

Gustavo Ferreira Lopes Gonçalves

**Análise de Emoções em Tweets Relacionados à
pandemia da Covid-19 no Estado do Rio de
Janeiro**

Niterói, RJ, Brasil

2021

Gustavo Ferreira Lopes Gonçalves

Análise de Emoções em Tweets Relacionados à pandemia da Covid-19 no Estado do Rio de Janeiro

Trabalho submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal Fluminense

Instituto de Computação

Departamento de Ciência da Computação

Orientador: Prof. Antonio Augusto de Aragão Rocha

Coorientadora: Profa. Aline Marins Paes Carvalho

Niterói, RJ, Brasil

2021

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

G635a Gonçalves, Gustavo Ferreira Lopes
Análise de Emoções em Tweets Relacionados à pandemia da Covid-19 no Estado do Rio de Janeiro / Gustavo Ferreira Lopes Gonçalves ; Antonio Augusto de Aragão Rocha, orientador ; Aline Marins Paes Carvalho, coorientadora. Niterói, 2021.
107 f. : il.

Trabalho de Conclusão de Curso (Graduação em Ciência da Computação)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2021.

1. Análise de Emoções. 2. Produção intelectual.I. Rocha, Antonio Augusto de Aragão, orientador. II. Carvalho, Aline Marins Paes, coorientadora. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD -

Gustavo Ferreira Lopes Gonçalves

Análise de Emoções em Tweets Relacionados à pandemia da Covid-19 no Estado do Rio de Janeiro

Trabalho submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Trabalho aprovado. Niterói, RJ, Brasil, 11 de Maio de 2021:


Prof. Antonio Augusto de Aragão
Rocha

Orientador


Profa. Aline Marins Paes Carvalho
Coorientadora


Prof. Alexandre Plastino de Carvalho
UFF


Profa. Simone de Lima Martins
UFF

Niterói, RJ, Brasil

2021

Agradecimentos

Gostaria de agradecer e dedicar esta dissertação às seguintes pessoas:

Minha avó Alda, minha mãe Rose, meu padrasto João e minha namorada Luísa, que sempre me deram todo o tipo de apoio nas horas difíceis.

Meus amigos da universidade Davi, Felipe, Daniel Zambrano, Gabriel Costa, Gabriel Fiuza, Marcos e Vitor, que me ajudaram a seguir com o curso até o fim.

Meus orientadores Aline e Guto, que prestaram todo o auxílio necessário para concluir esta dissertação.

Meus professores da universidade, com quem aprendi lições que certamente lembrei durante minha vida.

Meus colegas de trabalho, pessoas com as quais eu aprendo todos os dias.

À Universidade Federal Fluminense, por ter me dado a oportunidade de cursar e concluir o curso de Ciência da Computação.

Resumo

A pandemia da Covid-19 e suas medidas sanitárias como o isolamento social têm feito a presença de usuários em redes sociais, como Twitter, aumentarem em larga escala nos últimos dois anos. Tal comportamento pode ser explicado a partir das incertezas, inseguranças, preocupações com os graves efeitos da doença em si próprio e em seus entes queridos, aliados a problemas de ordem política no combate à pandemia. Assim, usuários buscam as redes sociais para relatarem suas angústias e compartilham seus pensamentos com outros usuários. Ao analisar tais publicações, é possível ver comportamentos distintos que estão relacionados ao tema da Covid-19 e essa variedade é o nosso objeto de estudo. Assim, essa monografia tem como objetivo coletar tweets e desenvolver análises a partir dos seu conteúdos, a partir de informação estatística e da criação de classificadores. A monografia informa os detalhes sobre como a coleta de tweets foi realizada, como tweets foram manualmente anotados, o uso de técnicas de pré-processamento e transformação de dados, e a criação de classificadores de emoções a partir de métodos de aprendizado de máquina. Em adição, a monografia inclui ainda uma análise do ponto de vista de modelagem de tópicos e observações dos conteúdos dos tweets ao longo do tempo. Esse estudo auxilia no entendimento dos comportamentos e emoções dos usuários do Twitter, resultando em evidências que podem ser usadas em possíveis situações similares à pandemia da Covid-19 que está afetando todos desde 2020.

Palavras-chaves:

Abstract

The Covid-19 pandemic and its sanitary measures such as social isolation have made the presence of users on social networks, such as Twitter, increase on a large scale in the past two years. The uncertainties, insecurities and concerns about the serious effects of the disease on themselves and their loved ones, allied to political issues to combat the pandemic, can explain such behaviour. Thus, users search social networks to report their anxieties and share their thoughts with other users. When analyzing these publications, it is possible to see different behaviours related to the theme of Covid-19, and this variety is our object of study. Thus, this monography aims to collect tweets and develop analyzes based on their content, including statistical analyzes and the creation of emotion classifiers. The monography informs how the collection of tweets was carried out, how tweets were manually annotated, the use of pre-processing and data transformation techniques, and the creation of emotion classifiers using machine learning methods. In addition, the monography also includes an analysis from the point of view of modeling topics and observations of the content of tweets over time. This study helps to understand the behaviours and emotions of Twitter users, resulting in evidence that can be used in possible situations similar to the Covid-19 pandemic that has been affecting everyone since 2020.

Keywords:

Lista de ilustrações

| | |
|---|----|
| Figura 1 – Funcionamento de uma API(HAT, 2020) | 5 |
| Figura 2 – Exemplos de textos a serem representados no Bag of Words(HUILGOL, 2020). | 6 |
| Figura 3 – Representação de Bag of Words para três exemplos distintos(HUILGOL, 2020). | 6 |
| Figura 4 – Funcionamento da técnica de Word Embedding(BORAH, 2021). | 8 |
| Figura 5 – Funcionamento do K-fold(RABELLO, 2019). | 9 |
| Figura 6 – Funcionamento do modelo LDA(BIANCHINI, 2018). | 13 |
| Figura 7 – Roda das Emoções proposta por Robert Plutchik(WIKIPEDIA, 2020a). | 14 |
| Figura 8 – Visão geral sobre o processo de análise de emoções feito nessa monografia. | 22 |
| Figura 9 – Usuário com a tag “Rio de Janeiro” definida em seu perfil. | 23 |
| Figura 10 – Campo localização que pode ser preenchido no perfil de um usuário do Twitter. | 24 |
| Figura 11 – Primeira tela do sistema de coleta com alguns dos Termos de Participação. | 26 |
| Figura 12 – Tabela Participantes. | 26 |
| Figura 13 – Tela 2 apresentando alguns <i>tweets</i> com emoções vinculadas pelo usuário. | 27 |
| Figura 14 – Tela 3 apresentando alguns <i>tweets</i> com emoções vinculadas pelo usuário. | 27 |
| Figura 15 – Tabelas <i>tweets</i> e <i>tweets</i> rotulados | 28 |
| Figura 16 – Frequência de cada emoção no conjunto de dados. | 29 |
| Figura 17 – Frequência de cada emoção no segundo conjunto de dados. | 31 |
| Figura 18 – Exemplo de matriz de confusão da classificação multiclasse no primeiro conjunto de dados. | 37 |
| Figura 19 – Exemplo de matriz de confusão da classificação multiclasse no segundo conjunto de dados. | 39 |
| Figura 20 – Contagem de <i>tweets</i> por semana. | 43 |
| Figura 21 – Nuvens de palavras para as Semanas 1 e 2. | 43 |
| Figura 22 – Nuvens de palavras para as Semanas 3 e 4. | 43 |
| Figura 23 – Nuvens de palavras para as Semanas 5 e 6. | 44 |
| Figura 24 – Nuvens de palavras para as Semanas 7 e 8. | 44 |
| Figura 25 – Nuvens de palavras para as Semanas 9 e 10. | 44 |
| Figura 26 – Nuvens de palavras para as Semanas 11 e 12. | 44 |
| Figura 27 – Contagem de <i>tweets</i> por data. | 47 |
| Figura 28 – Nuvens de palavras para os <i>tweets</i> anteriores e após a Data 1. | 47 |
| Figura 29 – Nuvens de palavras para os <i>tweets</i> anteriores e após as Datas 2 e 3. | 48 |
| Figura 30 – Nuvens de palavras para os <i>tweets</i> anteriores e após a Data 4. | 48 |
| Figura 31 – Nuvens de palavras para os <i>tweets</i> anteriores e após a Data 5. | 48 |

| | |
|--|----|
| Figura 32 – Nuvens de palavras para os <i>tweets</i> anteriores e após a Data 6. | 48 |
| Figura 33 – Nuvens de palavras para os <i>tweets</i> anteriores e após a Data 7. | 48 |
| Figura 34 – Distribuição de emoções para as Semanas 1 e 2. | 50 |
| Figura 35 – Distribuição de emoções para as Semanas 3 e 4. | 51 |
| Figura 36 – Distribuição de emoções para as Semanas 5 e 6. | 51 |
| Figura 37 – Distribuição de emoções para as Semanas 7 e 8. | 51 |
| Figura 38 – Distribuição de emoções para as Semanas 9 e 10. | 52 |
| Figura 39 – Distribuição de emoções para as Semanas 11 e 12. | 52 |
| Figura 40 – Distribuição de emoções para as datas 1 e 2/3. | 52 |
| Figura 41 – Distribuição de emoções para as datas 4 e 5. | 53 |
| Figura 42 – Distribuição de emoções para as datas 6 e 7. | 53 |
| Figura 43 – Contagem normalizada das emoções mais confundidas pelos participantes do processo de anotação manual de emoções. | 54 |
| Figura 44 – Tópicos gerados com o modelo LDA. | 56 |
| Figura 45 – Tópicos gerados com o modelo LDA Mallet. | 57 |

Listas de tabelas

| | |
|--|----|
| Tabela 1 – Identificadores associados a cada emoção do estudo. | 29 |
| Tabela 2 – Parâmetros variados nos classificadores do estudo. | 34 |
| Tabela 3 – Acurárias obtidas para a classificação multiclasse dos <i>tweets</i> do primeiro conjunto de dados. | 37 |
| Tabela 4 – Melhores acurárias obtidas com classificações binárias no primeiro conjunto de dados. | 38 |
| Tabela 5 – Acurárias obtidas para a classificação multiclasse dos <i>tweets</i> do segundo conjunto de dados. | 39 |
| Tabela 6 – Melhores acurárias obtidas com classificações binárias no segundo conjunto de dados. | 40 |
| Tabela 7 – Semanas do conjunto de dados. | 42 |
| Tabela 8 – Datas importantes do conjunto de dados. | 42 |
| Tabela 9 – Bigramas e trigramas de todas as semanas. | 45 |
| Tabela 10 – Bigramas e trigramas obtidos nos <i>tweets</i> antes e depois de cada data importante. | 49 |
| Tabela 11 – Tópicos gerados com o modelo CluWords. | 56 |
| Tabela 12 – Acurárias obtidas para a combinação de Raiva e Ansiedade no primeiro <i>dataset</i> | 70 |
| Tabela 13 – Acurárias obtidas para a combinação de Raiva e Desgosto no primeiro <i>dataset</i> | 70 |
| Tabela 14 – Acurárias obtidas para a combinação de Raiva e Medo no primeiro <i>dataset</i> | 71 |
| Tabela 15 – Acurárias obtidas para a combinação de Raiva e Alegria no primeiro <i>dataset</i> | 71 |
| Tabela 16 – Acurárias obtidas para a combinação de Raiva e Tristeza no primeiro <i>dataset</i> | 71 |
| Tabela 17 – Acurárias obtidas para a combinação de Raiva e Surpresa no primeiro <i>dataset</i> | 72 |
| Tabela 18 – Acurárias obtidas para a combinação de Raiva e Confiança no primeiro <i>dataset</i> | 72 |
| Tabela 19 – Acurárias obtidas para a combinação de Ansiedade e Desgosto no primeiro <i>dataset</i> | 72 |
| Tabela 20 – Acurárias obtidas para a combinação de Ansiedade e Medo no primeiro <i>dataset</i> | 73 |
| Tabela 21 – Acurárias obtidas para a combinação de Ansiedade e Alegria no primeiro <i>dataset</i> | 73 |

| | |
|---|----|
| Tabela 22 – Acurárias obtidas para a combinação de Ansiedade e Tristeza no primeiro <i>dataset</i> | 73 |
| Tabela 23 – Acurárias obtidas para a combinação de Ansiedade e Surpresa no primeiro <i>dataset</i> | 74 |
| Tabela 24 – Acurárias obtidas para a combinação de Ansiedade e Confiança no primeiro <i>dataset</i> | 74 |
| Tabela 25 – Acurárias obtidas para a combinação de Desgosto e Medo no primeiro <i>dataset</i> | 74 |
| Tabela 26 – Acurárias obtidas para a combinação de Desgosto e Alegria no primeiro <i>dataset</i> | 75 |
| Tabela 27 – Acurárias obtidas para a combinação de Desgosto e Tristeza no primeiro <i>dataset</i> | 75 |
| Tabela 28 – Acurárias obtidas para a combinação de Desgosto e Surpresa no primeiro <i>dataset</i> | 75 |
| Tabela 29 – Acurárias obtidas para a combinação de Desgosto e Confiança no primeiro <i>dataset</i> | 76 |
| Tabela 30 – Acurárias obtidas para a combinação de Medo e Alegria no primeiro <i>dataset</i> | 76 |
| Tabela 31 – Acurárias obtidas para a combinação de Medo e Tristeza no primeiro <i>dataset</i> | 76 |
| Tabela 32 – Acurárias obtidas para a combinação de Medo e Surpresa no primeiro <i>dataset</i> | 77 |
| Tabela 33 – Acurárias obtidas para a combinação de Medo e Confiança no primeiro <i>dataset</i> | 77 |
| Tabela 34 – Acurárias obtidas para a combinação de Alegria e Tristeza no primeiro <i>dataset</i> | 77 |
| Tabela 35 – Acurárias obtidas para a combinação de Alegria e Surpresa no primeiro <i>dataset</i> | 78 |
| Tabela 36 – Acurárias obtidas para a combinação de Alegria e Confiança no primeiro <i>dataset</i> | 78 |
| Tabela 37 – Acurárias obtidas para a combinação de Tristeza e Surpresa no primeiro <i>dataset</i> | 78 |
| Tabela 38 – Acurárias obtidas para a combinação de Tristeza e Confiança no primeiro <i>dataset</i> | 79 |
| Tabela 39 – Acurárias obtidas para a combinação de Surpresa e Confiança no primeiro <i>dataset</i> | 79 |
| Tabela 40 – Acurárias de teste obtidas com o BERT no primeiro <i>dataset</i> | 79 |
| Tabela 41 – Acurárias obtidas para a combinação de Raiva e Ansiedade no segundo <i>dataset</i> | 80 |

| | |
|--|----|
| Tabela 42 – Acurárias obtidas para a combinação de Raiva e Desgosto no segundo <i>dataset</i> | 80 |
| Tabela 43 – Acurárias obtidas para a combinação de Raiva e Medo no segundo <i>dataset</i> | 80 |
| Tabela 44 – Acurárias obtidas para a combinação de Raiva e Alegria no segundo <i>dataset</i> | 81 |
| Tabela 45 – Acurárias obtidas para a combinação de Raiva e Tristeza no segundo <i>dataset</i> | 81 |
| Tabela 46 – Acurárias obtidas para a combinação de Raiva e Surpresa no segundo <i>dataset</i> | 81 |
| Tabela 47 – Acurárias obtidas para a combinação de Raiva e Confiança no segundo <i>dataset</i> | 82 |
| Tabela 48 – Acurárias obtidas para a combinação de Ansiedade e Desgosto no segundo <i>dataset</i> | 82 |
| Tabela 49 – Acurárias obtidas para a combinação de Ansiedade e Medo no segundo <i>dataset</i> | 82 |
| Tabela 50 – Acurárias obtidas para a combinação de Ansiedade e Alegria no segundo <i>dataset</i> | 83 |
| Tabela 51 – Acurárias obtidas para a combinação de Ansiedade e Tristeza no segundo <i>dataset</i> | 83 |
| Tabela 52 – Acurárias obtidas para a combinação de Ansiedade e Surpresa no segundo <i>dataset</i> | 83 |
| Tabela 53 – Acurárias obtidas para a combinação de Ansiedade e Confiança no segundo <i>dataset</i> | 84 |
| Tabela 54 – Acurárias obtidas para a combinação de Desgosto e Medo no segundo <i>dataset</i> | 84 |
| Tabela 55 – Acurárias obtidas para a combinação de Desgosto e Alegria no segundo <i>dataset</i> | 84 |
| Tabela 56 – Acurárias obtidas para a combinação de Desgosto e Tristeza no segundo <i>dataset</i> | 85 |
| Tabela 57 – Acurárias obtidas para a combinação de Desgosto e Surpresa no segundo <i>dataset</i> | 85 |
| Tabela 58 – Acurárias obtidas para a combinação de Desgosto e Confiança no segundo <i>dataset</i> | 85 |
| Tabela 59 – Acurárias obtidas para a combinação de Medo e Alegria no segundo <i>dataset</i> | 86 |
| Tabela 60 – Acurárias obtidas para a combinação de Medo e Tristeza no segundo <i>dataset</i> | 86 |
| Tabela 61 – Acurárias obtidas para a combinação de Medo e Surpresa no segundo <i>dataset</i> | 86 |

| | |
|--|----|
| Tabela 62 – Acurárias obtidas para a combinação de Medo e Confiança no segundo <i>dataset</i> | 87 |
| Tabela 63 – Acurárias obtidas para a combinação de Alegria e Tristeza no segundo <i>dataset</i> | 87 |
| Tabela 64 – Acurárias obtidas para a combinação de Alegria e Surpresa no segundo <i>dataset</i> | 87 |
| Tabela 65 – Acurárias obtidas para a combinação de Alegria e Confiança no segundo <i>dataset</i> | 88 |
| Tabela 66 – Acurárias obtidas para a combinação de Tristeza e Surpresa no segundo <i>dataset</i> | 88 |
| Tabela 67 – Acurárias obtidas para a combinação de Tristeza e Confiança no segundo <i>dataset</i> | 88 |
| Tabela 68 – Acurárias obtidas para a combinação de Surpresa e Confiança no se- gundo <i>dataset</i> | 89 |
| Tabela 69 – Acurárias de teste obtidas com o BERT no segundo <i>dataset</i> | 89 |

Lista de abreviaturas e siglas

SVM *Support Vector Machines*

Sumário

| | | |
|--------------|---|-----------|
| 1 | INTRODUÇÃO | 1 |
| 1.1 | Organização do Texto | 3 |
| 2 | FUNDAMENTAÇÃO TEÓRICA | 4 |
| 2.1 | Introdução ao Aprendizado de Máquina | 4 |
| 2.2 | Coleta de dados | 4 |
| 2.3 | Pré-Processamento Textual | 5 |
| 2.4 | Treinamento dos Modelos de Aprendizado de Máquina | 8 |
| 2.4.1 | Separação de dados | 8 |
| 2.4.2 | Avaliação dos dados | 9 |
| 2.4.3 | Classificadores | 10 |
| 2.4.4 | Modelagem de tópicos | 12 |
| 2.5 | Definição dos Rótulos: Roda das Emoções | 13 |
| 3 | TRABALHOS RELACIONADOS | 15 |
| 3.1 | Trabalhos com temas relacionados à Covid-19 | 15 |
| 3.2 | Outros Trabalhos | 19 |
| 4 | ANÁLISE DE EMOÇÕES EM TWEETS SOBRE A COVID-19 | 21 |
| 4.1 | Coleta de tweets | 21 |
| 4.2 | Rotulação manual dos tweets | 24 |
| 4.2.1 | Expansão do dataset | 28 |
| 4.2.2 | Refinamento dos tweets rotulados | 29 |
| 4.3 | Seleção de emoções | 30 |
| 4.4 | Pré-processamento de dados | 31 |
| 4.4.1 | Separação dos dados | 32 |
| 4.5 | Transformação da representação textual para uma representação numérica | 32 |
| 4.6 | Treinamento do modelo | 33 |
| 4.7 | Avaliação do modelo | 35 |
| 5 | RESULTADOS OBTIDOS | 36 |
| 5.1 | Análise de emoções | 36 |
| 5.1.1 | Primeiro conjunto de dados | 36 |
| 5.1.2 | Segundo conjunto de dados | 39 |
| 5.2 | Análise do dataset | 41 |

| | | |
|----------|--|-----------|
| 5.2.1 | Agrupamento por semanas | 41 |
| 5.2.2 | Agrupamento por datas importantes | 46 |
| 5.3 | Análise das distribuições de emoções por período de tempo | 50 |
| 5.4 | Observações sobre a anotação manual de emoções | 54 |
| 5.5 | Modelagem de tópicos | 55 |
| 6 | CONCLUSÃO | 58 |
| 6.1 | Limitações | 59 |
| | REFERÊNCIAS | 61 |
| | APÊNDICES | 69 |
| | APÊNDICE A – RESULTADOS OBTIDOS NA ANÁLISE DE EMOÇÕES | 70 |

1 Introdução

Desde o fim do ano de 2019, o mundo inteiro se deparou com um problema grave: a Covid-19, que, ao que tudo indica até agora, surgiu na cidade de Wuhan, na China (WELLE, 2020a). No entanto, por ser uma doença altamente transmissível, a Covid-19 rapidamente se espalhou pelo mundo em 2020, chegando ao status de pandemia em 11 de Março de 2020 (OMS, 2020), atingindo primeiramente a Europa (WELLE, 2020b) e, após isso, os outros continentes do planeta. Diversas ações foram tomadas para impedir o avanço da doença como: o uso de máscaras e álcool em gel como formas de impedir a proliferação do vírus (OLIVEIRA, 2020), implementação de isolamento social (OLIVEIRA, 2020) e *lockdowns* para impedir a transmissão da Covid durante um certo período de tempo e frear a quantidade de casos (CNS, 2021), a criação de inúmeros leitos para o atendimento de casos graves da doença (JANEIRO, 2020b) e a proibição de serviços e atividades não-essenciais (JANEIRO, 2020a), como aulas presenciais em escolas e universidades, assistir filmes em cinemas e aglomerações em bares e festas. A situação mais parecida com a que vivemos atualmente foi a pandemia da gripe espanhola (NEUFELD, 2020), que ocorreu entre os anos de 1918 e 1920, em uma época que não existiam os recursos tecnológicos de hoje em dia. Como estamos vivendo uma situação praticamente inédita para toda uma geração, cada país reagiu à pandemia de sua própria maneira, mantendo inúmeras discussões sobre o que devemos fazer nessas circunstâncias. Apesar de não haver um consenso geral entre todos os países, existiu e ainda existe um esforço global para que todas as pessoas possam voltar às suas vidas normais o mais rápido possível, como o desenvolvimento recorde inferior a um ano de vacinas para combater a Covid-19 (SANTOS, 2021).

Nessa situação preocupante, nossa atenção está principalmente focada no país em que vivemos, o Brasil. Houve muitos debates sobre o que deveria ser priorizado: a economia do país ou a saúde da população (MARTINS, 2020). Com uma parte das pessoas devendo ficar em isolamento social como uma das medidas aplicadas para conter o avanço da Covid-19, o uso das redes sociais se intensificou ainda mais, seja para debater sobre inúmeros assuntos ou para manter contato com familiares e amigos. Entender como as pessoas estão se relacionando, seus estados emocionais e como a pandemia está afetando cada uma das pessoas é uma tarefa importante e vital para que possamos nos preparar para eventos futuros que possam ser similares ao que estamos vivendo, sendo esse o objetivo principal que o estudo dessa monografia deseja alcançar.

Para compreender os comportamentos dos usuários, existem diversos trabalhos relacionados à área de Aprendizado de Máquina que aplicam soluções como análise de emoções e modelagem de tópicos, como os estudos de Irene Li (LI et al., 2020), László Nemes

e Attila Kiss (NEMES; KISS, 2021), Shanthakumar (SHANTHAKUMAR; SEETHARAM; RAMESH, 2020), Pedro Brum (BRUM et al., 2020), Régis Ebeling (EBELING et al., 2020), Emily Chem (CHEN; LERMAN; FERRARA, 2020), Furqan Rustam (RUSTAM et al., 2021), Martin Müller (MÜLLER; SALATHÉ; KUMMERVOLD, 2020), Tiago de Melo (MELO; FIGUEIREDO, 2021), André Cristiani (CRISTIANI; LIEIRA; CAMARGO, 2020), Deho Oscar (Oscar Deho et al., 2018) e Liza Wikarsa (Indra; Wikarsa; Turang, 2016). É possível encontrar trabalhos anteriores, que foram feitos no Brasil, que analisam períodos de tempo importantes como as eleições presidenciais de 2018 (CRISTIANI; LIEIRA; CAMARGO, 2020) ou estudos que buscam compreender os assuntos mais falados nas redes sociais em um determinado período de tempo (EBELING et al., 2020). No entanto, existem poucos trabalhos que aplicam essas técnicas de Aprendizado de Máquina no contexto da pandemia da Covid-19 no Brasil, principalmente por ser uma situação muito recente para todos nós. Com isso em mente, o trabalho promovido nessa monografia busca aplicar técnicas de Aprendizado de Máquina para dados obtidos em redes sociais durante a pandemia da Covid-19, de forma a analisar o comportamento dos usuários da amostra de dados coletada.

A análise de emoções desenvolvida nessa monografia envolve dados coletados do Twitter na região do estado do Rio de Janeiro. Foram feitas diversas técnicas de transformação de dados em vários classificadores amplamente utilizados em outros estudos, comparando os desempenhos de cada classificador com cada técnica escolhida. Além disso, foram gerados múltiplos indicadores que exibem o comportamento dos usuários ao longo do tempo junto com uma tentativa de modelagem de tópicos para identificar os temas mais comentados pelos usuários do Twitter durante a pandemia da Covid.

Objetivos

Esse estudo tem como objetivo principal a criação de um mecanismo de classificação e análise de emoções para os *tweets* coletados no estado do Rio de Janeiro. Além disso, objetiva-se apresentar análises das características do *datasets*, de forma a entender comportamentos dos usuários do Twitter e ter indicativos de como esses comportamentos podem estar refletidos na população Brasileira de forma geral.

Metodologia

Para atingir os objetivos da monografia, primeiramente, torna-se necessário coletar os dados a serem estudados em mãos. Para isso, será utilizado um método de obtenção de *tweets*, limitando a busca para o estado do Rio de Janeiro. A partir dos *tweets* coletados, espera-se que o conjunto de dados contenha apenas *tweets* relacionados à pandemia da Covid-19.

Para que a análise de emoções pudesse ser feita, é preciso vincular os *tweets* com as emoções do estudo. As emoções serão definidas de acordo com a Roda das Emoções de Robert Plutchik ([WIKIPEDIA, 2020b](#)). Após a definição dos rótulos, procede-se para um processo de rotulação manual de *tweets*, a partir de divulgações em formulários abertos, de forma que qualquer pessoa disposta a ajudar possa anotar os dados. Com a finalização do processo de rotulação, obtém-se uma amostra de *tweets* vinculados a algum tipo de emoção, o suficiente para iniciar a próxima etapa do estudo.

Após a obtenção dos conjuntos de dados, procede-se para as duas próximas etapas do estudo: a análise dos dados de acordo com datas regulares e de acordo com datas associadas a eventos marcantes.

Por fim, procede-se à criação do classificador de emoções, tendo como base o estudo desenvolvido por Irene Zihui Li ([LI et al., 2020](#)). Como naquele trabalho, serão desenvolvidas diversas tarefas de classificação, combinando técnicas de transformação de dados distintas com vários classificadores, como o uso de técnicas clássicas como Bag-of-Words até métodos no estado-da-arte da área de aprendizado de representações para linguagem natural, como BERT .

