis) (análise de sentimento) (análise de sentimentos) (português) - open access	546
Google Scholar	(twitter) (sentimental analysis) (análise de sentimento) (análise de sentimentos) (português)(Paul Ekman)	49
Periódicos Capes	(coronavirus OR covid19 OR covid-19) AND (twitter OR social network OR redes sociais)	698
Science Direct	(sentimental analysis OR opinion mining OR análise de sentimento) AND (twitter) AND (portuguese OR português) - open access	79

2.1. Resultado da Revisão Sistemática da Literatura

Após a pesquisa realizada nas bases de dados IEEE XPLORE, Google Scholar, Science Direct e Periódicos Capes, 1561 artigos foram coletados. Destes, 1546 foram descartados por ao menos um dos seguintes motivos: estar duplicado, extrapolar o objetivo de pesquisa, falta de detalhamento dos procedimentos citados ou não fazer uso da teoria dos sentimentos de Paul Ekman. No total, 11 foram aprovados para a síntese final, dentre esses, 9 são trabalhos correlatos, pois respeitaram os critérios de inclusão dispostos na Tabela 1 e detalharam o procedimento realizado no projeto.

Esta pesquisa deu ênfase na coleta de trabalhos que realizaram a análise de textos em português, visto que a análise de sentimentos desenvolvida nesse projeto tem como foco os *tweets* dos brasileiros. Com isso, dentre os trabalhos correlatos, seis fizeram a análise de *tweets* em português e três em inglês.

Segundo [Drus and Khalid 2019], existem dois métodos principais de análise de sentimentos que foram identificados, um com abordagem de aprendizado de máquina e outro baseado em léxico. Aquele que faz uso do aprendizado de máquina é um modelo supervisionado e requer dados de treinamento para ser processado. Já o método baseado em léxico é conhecido como um modelo de aprendizado não supervisionado e não requer dados de treinamento, dependendo apenas do dicionário de registro de palavras no idioma para o qual foi desenvolvido.

Dentre os projetos de análise de *tweets* selecionados em português, 83% fizeram uso do método baseado em léxico. Uma das principais vantagens dessa abordagem é que ela não carece do tratamento e dos rótulos pré-estipulados, pois é baseada em tratamentos léxicos que envolvem o cálculo da polaridade de um texto a partir da orientação semântica

das palavras nele contidas [Araújo et al. 2013]. Das ferramentas utilizadas para aplicar esse modelo, o pacote *LexiconPT* aparece em três dos seis trabalhos analisados para *tweets* em português e o *NRC Emotion Lexicon*, em dois. Ambos os pacotes trazem dicionários que incluem informações relacionadas a diferentes tipos de sentimentos, permitindo a classificação de textos em determinado idioma. No caso dos pacotes *lexiconPT* e *NRC Emotion Lexicon*, em português e em inglês, respectivamente.

O pacote *lexiconPT* comporta as seguintes bibliotecas: a *OpLexicon V3.0* [Souza and Vieira 2012] e a *SentiLex-PT02* [Silva et al. 2010]. A primeira, é composta por 7014 termos com polaridades pré-definidas e a segunda contém 32191 termos, incluindo emoticons e hashtags. À cada frase, é atribuída uma classificação de sentimento a partir do somatório da polaridade atribuída em cada palavra contida nela. Assim, cada termo pertencente aos dicionários recebe -1 para sentimento negativo, 0 para neutro e 1 para positivo[3].

O pacote *NRC Emotion Lexicon* utiliza a biblioteca *Syuzhet* em *R* [Silva et al. 2022]. Esta fornece a análise de sentimento por cada frase do texto, sendo capaz de retornar 12 sentimentos diferentes, inclusive raiva, nojo, medo, felicidade, tristeza e surpresa, que são os principais sentimentos trazidos nos trabalhos de Paul Ekman. No entanto, é preciso fazer a conversão dos textos em português para inglês e, somente depois, classificá-los.

Quanto aos métodos de aprendizado de máquina, [Giachanou and Crestani 2016] afirmam que os classificadores *SVM* e *Naive Bayes* (NB) têm sido os mais prevalentes dentre as abordagens utilizadas. No caso do NB, os classificadores podem ser otimizados para distribuições de dados e tipos de recursos diferentes, por exemplo: *MultinomialNB* para valores inteiros; *BernoulliNB* para recursos binários; *GaussianNB* para recursos que seguem uma distribuição Gaussiana em todos os dados e *ComplementNB* para conjuntos de dados com diferentes quantidades de dados de treinamento para cada classe [Sittser 2020].

Conforme [Joaquim et al. 2021] outros modelos de classificadores podem utilizar princípios matemáticos, por exemplo, gradiente estocástico e regressão logística como funções de otimização para minimizar rotulações incorretas no treinamento.

2.2. Trabalhos Correlatos

[Pessanha et al. 2020] analisou o sentimento predominante dos brasileiros com relação à pandemia de Covid-19 no período de janeiro a junho de 2020. O pacote *lexiconPT* e a biblioteca *Sentilex-PT02* foram usados na transformação dos termos em símbolos (tokenização) para posterior leitura do algoritmo. Na análise de conteúdo e classificação dos sentimentos, foi adotada a técnica *Bag of Words* para conversão dos textos em vetores, para facilitar o estudo da frequência de todas as palavras distintas presentes no texto. Assim, foi possível dividir duas nuvens de palavras frequentes com teor negativo e positivo sobre a pandemia. No geral, o sentimento negativo foi predominante em todo o período analisado.

[Longo et al. 2020] utilizou o *lexiconPT* e as bibliotecas *OpLexicon V3.0* e a *Sentilex-PT02* para qualificar as impressões da população brasileira a respeito do ensino EAD em postagens do *Twitter*, considerando o contexto atual brasileiro frente à pandemia. A partir da análise de polaridade, [Longo et al. 2020] fizeram uma visualização em

cadeia dos assuntos centrais que mais apareciam nos *tweets* e quais os termos periféricos associados a eles. A conclusão foi que a implementação da educação à distância como ferramenta de emergência no contexto da COVID-19 tem impactado mais negativa do que positivamente tanto o professor quanto o aluno.

[Silva et al. 2022] teve como objetivo identificar o sentimento da população brasileira quanto ao SUS analisando os *tweets* publicados durante a pandemia da COVID-19, assim como realizar comparações dos sentimentos apresentados em diferentes momentos antes e durante a crise. Para realizar a classificação dos sentimentos, foi utilizado o pacote *NRC Sentiment* e a biblioteca *Syzhet* em *R*. Durante todo o período da pesquisa, os usuários sempre falaram em defender o SUS. Nos meses em que os sentimentos negativos foram superiores, esse fator correspondia ao medo demonstrado pelos usuários de que a capacidade de atuação do SUS não seria suficiente para atender a todos em um momento tão crítico.

[Joaquim et al. 2021] utilizaram o modelo Naive Bayes para analisar os sentimentos expressos pelos brasileiros no *Twitter* durante os meses de março a outubro de 2020. Eles aplicaram ainda técnicas de PNL na base de dados e a submeteram ao processamento de diversos algoritmos classificadores de aprendizado de máquina. Do total de *tweets*, 4491 foram classificados e utilizados para treino e teste dos classificadores. Os resultados alcançados foram: precisão de 0.78, revocação variando entre 0.56 e 0.82 e f1-score variando entre 0.57 e 0.80. O sentimento geral da população estudada não sofreu alteração significativa dentre os sentimentos analisados: desgosto, confiança, tristeza e alegria.

3. Material e Métodos

Este projeto utiliza uma estrutura de desenvolvimento baseada nos trabalhos analisados e é organizado em quatro fases:

1. classificação manual dos *tweets*;
2. pré-processamento da base de *tweets*;
3. construção do modelo de classificação; e
4. análise: interpretação e avaliação dos dados.

3.1. Características da base de dados

Esta análise de sentimentos é a continuação do trabalho apresentado em [Simões and Neves 2021] e utiliza a mesma base de dados, que foi minerada no período de 4 de fevereiro a 7 de março de 2021, resultando num total de 7.179 mil *tweets* tratados.

3.2. Classificação manual e pré-processamento da base de *tweets*

Após a classificação manual de sentimentos dos 7.179 mil *tweets*, foi observado que apenas uma parcela desses trazia algum conteúdo sobre Covid-19 ou possíveis sintomas da doença. Sendo assim, os *tweets* que não cumpriram esses requisitos foram removidos em um *Script Python*, restando 4.493 mil *tweets*.