1.1 Organização do Texto

O restante deste trabalho está organizado em seções de Fundamentação Teórica (Capítulo 2), que introduz conceitos usados ao decorrer do trabalho, Trabalhos Relacionados (Capítulo 3), correlacionando o trabalho atual com outros estudos similares, Análise de Emoções em *Tweets* Sobre a Covid-19 (Capítulo 4), que explicita o desenvolvimento feito nessa monografia, Resultados Obtidos (Capítulo 5), mostrando todos os indicadores gerados após o desenvolvimento desse estudo e a Conclusão (Capítulo 6).

2 Fundamentação Teórica

O estudo promovido nessa monografia abordou diversas técnicas para que seu objetivo fosse alcançado, como o uso de técnicas de coleta de dados, processamento de linguagem natural e recuperação de informações.

2.1 Introdução ao Aprendizado de Máquina

Machine Learning ou Aprendizado de Máquina é um campo de estudo da área de Inteligência Artificial que ensina computadores a analisar e classificar padrões de dados para fazer previsões. Pode-se resumir *Machine Learning* à simulação de um processo natural para seres humanos: aprender com a experiência.

O Aprendizado de Máquina é uma programação de sistemas para assimilar dados e classificar informações complexas, com caracterização da aprendizagem, para apresentar previsões e estimativas sobre esses dados.

Com *Machine Learning*, é possível aumentar a capacidade humana de resolver problemas e se antecipar a riscos, com base nos resultados levantados pelos programas. Esse campo de estudo possui usos que vão desde diagnósticos médicos, previsões do tempo e identificação de mudanças climáticas até análises e deduções sobre o mercado de ações (RIBEIRO, 2018).

2.2 Coleta de dados

Naturalmente, o primeiro passo em um processo de descoberta de conhecimento a partir de dados é a definição de um conjunto de dados. No caso dessa monografia, o foco será na descoberta de conhecimento a partir de publicações na rede social Twitter. Para tanto, faz-se o uso de uma API, conceito bastante utilizado e difundido em diversos softwares atualmente.

API (Application Programming Interface) é um conjunto de definições e protocolos usado no desenvolvimento e na integração de software de aplicações. Uma API é uma espécie de meio de comunicação com algum produto ou serviço sem a necessidade de entender sua implementação, simplificando o desenvolvimento de novas aplicações (HAT, 2020). Geralmente, APIs costumam apresentar documentações vinculadas para que algum interessado possa ser capaz de estruturar uma solicitação e enviá-la de acordo com o que foi especificado nas documentações e, com base nessa solicitação, receber uma resposta apropriada da API (HAT, 2020).

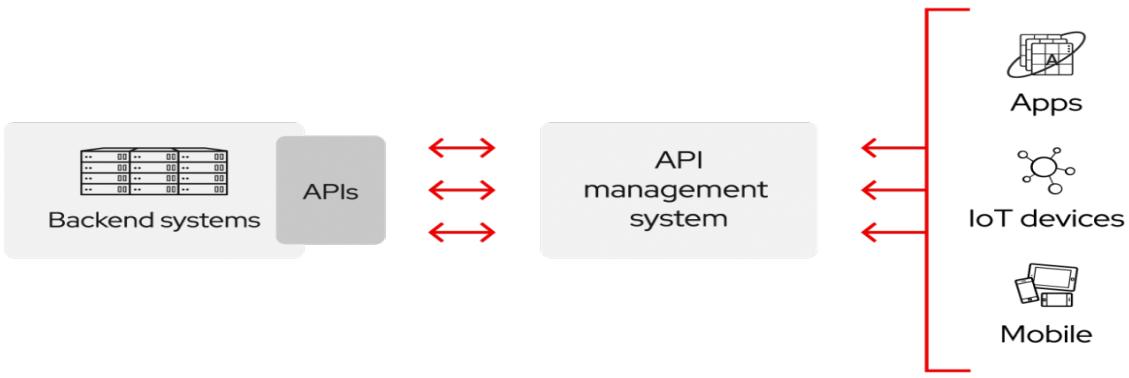


Figura 1 – Funcionamento de uma API([HAT, 2020](#)).

As APIs são uma maneira simplificada de conectar a própria infraestrutura por meio do desenvolvimento de aplicações. Além disso, elas também possibilitam o compartilhamento de dados com clientes e outros usuários externos. Esse compartilhamento é feito sem abrir mão da segurança e do controle, cabendo ao desenvolvedor da API determinar como gerenciar esses aspectos ([HAT, 2020](#)).

Existem três tipos de API: privada, de parceiros e pública. Uma API privada é usada apenas para usuários internos, enquanto uma API pública pode ser usada também por usuários externos. Já uma API de parceiros é compartilhada apenas com parceiros de negócio específicos ([HAT, 2020](#)).

A Figura 1 mostra que requisições podem ser feitas por várias origens diferentes, como aplicativos, dispositivos IoT e dispositivos móveis. Toda requisição passa por um sistema que faz o gerenciamento da API que, por sua vez, é desenvolvida em algum sistema *back-end*.

2.3 Pré-Processamento Textual

Para a extração e análise das informações relevantes para atingir o objetivo dessa monografia, é necessário citar a técnica de Pré-processamento de dados, sendo uma das fases mais importantes de um projeto de aprendizado de máquina ([WIKIPEDIA, 2018a](#)). Além de remover dados irrelevantes, essa técnica também auxilia na remoção de dados que possam impactar negativamente os resultados obtidos com o aprendizado de máquina. Pode-se dividir esse processo em diversos componentes: limpeza de dados, transformação de dados e redução de dados. A limpeza de dados envolve a remoção de caracteres especiais, falhas, etc. A transformação de dados está relacionada com a conversão dos dados originais para formatos mais apropriados, a partir de métodos de normalização ou discretização de valores numéricos, a tokenização de palavras, etc. Por fim, a redução de dados tem o objetivo de diminuir a complexidade do processamento computacional, facilitando a manipulação de grandes volumes de dados, como o uso de lematização ([GOMES, 2019](#)), a

- Review 1: This movie is very scary and long
- Review 2: This movie is not scary and is slow
- Review 3: This movie is spooky and good

Figura 2 – Exemplos de textos a serem representados no Bag of Words([HUILGOL, 2020](#)).

| | 1 This | 2 movie | 3 is | 4 very | 5 scary | 6 and | 7 long | 8 not | 9 slow | 10 spooky | 11 good | Length of the review(in words) |
|-------------|-----------|------------|---------|-----------|------------|----------|-----------|----------|-----------|--------------|------------|--------------------------------------|
| Review 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 7 |
| Review 2 | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 8 |
| Review 3 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 6 |

Figura 3 – Representação de Bag of Words para três exemplos distintos([HUILGOL, 2020](#)).

remoção de *stopwords* ([WIKIPEDIA, 2021b](#)) e de palavras muito ou pouco frequentes. As stopwords são palavras que não possuem valor semântico para uma determinada análise a ser feita nos dados e, normalmente, envolve palavras que são comumente usadas pelas pessoas, tornando essas palavras irrelevantes no contexto em que se encontram.

Transformação de dados

Para representar os textos de *tweets* em representações numéricas, é necessário utilizar técnicas de transformação aplicadas nos dados de entrada. O primeiro método escolhido foi o Bag-of-Words. No método Bag-of-Words ([SANTANA, 2020](#)), qualquer informação sobre a gramática ou a ordem das palavras é descartada, transformando o conjunto de palavras em um conjunto de termos com suas respectivas frequências. Para exemplificar essa técnica, podemos observar as Figuras 2 e 3.

Com essa técnica, cada *tweet* é separado em palavras, onde cada uma dessas palavras terá uma frequência calculada para todo o conjunto de *tweets*. Na tabela da Figura 3, é possível observar as frequências associadas a cada palavra em cada frase explicitada na Figura 2. A matriz é esparsa, fazendo com que ela apresente diversos valores 0 pois nem todas as palavras da matriz estarão presentes em cada um das revisões analisadas. Quanto mais revisões a serem representadas, maior será a matriz produzida pela técnica. Para implementar o Bag-of-Words, foi necessário utilizar o CountVectorizer ([LEARN, 2021e](#)) encontrado no scikit-learn ([PEDREGOSA et al., 2011](#)).

Além do Bag-Of-Words, outra técnica de transformação dos dados textuais para valores numéricos bastante utilizada é o TF-IDF. O TF-IDF, assim como Bag of Words, seria utilizado para medir a importância das palavras de cada *tweet*. No entanto, nessa representação, são utilizadas medidas estatísticas: *Term Frequency* (a frequência do termo), que mede a frequência com que um termo ocorre num documento, e *Inverse Document Frequency* (inverso da frequência nos documentos), que mede o quanto importante um termo é no contexto de todos os documentos (FONSECA, 2020).

$$TFIDF = TF(i, j) * IDF(i) \quad (2.1)$$

$$TF(i, j) = \frac{\text{Frequência do termo } i \text{ no documento } j}{\text{Total de palavras no documento } j} \quad (2.2)$$

$$IDF(i) = \log_2 \frac{\text{Total de documentos}}{\text{documentos com o termo } i} \quad (2.3)$$

As Equações 2.1 a 2.3 apresentam os cálculos feitos para *Term Frequency* e *Inverse Document Frequency*, onde TF é a frequência do termo (Term Frequency), IDF é o inverso da frequência nos documentos (Inverse Document Frequency), i é um termo e j é um documento. TF é o resultado da divisão entre a frequência de um termo *i* dentro de um documento *j* e o total de palavras do documento *j*. Já o IDF corresponde ao logaritmo, na base dois, da divisão entre o total de documentos e o número de documentos contendo o termo *i*.

Uma terceira e última técnica de transformação foi escolhida: Word Embedding. Aqui, as palavras do conjunto de dados são mapeadas como vetores densos de baixa dimensionalidade de números reais, fazendo com que palavras com o mesmo significado tenham vetores similares, enquanto palavras com significados bem distintos tenham vetores que representem essa discrepância (BROWNLEE, 2019).

A Figura 4 demonstra como a técnica de Word Embedding se comporta com diferentes dados de entrada. Para cada palavra, é construído um vetor (os embeddings) contendo diversos valores reais, onde cada dimensão pode ser vista como representando um atributo distinto para aquela palavra. Após a criação dos embeddings, o vetor pode ter suas dimensões reduzidas conforme a representação de dados for escolhida. No caso do exemplo, foi escolhida uma redução para duas dimensões para facilitar a visualização da similaridade entre as palavras. Com essa redução, podem-se visualizar os dados e observar que as palavras gato (*cat*) e gatinho (*kitten*) são bem mais próximas quando comparadas às palavras cachorro (*dog*) e casas (*houses*). As palavras gato e cachorro também são bem mais próximas, quando a comparação é feita entre essas palavras e a palavra casa. Na outra visualização, é possível notar que homem (*man*) e mulher (*woman*) possuem um

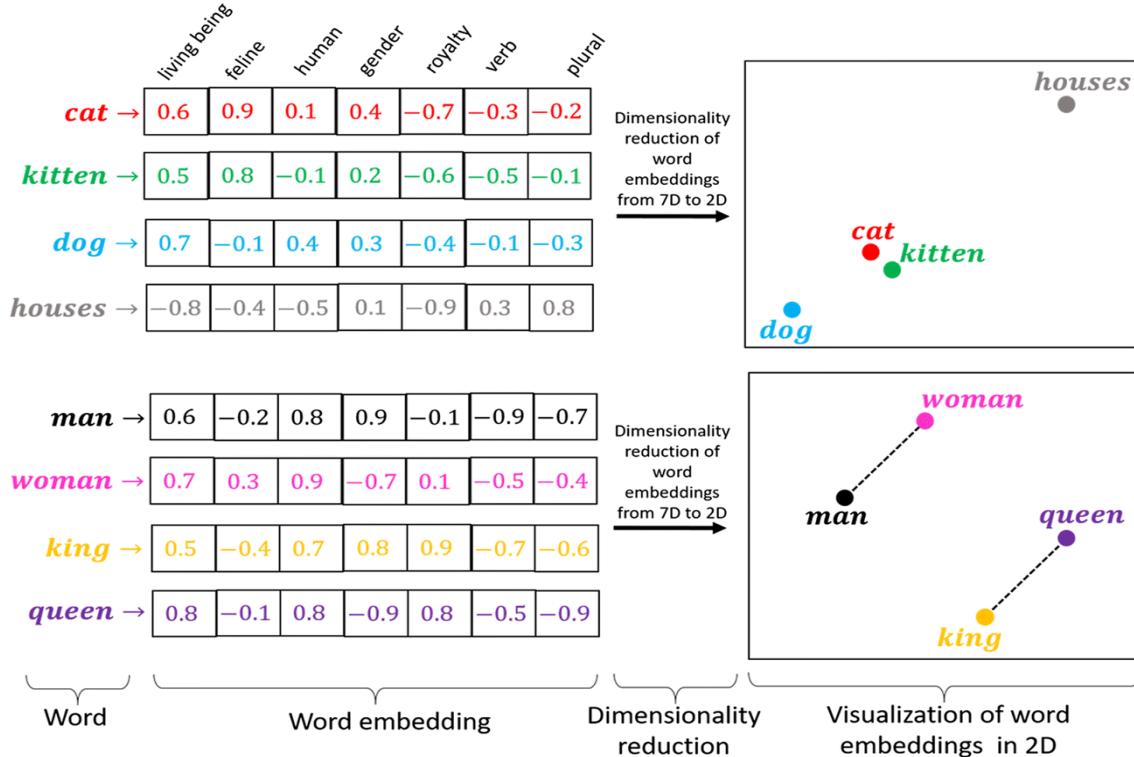


Figura 4 – Funcionamento da técnica de Word Embedding(BORAH, 2021).

grau de similaridade equivalente ao encontrado entre rei (*king*) e rainha (*queen*). Também é possível observar que homem e rei são tão próximos quanto mulher e rainha.

O algoritmo **BERT** (Bidirectional Encoder Representations from Transformers) (TEAM, 2021) é uma técnica de pré-treinamento de processamento de linguagem natural baseada em redes neurais criada pela Google e lançada em 2018. Com essa arquitetura, os modelos gerados processam o contexto completo de palavras pesquisadas, usando todos os termos em um texto, enquanto também pode apresentar resultados de mais qualidade para pesquisas com mais palavras (VOLPATO, 2019). Esse algoritmo se assemelha à técnica de Word Embeddings no sentido que utiliza vetores de números reais para representar os contextos de diversas palavras. A diferença é que esses vetores no BERT consideram o contexto das palavras mais próximas, enquanto os vetores de Word Embedding carregam significados para a palavra de forma isolada.

2.4 Treinamento dos Modelos de Aprendizado de Máquina

2.4.1 Separação de dados

Normalmente, os dados coletados são separados em duas categorias: dados de treinamento, que serão utilizados para o treinamento do modelo de Aprendizado de Máquina, e dados de teste, que serão utilizados para verificar seu desempenho sob condições

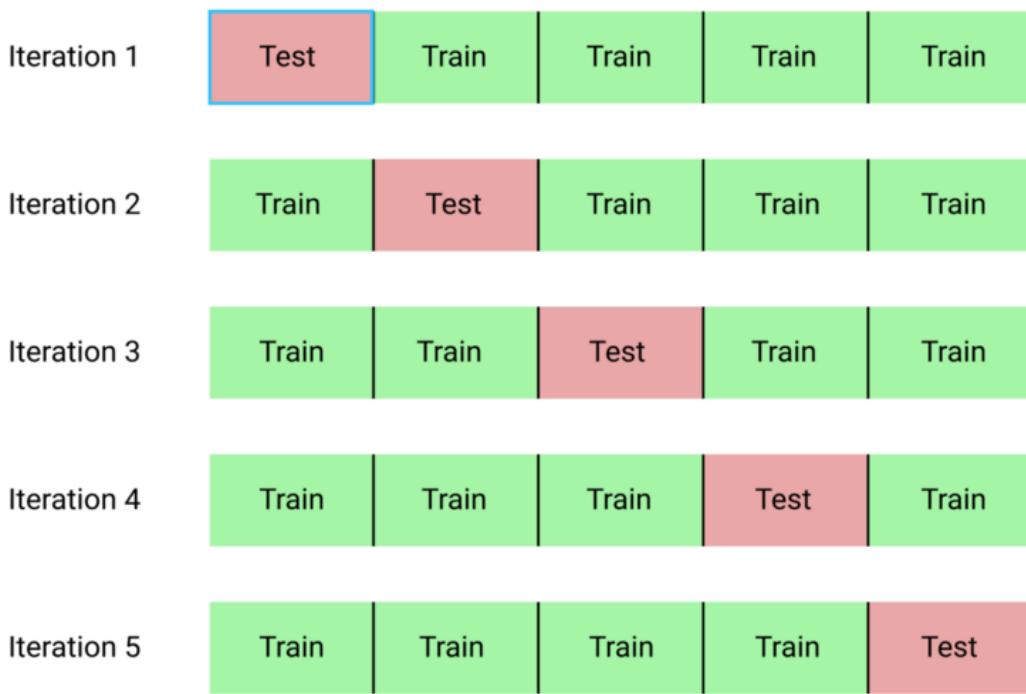


Figura 5 – Funcionamento do K-fold([RABELLO, 2019](#)).

reais de utilização. Além dessa divisão, pode-se usar também uma subdivisão do conjunto de treinamento, criando um conjunto de validação, utilizado para verificar a eficiência da rede quanto a sua capacidade de generalização durante o treinamento, e podendo ser empregado como critério de parada do treinamento ([CARVALHO, 2020](#)).

Para separar os dados em conjuntos de treinamento e de teste, é necessário escolher uma técnica que cuidará dessa parte em específico. Nesse trabalho, foi utilizada a *K-fold cross validation* ([LEARN, 2021i](#)). *K-fold* se baseia na divisão aleatória da base de dados em *K* grupos com uma quantidade de amostras similar em cada grupo. O valor *K* é definido na inicialização dessa técnica. Esses grupos são iterados, fazendo com que, a cada iteração, um deles seja escolhido como o conjunto de teste e com que os *K*-1 grupos restantes formem o conjunto de treinamento. Essas iterações garantem que todos os grupos serão utilizados como treinamento ou teste em algum momento da avaliação do modelo. A Figura 5 representa o funcionamento do *K-fold*.

2.4.2 Avaliação dos dados

Ao realizar uma tarefa de Aprendizado de Máquina e avaliar seus resultados, é possível que ocorram problemas de generalização no modelo desenvolvido. O primeiro deles é o *Underfitting*, que significa que o modelo utilizado não conseguiu aprender o suficiente com os dados de entrada. Os resultados do modelo apresentam um erro elevado tanto nos dados de treino quanto nos dados de teste. Outro problema encontrado é o *Overfitting*,

que ocorre quando o modelo aprende demais sobre os dados de treinamento dados como entrada. Neste caso, o modelo mostra-se adequado apenas para os dados de treino, como se o modelo tivesse apenas decorado os padrões desses dados e não fosse capaz de generalizar para outros dados nunca vistos antes. Quando isso acontece, os dados de treino apresentam resultados excelentes, enquanto que a performance do modelo cai drasticamente com os dados de teste (BRANCO, 2020).

2.4.3 Classificadores

Esse trabalho apresenta uma análise de emoções feita com diversos classificadores utilizando as técnicas citadas anteriormente, de forma a obter os mais variados resultados possíveis.

O primeiro classificador abordado é o **Random Forest** (LEARN, 2021d). O RF é um modelo *ensemble* que possui, como diferencial, uma combinação de diferentes modelos (no caso, árvores de decisão) para se obter um único resultado. No caso desse classificador, são usadas várias árvores de decisão, que são construídas com base em dois passos: seleção de amostras feitas com o método bootstrap, que é utilizado para a seleção de amostras de treino, e seleção de variáveis para cada nó da árvore, escolhendo-se essas variáveis de forma aleatória (TECH, 2020).

O segundo classificador usado é o método **kNN** (LEARN, 2021j). Quando esse método envolve tarefas de classificação, uma classe é calculada através da maior frequência das K instâncias mais similares à instância analisada. Para evitar casos de empates, o ideal é escolher um valor de K par quando o número de classes a serem classificados for ímpar ou escolher um K ímpar quando a quantidade de classes for par (ILEOH, 2018).

O algoritmo **Passive Aggressive** (LEARN, 2021g) é um algoritmo de aprendizado de máquina online e é bastante utilizado com grandes quantidades de dados. O aprendizado é feito a partir de cada dado de entrada (passo a passo). Caso a predição feita esteja correta, a estrutura do modelo é mantida e, caso contrário, mudanças são aplicadas para que a predição seja feita com sucesso (ALOKESH985, 2020).

O algoritmo **Naive Bayes** (LEARN, 2021b) é utilizado para tarefas de classificação e se baseia na probabilidade de eventos ocorrerem com base nas atributos que são extraídas de cada dado de entrada. É um algoritmo bem simples e que precisa de poucos ajustes para se obter uma acurácia satisfatória (TAMAIIS, 2019).

Já o algoritmo **Gradient Boosting** (LEARN, 2021c) tem como objetivo criar um *ensemble* de modelos fracos, geralmente árvores de decisão, onde cada um minimiza o erro do anterior, através de uma função de perda, até alcançar um resultado satisfatório. Para cada ajuste aplicado nesses modelos, multiplica-se uma taxa de aprendizagem, medindo o impacto de cada árvore no modelo final, onde valores pequenos possuem contribuições

pequenas no modelo gerado. Esse algoritmo pode ser usado em problemas de regressão e classificação (SILVA, 2020).

De forma similar ao **Gradient Boosting** (LEARN, 2021c), o algoritmo **XGBoost** (XGBOOST, 2021) também utiliza um conjunto de modelos fracos para alcançar um modelo final capaz de fornecer previsões a uma amostra de dados. No entanto, **XGBoost** possui uma funcionalidade a mais: cada amostra recebe pesos diferentes de acordo com os ajustes feitos no algoritmo. Além disso, a construção desses modelos fracos é feita de forma paralela, enquanto no **Gradient Boosting** é feita sequencialmente, permitindo com que a escalabilidade e a performance do **XGBoost** sejam melhores (RUSTAM et al., 2021).

Logistic Regression (LEARN, 2021f) é um algoritmo de classificação amplamente utilizado para estimar valores discretos, como valores binários, sim/não ou verdadeiro/falso se baseando em um grupo de variáveis independentes. Em outras palavras, o algoritmo prevê a probabilidade da ocorrência de um evento através de ajustes nos seus parâmetros, retornando saídas com valores esperados entre 0 e 1. Usam-se logaritmos para calcular as probabilidades de cada evento (ADMINVOOO, 2016).

Um **Multilayer Perceptron** (**MLP**) (LEARN, 2021k) é uma rede neural artificial composta por múltiplas camadas de Perceptrons. Esse modelo é composto por uma camada de entrada para receber o sinal, um número variável de camadas ocultas e uma camada de saída responsável por tomar uma decisão com base nos dados de entrada (TINÓS, 2018). Para aprender os pesos da rede neural, aplica-se o algoritmo de *Backpropagation*. A ideia do algoritmo *Backpropagation* é, com base no cálculo do erro ocorrido na camada de saída da rede neural, recalcular o valor dos pesos do vetor w da última camada de neurônios e assim proceder para as camadas anteriores, de trás para a frente, ou seja, atualizar todos os pesos w das camadas a partir da última até atingir a camada de entrada da rede (LEITE, 2018).

O algoritmo **Support Vector Machine** (**SVM**) (LEARN, 2021a) se baseia no princípio de separação ótima entre classes, de forma que caso as classes sejam separáveis, a solução escolhida separa as classes do problema o máximo possível. Para cada amostra de treinamento, uma representação vetorial é construída com uma classe associada. Esse processo envolve uma distribuição de probabilidade de onde os dados de treinamento são retirados, fazendo com que o classificador aprenda como mapear cada vetor e sua classe para alguma distribuição de probabilidades das amostras de treinamento (MOREIRA et al., 2014).

Para concluir a lista de classificadores usados, usou-se a camada de que se acopla ao **BERT** (TEAM, 2021). Nesse modelo, para realizar uma tarefa de classificação, é necessário colocar uma nova camada acima do modelo já construído. Então, é possível ajustar os pesos aprendidos anteriormente usando a tarefa de classificação e uma técnica chamada de *Fine-tuning*.

2.4.4 Modelagem de tópicos

Ao longo do estudo desenvolvido, foram feitas análises sobre o conjunto de dados obtido. Em um determinado momento, ocorreram tentativas de se utilizar algoritmos relacionados a área de Modelagem de Tópicos para que subconjuntos de temas pudessem ser identificados e caracterizados de forma apropriada.

O primeiro algoritmo usado foi o **Latent Dirichlet Allocation (LDA)**. O LDA é um modelo que se baseia em métodos estatísticos, onde cada tópico identificado é representado por um conjunto de palavras e um documento é representado por um conjunto de tópicos.

A Figura 6 representa como o algoritmo **LDA** funciona. Recebe-se uma quantidade **m** de documentos que, por sua vez, geram uma quantidade **n** de palavras. Essas palavras geram **k** tópicos, que podem ser entendidos como clusters de palavras. Os principais parâmetros para esse algoritmo são α e β . O parâmetro α especifica a quantidade de tópicos que compõem os documentos de entrada. Quanto maior a quantidade de tópicos escolhida, mais específica será a distribuição obtida pelo algoritmo. Já o β indica a quantidade de palavras que compõem um tópico. Quanto maior a quantidade de palavras escolhida, mais específica será a distribuição (CASTRO, 2020).

O outro algoritmo usado foi o CluWords (VIEGAS et al., 2019b). CluWords é uma representação de Word Embedding pré-treinada para uma fatoração de matriz não probabilística. O algoritmo explora as palavras mais próximas de um vetor de Word Embedding pré-treinado para gerar “palavras-meta” que são capazes de melhorar a representação de documentos, seja em termos sintáticos quanto semânticos. A combinação de evidências sintáticas (ocorrência de palavras em um documento) com a similaridade entre uma palavra e suas palavras vizinhas é feita através de um parâmetro α que balanceia os pesos de cada um desses aspectos.

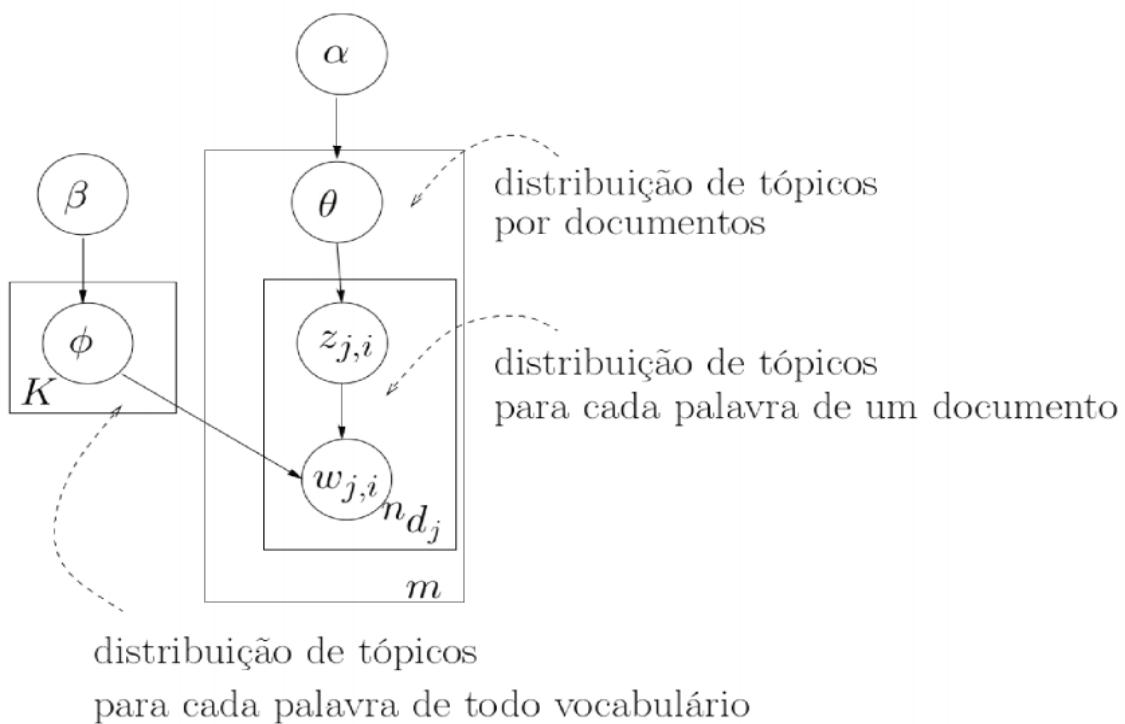


Figura 6 – Funcionamento do modelo LDA([BIANCHINI, 2018](#)).