Além disso, nos 4.493 mil *tweets* foi identificado que a maioria era neutro. Ou seja, o conteúdo não apresentou nenhum teor equivalente a negativo e positivo ou aos

sentimentos de Paul Ekman, já que estavam relacionados a pessoas que apresentavam sintomas da doença ou que faziam algum comentário sobre a Covid-19 sem nenhuma conotação positiva ou negativa.

Com base nisso, os *tweets* identificados como neutro na classificação de Paul Ekman foram removidos, pois eles não seriam relevantes para o algoritmo de treinamento, que visa identificar apenas os seis principais sentimentos da teoria do psicólogo. Após a remoção, restaram apenas 1.651 mil *tweets* na base de dados.

Depois dessa etapa, foi identificado que os *tweets* estavam poluídos com caracteres especiais e *Stopwords*, sendo necessário um novo processamento para a remoção desses elementos.

A ferramenta "Localizar e Substituir", do *Google Sheets*, foi utilizada para a remoção dos caracteres especiais. Após isso, todas as palavras com menos de três letras foram removidas em um *Script Python*. Essa limpeza é necessária para aumentar a precisão do algoritmo de treinamento durante a vetorização das palavras mais incidentes em cada *tweet* e também para gerar nuvens de palavras com termos que imprimam significado, como fazem os adjetivos e substantivos.

3.3. Construção do modelo de classificação

Para a análise sentimental dos *tweets*, foi escolhido o método baseado em aprendizado de máquina. A escolha se deve ao fato da minoria dos trabalhos correlatos analisados em português terem utilizado esta técnica. Para isso, foi utilizado o modelo *Naive Bayes* e a biblioteca *naive_bayes*, assim como o classificador *MultinomialNB*, que é adequado para a classificação com recursos discretos, como contagem de palavras para classificação de texto⁵.

A classificação de Polaridade e de Paul Ekman foram separadas em planilhas diferentes e cada uma delas veio acompanhada com os *tweets* correspondentes. Então, foram desenvolvidos dois modelos, um para a classificação de Polaridade e outro para a de Paul Ekman.

A vetorização dos corpos para ambos os modelos foi realizada com o *CountVec-torizer*, que segundo [Joaquim et al. 2021], é uma ferramenta da biblioteca *scikit-learn* que transforma um dado texto em um vetor com base na frequência em que cada palavra ocorre em um texto. O valor de cada célula se resume à contagem da palavra em uma dada amostra de texto, Essa implementação produz uma matriz esparsa das contagens.

No modelo *Naive Bayes* para treinamento da classificação de Polaridade, a base de dados foi dividida em 75% para treinamento e 25% para teste. A vetorização teve um *max-features* de 1000 para limitação do número de rótulos da base e na valização cruzada o número de *k* foi igual a 15.

No modelo de treinamento para os sentimentos de Paul Ekman, a base de dados foi dividida em 70% para treinamento e 30% para teste. A vetorização teve um *max-features* de 1000 para limitação do número de rótulos da base e na valização cruzada o *k* foi igual a 15.

⁵https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

4. Resultados e Discussão

A nuvem de palavras visualizada na Figura 1 apresenta os termos mais utilizados nos 1.651 mil *tweets* analisados. É possível observar um maior destaque para as referências de possíveis sintomas da Covid-19, com destaque para as palavras falta - que costumava estar associada a reclamações de falta de ar -, febre, tosse e covid. Assim, em um contexto geral dos *tweets* analisados, os usuários compartilharam na rede os sintomas que estavam apresentando e questionaram se estariam contaminados pela Covid-19.



Figura 1. Nuvem geral de *tweets*

Durante a classificação manual dos *tweets*, foi observado que o sentimento predominante entre os internautas era o de raiva. Desse modo, além de falarem sobre seus sintomas, os brasileiros usavam um tom de raiva nos textos, demonstrando que estavam irritados com o incômodo e o risco que aqueles sintomas ofereciam para sua saúde e qualidade de vida, além de trazerem a possibilidade de estarem contaminados pelo vírus da Covid-19.

A Figura 2 apresenta os principais termos utilizados nos *tweets* classificados como raiva, com destaque para os termos porra, caralho, inferno, ódio e chata, que foram fundamentais na identificação desse sentimento no texto.