2.5 Definição dos Rótulos: Roda das Emoções

Robert Plutchik foi um psicólogo norte-americano que desenvolveu um modelo das emoções baseando-se na teoria da psicologia evolutiva das emoções ([ALABAU, 2020](#)). Para ele, as emoções podem ser divididas em oito emoções básicas: alegria, tristeza, antecipação, surpresa, irritação, medo, confiança e nojo. Com base nesses conceitos, ele construiu a Roda das Emoções.

Segundo a roda apresentada na Figura 7, existem algumas combinações de emoções que são opostas entre si: alegria versus tristeza, raiva versus medo, confiança versus nojo e surpresa versus antecipação. As emoções primárias podem ser expressas em diferentes intensidades e podem se misturar entre si para formar emoções diferentes, como remorso, intimidação e amor ([WIKIPEDIA, 2020b](#)).

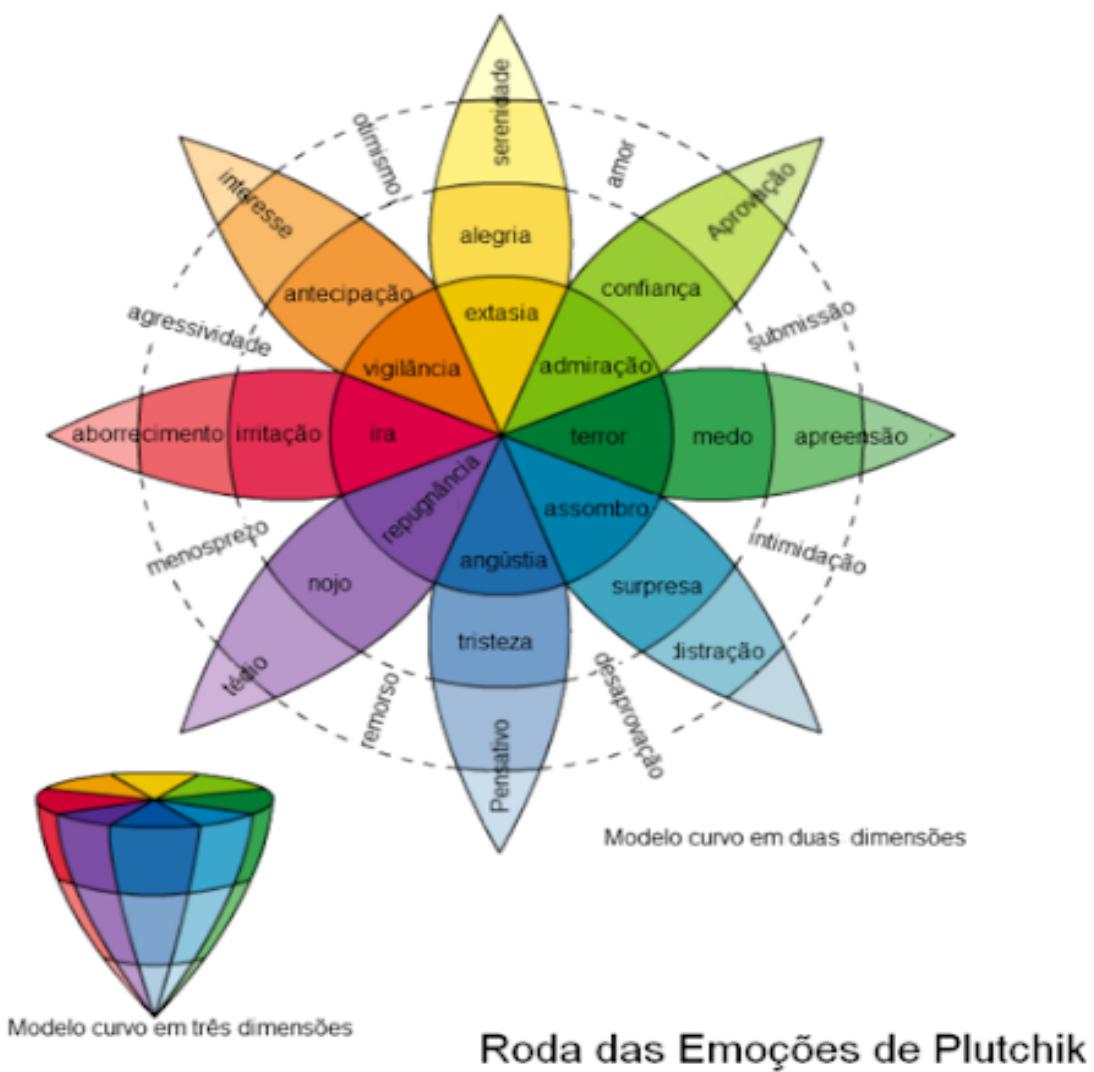


Figura 7 – Roda das Emoções proposta por Robert Plutchik ([WIKIPEDIA](#), 2020a).

3 Trabalhos Relacionados

Essa seção detalha estudos similares ao que será visto nesta monografia. Os trabalhos dessa seção foram divididos em dois grupos distintos: estudos com temas relacionados à Covid-19 e estudos com outros temas.

3.1 Trabalhos com temas relacionados à Covid-19

Na área de Aprendizado de Máquina, existem diversas pesquisas e estudos desenvolvidos no passado ou em processo de desenvolvimento atualmente, focando em assuntos como análises de emoções, modelagem de tópicos, processamento de imagens, etc. Os artigos separados nessa seção participam do mesmos campos interdisciplinares aplicados nessa monografia. Inicialmente, separaram-se os trabalhos mais similares utilizando dados relacionados à pandemia da Covid-19 vivida atualmente.

O principal trabalho relacionado ao estudo conduzido nessa monografia foi investigado em ([LI et al., 2020](#)). O artigo apresentou uma análise de emoções para uma base de dados com mais de oito milhões de *tweets* originados de diversos países do mundo, como os Estados Unidos, Índia, Inglaterra e Brasil. A coleta dos dados foi realizada através da API oficial do Twitter e consistiu em uma busca de *tweets* por uma série de palavras chaves como “coronavirus”, “covid19”, “confinamiento”, “flu”, “fever”, “cough”, “social distance”, “lockdown”, “pandemic”, “epidemic”, “conlabelious” e “infection”. Cada um dos *tweets* foi publicado entre os dias 24 e 26 de Março de 2020.

Com essa base, os autores sugeriram o treinamento de um classificador de oito emoções a partir dos embeddings de palavras encontrados pelo BERT. Além disso, também sugeriram a aplicação de um modelo BERT(ft) para que as performances das classificações de ambos os modelos pudessem ser comparadas. Os *tweets* foram classificados em, ao menos, uma das seguintes emoções: raiva, ansiedade, desgosto, medo, alegria, tristeza, surpresa e confiança. A anotação de classe dos *tweets* foi estruturada de duas formas diferentes. A primeira estrutura considerou exemplos com apenas um rótulo por exemplo. Nesse caso, cada uma das emoções recebeu um identificador de 0 a 7. Os resultados obtidos foram bastante satisfatórios, alcançando acuráncias acima de 95%. A segunda forma de anotação dos exemplos considerou um problema multirótulo. Nesse segundo caso, existiram três colunas de emoções associadas a cada *tweet*, onde cada uma dessas colunas foi preenchida com um identificador de 0 a 7. Por fim, os resultados obtidos não foram tão satisfatórios quanto os anteriores, alcançando acuráncias acima de 64%.

Os estudos de László Nemes e Attila Kiss ([NEMES; KISS, 2021](#)) também envolveram

a criação de um classificador de emoções para *tweets* relacionados a COVID-19, assumindo graus de sentimentos: positivo, fracamente positivo, fortemente positivo, neutro, fracamente negativo, fortemente negativo e negativo. Eles fizeram a coleta de *tweets* através da biblioteca tweepy que possui um vínculo com a API do Twitter. Todos os *tweets* foram classificados utilizando o modelo RNN (*Recurrent Neural Network*) (MONTREAL, 2016) e a biblioteca TextBlob. TextBlob é uma biblioteca usada para processamento textual. Ela disponibiliza uma API para diversas tarefas de PLN como a rotulação de part-of-speech, extração de substantivos e expressões, análise de sentimentos, classificação, tradução, etc (LORIA, 2018). Com os resultados das classificações, diferentes distribuições de sentimentos por amostras de *tweets* foram feitas em diversos meses do ano de 2020. Existiram exemplos em que a maior parte das emoções foram classificadas como positivas (fortemente positiva, positiva ou levemente positiva) e alguns exemplos que a quantidade de *tweets* neutros foi bastante considerável, alcançando patamares de quase 40%.

O artigo de Shanthakumar (SHANTHAKUMAR; SEETHARAM; RAMESH, 2020) classifica o seu *dataset* de mais de 500.000 *tweets*, obtido através da API do Twitter, em seis grupos diferentes que foram separados pelo uso de diferentes hashtags: assuntos gerais sobre Covid, quarentena, fechamento de escolas, compras feitas em momentos de pânico, lockdown e frustração/esperança. Todos os *tweets* foram coletados entre os dias 14 e 24 de Março de 2020. Diversas distribuições temporais desses grupos foram feitas, indicando pontos interessantes: os grupos de compras feitas em momentos de pânico e fechamento de escolas tiveram um pico de frequência de *tweets* nos primeiros dias estudados e, após isso, essas frequências apresentaram uma queda; diversos *tweets* mencionando quarentena, lockdown e assuntos gerais sobre Covid foram identificados ao longo do tempo, apresentando frequências altas; o grupo relacionado a frustração/esperança teve um grande crescimento ao longo do tempo, indicando que as pessoas começaram a ficar ainda mais frustradas ao decorrer dos dias.

No tema de Aprendizado de Máquina e tarefas relacionadas, foi feita uma análise de emoções utilizando uma variante do BERT desenvolvida pelo Facebook chamada “RoBERTa” (LIU et al., 2019), mais potente do que a versão tradicional do BERT, classificando os *tweets* entre diferentes graus de positividade ou negatividade de emoções. Nos grupos separados manualmente, a maior parte dos sentimentos identificados foram positivos, com exceção do grupo de compras feitas em momentos de pânico, onde o maior número de *tweets* com emoções negativas indica a frustração dos usuários em relação a esse tema. Por fim, foi feita uma modelagem de tópicos para efeitos de comparação com os grupos separados manualmente, utilizando-se o modelo “Seeded LDA” (JAGARLAMUDI; III; UDUPA, 2012). Seeded LDA provê um pequeno conjunto de palavras para guiar a descoberta de tópicos, influenciando tanto no tópico do documento quanto na distribuição de palavras por tópico. Essas palavras não precisam ser exaustivas, uma vez que o modelo é capaz de detectar outras palavras na mesma categoria através de ocorrências simultâneas

no *dataset*. Tentou-se gerar os tópicos com base nos grupos descritos anteriormente, com exceção do grupo frustração/esperança devido às naturezas polarizadas das palavras chaves que representavam esse grupo e também pela não existência de palavras chaves que representavam exclusivamente o grupo. A acurácia obtida pelo “Seeded LDA” foi de aproximadamente 69,11%.

Pedro Brum ([BRUM et al., 2020](#)) fez uma análise de *tweets* em português que estavam relacionados com a pandemia da Covid-19, realizando uma comparação de comportamentos dos usuários ao longo do tempo. Os *tweets* foram coletados através da API oficial do Twitter com o uso de diversas palavras-chave e pertencem ao período de tempo entre 23 de abril e 02 de julho de 2020. A primeira análise feita envolve o volume de *tweets* ao longo do tempo, considerando o número de *tweets* e *retweets* por dia. Para realizar uma análise textual, um pré-processamento foi aplicado no conjunto de *tweets* obtido para eliminar dados que não contribuíssem para a análise. Nessa etapa, foram construídas nuvens de palavras para *tweets* e *retweets*, onde foi possível identificar tópicos relacionados ao uso de algumas palavras. Em seguida, foi feita uma distribuição geográfica dos *tweets*, de forma que fosse possível separar o *dataset* em subconjuntos separados por regiões do Brasil. A quarta análise feita envolve o uso de URLs, hashtags e menções a usuários, gerando nuvens de palavras compostas por hashtags e uma distribuição da quantidade de hashtags, URLs e menções a usuários por semana. Por fim, analisaram-se os perfis dos usuários, com o estudo sendo capaz de inferir gêneros para cada um dos usuários, encontrar usuários jornalísticos e o uso de bots.

O trabalho produzido por Régis Ebeling ([EBELING et al., 2020](#)) busca analisar o comportamento de grupos com posicionamentos políticos opostos entre si durante a pandemia da Covid-19 no Brasil. O grupo “Quarenteners” engloba as pessoas que defendem a priorização da saúde durante esse momento grave enquanto o grupo “Cloroquiners” envolve todos que defendem a priorização da economia e medidas sem comprovação científica, como o uso de cloroquina. Todos os *tweets* foram obtidos através da API GetOldTweets ([JEFFERSON-HENRIQUE, 2018](#)) entre os dias 22 de março de 2020 e 7 de abril de 2020. Utilizou-se a API Botometer ([OSOME, 2021](#)) para eliminar *tweets* de bots e, após isso, aplicou-se um pré-processamento nos dados. Com os dados tratados, foram feitas uma modelagem de tópicos utilizando o modelo *Latent Dirichlet Allocation* (LDA), uma análise da polarização política e uma exploração de aspectos psicológicos. Concluiu-se que: os “Cloroquiners” são polarizados com movimentos políticos de direita, enquanto os “Quarenteners” pertencem a movimentos de esquerda; os temas que evidenciam as diferenças entre os grupos estão estritamente ligados ao apoio/rejeição ao presidente ao estabelecer o dilema entre saúde e economia; ambos os grupos possuem emoções negativas semelhantes que surgem com base no seu descontentamento com a pandemia; a baixa sofisticação cognitiva é mais influente na percepção sobre a pandemia do que a orientação política.

O artigo de Emily Chen ([CHEN; LERMAN; FERRARA, 2020](#)) propôs a criação de um *dataset* de *tweets* relacionados à pandemia de Covid-19 no mundo inteiro, envolvendo o período de tempo entre 21 de Janeiro e 21 de Dezembro de 2020 até o momento da escrita dessa seção. A coleta dos *tweets* foi feita utilizando a API oficial do Twitter e a biblioteca Tweepy, com a busca de *tweets* sendo feita por meio do uso de palavras chave, como “coronavirus”, “covid19”, “outbreak”, “pandemic”, etc. O conjunto de dados fornecido envolve *tweets* em diversas linguagens, como inglês, português, espanhol, francês, etc, e está sendo constantemente atualizado pelos autores em um repositório do GitHub ([ECHEN102, 2020](#)).

O artigo produzido por Furqan Rustam ([RUSTAM et al., 2021](#)) se propõe a fazer uma comparação de desempenho entre diversos classificadores com o objetivo de fazer uma análise de sentimentos de *tweets*, onde cada *tweet* pode ser classificado como positivo, neutro ou negativo. Os *tweets* foram obtidos através de um *dataset* relacionado à Covid-19 disponibilizado no site IEEE DataPort, onde existe uma pontuação de emoção vinculada a cada *tweet*. Como primeiro passo do trabalho, o conjunto de dados passa por um pré-processamento e, após isso, outras pontuações de emoção são obtidas para cada *tweet* utilizando a biblioteca TextBlob. Para extrair as características dos *tweets*, foram usadas as técnicas de TF-IDF, Bag of Words e uma combinação entre essas duas técnicas. Os classificadores utilizados foram Random Forest, XGBoost, SVC, Extra Trees e Decision Tree, utilizando as pontuações de acurácia, precisão, recall e f1 como parâmetros para comparar as performances de cada classificador. As melhores acuráncias obtidas foram com o classificador Extra Trees, alcançando acuráncias 88% usando as pontuações do *dataset* e 92% usando as pontuações calculadas com as técnicas TF-IDF e Bag of Words. Os outros classificadores também apresentaram bons resultados, com acuráncias sempre acima de 80%.

Martin Müller ([MÜLLER; SALATHÉ; KUMMERVOLD, 2020](#)) apresentou, em seu trabalho, a criação de um novo modelo BERT chamado CT-BERT, utilizando o modelo pré-treinado BERT-LARGE como base. CT-BERT é um modelo pré-treinado com um grande conjunto de dados de *tweets* relacionados a Covid-19, alcançando uma melhora de 10 a 30% nos resultados obtidos em comparação com os resultados utilizando o modelo BERT-LARGE. Para validar o funcionamento de seu modelo, Martin Müller criou cinco conjuntos de treinamento multirótulo independentes entre si: Categoria de Covid-19, Sentimento sobre vacina, Posição relacionada ao uso de vacinas, emoções no Twitter e Banco de emoções de Stanford. O modelo CT-BERT foi treinado com o uso de fine-tuning usando *tweets* relacionados ao coronavírus obtidos através da plataforma Crowdbreaks. Após concluir a etapa de treinamento do CT-BERT, classificações foram feitas para cada um dos *datasets*, usando a métrica F1 para avaliar os resultados obtidos. Pôde-se observar que os melhores resultados estavam mais relacionados aos *datasets* que tinham temáticas parecidas com os dados usados no treinamento (Categoria de Covid-19,

Sentimento sobre vacina e Posição relacionada ao uso de vacinas). Mesmo assim, em todos os casos, houve uma melhora nas métricas obtidas com o CT-BERT quando comparados com as classificações feitas usando o modelo BERT-LARGE.

3.2 Outros Trabalhos

Os artigos a seguir também abordam temas similares ao estudo dessa monografia mas são aplicados em outros tipos de dados, que não estão relacionados com a Covid.

Tiago de Melo ([MELO; FIGUEIREDO, 2021](#)) apresentou, em seu artigo, uma modelagem de tópicos baseada em dois *datasets* diferentes, com o primeiro sendo composto de notícias produzidas pelo portal UOL e o segundo sendo composto por *tweets* coletados utilizando a biblioteca TwitterScraper em Python. Todos os dados, em português, foram obtidos entre Janeiro e Maio de 2020. Tiago desenvolveu um modelo MALLET (Machine Learning for Language Toolkit) que é uma implementação da técnica LDA para poder criar os diversos tópicos que representam seus dados. Para poder aplicar uma pontuação de emoções em cada um dos textos obtidos, os dados foram traduzidos para Inglês pela escassez de ferramentas poderosas de análise de emoções em Português. Assim, cada uma das notícias ou *tweets* receberam valores numéricos representando a negatividade ou positividade do dado analisado. A análise de emoções foi feita com a ferramenta VADER (Valence Aware Dictionary and Sentiment Reasoner). Além disso, Tiago também criou um modelo de Reconhecimento de entidade nomeada com um novo modelo treinado com a biblioteca SpaCy, onde ele foi capaz de discernir palavras em categorias como Pessoa, Organização, Doença, Sintomas e Drogas.

André Cristiani ([CRISTIANI; LIEIRA; CAMARGO, 2020](#)) criou uma análise de sentimentos aplicada em *tweets* relacionados às eleições a presidente do Brasil em 2018, dividindo os sentimentos em três tipos: positivo, negativo e neutro. A coleta dos *tweets* foi feita através da API oficial do Twitter e foi feita durante a ocorrência de diversos eventos como debates, entrevistas e dias de votação. Uma amostra de 600 *tweets* passou por um processo manual de anotação de sentimentos e, após isso, os *tweets* foram agrupados entre os candidatos à presidência. Após o pré-processamento do conjunto de dados, André comparou o desempenho dos classificadores Naive Bayes e SVM utilizando a técnica TF-IDF, obtendo melhores acurácia para o SVM (cerca de 66,66%). Com esses resultados em mãos, o SVM foi o escolhido para classificar todo o *dataset* e, por fim, identificar como cada sentimento estava distribuído para cada candidato nos primeiro e segundo turnos.

O trabalho de Deho Oscar ([Oscar Deho et al., 2018](#)) envolveu uma análise de sentimentos de *tweets* utilizando a técnica de Word Embedding através do modelo Skip-Gram. O conjunto de dados coletado envolve *tweets* relacionados à estabilização de bases militares dos EUA em Gana, com essa obtenção de dados sendo feita através da biblioteca

Tweepy. Para determinar as polaridades de cada *tweet* (positivo ou negativo), foi utilizada a ferramenta VADER. Uma vez que os *tweets* foram anotados com alguma polaridade de emoção, o classificador Random Forest foi aplicado nesse conjunto de dados, obtendo uma acurácia de aproximadamente 81%. Deho Oscar criou uma distribuição das emoções obtidas com o classificador e identificou que grande parte dos *tweets* (66,5%) é negativo, mostrando que muitas pessoas eram contra esse movimento de bases militares dos EUA em Gana.

Liza Wikarsa ([Indra; Wikarsa; Turang, 2016](#)) utilizou o modelo de Logistic Regression para efetuar classificações de seu *dataset*, que dividia seus *tweets* em quatro categorias distintas: saúde, música, esporte e tecnologia. Cada um dos *tweets* foi obtido através da API oficial do Twitter em uma busca com filtros de palavras chave. Após a coleta dos dados, todos os *tweets* foram pré-processados, de forma a excluir informações irrelevantes para a classificação. Liza utilizou a técnica de Bag-of-Words para poder representar os tokens encontrados nos *tweets*, enviando essa representação de dados para o modelo de Logistic Regression. No geral, as acurárias obtidas para o *dataset* foram de aproximadamente 90%, apresentando métricas bastante satisfatórias.

Finalmente, os trabalhos citados até então possuem diversos pontos em comum com essa monografia: a coleta de *tweets* através do uso de APIs, o uso de técnicas de transformação de dados como TF-IDF, Bag-of-Words e Word Embedding, a construção de um *dataset* com esses *tweets*, a construção de um classificador de análise de sentimentos com o uso de diversos classificadores (SVM, Logistic Regression, Random Forest, BERT, etc), a comparação entre os desempenhos de cada classificador e a modelagem de tópicos baseada no conjunto de *tweets*. O diferencial dessa monografia é que essas técnicas são aplicadas em um conjunto de dados que envolvem o estado do Rio de Janeiro, com várias análises feitas para uma melhor compreensão dos dados obtidos para o estudo. Além disso, a análise de emoções, assim como feita por Irene Zihui Li ([LI et al., 2020](#)), se diferencia da maioria por possuir diversas emoções como classes, ao invés de aplicar classificações em emoções com algum grau de positividade ou negatividade.

4 Análise de Emoções em *Tweets* Sobre a Covid-19

A metodologia proposta para o desenvolvimento desse trabalho pode ser resumida na Figura 8. Primeiramente, os *tweets* são coletados, gerando um conjunto de dados pertencentes ao estado do Rio de Janeiro, através das posições geográficas informadas pelo usuário em seu perfil. Após isso, alguns *tweets* passam por um processo de anotação manual de emoções, onde cada um dos *tweets* é vinculado a uma emoção do estudo. Então, duas emoções são escolhidas para criar um *dataset* menor, que passará pelo processo de treinamento com um método de aprendizado de máquina. Os dados são pré-processados e separados em conjuntos de treinamento e de teste através da K-Fold cross validation (LEARN, 2021i). Com o uso de técnicas como TF-IDF e Bag-of-Words, a extração de características é feita e o processo de classificação se inicia. O modelo passa por uma etapa de treinamento com o conjunto correspondente e, uma vez que esse passo seja concluído, o conjunto de teste passa pelo processo de classificação, gerando métricas de como o classificador realizou as previsões. Para encontrar os melhores parâmetros do classificador utilizado, foi utilizado o módulo GridSearchCV (LEARN, 2021h) da biblioteca sklearn (PEDREGOSA et al., 2011). Cada um desses passos será abordado em uma seção própria nessa monografia.

4.1 Coleta de *tweets*

A primeira decisão passou pelos diferentes métodos de coleta de *tweets*. O primeiro método pensado foi a utilização da API oficial do Twitter (TWITTER, 2021c), por já ser uma API com uma boa documentação, bem consolidada e disponibilizada pela própria empresa. O principal problema encontrado foi que a coleta de *tweets* antigos é limitada em até uma semana anterior à data atual. A coleta começou a ser feita em meados de Maio de 2020 e o intuito da pesquisa era obter dados desde o início de Março de 2020, impossibilitando a utilização dessa abordagem. Para encontrar caminhos distintos, foram observadas outras maneiras de coleta, como a utilização da API Premium do Twitter (TWITTER, 2021b) ou do Historical PowerTrack (TWITTER, 2021a). No entanto, ambas as soluções eram pagas e precisávamos de uma solução gratuita. Por fim, chegou-se à solução que foi usada para obter esses *tweets* e criar um *dataset* de mais de 200.000 registros, no total.