Figura 2. Nuvem de *tweets*: raiva

As duas nuvens de palavras visualizadas na Figura 3 apresentam os principais termos utilizados nos *tweets* classificados como negativos e positivos. Na nuvem de termos mais incidentes nos *tweets* negativos, há maior ocorrência de palavrões e reclamações como ódio, mal, merda e invocações a Deus, geralmente para pedidos de ajuda e força no enfrentamento dos sintomas. Já na nuvem de termos mais incidentes nos *tweets* positivos, é possível observar que os internautas compartilhavam, majoritariamente, a melhora de seus sintomas ou dos de algum conhecido. Assim, as seguintes palavras aparecem com maior frequência: bem, melhor, finalmente, passou, menos, bom, boa e diminuiu.



A Tabela 3 apresenta a validação do modelo de Paul Ekman. O melhor *F1-Score* apresentado foi de 0.67 para a classe raiva e o pior foi de 0.00, para nojo. Como raiva foi o sentimento com o maior campo amostral da base de *tweets* utilizada para o treinamento, o modelo se apresentou mais eficaz na identificação desse sentimento. O contrário aconteceu com o sentimento de nojo, que apresentou o menor campo amostral da base de *tweets*.

Tabela 3. Validação dos Resultados: modelo de Paul Ekman

Modelo	<i>F1-score</i>	<i>Suporte</i>
<i>Felicidade</i>	0.29	144
<i>Medo</i>	0.39	241
<i>Nojo</i>	0.00	11
<i>Raiva</i>	0.67	602
<i>Surpresa</i>	0.07	108
<i>Tristeza</i>	0.58	545

O gráfico na Figura 4 traz a ordem da contagem classificatória de Paul Ekman. O sentimento de raiva aparece 602 vezes e o de nojo apenas 11 vezes. Logo, quanto maior o número de amostras, mais o *F1-Score* de uma classe tende a melhorar, pois o número de termos associados a essa classe aumenta e o modelo se torna mais capaz de identificá-la.

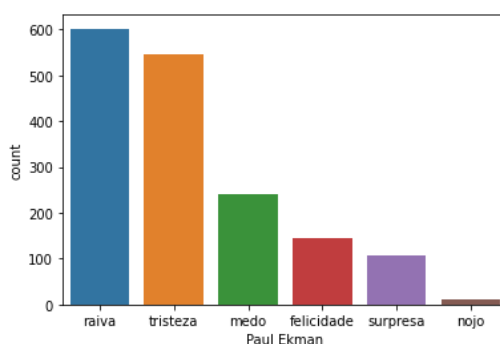


Figura 4. Contagem classificatória de Paul Ekman

A Tabela 4 apresenta a validação do modelo de Polaridade. O melhor *F1-Score* apresentado foi o de 0.69 para a classe negativo e o pior foi de 0.17, para positivo. O desbalanço na precisão do modelo de Polaridade também ocorreu, pois a base amostral dos

tweets com classificação negativa é muito superior aos de classificação positiva. Diferente do modelo de Paul Ekman, que apresentou 6 classes, o de Polaridade tem apenas 3, de modo a acentuar ainda mais o desbalanço do modelo.

Tabela 4. Validação dos Resultados: modelo de Polaridade

Modelo	<i>F1-score</i>	<i>Suporte</i>
<i>Negativo</i>	0.69	938
<i>Neutro</i>	0.47	628
<i>Positivo</i>	0.17	85

O gráfico na Figura 5 traz a ordem da contagem classificatória de Polaridade. Os *tweets* rotulados como negativos são a maioria, aparecendo 938 vezes, já os positivos aparecem apenas 85 vezes.

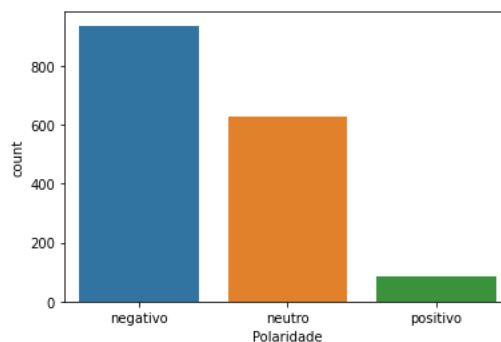


Figura 5. Contagem classificatória de Paul Ekman

5. Considerações Finais e Trabalhos Futuros

O presente trabalho foi fundamentado na revisão sistemática da literatura sobre análises de sentimentos dos brasileiros no *Twitter* com relação à Covid-19.