A coleta do *dataset* de *tweets* foi feita utilizando a ferramenta de busca encontrada na interface do próprio Twitter. Na barra de busca, podemos procurar por inúmeros filtros

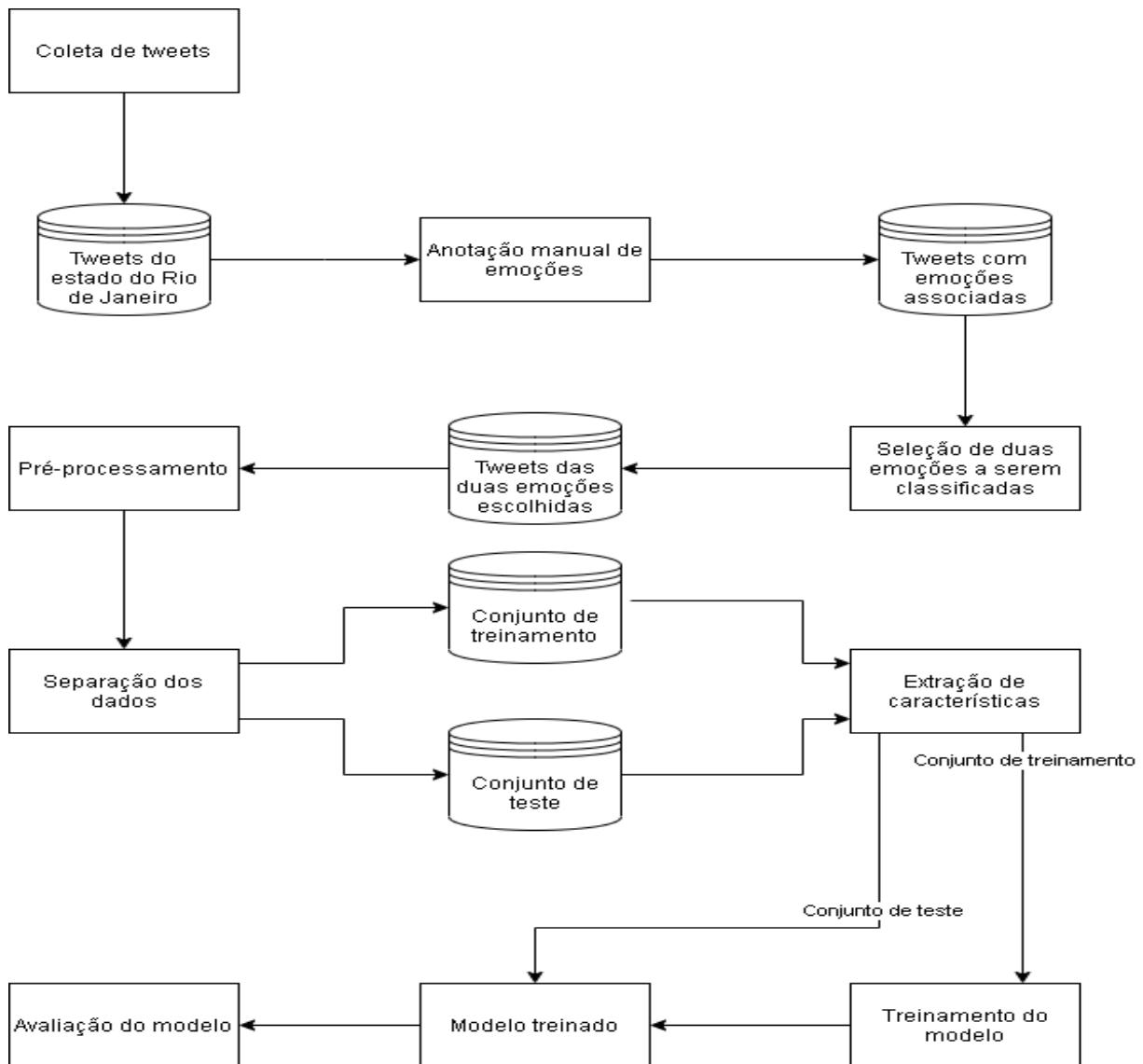


Figura 8 – Visão geral sobre o processo de análise de emoções feito nessa monografia.

utilizando a busca avançada proposta no sistema: palavra chave, hashtags, linguagem, usuários, menções a usuários, data, localização, etc.

Para efetuar a busca de *tweets* e, ao mesmo tempo, armazenar suas informações em arquivos, foi necessário utilizar um script disponibilizado em um repositório do GitHub ([JEFFERSON-HENRIQUE, 2018](#)) que utiliza uma API própria para simular uma busca feita dentro do próprio Twitter. Nas instruções do script, podemos utilizar os diversos parâmetros de busca utilizados no Twitter e, para a busca implementada, foram utilizados filtros de palavra chave, hashtags, nomes de usuários, data e localização.

Nos parâmetros de palavras chaves, hashtags e nomes de usuários, os termos pesquisados foram “**coronavirus**”, “**covid-19**”, “**isolamento**”, “**pandemia**”, e “**quarentena**” que são termos diretamente relacionados à situação vivida pelo mundo inteiro



Figura 9 – Usuário com a tag “Rio de Janeiro” definida em seu perfil.

com a Covid-19.

Para filtrar a localização de cada *tweet*, duas tentativas foram feitas: geolocalização com coordenadas de latitude e longitude anexadas a cada *tweet* ou busca pela *tag* de localização definida nos perfis de cada usuário do Twitter. A primeira abordagem não se mostrou bem sucedida pois o número de amostras obtido foi muito pequeno. A segunda abordagem provou ser mais eficiente, recuperando todos os *tweets* de usuários que possuem a *tag* “Rio de Janeiro” definida na localização de seu perfil. No entanto, essa abordagem tem uma desvantagem: não conseguimos diferenciar se o usuário é residente da cidade do Rio de Janeiro ou do estado do Rio de Janeiro pois o campo que define esse valor acaba sendo um campo de valor aberto, permitindo com que o usuário faça infinitas combinações com a palavra Rio de Janeiro. A Figura 9 exibe aonde pode-se visualizar a localização geográfica definida pelo usuário no Twitter, enquanto a Figura 10 mostra como essa mesma localização pode ser definida.

Por fim, os *tweets* coletados pertencem ao período de tempo entre os dias 01/03/2020 e 23/05/2020, inclusive. Os dados foram coletados semanalmente, com um total de 12 semanas definidas.

O script, em Python, responsável por coletar os *tweets* foi executado em um computador utilizando o sistema operacional Ubuntu e, para automatizar sua coleta, foi utilizado um código em Shell Script que executou a busca diversas vezes até que toda a

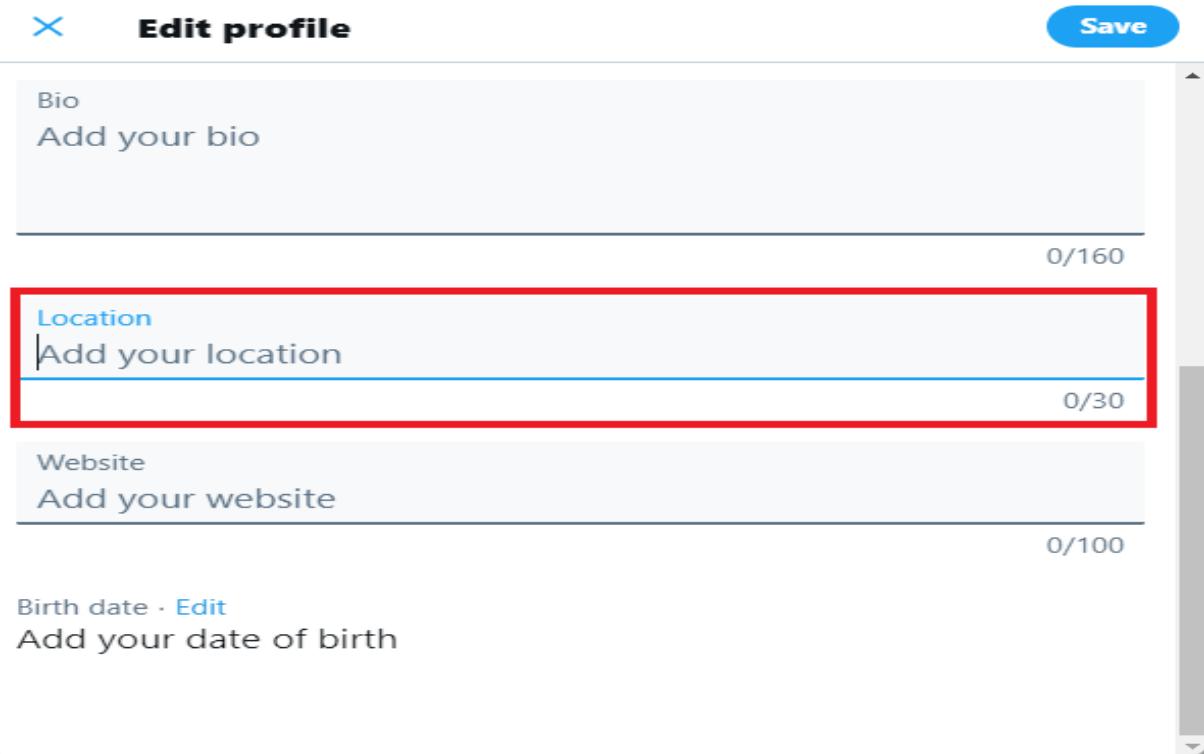


Figura 10 – Campo localização que pode ser preenchido no perfil de um usuário do Twitter.

faixa de tempo fosse coberta e seus *tweets* armazenados.

Realizar a coleta utilizando esse método possui uma desvantagem: o limite de *tweets* que podem ser coletados em uma faixa de tempo. A coleta foi bastante demorada, pois o sistema do Twitter entende inúmeras requisições feitas como um possível ataque contra a segurança de seu sistema, fazendo com que a busca por *tweets* fosse interrompida diversas vezes em seu processo, tornando-a cansativa. Por causa dessa imposição feita pelo Twitter, é bem possível e provável que nem todos os *tweets* relacionados aos filtros utilizados tenham sido encontrados e armazenados no *dataset* desse estudo.

4.2 Rotulação manual dos *tweets*

Para treinar os classificadores para distinguirem as emoções presentes nos *tweets*, é necessário rotulá-los. Essa rotulação, no entanto, apresenta um grande desafio: todo rótulo atribuído a um *tweet* precisa ser feito manualmente. Esse processo busca definir possíveis emoções atreladas aos *tweets* coletados na região do estado do Rio de Janeiro e, para que uma emoção seja vinculada a um *tweet*, é necessária a análise de seres humanos que podem entender que um mesmo *tweet* pode apresentar uma ou mais emoções. As emoções escolhidas para serem rotuladas são as mesmas utilizadas no artigo produzido por Irene Li (LI et al., 2020), com a adição de duas opções (*Outra* e *Neutra*), para o caso de um *tweet*

que não possa ser relacionado com nenhuma das emoções estabelecidas. São elas: Raiva, Ansiedade, Desgosto, Medo, Alegria, Tristeza, Surpresa, Confiança, Outra e Neutra.

Ao compararmos essas emoções escolhidas com a Roda das Emoções vista na Figura 7, pode-se notar diversas semelhanças entre o trabalho proposto por Robert Plutchik ([WIKIPEDIA, 2020b](#)) e o estudo dessa monografia. As diferenças observadas se encontram nas emoções de Ansiedade e Desgosto que não são utilizadas na teoria de Plutchik. No entanto, podemos considerá-las próximas às emoções de Antecipação e Nojo, respectivamente, devido às emoções possuírem um certo grau de similaridade entre si.

De forma a obter visões de pessoas diferentes, foi decidido que a rotulação de *tweets* precisaria ser feita por meio de um formulário preenchido com 20 *tweets* aleatórios e que o mesmo seria divulgado em diversos meios de comunicação, para que vários participantes pudessem participar desse processo voluntariamente.

Inicialmente, tornou-se necessário definir como o formulário seria desenvolvido e entregue aos possíveis participantes dessa pesquisa. O uso do Google Forms foi cogitado mas, devido à sua complexidade para a escolha aleatória de 20 *tweets*, a alternativa escolhida foi a implementação própria de um formulário. A implementação do formulário foi feita utilizando a plataforma low code OutSystems ([OUTSYSTEMS, 2021](#)), que é uma ferramenta bastante utilizada na indústria.

O formulário foi dividido em quatro telas diferentes: Termos de Participação, primeiros 10 *tweets* a serem rotulados, últimos 10 *tweets* a serem rotulados e fim da pesquisa. A primeira tela, exibida na Figura 11, possui o termo de consentimento e confidencialidade, campos de nome e e-mail do participante e informações do orientando e orientadores. As informações do usuário foram armazenadas em uma tabela chamada **Participantes**, que possui atributos id, nome e e-mail, assim como a Figura 12 mostra. Os dados dessa tabela não foram atrelados aos *tweets* rotulados e têm seu propósito voltado apenas para quantificar os participantes do processo de rotulação. Outro ponto de atenção é que o registro dos participantes não impediu que os mesmos participantes refizessem a pesquisa por vontade própria.

As duas telas seguintes são responsáveis por selecionar *tweets* do *dataset* aleatoriamente conforme exibem as Figuras 13 e 14. Esse procedimento foi feito de forma aleatória para que o conjunto de *tweets* rotulados pudesse ser expandido o máximo possível, independente do autor ou do dia da publicação do *tweet*. Em cada uma dessas telas, o usuário pode atribuir uma ou mais emoções a cada *tweet* selecionado. Após os 10 *tweets* de cada tela serem rotulados, eles são armazenados em uma outra tabela. Os *tweets* a serem selecionados estavam definidos em uma tabela chamada ***tweets***, que possui atributos id, *tweet* e um atributo booleano para cada emoção definida para essa pesquisa. No caso desses *tweets* não-rotulados, cada atributo booleano de emoção é definido como *falso*. Os *tweets* com emoções vinculadas são armazenados na tabela ***tweets rotulados***, que possui

uff

Tweets relacionados a Covid-19

Esse é o termo de consentimento livre e esclarecido para solicitar a participação de voluntários na tarefa de classificação de emoções em tweets relacionados à pandemia da Covid-19.

| | |
|--------------------|----------------------|
| Nome completo | <input type="text"/> |
| Endereço de e-mail | <input type="text"/> |

OBJETIVOS DA PESQUISA
 Esta pesquisa visa classificar emoções em uma amostra de tweets postados por usuários que estejam no estado do Rio de Janeiro e que possuem alguma relação com à pandemia da Covid-19 que afeta o mundo inteiro.

INSTRUÇÕES
 A página irá carregar 20 tweets para serem classificados como expressões de oito possíveis emoções selecionadas para o estudo: Medo, Alegria, Tristeza, Ansiedade, Raiva, Desgosto, Surpresa e Confiança. Caso o tweet não se encaixe em um desses sentimentos, existem as opções Outra e Neutra que podem ser preenchidas pelo usuário.
 O participante da pesquisa pode definir várias emoções para cada tweet do formulário.

IDADE
 Eu declaro ter mais de 18 (dezoito) anos de idade e concordo em participar de um estudo conduzido por Gustavo Ferreira Lopes Gonçalves, sob a orientação dos profs. Antonio Augusto de Aragão Rocha (IC/UFF) e Aline Marins Paes Carvalho (IC/UFF).

CONFIDENCIALIDADE
 Eu estou ciente de que meu nome não será divulgado em hipótese alguma. Também estou ciente de que os dados obtidos por meio deste estudo serão mantidos sob confidencialidade, e os

Figura 11 – Primeira tela do sistema de coleta com alguns dos Termos de Participação.

|  Participants |
|--|
|  Id |
|  Name |
|  Email |

Figura 12 – Tabela Participantes.

os mesmos atributos de *tweets*, possuindo a diferença de que os atributos booleanos são preenchidos com *verdadeiro* caso o checkbox da emoção seja marcado pelo usuário. As definições das duas tabelas se encontram na Figura 15.

Por fim, a última tela exibe apenas um texto de agradecimento ao usuário por ter participado do processo de forma voluntária.

A divulgação desse formulário se deu em duas etapas distintas: expansão do *dataset* de *tweets* rotulados e refinamento dos *tweets* rotulados.

uff

Classifique os tweets abaixo

1) 6 dia de quarentena <pic.twitter.com/QmFytLy0Mo>

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

2) Essa quarentena tá acabando com a minha vida kkk

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

3) Defender a ditadura promove a desordem e é inconstitucional. Só com democracia construiremos um país menos desigual e mais eficiente e afetivo. Neste momento, deveríamos investir nosso tempo em salvar vidas e combater a pandemia e suas consequências.

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

4) parabéns p mimmm vinte na quarentena

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

5) Quarentena

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

Figura 13 – Tela 2 apresentando alguns *tweets* com emoções vinculadas pelo usuário.

uff

Classifique os tweets abaixo

11) Aqui está um programa para você... Editorial: A saída de Nelson Teich e a politização do coronavírus Editorial - Gazeta do Povo <https://open.spotify.com/episode/1B7RgvEVptciTN7CWt50?si=vp7brfESEkMWCblCvhhg>

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

12) Entre no elevador distraída no meu andar e tinha uma pessoa dentro, sem máscara. Me assustei mas a porta já tinha fechado. Ele: -fica tranquila porque não tenho covid19 . Eu: -Eu tenho. Ele se encolheu no cantinho. Pra deixar de ser idiota.

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

13) Esse bofe é realmente bonito, ou é a quarentena que tá me deixando louca?

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

14) Cidade do Rio de Janeiro. Medida preventiva de enfrentamento ao avanço do coronavírus (SARS-CoV-2: COVID-19). <https://odia.ig.com.br/rio-de-janeiro/2020/04/5902066-prefeitura-do-rio-vai-instalar-cabines-de-desinfeccao-para-a-populacao-em-pontos-da-cidade.html...>

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

15) acabar essa quarentena , marcar uma pelada de 12:00 até 00:00, fodase kkkkkk

Medo Alegria Tristeza Ansiedade Raiva Desgosto Surpresa Confiança Outra Neutra

Figura 14 – Tela 3 apresentando alguns *tweets* com emoções vinculadas pelo usuário.

| Tweets | | LabeledTweets | |
|--------------|--|---------------|--|
| Id | | Id | |
| Tweet | | Tweet | |
| Anger | | Anger | |
| Anticipation | | Anticipation | |
| Disgust | | Disgust | |
| Fear | | Fear | |
| Joy | | Joy | |
| Sadness | | Sadness | |
| Surprise | | Surprise | |
| Trust | | Trust | |
| Other | | Other | |
| Neutral | | Neutral | |

Figura 15 – Tabelas *tweets* e *tweets* rotulados

4.2.1 Expansão do *dataset*

Nessa etapa, o objetivo da pesquisa era obter o número máximo possível de *tweets* rotulados pelos participantes, removendo os *tweets* já rotulados da tabela que continha os *tweets* não rotulados. Essa primeira etapa gerou cerca de 1300 *tweets* com alguma emoção vinculada e com a participação de, aproximadamente, 65 participantes.

Nessa parte do estudo, a análise foi feita com *tweets* com uma emoção prevalente apenas. Como um usuário pode preencher uma ou várias possíveis emoções que expressam um *tweet*, foi necessário escolher um método para que apenas uma emoção fosse selecionada. Dessa forma, os *tweets* foram rotulados com base na ordem dos identificadores de cada emoção. Por exemplo, o algoritmo verificaria se o *tweet* é pertencente à primeira emoção Raiva. Caso essa emoção não representasse esse *tweet*, a próxima emoção a ser analisada seria a segunda emoção Ansiedade. E esse passo era repetido até chegar ao final da lista de emoções, representada na Tabela 1. Caso o usuário marcasse a opção “Outra” ou “Neutra”, o *tweet* era descartado nesse processo.

Com o novo conjunto de dados obtido, pôde-se analisar como as emoções estavam distribuídas e se ocorreu algum desbalanceamento. A distribuição das emoções se encontra na Figura 16. Nota-se que há uma maior concentração dos *tweets* nas duas primeiras emoções: Raiva e Ansiedade. Também é possível observar que há poucas amostras de *tweets*

| Identificador | Emoção |
|---------------|-----------|
| 0 | Raiva |
| 1 | Ansiedade |
| 2 | Desgosto |
| 3 | Medo |
| 4 | Alegria |
| 5 | Tristeza |
| 6 | Surpresa |
| 7 | Confiança |

Tabela 1 – Identificadores associados a cada emoção do estudo.

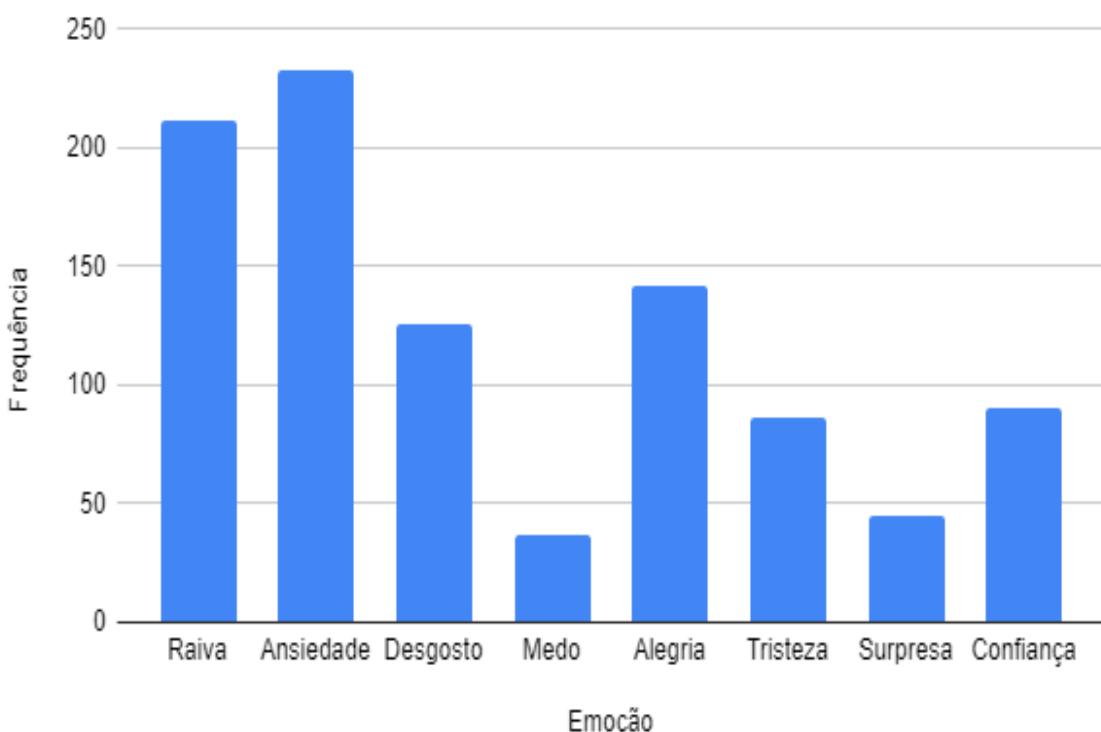


Figura 16 – Frequência de cada emoção no conjunto de dados.

que representem Medo e Surpresa. Uma desvantagem desse método de seleção escolhido é que podem existir diversas emoções similares para um determinado *tweet* e o algoritmo escolhe apenas a primeira que encontra, o que pode causar esse desbalanceamento nos dados.

4.2.2 Refinamento dos tweets rotulados

Além das frequências de emoções desiguais, um outro problema originado dessa etapa inicial é que a emoção vinculada a cada *tweet* se originou da análise de apenas uma única pessoa, fazendo com que a emoção atrelada a um *tweet* fosse muito subjetiva. Para minimizar esse problema, a segunda etapa do processo de rotulação foi iniciada.

Dessa vez, o *dataset* de *tweets* a serem rotulados foi composto dos 1300 *tweets* rotulados anteriormente, de forma que diversos usuários pudessem avaliar as emoções vinculadas à uma mesma amostra, aumentando o grau de confiança de que uma emoção em comum entre as análises está realmente vinculado a um *tweet*. Foram obtidas mais 1600 rotulações que, ao serem somadas com o que foi obtido na fase anterior, formaram um total de cerca de 2900 rotulações vinculadas aos 1300 *tweets* obtidos na primeira etapa. Para identificar a principal emoção de um *tweet*, considerou-se a emoção que foi mais selecionada pelos participantes da pesquisa.

Em caso de empate, o *tweet* foi replicado para cada uma das emoções mais votadas, contanto que elas tivessem alguma similaridade ou pudessem ocorrer em alguma situação. Esses casos de empate ocorreram para as seguintes combinações:

1. Raiva, Desgosto e Tristeza: uma pessoa pode estar triste e, simultaneamente, com raiva de algo que aconteceu ou possa acontecer.
2. Desgosto, Medo e Tristeza: esses comportamentos são normais para casos de pessoas com alguma melancolia e que tenham medo do que pode acontecer.
3. Ansiedade e Medo: existem vezes que a pessoa está ansiosa por algo por ter medo do que pode acontecer.
4. Alegria e Surpresa: existem situações que podem deixar uma pessoa feliz e surpresa.
5. Tristeza e Surpresa: existem situações que podem deixar uma pessoa triste e surpresa.

Após a aplicação desse algoritmo nas 2900 rotulações obtidas, 700 *tweets* foram gerados com suas respectivas emoções vinculadas. A distribuição de cada um dos sentimentos se encontra no gráfico da Figura 17.

Percebe-se que a distribuição das emoções ficou mais balanceada nesse conjunto de dados do que no conjunto de dados obtido na primeira etapa da rotulação manual. No entanto, ainda é possível observar que Medo é vinculado a uma quantidade relativamente menor do que as outras emoções. De acordo com o dicionário Michaelis [34], um dos significados de ansiedade é “Sentimento e sensação de inquietação, medo ou receio”, explicitando a semelhança entre as emoções de Ansiedade e Medo, podendo ser uma causa suficiente para explicar esse fenômeno no conjunto de dados obtido nessa segunda etapa.

4.3 Seleção de emoções

A partir do conjunto de *tweets* com as oito emoções do estudo associadas, selecionaram-se duas emoções para iniciar o processo de classificação. Todas as emoções foram combinadas, fazendo com que esse processo fosse repetido várias vezes até que todas as combinações

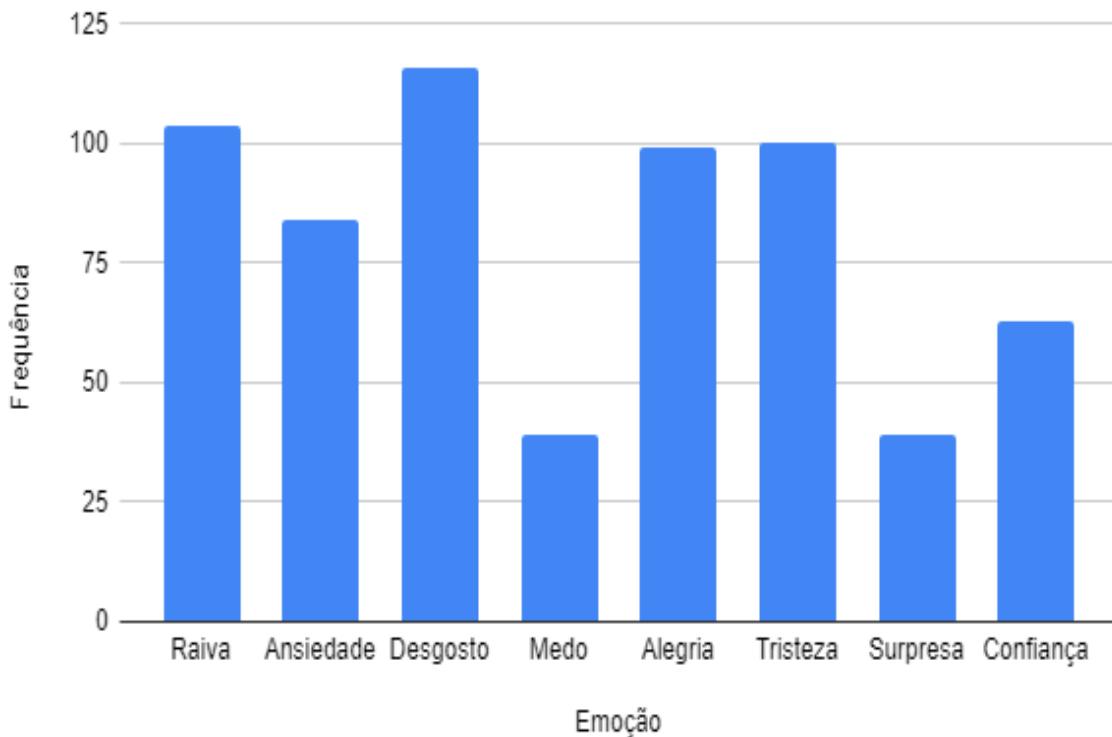


Figura 17 – Frequência de cada emoção no segundo conjunto de dados.

possuíssem um conjunto de dados correspondente. Por exemplo, caso Alegria e Tristeza fossem selecionadas, o *dataset* obtido envolveria apenas os *tweets* de Alegria e Tristeza. Todas as combinações feitas são: (Raiva e Ansiedade), (Raiva e Desgosto), (Raiva e Medo), (Raiva e Alegria), (Raiva e Tristeza), (Raiva e Surpresa), (Raiva e Confiança), (Ansiedade e Desgosto), (Ansiedade e Medo), (Ansiedade e Alegria), (Ansiedade e Tristeza), (Ansiedade e Surpresa), (Ansiedade e Confiança), (Desgosto e Medo), (Desgosto e Alegria), (Desgosto e Tristeza), (Desgosto e Surpresa), (Desgosto e Confiança), (Medo e Alegria), (Medo e Tristeza), (Medo e Surpresa), (Medo e Confiança), (Alegria e Tristeza), (Alegria e Surpresa), (Alegria e Confiança), (Tristeza e Surpresa), (Tristeza e Confiança) e (Surpresa e Confiança).

Além da lista acima, usou-se também o conjunto de dados inteiro para que as performances dos classificadores binários e do classificador com oito classes pudessem ser comparadas entre si.

4.4 Pré-processamento de dados

Essa etapa foi responsável pelo tratamento do conjunto de dados, excluindo diversos elementos de um *tweet* que são desnecessários para a análise de emoções a ser feita nesse estudo. Como forma de auxílio, foi utilizada a biblioteca Ekphrasis ([BAZIOTIS; PELEKIS; DOULKERIDIS, 2017](#)) que possui um pré-processamento de termos adaptados a tudo que

é usado em redes sociais como o Twitter. Os seguintes elementos foram removidos nessa etapa de pré-processamento:

1. Links, nomes de usuários, e-mails e *retweets*: Cada um desses elementos não possui qualquer impacto na análise de emoções e, por isso, foram removidos. No caso dos textitretweets, apenas os tokens “*RT*” foram excluídos.
2. Hashtags e conversão para letras minúsculas: Todas as hashtags encontradas tiveram o caracter “#” removido e todos os termos tiveram suas letras convertidas para minúsculas, de forma a unificar os tokens obtidos.
3. Números: valores numéricos encontrados de forma isolada também foram removidos. Expressões compostas por letras e números não se encaixam nessa categoria, permanecendo no conjunto de dados pré-processado.
4. Stopwords: Como são palavras que normalmente não agregam nenhum valor semântico, as stopwords também foram excluídas. Essa exclusão foi feita com base na lista disponibilizada pela biblioteca SpaCy ([HONNIBAL et al., 2020](#)) no seu módulo `pt_core_news_md` ([EXPLOSION, 2020](#)).
5. Palavras chaves: todas as palavras chaves utilizadas na etapa de coleta de *tweets* foram desconsideradas, uma vez que suas altas frequências poderiam causar impactos negativos na análise de emoções.

4.4.1 Separação dos dados

Para treinar os classificadores e também verificar suas habilidades de generalização, é necessário considerar dois conjuntos: o de treinamento e o de teste. O primeiro grupo será responsável em treinar o classificador para que ele seja capaz de aprender padrões nos *tweets* que estejam relacionados a cada emoção selecionada anteriormente. Após a etapa de treinamento, o conjunto de teste é utilizado para que as métricas do algoritmo de classificação sejam obtidas, identificando se o classificador é capaz ou não de identificar as emoções em um conjunto de dados não visto durante a etapa de treinamento. Para fazer a separação entre esses dois conjuntos, foi utilizada a técnica de K-Fold cross-validation, com K sendo igual a cinco, com o conjunto de treinamento sendo composto por 80% dos *tweets* e o conjunto de teste pelos 20% restantes a cada iteração.

4.5 Transformação da representação textual para uma representação numérica

Para que o processo de aprendizado de máquina esteja apto para começar, é necessário representar os conjuntos de treinamento e de teste em formatos que os classificadores

possam entender. Como o *dataset* é composto apenas por textos, é necessário convertê-los para formatos numéricos para que o aprendizado de máquina ocorra com sucesso. Para isso, foram usadas as seguintes técnicas de extração de características:

1. *Bag-of-Words*: É uma técnica bastante utilizada para a classificação de documentos. Um texto é representado como se fosse um “saco” de palavras, onde a ocorrência de cada palavra é usada como característica ([WIKIPEDIA, 2021a](#)). Foi utilizada a implementação de CountVectorizer fornecida pela biblioteca sklearn ([PEDREGOSA et al., 2011](#)).
2. TF-IDF: O valor tf-idf é uma medida estatística que indica a importância de uma palavra de um documento em relação a uma coleção de documentos, sendo frequentemente usada na recuperação de informações e na mineração de dados ([WIKIPEDIA, 2018b](#)). Para aplicar essa técnica, utilizou-se a implementação de TfidfVectorizer disponibilizada pela biblioteca sklearn ([PEDREGOSA et al., 2011](#)).
3. Word Embedding: É uma representação de palavras no formato de vetores com valores reais, onde cada valor possui um significado para a palavra. É uma técnica bastante usada em análises sintáticas e análises de emoções ([WIKIPEDIA, 2021c](#)). Para a implementação de Word Embedding nessa monografia, foi utilizado o conjunto de vetores de palavras em português fornecido pela biblioteca fastText ([JOULIN et al., 2016](#)).
4. BERT: Utiliza duas estruturas internas na forma de um encoder e um decoder, sendo capaz de representar o input da rede (no caso, textos) em um espaço multidimensional complexo. Através dos encoders e decoders, é possível capturar as inter-relações entre as diferentes partes do texto que estão sendo processadas. Nessa etapa de modelagem da linguagem, o BERT é treinado para prever algumas palavras que são mascaradas aleatoriamente, com uma probabilidade fixada e para prever se uma sentença é sequência lógica da sentença anterior. Essas sentenças são concatenadas em estruturas maiores como frases e parágrafos para que a técnica funcione apropriadamente. Assim, o BERT é forçado a usar as palavras na proximidade da palavra mascarada, e o contexto das duas sentenças apresentadas ao modelo em sequência, para fazer a predição correspondente, aprendendo a inferir uma palavra através do contexto da sentença onde está inserida, assim como entender as relações entre duas sentenças sequenciais ([CECCON, 2020](#)).

4.6 Treinamento do modelo

Para encontrar os melhores parâmetros possíveis para cada classificador utilizado nesse estudo, o módulo GridSearchCV ([LEARN, 2021h](#)) foi escolhido. Esse módulo é

| Classificador | Parâmetros | Valores |
|---------------------|--------------------|--|
| Random Forest | n_estimators | [1, 5, 10] |
| kNN | n_neighbors | [1, 3, 5, 7, 9, 11] |
| | weights | ["uniform", "distance"] |
| | metric | ["euclidean", "manhattan"] |
| Passive Aggressive | C | [0.003, 0.01, 0.03, 0.1] |
| | loss | ["hinge", "squared_hinge"] |
| Gradient Boosting | learning_rate | [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2] |
| | max_depth | [3,5,8] |
| | max_features | ["log2", "sqrt"] |
| | criterion | ["friedman_mse", "mae"] |
| | n_estimators | [1, 5, 10] |
| XGB | objective | ["binary:logistic", "binary:logitraw", "binary:hinge"] |
| | learning_rate | [0.01, 0.05, 0.1, 0.3, 0.5] |
| | max_depth | [1, 2, 4, 6] |
| MLP | hidden_layer_sizes | [(50,50,50), (50,100,50), (100,)] |
| | activation | ["tanh", "relu"] |
| | solver | ["sgd", "adam"] |
| | alpha | [0.0001, 0.05] |
| | learning_rate | ["constant", "adaptive"] |
| Logistic Regression | penalty | ["l1", "l2"] |
| | C | np.logspace(-3,3,7) |
| SVC | C | [1, 10, 100, 1000] |
| | gamma | [1, 0.1, 0.001, 0.0001] |
| | kernel | ["linear", "rbf"] |

Tabela 2 – Parâmetros variados nos classificadores do estudo.

uma busca exaustiva feita a partir dos parâmetros e de seus valores especificados como entrada. A Tabela 2 exibe todas as variações utilizadas para cada um dos classificadores. Os classificadores que não receberam esse tratamento foram o Naive Bayes, que teve seus parâmetros definidos como os valores padrão, e o BERT, que teve seus parâmetros fixados a partir do estudo feito por Irene Zihui Li (LI et al., 2020). É importante ressaltar que o classificador do BERT é apenas uma camada de classificação localizada acima do modelo BERT definido, fazendo com que o classificador não seja um modelo por si só, como são outros classificadores como Random Forest e kNN. No caso do modelo desse estudo, utilizou-se fine tuning com um modelo pré-treinado chamado *neuralmind/bert-base-portuguese-cased* (SOUZA; NOGUEIRA; LOTUFO, 2020). O conjunto de treinamento foi utilizado como entrada para que o processo de classificação pudesse ser iniciado com o treinamento de cada classificador.

4.7 Avaliação do modelo

Após obter o modelo treinado na etapa anterior, utilizou-se o conjunto de teste para entender como cada classificador se tornou capaz de reconhecer padrões e, a partir deles, determinar corretamente quais são as emoções vinculadas para os *tweets* do conjunto de teste.

A principal métrica utilizada para avaliar as performances dos classificadores foi a acurácia. A acurácia representa a fração das previsões feitas corretamente, podendo alcançar um valor máximo de 1, quando todos os dados são previstos corretamente, ou um valor mínimo de 0, quando nenhum dado é previsto corretamente. A fórmula de acurácia é representada pela Equação 4.1.

$$\text{acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

TP (True Positive) representa todos os dados definidos corretamente para uma classe. TN (True Negative) envolve todos os dados que não foram definidos corretamente para uma classe. FP (False Positive) representa todos os dados definidos erroneamente para uma classe. FN (False Negative) envolve todos os dados que não foram definidos erroneamente para uma classe.

Após obter as métricas para todos os classificadores desse estudo, é possível analisar seus resultados e determinar o quanto eficiente foi a análise de emoções promovida nessa monografia.

5 Resultados Obtidos

Os resultados a serem exibidos nessa Seção envolvem a análise de emoções desenvolvida nesse estudo, algumas outras análises feitas no *dataset* coletado para entender as suas características, análises sobre a anotação manual de emoções e outras tentativas realizadas.

5.1 Análise de emoções

Essa Seção apresenta todos os resultados obtidos com a análise de emoções feita para os dois *datasets* anotados na Seção 4.2. O primeiro conjunto de dados representa os dados obtidos na etapa de expansão do *dataset* presente na Seção 4.2.1, enquanto o segundo é o conjunto obtido na etapa de refinamento de *tweets* localizada na Seção 4.2.2. Cada um dos *datasets* possui suas próprias características, que estão listadas abaixo.

1. Todos os *tweets* do segundo *dataset* existem no primeiro *dataset* mas não o contrário. Isso se deve ao processo de refinamento de *tweets*, que utilizou os *tweets* obtidos na etapa de expansão dos dados.
2. O segundo *dataset* é composto apenas por *tweets* que tiveram alguma emoção predominante no processo de rotulação ou que atenderam os critérios de desempate estabelecidos na Seção 4.2.2.
3. Um *tweet* presente nos dois *datasets* pode não possuir o mesmo rótulo devido aos critérios usados na construção de cada *dataset*.

Todas as acurárias apresentadas nesse capítulo foram obtidas através da média das cinco execuções definidas na técnica de K-fold cross validation (LEARN, 2021i), conforme explicado na Seção 4.4.1.

5.1.1 Primeiro conjunto de dados

No intuito de replicar os estudos feitos por Irene Zihui Li (LI et al., 2020), a primeira tentativa feita foi classificar todas as oito emoções em um classificador multiclasse, utilizando o *dataset* coletado na Seção 4.1. Após aplicar todo o processo de aprendizado de máquina nos conjuntos de treinamento e de teste, obteve-se os resultados da Tabela 3. A Tabela inclui os resultados obtidos com os métodos Word Embedding com o uso de fastText, Count Vectorizer e TF-IDF, exibindo os valores obtidos nos conjuntos de treinamento e teste para o primeiro conjunto de *tweets*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|--------|--------|--------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 98.97% | 25.64% | 91.58% | 25.75% | 90.68% | 25.02% |
| kNN | 100.00% | 24.30% | 36.15% | 23.38% | 38.80% | 26.98% |
| Passive Aggressive | 56.49% | 25.65% | 99.97% | 22.86% | 99.97% | 25.54% |
| Naive Bayes | 0.00% | 0.00% | 80.90% | 30.89% | 50.72% | 29.25% |
| Gradient Boosting | 75.28% | 30.49% | 92.04% | 26.98% | 97.79% | 26.57% |
| XGB | 0.00% | 0.00% | 42.17% | 27.91% | 45.96% | 30.48% |
| MLP | 24.64% | 24.1% | 98.89% | 27.70% | 49.90% | 28.22% |
| Logistic Regression | 54.45% | 31.82% | 97.81% | 30.48% | 64.21% | 31.20% |
| SVC | 40.73% | 32.13% | 89.03% | 30.27% | 31.15% | 29.35% |

Tabela 3 – Acurárias obtidas para a classificação multiclasse dos *tweets* do primeiro conjunto de dados.

| | Raiva | Ansiedade | Desgosto | Medo | Alegria | Tristeza | Surpresa | Confiança |
|-----------|-------|-----------|----------|------|---------|----------|----------|-----------|
| Raiva | 26 | 14 | 10 | 3 | 6 | 9 | 3 | 3 |
| Ansiedade | 12 | 20 | 8 | 3 | 6 | 5 | 4 | 7 |
| Desgosto | 2 | 4 | 5 | 1 | 1 | 2 | 0 | 3 |
| Medo | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Alegria | 0 | 3 | 1 | 0 | 12 | 1 | 1 | 4 |
| Tristeza | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| Surpresa | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Confiança | 1 | 2 | 1 | 0 | 2 | 0 | 1 | 1 |

Figura 18 – Exemplo de matriz de confusão da classificação multiclasse no primeiro conjunto de dados.

Pode-se observar que existem vários casos de overfitting e underfitting para os classificadores desse estudo. Overfitting indica que o modelo treinado aprendeu boa parte dos padrões e ruídos do conjunto de treinamento mas, ao aplicar o mesmo aprendizado para o conjunto de testes, as predições não funcionam como o esperado, pois o modelo acabou se adaptando demais ao conjunto de treinamento e não consegue generalizar o comportamento para outros casos. Já o underfitting mostra que o modelo treinado não conseguiu identificar esses padrões e ruídos de forma eficaz, obtendo resultados ruins para as etapas de treinamento e de teste. Tentou-se diversas variações distintas de parâmetros para os classificadores do estudo, mas nenhuma conseguiu solucionar os problemas encontrados.

Além disso, foi extraída uma matriz de confusão de um dos classificadores do estudo na Figura 18. É possível notar que existe muita confusão entre alguns pares de emoções: Raiva/Ansiedade, Raiva/Desgosto, Raiva/Tristeza e Ansiedade/Desgosto. As emoções de Raiva e Ansiedade podem estar entre as mais confundidas por possuir um número de exemplos consideravelmente maior em relação as outras emoções.

| Sentimentos | Acurácia de teste | Classificador |
|----------------------|-------------------|--|
| Raiva e Alegria | 81,57% | BERT-classifier |
| Desgosto e Alegria | 72,24% | BERT-classifier |
| Raiva e Ansiedade | 71,61% | BERT-classifier |
| Raiva e Tristeza | 71,49% | Gradient Boosting (Word Embedding c/ fastText) |
| Alegria e Tristeza | 70,17% | BERT-classifier |
| Ansiedade e Desgosto | 69,11% | BERT-classifier |
| Desgosto e Confiança | 67,78% | BERT-classifier |

Tabela 4 – Melhores acuráncias obtidas com classificações binárias no primeiro conjunto de dados.

Como os resultados não foram tão bons com esse primeiro conjunto de dados, decidiu-se por criar classificadores binários. Essa solução diminuiria a complexidade do problema ao diminuir a quantidade de emoções a serem classificadas, reduzindo um total de oito emoções para apenas duas. Para construir os classificadores binários, combinaram-se todas as emoções do estudo em diferentes pares, assim como explicado na Seção 4.3. A partir dessas combinações, os classificadores foram treinados e avaliados para identificar os casos com melhores acuráncias de treinamento e de teste. Os melhores resultados obtidos estão listados na Tabela 4.

Como existem 28 combinações de emoções distintas, conforme visto na Seção 4.3, foi necessário selecionar apenas os melhores resultados para facilitar a visualização das performances dos classificadores. Todos os resultados obtidos para esse conjunto de dados se encontram no Apêndice A. Em relação ao BERT, obtiveram-se resultados apenas de combinações que poderiam gerar resultados interessantes, devido a demora de quase 1 hora para o treinamento das cinco execuções do algoritmo. Os resultados obtidos foram bastante satisfatórios para algumas combinações de emoções. A combinação de Alegria/Tristeza gerou acuráncias acima de 70%, indicando que o classificador consegue diferenciar duas emoções opostas. Ou seja, em geral, pode-se observar que as melhores acuráncias obtidas possuem sentimentos bem contrastantes entre si, como Alegria/Tristeza e Raiva/Alegria. Percebeu-se também que as combinações de emoções que tinham os melhores resultados nos outros classificadores, na maioria dos casos, tiveram esses resultados superados pelo BERT. Um ponto observado no estudo é que a combinação de qualquer outra emoção com Medo ou Surpresa gerou acuráncias de treinamento e de teste altas. No entanto, isso foi resultado da distribuições desbalanceadas das emoções, pois todos os classificadores tendiam a classificar a emoção com a frequência maior da combinação, havendo poucos casos relacionados a Medo e/ou Surpresa, impedindo com que os classificadores aprendessem algum padrão nesses *tweets*. O mesmo fenômeno se repetiu para as combinações de Raiva/Confiança, Ansiedade/Tristeza e Ansiedade/Confiança, onde a primeira emoção do par possui uma concentração de *tweets* consideravelmente maior do que a segunda emoção, gerando

| | Raiva | Ansiedade | Desgosto | Medo | Alegria | Tristeza | Surpresa | Confiança |
|-----------|-------|-----------|----------|------|---------|----------|----------|-----------|
| Raiva | 1 | 3 | 2 | 2 | 3 | 3 | 2 | 1 |
| Ansiedade | 0 | 2 | 1 | 0 | 3 | 0 | 2 | 1 |
| Desgosto | 14 | 3 | 9 | 4 | 2 | 8 | 1 | 2 |
| Medo | 0 | 1 | 2 | 1 | 0 | 1 | 0 | 1 |
| Alegria | 2 | 3 | 5 | 0 | 6 | 4 | 1 | 5 |
| Tristeza | 4 | 4 | 3 | 0 | 3 | 4 | 2 | 2 |
| Surpresa | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| Confiança | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |

Figura 19 – Exemplo de matriz de confusão da classificação multiclasse no segundo conjunto de dados.

acuráncias altas por acertar vários casos da emoção que aparecem com mais frequência. Além disso, nota-se a ocorrência da combinação Alegria/Tristeza, que, na teoria de Robert Plutchik ([WIKIPEDIA, 2020b](#)), são emoções opostas entre si, indicando que a teoria, para esse conjunto de emoções, se tornou válida para os classificadores desse estudo, indicando que eles também conseguem distinguir essas emoções opostas de forma satisfatória.

5.1.2 Segundo conjunto de dados

Para o segundo conjunto de *tweets* anotados após o refinamento das anotações nos *tweets*, inicialmente, tentou-se aplicar a classificação utilizando todas as oito emoções do estudo juntas da mesma forma feita para o primeiro *dataset* anotado. A Tabela 5 exibe os resultados obtidos nessa etapa.

| Classificador | Word Embedding (fastText) | | Count Vectorizer | | TF-IDF | |
|---------------------|---------------------------|--------|------------------|--------|--------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 86.78% | 15.66% | 80.66% | 14.97% | 82.17% | 13.73% |
| | kNN | 87.39% | 13.73% | 28.92% | 14.56% | 33.83% |
| Passive Aggressive | 49.04% | 19.64% | 87.08% | 17.85% | 87.39% | 17.44% |
| | Naive Bayes | 0.00% | 0.00% | 79.43% | 20.46% | 67.68% |
| Gradient Boosting | 71.36% | 17.30% | 83.14% | 16.20% | 87.05% | 17.99% |
| | XGB | 0.00% | 0.00% | 40.07% | 18.68% | 46.12% |
| Logistic Regression | 16.65% | 16.62% | 86.54% | 20.74% | 16.07% | 16.21% |
| | SVC | 55.60% | 22.12% | 86.26% | 21.70% | 70.74% |
| | 33.17% | 20.87% | 82.31% | 20.33% | 17.20% | 15.80% |

Tabela 5 – Acurárias obtidas para a classificação multiclasse dos *tweets* do segundo conjunto de dados.

Os resultados obtidos na Tabela 5 não foram tão bons ao analisarmos os resultados de forma absoluta, apresentando acurárias inferiores a 23% e, para alguns casos como Random Forest, kNN e Gradient Boosting, apresentando características de overfitting

| Sentimentos | Acurácia de teste | Classificador |
|----------------------|-------------------|--|
| Raiva e Alegria | 78,27% | Logistic Regression (Word Embedding c/ fastText) |
| Raiva e Confiança | 78,09% | BERT-classifier |
| Alegria e Tristeza | 76,80% | BERT-classifier |
| Desgosto e Alegria | 75,19% | BERT-classifier |
| Ansiedade e Desgosto | 72,40% | BERT-classifier |
| Alegria e Confiança | 72,16% | Logistic Regression (Word Embedding c/ fastText) |
| Raiva e Ansiedade | 70,74% | Logistic Regression (TF-IDF) |
| Desgosto e Confiança | 70,67% | BERT-classifier |

Tabela 6 – Melhores acuráncias obtidas com classificações binárias no segundo conjunto de dados.

e, para outros classificadores como Passive Aggressive, MLP, Logistic Regression e SVC, underfitting. Mesmo utilizando outro método para a anotação de *tweets*, pode-se concluir que ambos os *datasets* apresentaram resultados ruins para esse tipo de problema. Esse comportamento indica que a classificação de todas as emoções juntas é uma tarefa bastante complexa, elevando muito a dificuldade de encontrar classificadores que sejam capazes de realizar uma análise de emoções apropriada para essa combinação de classes.

Também foi recuperada uma matriz de confusão de um dos classificadores do estudo na Figura 19. Aqui, pode-se notar muita confusão entre Raiva/Desgosto e Desgosto/Tristeza. Uma possibilidade para esses dois pares de emoções serem bastante confundidos é que as emoções possuem um certo grau de similaridade entre si, o que pode causar dificuldades em distingui-las.

Assim como feito no primeiro *dataset*, foi realizado um levantamento com todas as combinações possíveis de emoções para identificar os casos com melhores acuráncias de treinamento e de teste. Os melhores resultados obtidos estão listados na Tabela 6.

Assim como foi feito nos resultados do primeiro *dataset*, foi necessário selecionar apenas os melhores resultados para facilitar a visualização das performances dos classificadores. Todos os resultados obtidos para esse conjunto de dados também se encontram no Apêndice dessa monografia. Ocorreram diferenças notáveis entre as acuráncias obtidas entre cada um dos conjuntos de dados de cada etapa da rotulação manual. Na segunda parte, as oito melhores combinações de emoções apresentaram acuráncias de teste superiores a 70% enquanto, na primeira parte, apenas quatro resultados possuíram esse mesmo comportamento. A maior acurácia obtida na segunda fase foi ligeiramente inferior para a maior acurácia da primeira fase e, coincidentemente, refletem a mesma combinação de emoções: Raiva e Alegria. Isso indica que ambas as emoções são as mais divergentes entre si dentre todas as combinações possíveis e implicam em uma facilidade maior para os classificadores entenderem suas diferenças e aplicarem suas predições de forma mais coerente com a realidade. Também pode-se observar que a maior parte das combinações de

sentimentos foram similares em ambos os conjuntos de dados, com exceção da combinação de Raiva e Tristeza que não apresentou resultados muito bons no segundo conjunto de dados. Esse comportamento acabou sendo esperado por serem emoções que podem estar presentes em uma mesma situação, gerando uma dificuldade maior na diferenciação das emoções, como observado no primeiro critério adotado na Seção 4.2.2. Diferentemente dos resultados obtidos no primeiro conjunto de dados, existem casos em que o classificador Logistic Regression, utilizando a técnica de Word Embedding com a biblioteca fastText, que fornece representações de texto no formato dessa técnica, acaba apresentando melhores resultados do que o BERT. Para o caso de Raiva/Ansiedade, o uso do classificador Logistic Regression com a técnica de TF-IDF obteve a maior acurácia de teste.

Levando em consideração a Roda das Emoções de Robert Plutchik ([ALABAU, 2020](#)), é possível perceber a presença de combinações de emoções opostas entre si nos resultados obtidos, indicando uma maior facilidade na diferenciação dessas emoções por parte dos classificadores. Alegria/Tristeza obteve uma acurácia ainda melhor no segundo conjunto de dados, estando presente nos melhores resultados de ambos os *datasets* construídos. A presença da combinação Desgosto/Confiança ocorre apenas aqui, sendo o equivalente a uma possível combinação de Nojo/Confiança. No entanto, as combinações envolvendo Medo e Surpresa acabaram não obtendo resultados satisfatórios devido à presença de poucas amostras dessas emoções, impossibilitando a validação completa da teoria de Robert Plutchik ([WIKIPEDIA, 2020b](#)).

5.2 Análise do *dataset*

Com o objetivo de entender as características do conjunto de *tweets* coletado para esse estudo, estabeleceram-se divisões distintas para agrupar o *dataset* em subconjuntos. A primeira divisão separou todos os *tweets* em semanas, com a representação de cada semana indicada na Tabela 7. A segunda divisão foi feita utilizando datas de acontecimentos importantes na pandemia para o estado do Rio de Janeiro. Foram levantadas sete datas distintas e, a partir dessas datas, *tweets* foram separados em dois subconjuntos distintos: dois dias anteriores ao acontecimento e dois dias após o acontecimento. Cada uma das datas levantadas encontra-se na Tabela 8. A partir de cada uma dessas divisões, foram feitos indicadores para visualizar as particularidades de seus dados.

5.2.1 Agrupamento por semanas

O primeiro indicador gerado foi a distribuição de *tweets* coletados pela semanas cobertas pelo *dataset*. Essa distribuição pode ser visualizada na Figura 20.

Pode-se observar que a quantidade de *tweets* coletados nas semanas 3, 4 e 12 foi bem maior do que a quantidade coletada nas outras semanas. Com exceção da semana

| Semana | Faixa de tempo |
|--------|-------------------------|
| 1 | 01/03/2020 a 07/03/2020 |
| 2 | 08/03/2020 a 14/03/2020 |
| 3 | 15/03/2020 a 21/03/2020 |
| 4 | 22/03/2020 a 28/03/2020 |
| 5 | 29/03/2020 a 04/04/2020 |
| 6 | 05/04/2020 a 11/04/2020 |
| 7 | 12/04/2020 a 18/04/2020 |
| 8 | 19/04/2020 a 25/04/2020 |
| 9 | 26/04/2020 a 02/05/2020 |
| 10 | 03/05/2020 a 09/05/2020 |
| 11 | 10/05/2020 a 16/05/2020 |
| 12 | 17/05/2020 a 23/05/2020 |

Tabela 7 – Semanas do conjunto de dados.

| Data | Significado |
|------------|---|
| 05/03/2020 | Primeiro caso de Covid no Rio de Janeiro |
| 16/03/2020 | Primeiro estado de emergência declarado |
| 17/03/2020 | Decreto de isolamento social |
| 30/03/2020 | Primeira prorrogação do isolamento social |
| 13/04/2020 | Segunda prorrogação do isolamento social |
| 30/04/2020 | Terceira prorrogação do isolamento social |
| 08/05/2020 | Quarta prorrogação do isolamento social |

Tabela 8 – Datas importantes do conjunto de dados.