Esta pesquisa se justifica devido ao aumento dos conflitos sociais que surgiram no Brasil como efeito colateral da pandemia, tornando necessário entender como a população reage aos acontecimentos e às medidas, principalmente governamentais, tomadas frente a eles.

A partir dos 1.546 mil artigos coletados durante a revisão sistemática, 9 foram incluídos na síntese final. Os trabalhos selecionados nortearam a definição das melhores técnicas para a realização de uma análise de sentimentos.

Foi observado que o sentimento predominante dos brasileiros com relação à Covid-19 foi o de raiva, pela classificação sentimental de Paul Ekman, e o negativo, pela classificação de Polaridade. A predominância de tais sentimentos foi muito superior a dos demais, o que tornou a base de dados desbalanceada. Ademais, o sentimento de raiva esteve principalmente associado ao incômodo das pessoas por estarem sentindo algum sintoma da doença.

Devido à base de *tweets* utilizada para o treinamento dos modelos ter sido pequena e desbalanceada, a precisão deles ficou ruim, apresentando o melhor resultado com o *F1-score* de 0.67 para a classificação de Paul Ekman e de 0.69 para a de Polaridade, ainda que tendo sido utilizado o modelo *Naive Bayes*, que apresenta um bom desempenho em análises de sentimentos, conforme investigado na fase de revisão da literatura.

Sendo assim, esta pesquisa colaborou na identificação do perfil sentimental dos brasileiros quanto à pandemia da Covid-19 no período de 4 de fevereiro a 7 de março de 2021, além de pontuar as principais causas desse perfil. Ademais, a construção dos modelos de classificação sentimental permite que sejam desenvolvidas outras análises de sentimentos com maior agilidade.

Ressalta-se a dificuldade na busca por trabalhos de análise de sentimentos utilizando léxico para português, o que influenciou na escolha do uso do aprendizado de máquina na realização desse projeto. Além disso, foram encontradas dificuldades na limpeza da base de *tweets* utilizada, que apresentou muita poluição em seu texto, dificultando o julgamento do teor da mensagem e, posteriormente, a qualidade do modelo de classificação.

Para pesquisas futuras, é sugerido o uso de bases de dados com maior volume de *tweets*, bem como a aplicação de outras técnicas e parâmetros do modelo de classificação *Naive Bayes* para treinamento dos modelos. Assim, será possível comparar os resultados das métricas obtidas após a aplicação desses diferentes parâmetros e técnicas, o que poderá contribuir para a melhoria dos resultados.

Referências

- Araújo, M., Gonçalves, P., and Benevenuto, F. (2013). Métodos para análise de sentimentos no twitter.
- Drus, Z. and Khalid, H. (2019). Análise de sentimento em mídias sociais e sua aplicação: sistemática.
- Giachanou, A. and Crestani, F. (2016). Like it or not: a survey of twitter sentiment analysis methods.
- Joaquim, C. E. L., Barbosa, C. H. M., and Ishikawa, E. (2021). Análise de sentimentos da população brasileira durante a pandemia de covid-19 como ferramenta de exploração da expressão psicossocial no espaço cibernético.
- Longo, C. T. Y., Canovas, D. P. S., Fontes, A. S., Bessegato, L. F., and Santos, A. A. M. (2020). Desafios e repercussões da educação à distância no brasil em tempos de pandemia: um olhar à luz da análise de sentimentos no twitter.
- Pessanha, G. R. G., Fidelis, T. O., Freire, C. D., and Soares, E. A. (2020). fiqueemcasa: análise de sentimento dos usuários do twitter em relação ao covid-19.
- Silva, H., Andrade, E., Araújo, D., and Dantas, J. (2022). Sentiment analysis of tweets related to sus before and during covid-19 pandemic.
- Silva, M. J., Carvalho, P., Costa, C., and Sarmento, L. (2010). Automatic expansion of a social judgment lexicon for sentimento analysis.
- Simões, M. D. and Neves, A. R. M. (2021). Mineração de texto no twitter: uma ferramenta auxiliar na detecção de epidemias.

Sittser, K. (2020). Classification of futuristic technologies described in speculative fiction novels.

Souza, M. and Vieira, R. (2012). Sentiment analysis on twitter data for portuguese language.