1, todas as outras semanas possuem 5000 a 7500 *tweets*. Essa grande discrepância para as semanas 3 e 12 pode ser explicada pelas limitações impostas na coleta dos dados pelo Twitter, conforme exposto na Seção 4.1. No entanto, a semana 1 apresenta uma quantidade bem inferior de *tweets* relacionados a Covid-19 e é bem plausível supor que esses dados não foram afetados por essa limitação, uma vez que o primeiro acontecimento importante ocorreu apenas no dia 05/03/2020, conforme demonstrado na Tabela 8. É possível que *tweets* relacionados a Covid-19 não tenham aparecido na mesma frequência das outras semanas, pois o assunto somente começou a se tornar viral no final da primeira semana.

Além da distribuição das frequências de *tweets*, nuvens de palavras foram construídas para demonstrar quais foram as palavras mais usadas para cada um dos períodos desse conjunto de dados.

Para todas as nuvens de palavras geradas por semana nas Figuras 21 a 26, foram encontradas várias palavras que aparecem em mais de uma semana dessa divisão: “brasil”, “mundo”, “gente”, “pessoas”, “todo”, “ter”, “pq”, “agora”, “ser”, “pode”, “casa”, “vou”, “dia”, “tudo”, “ainda”, “bolsonaro”, “acabar”, “aqui”, “nada”, “ficar” e “fazer”. É possível

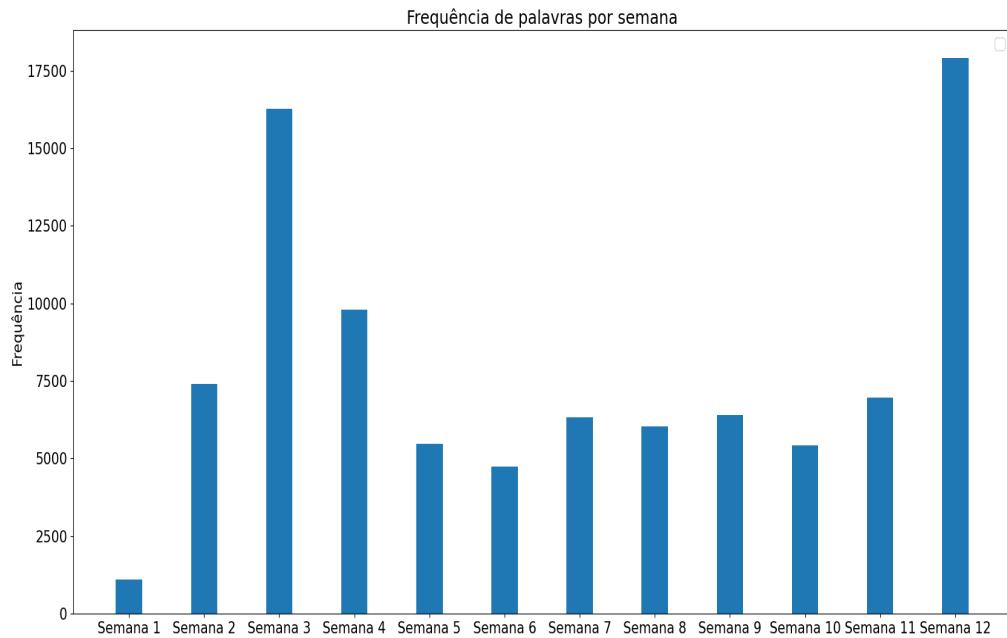


Figura 20 – Contagem de *tweets* por semana.



Figura 21 – Nuvens de palavras para as Semanas 1 e 2.



Figura 22 – Nuvens de palavras para as Semanas 3 e 4.

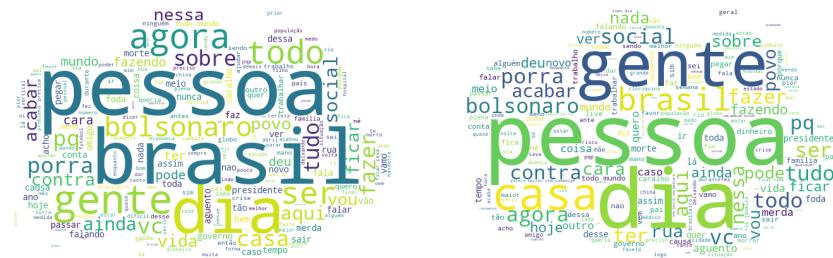


Figura 23 – Nuvens de palavras para as Semanas 5 e 6.



Figura 24 – Nuvens de palavras para as Semanas 7 e 8.



Figura 25 – Nuvens de palavras para as Semanas 9 e 10.



Figura 26 – Nuvens de palavras para as Semanas 11 e 12.

| Subconjunto | Bigramas/Trigramas |
|-------------|--|
| Semana 1 | passo álcool, fake news, nenhum caso, tá com medo, Brasil em 2020, álcool em gel |
| Semana 2 | muita gente, corro risco, milhões de pessoas, Álcool em gel, PAROU o mundo, viaja e traz, acima de 60 |
| Semana 3 | casos confirmados, fake news, vou morrer, Muita gente, ambiente fechado, vc ficar, álcool em gel, M da Saúde, sair de casa, NAO E BRINCADEIRA |
| Semana 4 | sendo fechadas, SACO CHEIO, vida normal, 15 dias, ficar em casa, Casos no Brasil, entra em colapso, cuidar das pessoas, álcool em gel, presidente da República, DECRETO RIO N° |
| Semana 5 | 15 dias, SECRETARIA MUNICIPAL, Taxa de letalidade, MUNICIPAL DE SAÚDE, Funcionários da Saúde, ficar em casa, higienização das Ruas, Tá muito difícil, Declaração de Emergência |
| Semana 6 | atitude Irresponsavel, hospital Icaraí, efeitos colaterais, vc se sair, dor de garganta |
| Semana 7 | vou morrer, ninguém morra, levem a sério, ficar em casa, ficar 14 dias, trancada em casa, manter a empresa |
| Semana 8 | muitas pessoas, 3 dias, vou morrer, dor de cabeça, calados na Crise, álcool em gel, o ÚNICO país |
| Semana 9 | saúde pública, nenhum país, PRESIDENTE BOLSONARO, dor de cabeça, homem em casa |
| Semana 10 | saúde pública, ngm respeita, fazendo churrasco, Ficar em casa, FICA EM CASA, tira a máscara, gripes e resfriados |
| Semana 11 | saúde mental, pessoas morreram hoje, tossir na cara, pessoas com comorbidades |
| Semana 12 | casos confirmados, gente morrendo, nao aguento, ficar em casa, sair de casa, medo de morrer, passadores de pano |

Tabela 9 – Bigramas e trigramas de todas as semanas.

notar que um tema comum entre essas palavras é uma preocupação geral sobre a pandemia no Brasil e no mundo todo para os usuários do Twitter desse estudo.

A Tabela 9 exibe bigramas e trigramas utilizados em cada semana desse conjunto de dados. É possível observar cinco assuntos predominantes: ceticismo com a Covid-19 e seus possíveis impactos na sociedade, preocupação com a pandemia, governo e suas ações durante a crise da Covid-19, informações gerais sobre os impactos da doença no Brasil e ações relacionadas ao isolamento social.

O ceticismo pode ser encontrado na primeira semana do estudo, através de termos como “fake news”, “nenhum caso” e “tá com medo”. Ao decorrer das semanas, pode-se notar que esse ceticismo desaparece, com exceção da Semana 3 que apresenta alguns resquícios desse comportamento, indicando que os usuários do Twitter entenderam a real situação que o Brasil e o mundo estavam naquele momento.

A preocupação com a Covid-19 e a pandemia pode ser bem visualizada em todas

as semanas. É interessante observar que, principalmente, nas primeiras quatro semanas, havia diversas menções ao uso de álcool em gel, através do uso de palavras como “passo álcool” e “álcool em gel”. Essas menções podem estar ocorrendo devido a campanhas de conscientização no Twitter sobre os cuidados necessários para evitar a doença em uma época em que se havia muita incerteza de como lidar com a situação. Também existiu uma preocupação dos usuários com sua própria saúde e das pessoas no mundo através do uso de termos como “muita gente”, “corro risco”, “milhões de pessoas”, “vou morrer”, “ninguém morra”, “saúde mental”, etc. Esse comportamento pode ser observado em todas as semanas a partir da segunda, que foi a época que os primeiros casos de Covid-19 surgiram no Brasil.

Menções aos governos federais e municipais e suas ações na pandemia podem ser observadas entre as Semanas 3 e 9. Há um maior foco nos governos municipais nas Semanas 4 a 6, onde é possível observar menções a órgãos como uma Secretaria Municipal de Saúde e “hospital Icaraí” e ações em cada cidade como “DECRETO RIO Nº”, “higienização das Ruas” e “Declaração de Emergência”. Já o governo federal é citado esporadicamente durante as semanas como referências ao Ministério da Saúde e ao presidente Jair Bolsonaro.

Tweets contendo informações gerais sobre a pandemia parecem ter sido encontrados em algumas semanas, sem apresentar um padrão definido. Termos como “casos confirmados”, “Casos no Brasil”, “15 dias” e “Taxa de letalidade” tendem a indicar conteúdos informativos à população, sendo possivelmente feitos por usuários jornalísticos ou por usuários especializados na área médica.

Por fim, pode-se notar diversas menções a atitudes relacionadas ao isolamento social em grande parte das semanas a partir da Semana 2. Existem palavras favoráveis ao isolamento como o uso de “ficar em casa” e outras ocorrências desfavoráveis ao isolamento como “sair de casa”, “ngm respeita” e “fazendo churrasco”. É interessante observar que essas reações contrárias apenas começam a ser vistas a partir das últimas três semanas do *dataset*, o que indica uma certa perda de paciência com o isolamento social por parte de alguns usuários.

5.2.2 Agrupamento por datas importantes

Assim como foi realizado no agrupamento anterior, fez-se uma distribuição dos *tweets* separados pelas datas identificadas na Tabela 8. A distribuição pode ser vista na Figura 27.

Da mesma maneira que pode-se observar uma concentração de *tweets* em algumas semanas específicas, aqui ocorre uma concentração de *tweets* na Data 4. Essa concentração pode ser explicada pois a Data 4 possui *tweets* das Semanas 3 e 4, onde a Semana 3 apresenta uma alta quantidade de *tweets*. As Datas 5 a 7 apresentam quantidades parecidas, enquanto há menos *tweets* nas Datas 1 e 2. A Data 2 inclui a Data 3 no mesmo subconjunto

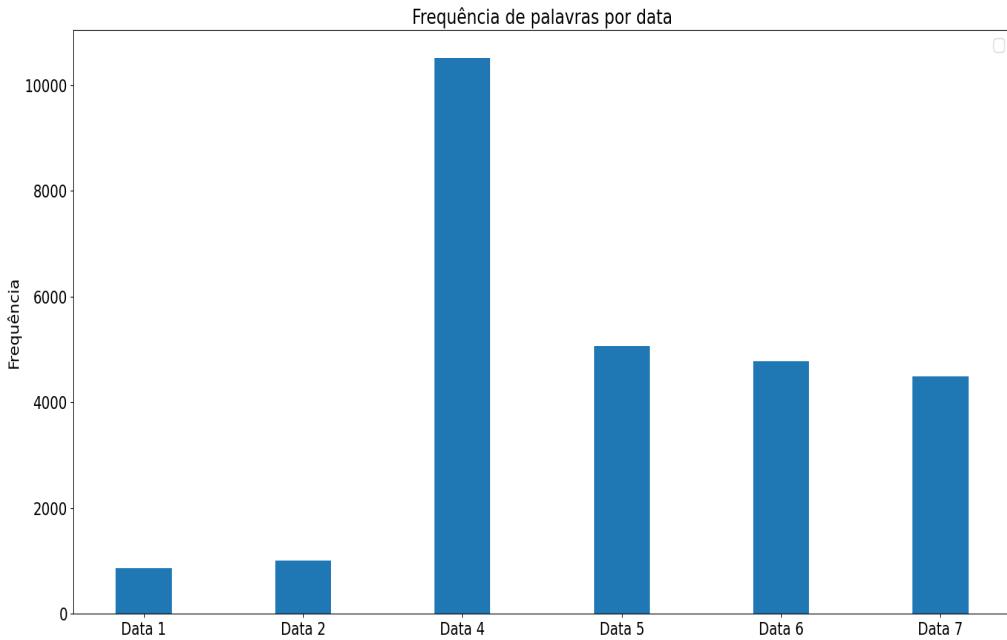


Figura 27 – Contagem de *tweets* por data.

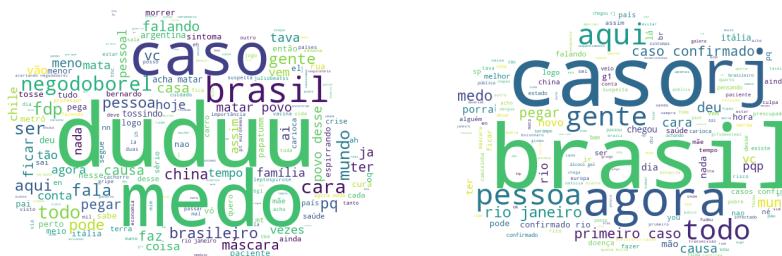


Figura 28 – Nuvens de palavras para os *tweets* anteriores e após a Data 1.

de *tweets* por causa da diferença de apenas um dia entre os acontecimentos. A baixa concentração de *tweets* nas Datas 1, 2 e 3 pode ser efeito da limitação na coleta dos *tweets*, recuperando poucos dados nos dias dos subconjuntos correspondentes.

Também foram construídos indicadores de palavras mais usadas pelos usuários através de nuvens de palavras. As nuvens de palavras de cada data estão representados nas Figuras 28 a 33. Todas as palavras chaves utilizadas na coleta dos *tweets* foram desconsideradas na geração dessas métricas por ser claro de que iriam aparecer como um dos assuntos mais comentados, não sendo relevante para as observações a serem feitas.

Após analisar todas essas Figuras, verificou-se diversos termos em comum com as datas dessa divisão: “agora”, “brasil”, “gente”, “ter”, “casa”, “ficar”, “pessoas”, “pq”, “tudo”, “vou”, “aqui”, “acabar”, “bolsonaro”, “dia”, “fazer”, “fazendo”, “hoje”, “mundo”,



Figura 29 – Nuvens de palavras para os *tweets* anteriores e após as Datas 2 e 3.



Figura 30 – Nuvens de palavras para os *tweets* anteriores e após a Data 4.



Figura 31 – Nuvens de palavras para os *tweets* anteriores e após a Data 5.



Figura 32 – Nuvens de palavras para os tweets anteriores e após a Data 6.



Figura 33 – Nuvens de palavras para os *tweets* anteriores e após a Data 7.

| Subconjunto | Bigramas/Trigramas |
|----------------------|---|
| Data 1 - Antes | pessoal viaja, fake news, tá com medo, álcool em gel |
| Data 1 - Depois | tá com medo, Brasil em 2020 |
| Datas 2 e 3 - Antes | corro risco, milhões de pessoas |
| Datas 2 e 3 - Depois | casos confirmados, álcool em gel, hora de ficarmos, M da Saúde, culpa do BR, entrar em Pânico, garrafas de álcool, hospital de gente |
| Data 4 - Antes | Vitamina D, 11 letras, presidente da República, DECRETO RIO N°, a crise econômica, FICAR EM CASA |
| Data 4 - Depois | 15 dias, SECRETARIA MUNICIPAL, sair de casa, MUNICIPAL DE SAÚDE, Taxa de letalidade |
| Data 5 - Antes | atitude irresponsável, ninguém morra, sociedade civil, sair a noite hospital Icaraí, isolado n° 1, tô me sentindo, ditadura do Witzel |
| Data 5 - Depois | saco cheio, vou ficar, ficar em casa, vontade de sair |
| Data 6 - Antes | seguro desemprego, dor de cabeça, Fica em casa, monte de gente |
| Data 6 - Depois | 91.589 casos, controle de acesso, sair de casa, resultado do exame |
| Data 7 - Antes | Vc sobreviver, casos confirmados, 1 saco, gripes e resfriados, roda de samba, ficar em casa, índice de letalidade, resolver minha vida, |
| Data 7 - Depois | fazendo churrasco, LOTAR HOSPITAL, muita gente, tira a máscara |

Tabela 10 – Bigramas e trigramas obtidos nos *tweets* antes e depois de cada data importante.

“todo” e “ser”. Com essas palavras, pode-se concluir que, de forma geral, houve uma grande preocupação do usuário sobre a situação da pandemia no Brasil e no mundo durante a época que os *tweets* foram publicados.

Para complementar o que foi observado nas nuvens de palavras, foram coletados os bigramas e trigramas de cada subconjunto pertencente às datas levantadas nessa divisão. Um bigrama é uma sequência de dois elementos adjacentes de uma sequência de tokens, que normalmente são letras, sílabas ou palavras, enquanto um trígrama é uma sequência de três elementos próximos. A Tabela 10 exibe alguns bigramas e trigramas relevantes nesse estudo para cada acontecimento.

Com as nuvens de palavras, os bigramas e os trigramas, novas conclusões puderam ser feitas para uma melhor compreensão do conjunto de dados estudado.

Nos períodos de tempo próximos à Data 1 (primeiro caso confirmado de Covid no RJ), a reação das pessoas estava tendendo a levantar dúvidas sobre a gravidade da doença e/ou se ela realmente tinha chegado no Brasil, como a presença de expressões como “pessoal viaja” e “tá com medo”.

Já nas Datas 2 e 3, o comportamento das pessoas mudou drasticamente. Antes dessas duas datas, pouco se discutia sobre a pandemia da Covid-19, apresentando poucos bigramas e trigramas. Após essas duas datas, as pessoas começaram a ficar preocupadas com a doença e como ela poderia afetar o Rio de Janeiro e o Brasil, podendo observar vários

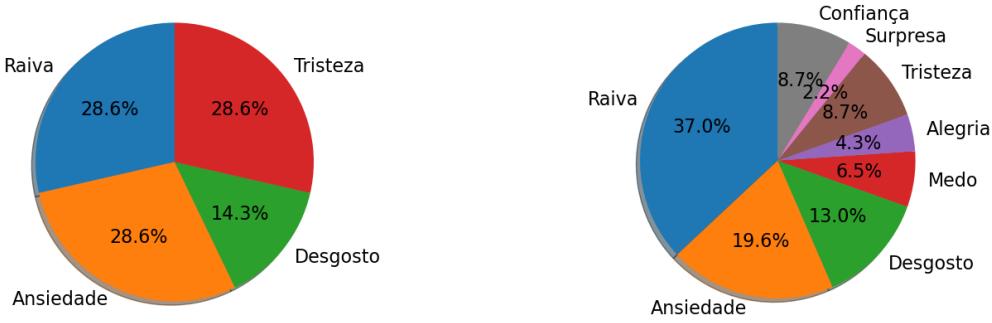


Figura 34 – Distribuição de emoções para as Semanas 1 e 2.

trigramas relacionados a isso como “hora de ficarmos”, “entrar em Pânico” e “garrafas de álcool”.

Nas datas seguintes, pôde-se notar alguns padrões: discussões sobre aspectos sanitários, políticos e econômicos relacionados à pandemia e reações contrárias ao isolamento social. É interessante notar que aparecem bigramas e trigramas relacionados à essas reações contrárias justamente após as Datas 4 a 7 que são prorrogações do isolamento social, indicando que as pessoas já estavam começando a ficar saturadas com a pandemia e desejando voltar às suas vidas normais. Por exemplo: “sair de casa” na Data 4, “vontade de sair” na Data 5, “sair de casa” na Data 6 e “tira a máscara” na Data 7. Outro indicativo desse comportamento é a presença da palavra *sair* nas nuvens de palavras das Datas 4 a 7.

5.3 Análise das distribuições de emoções por período de tempo

Para observar como as emoções estão distribuídas ao longo do tempo, criaram-se distribuições de emoções por período de tempo a partir do conjunto de dados obtido após o refinamento dos *tweets* da Seção 4.2. As distribuições podem ser vistas nas Figuras 34 a 39. As distribuições por data agruparam os *tweets* anteriores e posteriores em um único conjunto para facilitar a visão dos dados e se encontram nas Figuras 40 a 42.

Semana 1 e *Data 1* tiveram poucos *tweets*, então é possível que as emoções vistas não refletem necessariamente as emoções dos usuários do Twitter. Por isso, esses períodos de tempo serão descartados da análise a ser feita.

Na maioria das semanas, Raiva, Ansiedade e Desgosto foram as três principais emoções dominantes. Com exceção da semana 10, Raiva foi a emoção que mais apareceu nos *tweets* rotulados pelos participantes da pesquisa. Curiosamente, a semana 10 apresentou Confiança como a emoção majoritária, sendo seguido pelas outras três emoções citadas anteriormente. As semanas 2, 3, 4, 5, 9 e 12 tiveram Ansiedade como a segunda emoção com mais ocorrências, enquanto as semanas 6, 7, 8, 10 e 11 mostraram que Desgosto foi a



Figura 35 – Distribuição de emoções para as Semanas 3 e 4.

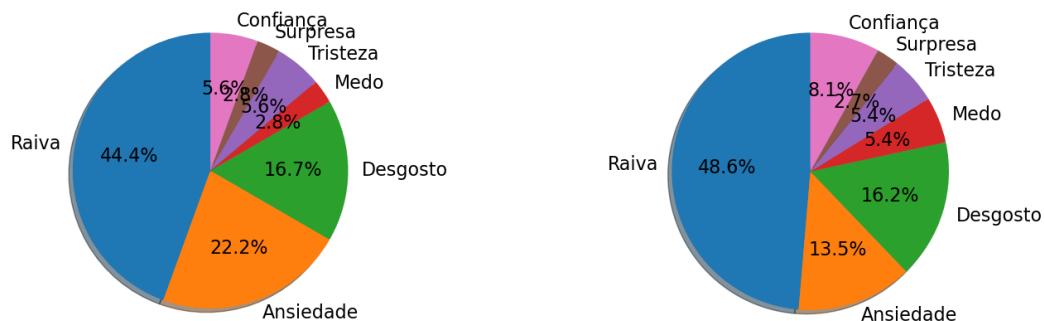
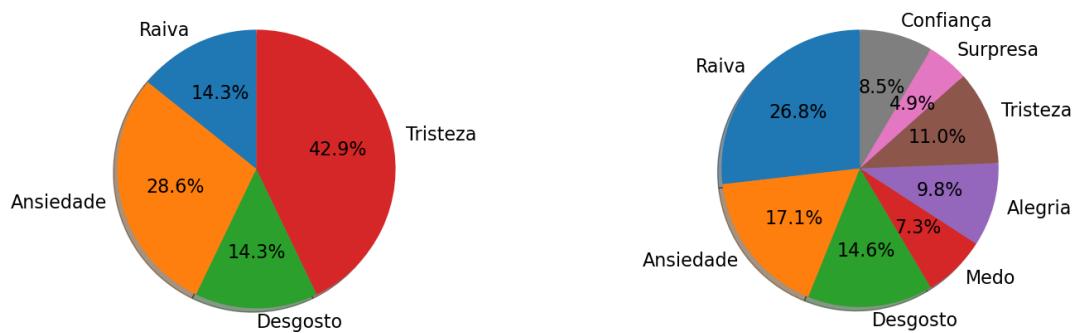
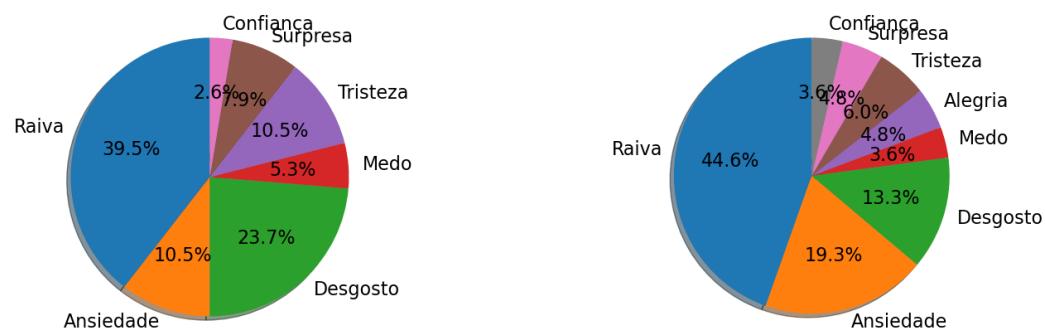
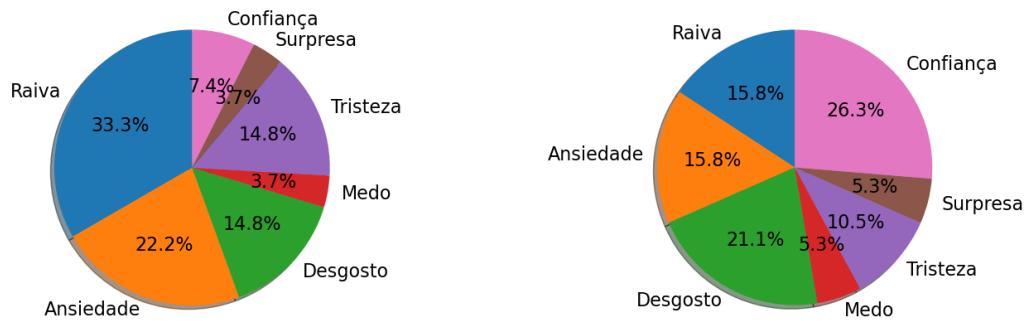


Figura 36 – Distribuição de emoções para as Semanas 5 e 6.



Figura 37 – Distribuição de emoções para as Semanas 7 e 8.



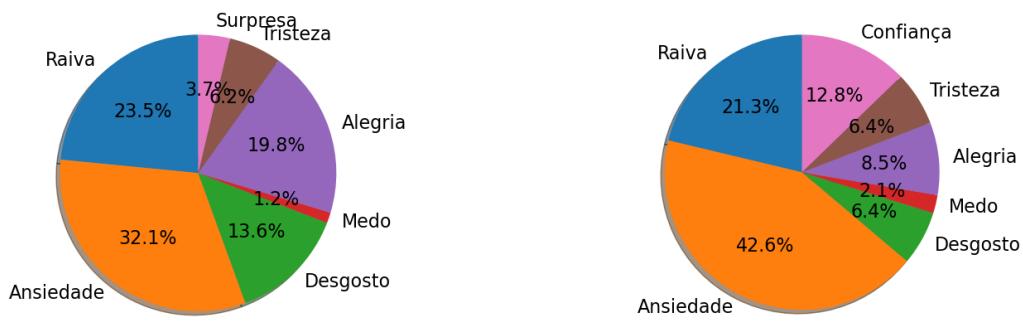


Figura 41 – Distribuição de emoções para as datas 4 e 5.

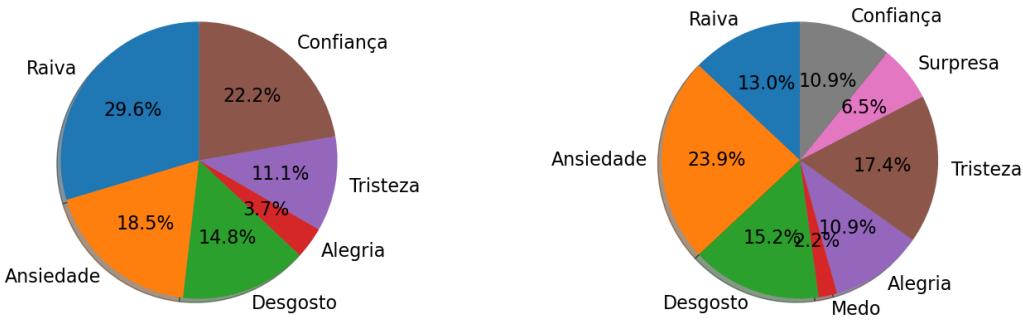


Figura 42 – Distribuição de emoções para as datas 6 e 7.

segunda emoção que mais apareceu. Isso pode indicar, em uma tentativa de generalização, que as pessoas estavam mais ansiosas no começo da pandemia e começaram a sentir desgosto a partir da segunda metade do período coletado para esse estudo.

Ao analisar as distribuições de emoções por data, é possível observar ocorrências parecidas de todas as emoções, com exceção de Surpresa e Medo que aparecem pouco, igualmente visto nas frequências de emoções da Seção 4.2.2. Raiva predominou nas datas 2/3 e 6, ocorrendo um grande descontentamento quando as primeiras medidas emergenciais foram tomadas nas datas 2/3 e quando houve uma terceira prorrogação do isolamento social, podendo indicar que as pessoas estavam começando a ficar saturadas com essa medida sanitária. Ansiedade apareceu em primeiro lugar nas datas 4, 5 e 7, indicando que as pessoas estavam ficando mais ansiosas com o fim da pandemia quando as primeiras duas prorrogações do isolamento e a quarta prorrogação foram decretadas. Um ponto interessante de notar é que Confiança apareceu como a segunda emoção mais predominante na data 6. Ao compararmos o comportamento observado nas distribuições de ambos os *datasets*, o alto grau de confiança da data 6 não está relacionado ao da semana 10, uma vez que a data 6 é o dia 30/04/2020 que pertence à semana 9 (26/04/2020 a 02/05/2020). Isso indica que não há correlação do alto grau de confiança com o fato que ocorreu na

| | Raiva | Ansiedade | Desgosto | Medo | Alegria | Tristeza | Surpresa | Confiança |
|-----------|-------|--------------|--------------|---------------|--------------|--------------|--------------|---------------|
| Raiva | 0 | 0,4035087719 | 1 | 0,08771929825 | 0,1578947368 | 0,5438596491 | 0,1403508772 | 0,05263157895 |
| Ansiedade | 0 | 0 | 0,5789473684 | 0,5438596491 | 0,6140350877 | 0,8947368421 | 0,4210526316 | 0,5438596491 |
| Desgosto | 0 | 0 | 0 | 0,1754385965 | 0,1929824561 | 0,8771929825 | 0,4210526316 | 0,2631578947 |
| Medo | 0 | 0 | 0 | 0 | 0 | 0,5087719298 | 0,1403508772 | 0,05263157895 |
| Alegria | 0 | 0 | 0 | 0 | 0 | 0,1578947368 | 0,350877193 | 0,5263157895 |
| Tristeza | 0 | 0 | 0 | 0 | 0 | 0 | 0,3333333333 | 0,1228070175 |
| Surpresa | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,1403508772 |
| Confiança | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figura 43 – Contagem normalizada das emoções mais confundidas pelos participantes do processo de anotação manual de emoções.

data 6 (terceira prorrogação do isolamento social).

5.4 Observações sobre a anotação manual de emoções

Ao longo da divulgação do formulário da Seção 4.2, muitas pessoas reportaram dificuldades na realização desse processo manual de emoções. Algumas dificuldades relatadas foram a dificuldade em diferenciar emoções e a dificuldade em escolher uma emoção em *tweets* contendo ironia ou sarcasmo. Para expressar essas dificuldades em números, foram identificados todos os *tweets* em que mais de uma emoção tivesse recebido o maior número de anotações pelos usuários. Contabilizaram-se todas as combinações dessas emoções mais destacadas pelos usuários para que seja possível observar quais combinações de emoções mais foram confundidas e difíceis de diferenciar pelos usuários da pesquisa. A contagem das combinações foi normalizada para facilitar a visualização dos dados e se encontra na Figura 43.

A Figura 43 possui valores destacados por cores que representam níveis de dificuldade de diferenciação para as emoções destacadas, onde verde é um nível baixo, amarelo representa um nível médio, laranja é um nível alto e vermelho representa um nível muito alto. As células com o fundo na cor branca devem ser desconsideradas da análise. Pode-se notar que a combinação das emoções de Raiva e Desgosto foi a mais difícil de ser diferenciada pelos participantes da pesquisa, sendo acompanhada de perto pelas combinações de Ansiedade/Tristeza e Desgosto/Tristeza. Coincidemente, as combinações Raiva/Desgosto e Desgosto/Tristeza foram utilizados como critério de desempate para a construção do segundo conjunto de dados, conforme relatado na Seção 4.2. Essas combinações apresentaram performances ruins nos classificadores do segundo dataset na análise de emoções desenvolvida na seção 5.1.2, indicando uma certa dificuldade em diferenciar esses pares de emoções.

Outro aspecto interessante é correlacionar as emoções da Roda das Emoções de Robert Plutchik ([WIKIPEDIA, 2020b](#)) com as combinações vistas aqui. Segundo a teoria proposta por Robert, existem algumas combinações de emoções que são opostas entre si:

Antecipação/Surpresa, Alegria/Tristeza, Confiança/Nojo, Raiva/Medo. As combinações Alegria/Tristeza e Raiva/Medo foram umas das menos confundidas pelos usuários, indicando que, de fato, essas emoções possuem um certo grau de divergência entre si, facilitando na sua diferenciação por uma pessoa. As combinações equivalentes a Antecipação/Surpresa e Confiança/Nojo seriam Ansiedade/Surpresa e Confiança/Desgosto. É possível que por causa das emoções Ansiedade/Antecipação e Desgosto/Nojo serem um pouco distintas entre si, o grau de confusão por parte dos usuários tenha sido um pouco maior para esses casos. No entanto, essas combinações ainda assim não foram as mais problemáticas para os participantes do processo de rotulação.

5.5 Modelagem de tópicos

Com base nas análises dos dados feitas até então, implementações relacionadas a modelagem de tópicos foram feitas para identificar os assuntos predominantes dentro do conjunto de *tweets* estudados. Utilizaram-se três implementações distintas: **LDA** (Latent Dirichlet Allocation), **LDA Mallet** e **CluWords**.

Com a implementação do modelo **LDA** ([GENSIM, 2020a](#)), identificaram-se três tópicos distintos: preocupação com a Covid-19 e seus efeitos no mundo, discussões sobre o isolamento social e hobbies no dia a dia. Mesmo que o número de tópicos fosse aumentado para um valor acima de três nas configurações do modelo **LDA**, os tópicos gerados sempre se focavam nesses tópicos citados. Para o conjunto de *tweets* anteriores ao acontecimento da Data 7, obtiveram-se a distribuição da Figura 44 com a configuração de tópicos definida como cinco.

Para o modelo **LDA Mallet** ([GENSIM, 2020b](#)), foram gerados alguns tópicos distintos mas não foi possível generalizar um tema para qualquer um desses tópicos. O mesmo comportamento observado no modelo LDA se repetiu aqui: mesmo aumentando consideravelmente o número de tópicos a serem gerados no modelo, os tópicos sempre ficam concentrados nos tópicos citados anteriormente. Para o conjunto de *tweets* anteriores ao acontecimento da Data 7, obteve-se a distribuição da Figura 45 com a configuração de tópicos definida como cinco.

Por fim, foi utilizado o modelo **CluWords** ([VIEGAS et al., 2019a](#)), que é melhor adaptado para textos curtos como os *tweets* desse estudo. O resultado da implementação desse modelo gerou diversos tópicos relacionados às classes sintáticas das palavras, como a criação de tópicos para verbos, nomes, etc. Como a intenção desse estudo é a identificação de assuntos predominantes no *dataset* de *tweets*, esse modelo foi descartado. A Tabela 11 exemplifica um resultado obtido utilizando o **CluWords** com o número de tópicos definidos como cinco para o conjunto de *tweets* anteriores ao acontecimento da Data 7.

Por fim, o modelo **LDA** foi o que obteve os melhores resultados até então, mas

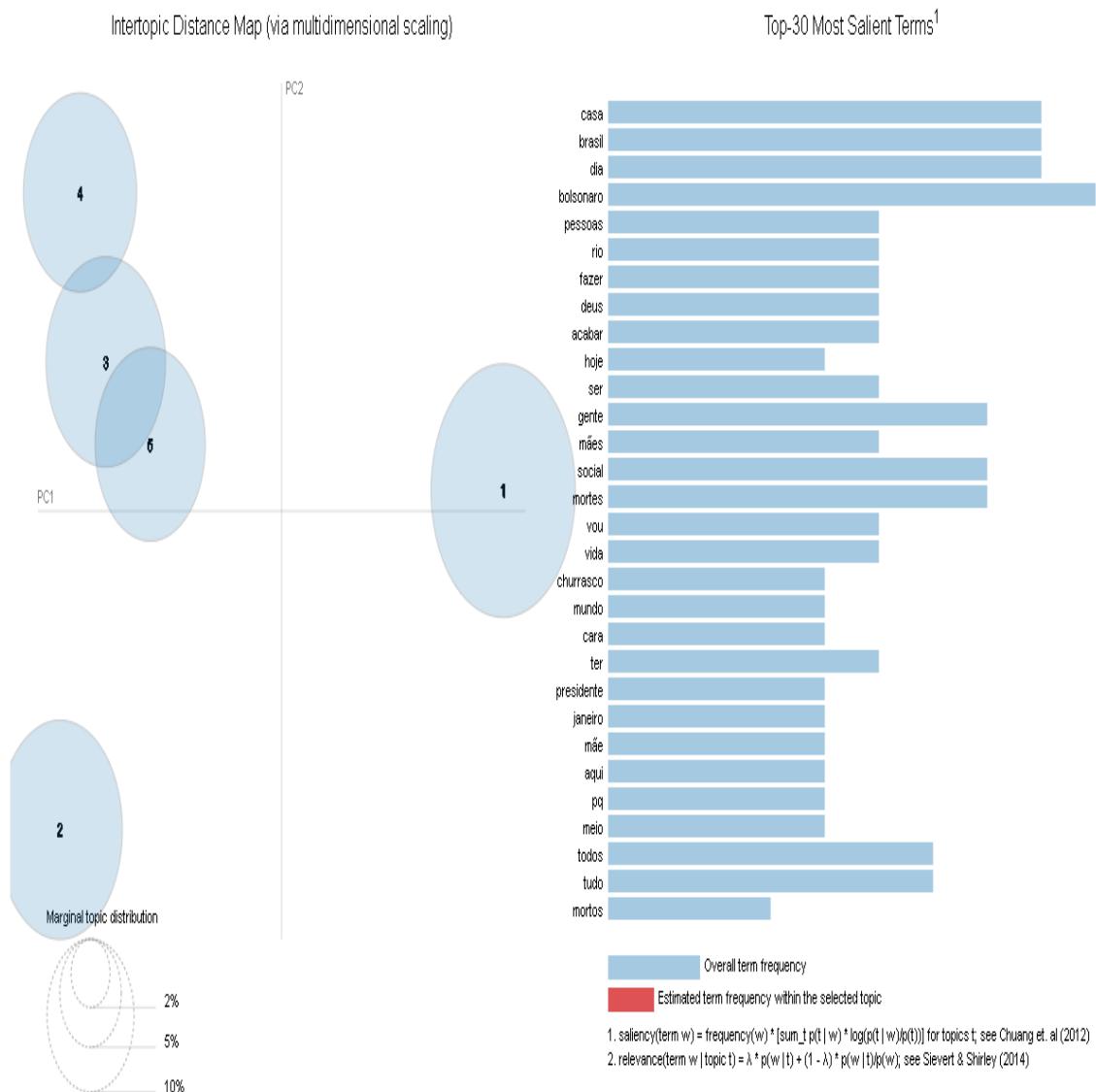


Figura 44 – Tópicos gerados com o modelo LDA.

| Tópico | Palavras relacionadas |
|--------|---|
| 1 | tentar, fazer, utilizar, levar, procurar |
| 2 | fazendo, percebendo, encontrando, colocando, tentando |
| 3 | cezar, rogério, amorim, joyce, marcelo |
| 4 | porque, sim, obviamente, óbvio, pois |
| 5 | conseguiram, deixaram, passaram, vão, fizeram |

Tabela 11 – Tópicos gerados com o modelo CluWords.

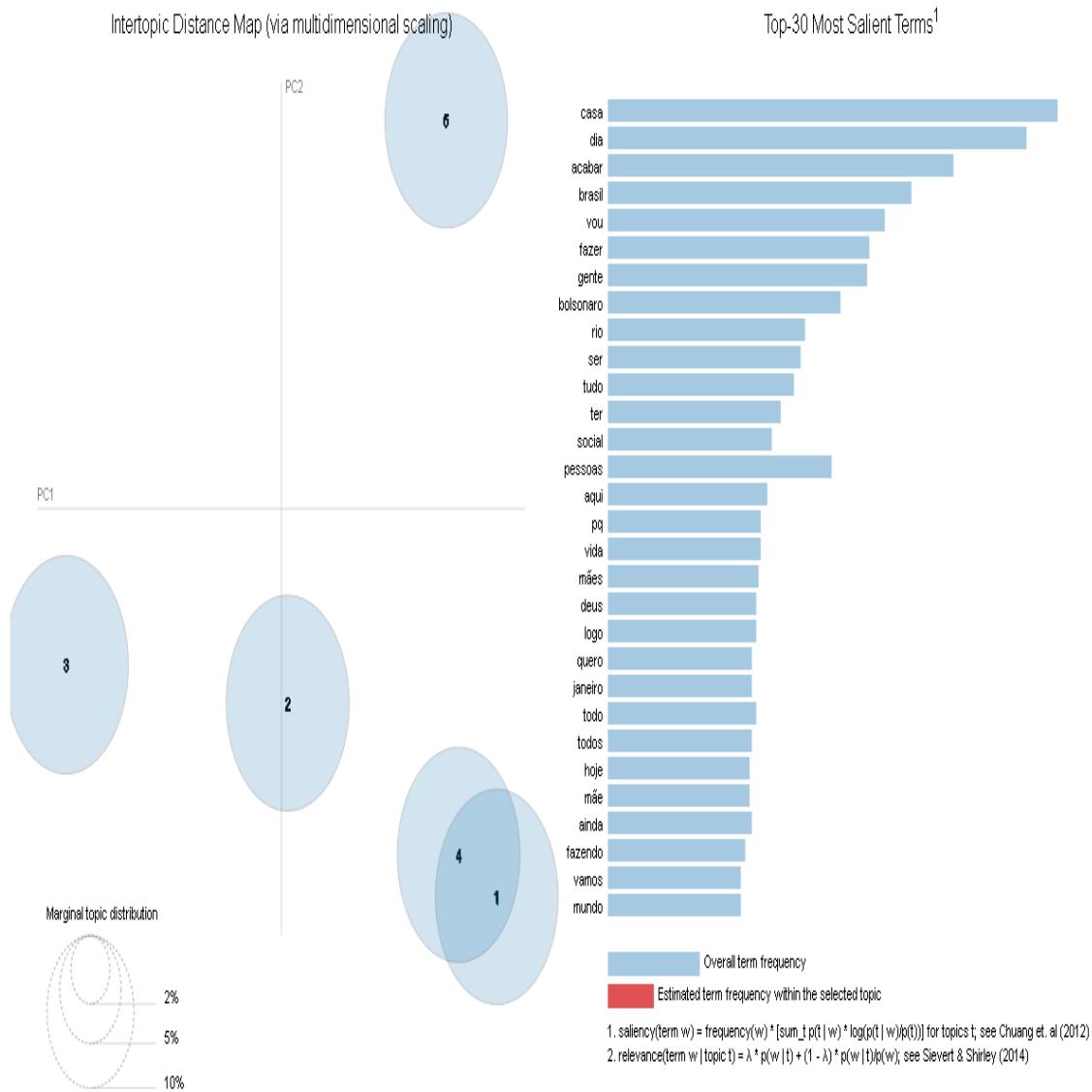


Figura 45 – Tópicos gerados com o modelo LDA Mallet.

acredita-se que é possível chegar a resultados ainda melhores. No entanto, esse estudo acabou não sendo mais aprofundado, deixando possíveis trabalhos futuros a serem feitos na área de Modelagem de Tópicos.

6 Conclusão

Esse trabalho investigou os sentimentos e características de *tweets* coletados no Estado do Rio de Janeiro durante a pandemia de COVID-19 que assolou o mundo nos anos de 2020 e 2021. Para tanto, os *tweets* passaram por um processo de coleta de dados seguido de uma análise das emoções usando várias técnicas de Aprendizado de Máquina, pré-processamento dos textos e a transformação da representação textual para uma representação numérica. Além da análise de emoções, foram gerados diversos indicadores para que fosse possível analisar as características do conjunto de dados coletados, como a criação de nuvens de palavras, distribuições temporais e a identificação de bigramas e trigramas. Espera-se que com tais análises em mãos seja possível compreender melhor como os usuários do Twitter estão reagindo à pandemia da Covid-19 e, com base em tudo o que foi observado, tomar atitudes que possam ajudar a população em outras crises similares à que vivemos atualmente.

Com os resultados da análise de emoções feita nos conjuntos de dados obtidos nas etapas de expansão do *dataset* e refinamento dos *tweets* da Seção 4.2, obtiveram-se resultados para classificadores contendo as oito emoções e para classificadores binários. Os experimentos iniciais dos classificadores com todas as emoções equivalentes às vistas na Roda das Emoções mostraram que os classificadores eram ineficientes na diferenciação das emoções do estudo. Para obter resultados diferentes, dividiram-se as oito emoções em combinações binárias até que todas as emoções fossem combinadas entre si, sendo possível identificar quais emoções foram diferenciadas com mais facilidade e quais as outras emoções mais complicadas de se diferenciar. Com as acurácia dos classificadores binários em mãos, é possível concluir que a decisão tomada sobre o foco maior na construção desses classificadores mostrou-se acertada, deixando clara a diferença nos resultados entre as combinações binárias de emoções e a combinação de todas elas juntas. Também foi possível observar que, ao reunir todas as oito emoções em uma mesma tarefa de classificação, a complexidade do problema se torna muito elevada, resultando na incapacidade dos classificadores desenvolvidos nessa monografia em realizar sua classificação de forma satisfatória.

Ao efetuar a análise dos dados com duas distribuições distintas, foi possível encontrar temas em comum e achar possíveis explicações para os comportamentos observados. O ceticismo de parte da população com a Covid-19 ocorreu em uma época próxima a divulgação do primeiro caso da doença no Brasil, quando a emergência sanitária no país e no estado do Rio de Janeiro ainda não haviam sido decretadas. Preocupações com a Covid-19 e seus impactos na sociedade podem ser observados na maior parte do *dataset* e indica que, desde que o estado de emergência foi decretado, há uma grande atenção por

parte da sociedade nesse assunto. Por fim, reações favoráveis ao isolamento podem ser bastante vistas no conjunto de dados, enquanto o surgimento de reações desfavoráveis provavelmente está diretamente vinculado às diversas prorrogações do isolamento social, gerando reações de inquietação em diversos usuários do Twitter.

Foi possível notar também que, no processo de rotulação de emoções, existiram emoções que possuíam similaridades entre si como Desgosto/Tristeza e Raiva/Desgosto, explicando algumas dificuldades encontradas pelos participantes desse processo e também os resultados não tão bons obtidos com essas combinações na análise de emoções promovida nessa monografia.

Por fim, implementações de Modelagem de Tópicos foram feitas, mostrando que os resultados obtidos com o modelo LDA condizem com o que foi observado no conjunto de dados. No entanto, ainda há a necessidade de aplicar modelos mais robustos e entender se o mesmo comportamento apresentado pelo algoritmo LDA seria replicado nesses modelos para que uma conclusão mais precisa pudesse ser feita na Modelagem de Tópicos desse conjunto de dados.

6.1 Limitações

Um dos problemas encontrados na análise de emoções desenvolvida nesse trabalho foi a falta de exemplos para as emoções de Medo e Surpresa. Parte desse problema ocorreu por causa da relativamente baixa quantidade de pessoas que participaram do processo de rotulação manual de emoções. É possível também que o *dataset* coletado realmente possua poucos *tweets* relacionados a essas duas emoções. Com mais tempo de coleta e mais tempo para anotação dos dados, esse problema poderia ser sanado.

Uma possível análise de emoções mais robusta pode ser feita para todas as oito emoções juntas em um mesmo processo de classificação. Os resultados obtidos nesse estudo foram ruins, levando a um foco maior nos classificadores binários desenvolvidos. Um ponto de atenção é que esse problema de classificação multiclasse é razoavelmente mais complexo do que o problema binário mas, certamente, os resultados obtidos seriam interessantes. Ainda em relação ao processo de classificação, uma abordagem instigante seria analisar por quanto tempo sua taxa de acertos perdura. Caso o classificador comece a perder sua eficácia ao longo do tempo, poderiam ser criados mecanismos para atualização do modelo treinado, sem que fosse necessário proceder com todo o processo de rotulação e aprendizado do zero.

Um outro caminho pouco explorado dessa dissertação foi a área de Modelagem de Tópicos. Da mesma forma que foi feita uma análise das palavras mais usadas e os estudos iniciais com os modelos LDA, LDA Mallet e Cluwords, é possível que um algoritmo mais robusto possa identificar assuntos inexplorados por esse estudo, ampliando a visão obtida

nesse *dataset* ou em *datasets* similares.

Referências

- ADMINVOOO. *Fundamentos dos Algoritmos de Machine Learning (com código Python e R)*. 2016. Disponível em: <<https://www.vooo.pro/insights/fundamentos-dos-algoritmos-de-machine-learning-com-codigo-python-e-r/>>. Citado na página 11.
- ALABAU, I. *A roda das emoções de Robert Plutchik*. 2020. Disponível em: <<https://br.psicologia-online.com/a-roda-das-emocoes-de-robert-plutchik-237.html>>. Citado 2 vezes nas páginas 13 e 41.
- ALOKESH985. *Passive Aggressive Classifiers*. 2020. Disponível em: <<https://www.geeksforgeeks.org/passive-aggressive-classifiers/>>. Citado na página 10.
- BAZIOTIS, C.; PELEKIS, N.; DOULKERIDIS, C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 747–754. Citado na página 31.
- BIANCHINI, L. *Análise exploratória dos tópicos no Stack Overflow usando LDA (Latent Dirichlet Allocation)*. 2018. Disponível em: <<https://rd.uff.br/bitstream/prefix/2096/1/BIANCHINI.pdf>>. Citado 2 vezes nas páginas ix e 13.
- BORAH, A. *Word Embeddings: How do organizations use them for building recommendation systems?* 2021. Disponível em: <<https://medium.com/mlearning-ai/word-embeddings-how-do-organizations-use-them-for-building-recommendation-systems-e0341cf5e638>>. Citado 2 vezes nas páginas ix e 8.
- BRANCO, H. *Redes Neurais Artificiais*. 2020. Disponível em: <<https://abraed.org/overfitting-e-underfitting-em-machine-learning/>>. Citado na página 10.
- BROWNLEE, J. *What Are Word Embeddings for Text?* 2019. Disponível em: <<https://machinelearningmastery.com/what-are-word-embeddings/>>. Citado na página 7.
- BRUM, P. V. et al. A characterization of portuguese tweets regarding the covid-19 pandemic. In: *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. Porto Alegre, RS, Brasil: SBC, 2020. p. 177–184. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/kdmile/article/view/11974>>. Citado 2 vezes nas páginas 2 e 17.
- CARVALHO, A. C. P. de Leon Ferreira de. *Redes Neurais Artificiais*. 2020. Disponível em: <<https://sites.icmc.usp.br/andre/research/neural/desenv.htm>>. Citado na página 9.
- CASTRO, B. Y. S. de. *Como modelar tópicos através de Latent Dirichlet Allocation (LDA) através da biblioteca Gensim*. 2020. Disponível em: <<https://medium.com/somos-tera/como-modelar-t%C3%B3picos-atrav%C3%A9s-de-latent-dirichlet-allocation-lda-atrav%C3%A9s-da-biblioteca-gensim-1fa17357ad4b>>. Citado na página 12.

- CECCON, D. *BERT, o modelo de processamento de linguagem natural que revolucionou a área.* 2020. Disponível em: <<https://iaexpert.academy/2020/04/27/bert-o-modelo-de-processamento-de-linguagem-natural-que-revolucionou-a-area/>>. Citado na página 33.
- CHEN, E.; LERMAN, K.; FERRARA, E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, JMIR Publications Inc., v. 6, n. 2, p. e19273, May 2020. ISSN 2369-2960. Disponível em: <<http://dx.doi.org/10.2196/19273>>. Citado 2 vezes nas páginas 2 e 18.
- CNS, C. N. de S. *Lockdown: CNS defende distanciamento social mais rigoroso diante do momento mais grave da pandemia.* 2021. Disponível em: <[http://conselho.saude.gov.br/ultimas-noticias-cns/1628-lockdown-cns-defende-distanciamento-social-mais-rigoroso-diante-do-momento-mais-grave-da-pa~:text=A%20recomenda%C3%A7%C3%A3o%20segue%20ainda%20mais,dos%20servi%C3%A7os%20tingido%20n%C3%ADos%20cr%C3%ADticos%E2%80%9D](http://conselho.saude.gov.br/ultimas-noticias-cns/1628-lockdown-cns-defende-distanciamento-social-mais-rigoroso-diante-do-momento-mais-grave-da-pa~:text=A%20recomenda%C3%A7%C3%A3o%20segue%20ainda%20mais,dos%20servi%C3%A7os%20atingido%20n%C3%ADos%20cr%C3%ADticos%E2%80%9D)> Citado na página 1.
- CRISTIANI, A.; LIEIRA, D.; CAMARGO, H. A sentiment analysis of brazilian elections tweets. In: *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning.* Porto Alegre, RS, Brasil: SBC, 2020. p. 153–160. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/kdmile/article/view/11971>>. Citado 2 vezes nas páginas 2 e 19.
- EBELING, R. et al. Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In: *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning.* Porto Alegre, RS, Brasil: SBC, 2020. p. 89–96. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/kdmile/article/view/11963>>. Citado 2 vezes nas páginas 2 e 17.
- ECHEN102. *COVID-19-TweetIDs.* 2020. Disponível em: <<https://github.com/echen102/COVID-19-TweetIDs>>. Citado na página 18.
- EXPLOSION. *ptcorenewsmd – 2.3.0.2020.* Disponível em : <>. Citado na página 32.
- FONSECA, C. *Introdução a Bag of Words e TF-IDF.* 2020. Disponível em: <<https://medium.com/turing-talks/introdu%C3%A7%C3%A3o-a-bag-of-words-e-tf-idf-43a128151ce9>>. Citado na página 7.
- GENSIM. *models.ldamodel – Latent Dirichlet Allocation.* 2020. Disponível em: <<https://radimrehurek.com/gensim/models/ldamodel.html>>. Citado na página 55.
- GENSIM. *models.wrappers.ldamallet – Latent Dirichlet Allocation via Mallet.* 2020. Disponível em: <https://radimrehurek.com/gensim_3.8.3/models/wrappers/ldamallet.html>. Citado na página 55.
- GOMES, P. C. T. *Conheça as Etapas do Pré-Processamento de dados.* 2019. Disponível em: <<https://www.datageeks.com.br/pre-processamento-de-dados/>>. Citado na página 5.

- HAT, R. *O que é API?* 2020. Disponível em: <<https://www.redhat.com/pt-br/topics/api/what-are-application-programming-interfaces>>. Citado 3 vezes nas páginas ix, 4 e 5.
- HONNIBAL, M. et al. *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo, 2020. Disponível em: <<https://doi.org/10.5281/zenodo.1212303>>. Citado na página 32.
- HUILGOL, P. *Quick Introduction to Bag-of-Words (BoW) and TF-IDF for Creating Features from Text*. 2020. Disponível em: <<https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>>. Citado 2 vezes nas páginas ix e 6.
- ILEOH. *O Algoritmo K-Nearest Neighbors (KNN) Em Machine Learning*. 2018. Disponível em: <<https://portaldatascience.com/o-algoritmo-k-nearest-neighbors-knn-em-machine-learning/>>. Citado na página 10.
- Indra, S. T.; Wikarsa, L.; Turang, R. Using logistic regression method to classify tweets into the selected topics. In: *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. [S.l.: s.n.], 2016. p. 385–390. Citado 2 vezes nas páginas 2 e 20.
- JAGARLAMUDI, J.; III, H. D.; UDUPA, R. Incorporating lexical priors into topic models. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, 2012. p. 204–213. Disponível em: <<https://www.aclweb.org/anthology/E12-1021>>. Citado na página 16.
- JANEIRO, P. da Cidade do Rio de. *Combate à Covid: Prefeitura determina fechamento dos serviços não essenciais por dez dias*. 2020. Disponível em: <<https://prefeitura.rio/cidade/prefeitura-do-rio-determina-fechamento-dos-servicos-nao-essenciais-por-dez-dias/>>. Citado na página 1.
- JANEIRO, P. da Cidade do Rio de. *Coronavírus: Prefeitura do Rio amplia reserva de leitos para pacientes infectados*. 2020. Disponível em: <<https://prefeitura.rio/cidade/coronavirus-prefeitura-do-rio-amplia-reserva-de-leitos-para-pacientes-infectados/>>. Citado na página 1.
- JEFFERSON-HENRIQUE. *GetOldTweets-python*. 2018. Disponível em: <<https://github.com/Jefferson-Henrique/GetOldTweets-python>>. Citado 2 vezes nas páginas 17 e 22.
- JOULIN, A. et al. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016. Citado na página 33.
- LEARN scikit. 1.4. *Support Vector Machines*. 2021. Disponível em: <<https://scikit-learn.org/stable/modules/svm.html>>. Citado na página 11.

LEARN scikit. 1.9. *Naive Bayes*. 2021. Disponível em: <https://scikit-learn.org/stable/modules/naive_bayes.html>. Citado na página 10.

LEARN scikit. *sklearn.ensemble.GradientBoostingClassifier*. 2021. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>>. Citado 2 vezes nas páginas 10 e 11.

LEARN scikit. *sklearn.ensemble.RandomForestClassifier*. 2021. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>>. Citado na página 10.

LEARN scikit. *sklearn.feature_extraction.text.CountVectorizer*. 2021. Disponível em : <>. Citado na página 6.

LEARN scikit. *sklearn.linear_model.LogisticRegression*. 2021. Disponível em : <>. Citado na página 11.

LEARN scikit. *sklearn.linear_model.PassiveAggressiveClassifier*. 2021. Disponível em : <>. Citado na página 10.

LEARN scikit. *sklearn.model_selection.GridSearchCV*. 2021. Disponível em : <>. Citado 2 vezes nas páginas 21 e 33.

LEARN scikit. *sklearn.model_selection.KFold*. 2021. Disponível em : <>. Citado 3 vezes nas páginas 9, 21 e 36.

LEARN scikit. *sklearn.neighbors.KNeighborsClassifier*. 2021. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>>. Citado na página 10.

LEARN scikit. *sklearn.neural_network.MLPClassifier*. 2021. Disponível em : <>. Citado na página 11.

LEITE, T. M. *Redes Neurais, Perceptron Multicamadas e o Algoritmo Backpropagation*. 2018. Disponível em: <<https://medium.com/ensina-ai/redes-neurais-perceptron-multicamadas-e-o-algoritmo-backpropagation-4a2f3a2a2a>>. Citado na página 11.

LI, I. et al. *What are We Depressed about When We Talk about COVID19: Mental Health Analysis on Tweets Using Natural Language Processing*. 2020. Citado 7 vezes nas páginas 1, 3, 15, 20, 24, 34 e 36.

LIU, Y. et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. Citado na página 16.

LORIA, S. *textblob documentation. Release 0.15*, v. 2, 2018. Citado na página 16.

- MARTINS, P. *O aparente dilema implicado pela pandemia da COVID-19: salvar vidas ou a economia?* 2020. Disponível em: <<https://www.abrasco.org.br/site/noticias/o-aparente-dilema-implicado-pela-pandemia-da-covid-19-salvar-vidas-ou-a-economia-artigo/47221/>>. Citado na página 1.
- MELO, T. de; FIGUEIREDO, C. M. S. Comparing news articles and tweets about covid-19 in brazil: Sentiment analysis and topic modeling approach. *JMIR Public Health Surveill*, v. 7, n. 2, p. e24585, Feb 2021. ISSN 2369-2960. Disponível em: <<http://publichealth.jmir.org/2021/2/e24585/>>. Citado 2 vezes nas páginas 2 e 19.
- MONTREAL, U. de. *Recurrent neural networks*. 2016. Disponível em: <<https://readthedocs.io/en/latest/rnn.html>>. Citado na página 16.
- MOREIRA, A. et al. O algoritmo support vector machine aplicado ao mapeamento do uso e ocupação do solo (the support vector machine algorithm applied to mapping and land use). *Revista Brasileira de Geografia Física*, v. 07, p. 291–303, 02 2014. Citado na página 11.
- MÜLLER, M.; SALATHÉ, M.; KUMMERVOLD, P. E. *COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter*. 2020. Citado 2 vezes nas páginas 2 e 18.
- NEMES, L.; KISS, A. Social media sentiment analysis based on covid-19. *Journal of Information and Telecommunication*, Taylor Francis, v. 5, n. 1, p. 1–15, 2021. Disponível em: <<https://doi.org/10.1080/24751839.2020.1790793>>. Citado 2 vezes nas páginas 2 e 15.
- NEUFELD, P. M. *Memória médica: a Gripe Espanhola de 1918*. 2020. Disponível em: <<http://www.rbac.org.br/artigos/memoria-medica-gripe-espanhola-de-1918/>>. Citado na página 1.
- OLIVEIRA, P. I. *Veja as dicas da OMS para se proteger do novo coronavírus*. 2020. Disponível em: <<https://agenciabrasil.ebc.com.br/saude/noticia/2020-02/veja-dicas-da-oms-para-se-proteger>>. Citado na página 1.
- OMS, O. M. da S. *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. 2020. Disponível em: <<https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>>. Citado na página 1.
- Oscar Deho, B. et al. Sentiment analysis with word embedding. In: *2018 IEEE 7th International Conference on Adaptive Science Technology (ICAST)*. [S.l.: s.n.], 2018. p. 1–4. Citado 2 vezes nas páginas 2 e 19.

OSOME. *Botometer Pro API Documentation*. 2021. Disponível em: <<https://rapidapi.com/OSoMe/api/botometer-pro>>. Citado na página 17.

OUTSYSTEMS. *Build Applications Fast, Right, and for the Future*. 2021. Disponível em: <<https://www.outsystems.com/>>. Citado na página 25.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 3 vezes nas páginas 6, 21 e 33.

RABELLO, E. B. *Cross Validation: Avaliando seu modelo de Machine Learning*. 2019. Disponível em: <<https://medium.com/@edubrazrabello/cross-validation-avaliando-seu-modelo-de-machine-learning-4a2a2a2a2a2a>>. Citado 2 vezes nas páginas ix e 9.

RIBEIRO, D. *O que é Machine Learning? Tecnologia permite 'adivinhar' o que você quer*. 2018. Disponível em: <<https://abracd.org/overfitting-e-underfitting-em-machine-learning/>>. Citado na página 4.

RUSTAM, F. et al. A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. *PLoS one*, v. 16, n. 2, p. e0245909, 2021. ISSN 1932-6203. Disponível em: <<https://europepmc.org/articles/PMC7906356>>. Citado 3 vezes nas páginas 2, 11 e 18.

SANTANA, R. *Como preparar Dados de Texto para Machine Learning*. 2020. Disponível em: <<https://minerandodados.com.br/como-preparar-dados-de-texto-para-machine-learning/>>. Citado na página 6.

SANTOS, M. T. *Como as vacinas para a Covid-19 ficaram prontas tão rápido?* 2021. Disponível em: <<https://saude.abril.com.br/medicina/como-as-vacinas-para-a-covid-19-ficaram-prontas-tao-rapido#:~:text=Pode%20confiar!,m%C3%A3o%20de%20seguran%C3%A7a%20e%20efic%C3%A1cia.&text=Em%20menos%20de%20doze%20meses,aplicadas%20em%20dezenas%20de%20pa%C3%ADses.>> Citado na página 1.

SHANTHAKUMAR, S. G.; SEETHARAM, A.; RAMESH, A. *Analyzing Societal Impact of COVID-19: A Study During the Early Days of the Pandemic*. 2020. Citado 2 vezes nas páginas 2 e 16.

SILVA, J. *Uma breve introdução ao algoritmo de Machine Learning Gradient Boosting utilizando a biblioteca Scikit-Learn*. 2020. Disponível em: <<https://medium.com>equals-lab/uma-breve-introdu%C3%A7%C3%A3o-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-scikit-learn-5a2a2a2a2a2a>>. Citado na página 11.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. [S.l.: s.n.], 2020. Citado na página 34.

- TAMAIS, A. L. M. *Modelos de Predição / Naive Bayes*. 2019. Disponível em: <<https://medium.com/turing-talks/turing-talks-16-modelo-de-pred%C3%A7%C3%A3o-naive-bayes-6a3e744e7>>. Citado na página 10.
- TEAM, T. H. F. *BERT*. 2021. Disponível em: <https://huggingface.co/transformers/model_doc/bert.html>. Citado 2 vezes nas páginas 8 e 11.
- TECH, D. *O que é e como funciona o algoritmo RandomForest*. 2020. Disponível em: <<https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/>>. Citado na página 10.
- TINÓS, R. *Perceptron Multicamadas*. 2018. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4445475/mod_resource/content/1/rn_5_mlp_1.pdf>. Citado na página 11.
- TWITTER. *Historical PowerTrack API*. 2021. Disponível em: <<https://developer.twitter.com/en/docs/twitter-api/enterprise/historical-powertrack-api/overview>>. Citado na página 21.
- TWITTER. *Introducing Twitter premium APIs*. 2021. Disponível em: <<https://developer.twitter.com/en/products/twitter-api/premium-apis>>. Citado na página 21.
- TWITTER. *Twitter API*. 2021. Disponível em: <<https://developer.twitter.com/en/docs/twitter-api>>. Citado na página 21.
- VIEGAS, F. et al. *CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling*. 2019. Disponível em: <<https://github.com/feliperviegas/cluwords>>. Citado na página 55.
- VIEGAS, F. et al. Cluwords: Exploiting semanticword clustering representation for enhanced topic modeling. 2019. Citado na página 12.
- VOLPATO, B. *Google anuncia BERT, seu novo algoritmo de pesquisa*. 2019. Disponível em: <<https://resultadosdigitais.com.br/blog/google-bert/>>. Citado na página 8.
- WELLE, D. *China tem 1ª morte por misteriosa pneumonia viral*. 2020. Disponível em: <<https://g1.globo.com/mundo/noticia/2020/01/11/china-tem-1a-morte-por-misteriosa-pneumonia-viral.html>>. Citado na página 1.
- WELLE, D. *Europa confirma primeiros casos de coronavírus*. 2020. Disponível em: <<https://www.dw.com/pt-br/europa-confirma-primeiros-casos-de-coronav%C3%ADrus/a-52145128>>. Citado na página 1.
- WIKIPEDIA. *Pré-processamento de dados*. 2018. Disponível em: <https://pt.wikipedia.org/wiki/Pr%C3%A9-%C3%A9-processamento_de_dados>. Citado na página 5.

- WIKIPEDIA. *Tf–idf*. 2018. Disponível em: <<https://pt.wikipedia.org/wiki/Tf%E2%80%93idf>>. Citado na página 33.
- WIKIPEDIA. *Robert Plutchik*. 2020. Disponível em: <https://pt.wikipedia.org/wiki/Robert_Plutchik>. Citado 2 vezes nas páginas ix e 14.
- WIKIPEDIA. *Robert Plutchik*. 2020. Disponível em: <https://pt.wikipedia.org/wiki/Robert_Plutchik>. Citado 6 vezes nas páginas 3, 13, 25, 39, 41 e 54.
- WIKIPEDIA. *Bag-of-words model*. 2021. Disponível em: <https://en.wikipedia.org/wiki/Bag-of-words_model>. Citado na página 33.
- WIKIPEDIA. *Stop word*. 2021. Disponível em: <https://en.wikipedia.org/wiki/Stop_word>. Citado na página 6.
- WIKIPEDIA. *Word embedding*. 2021. Disponível em: <https://en.wikipedia.org/wiki/Word_embedding>. Citado na página 33.
- XGBOOST. *XGBoost Documentation*. 2021. Disponível em: <<https://xgboost.readthedocs.io/en/latest/>>. Citado na página 11.

Apêndices

APÊNDICE A – Resultados obtidos na análise de emoções

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 98.48% | 60.90% | 96.07% 57.75% 95.90% 57.33% |
| kNN | 100.00% | 60.67% | 63.48% 52.13% 69.78% 60.45% |
| Passive Aggressive | 81.29% | 62.02% | 100.00% 65.39% 100.00% 65.62% |
| Naive Bayes | 0.00% | 0.00% | 95.73% 62.25% 93.82% 62.02% |
| Gradient Boosting | 95.22% | 68.54% | 89.83% 59.55% 93.48% 61.35% |
| XGB | 0.00% | 0.00% | 70.96% 62.47% 75.28% 62.70% |
| MLP | 54.10% | 51.46% | 98.43% 63.37% 53.43% 53.48% |
| Logistic Regression | 80.84% | 67.64% | 97.92% 64.27% 87% 63.82% |
| SVC | 76.12% | 67.64% | 96.97% 64.72% 67.98% 64.04% |

Tabela 12 – Acurárias obtidas para a combinação de Raiva e Ansiedade no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 97.93% | 63.03% | 94.08% 58.88% 94.60% 60.65% |
| kNN | 100.00% | 61.83% | 64.35% 59.47% 66.57% 58.30% |
| Passive Aggressive | 86.32% | 61.85% | 100.00% 56.50% 100.00% 60.66% |
| Naive Bayes | 0.00% | 0.00% | 95.49% 58.58% 77.37% 63.02% |
| Gradient Boosting | 92.60% | 60.96% | 85.43% 64.21% 88.76% 63.93% |
| XGB | 0.00% | 0.00% | 67.46% 63.32% 73.08% 63.92% |
| MLP | 62.72% | 62.72% | 99.70% 60.67% 62.72% 62.72% |
| Logistic Regression | 77.88% | 60.96% | 99.56% 59.50% 78% 62.44% |
| SVC | 67.60% | 61.25% | 97.78% 58.31% 62.72% 62.72% |

Tabela 13 – Acurárias obtidas para a combinação de Raiva e Desgosto no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.59% | 84.74% | 96.69% |
| kNN | 100.00% | 83.93% | 85.24% |
| Passive Aggressive | 95.38% | 81.12% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 91.57% |
| Gradient Boosting | 94.38% | 79.10% | 96.08% |
| XGB | 0.00% | 0.00% | 86.25% |
| MLP | 85.14% | 85.14% | 86.85% |
| Logistic Regression | 86.85% | 85.14% | 96.08% |
| SVC | 85.14% | 85.14% | 95.08% |
| | | | 85.14% |
| | | | 85.14% |

Tabela 14 – Acurárias obtidas para a combinação de Raiva e Medo no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 98.73% | 68.62% | 96.68% |
| kNN | 100.00% | 67.21% | 69.42% |
| Passive Aggressive | 95.91% | 79.08% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 98.87% |
| Gradient Boosting | 97.46% | 72.87% | 92.59% |
| XGB | 0.00% | 0.00% | 74.44% |
| MLP | 59.89% | 59.89% | 99.29% |
| Logistic Regression | 87.50% | 75.99% | 99.43% |
| SVC | 80.93% | 73.16% | 98.16% |
| | | | 65.83% |
| | | | 65.54% |
| | | | 61.30% |

Tabela 15 – Acurárias obtidas para a combinação de Raiva e Alegria no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.06% | 69.45% | 94.55% |
| kNN | 100.00% | 66.76% | 71.39% |
| Passive Aggressive | 86.99% | 62.42% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 96.81% |
| Gradient Boosting | 91.78% | 71.49% | 90.52% |
| XGB | 0.00% | 0.00% | 72.65% |
| MLP | 71.14% | 71.14% | 99.25% |
| Logistic Regression | 78.61% | 71.15% | 99.16% |
| SVC | 72.57% | 70.47% | 97.48% |
| | | | 68.80% |
| | | | 71.14% |
| | | | 71.14% |

Tabela 16 – Acurárias obtidas para a combinação de Raiva e Tristeza no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|----------------|
| | treino | teste | treino teste |
| Random Forest | 97.96% | 82.10% | 95.72% 82.09% |
| kNN | 100.00% | 83.66% | 82.49% 82.39% |
| Passive Aggressive | 95.62% | 82.87% | 100.00% 80.55% |
| Naive Bayes | 0.00% | 0.00% | 97.96% 82.49% |
| Gradient Boosting | 95.04% | 79.80% | 97.76% 99.51% |
| XGB | 0.00% | 0.00% | 83.17% 82.88% |
| MLP | 82.49% | 82.49% | 82.49% 82.49% |
| Logistic Regression | 87.26% | 82.11% | 98.44% 82.49% |
| SVC | 83.76% | 82.10% | 97.76% 82.49% |

Tabela 17 – Acurárias obtidas para a combinação de Raiva e Surpresa no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-----------------|
| | treino | teste | treino teste |
| Random Forest | 98.18% | 73.85% | 96.19% 94.70% |
| kNN | 100.00% | 67.89% | 74.01% 74.75% |
| Passive Aggressive | 92.72% | 70.84% | 100.00% 100.00% |
| Naive Bayes | 0.00% | 0.00% | 96.61% 76.24% |
| Gradient Boosting | 94.62% | 71.87% | 91.14% 92.47% |
| XGB | 0.00% | 0.00% | 76.16% 77.15% |
| MLP | 70.20% | 70.20% | 98.84% 70.20% |
| Logistic Regression | 83.03% | 75.19% | 98.43% 73.68% |
| SVC | 76.16% | 72.19% | 96.61% 70.20% |

Tabela 18 – Acurárias obtidas para a combinação de Raiva e Confiança no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-----------------|
| | treino | teste | treino teste |
| Random Forest | 97.84% | 63.51% | 95.26% 93.94% |
| kNN | 100.00% | 59.87% | 66.51% 68.38% |
| Passive Aggressive | 89.13% | 63.77% | 100.00% 100.00% |
| Naive Bayes | 0.00% | 0.00% | 94.99% 77.02% |
| Gradient Boosting | 92.41% | 62.66% | 85.10% 87.19% |
| XGB | 0.00% | 0.00% | 71.31% 73.54% |
| MLP | 64.90% | 64.90% | 99.37% 64.90% |
| Logistic Regression | 80.01% | 67.12% | 98.19% 79.04% |
| SVC | 67.97% | 65.74% | 95.40% 64.90% |

Tabela 19 – Acurárias obtidas para a combinação de Ansiedade e Desgosto no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|--------|
| | treino | teste | treino |
| Random Forest | 98.15% | 83.70% | 97.22% |
| kNN | 100.00% | 80.37% | 86.30% |
| Passive Aggressive | 86.94% | 82.22% | 99.81% |
| Naive Bayes | 0.00% | 0.00% | 82.22% |
| Gradient Boosting | 94.07% | 82.96% | 97.87% |
| XGB | 0.00% | 0.00% | 86.67% |
| MLP | 86.30% | 86.30% | 85.56% |
| Logistic Regression | 86.85% | 85.93% | 86.30% |
| SVC | 86.30% | 86.30% | 86.30% |

Tabela 20 – Acurárias obtidas para a combinação de Ansiedade e Medo no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 98.73% | 63.47% | 94.80% |
| kNN | 100.00% | 57.87% | 72.40% |
| Passive Aggressive | 84.40% | 62.40% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 60.53% |
| Gradient Boosting | 94.67% | 64.80% | 97.00% |
| XGB | 0.00% | 0.00% | 62.67% |
| MLP | 62.13% | 62.13% | 70.93% |
| Logistic Regression | 80.67% | 65.60% | 89.00% |
| SVC | 75.40% | 65.07% | 84.81% |

Tabela 21 – Acurárias obtidas para a combinação de Ansiedade e Alegría no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 96.55% | 70.21% | 95.69% |
| kNN | 100.00% | 64.28% | 72.42% |
| Passive Aggressive | 83.18% | 64.91% | 73.35% |
| Naive Bayes | 0.00% | 0.00% | 100.00% |
| Gradient Boosting | 90.36% | 64.24% | 66.13% |
| XGB | 0.00% | 0.00% | 87.62% |
| MLP | 73.04% | 73.04% | 72.73% |
| Logistic Regression | 79.62% | 73.35% | 98.67% |
| SVC | 73.43% | 73.04% | 71.17% |

Tabela 22 – Acurárias obtidas para a combinação de Ansiedad e Tristeza no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.93% | 81.29% | 96.40% |
| kNN | 100.00% | 84.16% | 83.81% |
| Passive Aggressive | 90.47% | 74.82% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 98.02% |
| Gradient Boosting | 92.54% | 78.40% | 98.56% |
| XGB | 0.00% | 0.00% | 83.99% |
| MLP | 83.81% | 83.81% | 98.02% |
| Logistic Regression | 85.97% | 81.66% | 97.66% |
| SVC | 83.90% | 83.45% | 97.12% |
| | | | 83.45% |
| | | | 83.81% |
| | | | 83.81% |
| | | | 83.81% |
| | | | 83.81% |

Tabela 23 – Acurárias obtidas para a combinação de Ansiedade e Surpresa no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 98.22% | 69.99% | 95.36% |
| kNN | 100.00% | 69.35% | 76.70% |
| Passive Aggressive | 88.08% | 65.66% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 94.12% |
| Gradient Boosting | 91.10% | 66.86% | 90.48% |
| XGB | 0.00% | 0.00% | 78.79% |
| MLP | 72.14% | 72.13% | 98.06% |
| Logistic Regression | 79.72% | 71.23% | 96.90% |
| SVC | 74.61% | 71.83% | 95.20% |
| | | | 72.76% |
| | | | 79.64% |
| | | | 71.51% |
| | | | 72.13% |
| | | | 72.13% |
| | | | 72.75% |
| | | | 72.13% |

Tabela 24 – Acurárias obtidas para a combinação de Ansiedade e Confiança no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|--------|
| | treino | teste | treino |
| Random Forest | 97.39% | 77.33% | 96.01% |
| kNN | 100.00% | 71.16% | 76.84% |
| Passive Aggressive | 88.63% | 60.23% | 99.85% |
| Naive Bayes | 0.00% | 0.00% | 96.32% |
| Gradient Boosting | 93.71% | 68.20% | 97.70% |
| XGB | 0.00% | 0.00% | 81.29% |
| MLP | 77.30% | 77.31% | 95.86% |
| Logistic Regression | 81.44% | 77.35% | 97.39% |
| SVC | 77.45% | 77.31% | 96.01% |
| | | | 76.10% |
| | | | 83.13% |
| | | | 78.54% |
| | | | 77.30% |
| | | | 77.31% |
| | | | 77.31% |
| | | | 77.31% |

Tabela 25 – Acurárias obtidas para a combinação de Desgosto e Medo no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 99.16% | 58.99% | 95.33% | 57.09% | 95.52% | 55.21% |
| kNN | 100.00% | 62.28% | 63.33% | 52.23% | 66.23% | 59.35% |
| Passive Aggressive | 89.93% | 67.13% | 100.00% | 57.83% | 100.00% | 60.06% |
| Naive Bayes | 0.00% | 0.00% | 96.83% | 63.79% | 94.49% | 62.33% |
| Gradient Boosting | 97.95% | 69.79% | 96.55% | 60.06% | 98.32% | 62.33% |
| XGB | 0.00% | 0.00% | 71.27% | 60.45% | 73.32% | 60.06% |
| MLP | 53.36% | 55.57% | 99.35% | 60.45% | 52.89% | 50.75% |
| Logistic Regression | 86.57% | 68.29% | 99.07% | 61.56% | 91% | 63.81% |
| SVC | 81.53% | 69.41% | 97.95% | 62.31% | 61.57% | 53.72% |

Tabela 26 – Acurárias obtidas para a combinação de Desgosto e Alegria no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 98.47% | 57.04% | 93.87% | 58.95% | 93.51% | 58.01% |
| kNN | 100.00% | 49.97% | 65.93% | 60.38% | 63.91% | 50.01% |
| Passive Aggressive | 87.49% | 58.48% | 100.00% | 58.99% | 100.00% | 58.97% |
| Naive Bayes | 0.00% | 0.00% | 99.17% | 59.40% | 91.27% | 59.91% |
| Gradient Boosting | 96.93% | 46.69% | 97.52% | 57.08% | 99.18% | 61.31% |
| XGB | 0.00% | 0.00% | 72.17% | 61.77% | 77.12% | 61.34% |
| MLP | 59.43% | 58.96% | 99.65% | 55.66% | 59.43% | 59.44% |
| Logistic Regression | 81.72% | 55.16% | 99.76% | 57.57% | 83.72% | 58.49% |
| SVC | 64.27% | 58.94% | 98.47% | 57.56% | 59.43% | 59.44% |

Tabela 27 – Acurárias obtidas para a combinação de Desgosto e Tristeza no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 97.66% | 70.20% | 93.86% | 72.50% | 93.86% | 71.93% |
| kNN | 100.00% | 75.43% | 73.68% | 73.68% | 73.24% | 73.11% |
| Passive Aggressive | 94.01% | 72.55% | 100.00% | 70.76% | 100.00% | 69.58% |
| Naive Bayes | 0.00% | 0.00% | 98.39% | 62.59% | 74.71% | 73.68% |
| Gradient Boosting | 96.05% | 68.37% | 98.54% | 73.11% | 99.85% | 71.93% |
| XGB | 0.00% | 0.00% | 76.03% | 68.44% | 77.05% | 70.20% |
| MLP | 73.68% | 73.68% | 98.25% | 73.11% | 73.68% | 73.68% |
| Logistic Regression | 81.29% | 73.09% | 98.25% | 71.93% | 73.68% | 73.68% |
| SVC | 76.02% | 73.09% | 97.08% | 70.17% | 73.68% | 73.68% |

Tabela 28 – Acurárias obtidas para a combinação de Desgosto e Surpresa no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 97.92% | 60.19% | 94.21% | 55.13% | 92.25% | 60.69% |
| kNN | 100.00% | 61.54% | 64.22% | 51.88% | 67.24% | 54.68% |
| Passive Aggressive | 90.28% | 62.02% | 100.00% | 56.55% | 100.00% | 58.89% |
| Naive Bayes | 0.00% | 0.00% | 96.53% | 57.45% | 92.02% | 61.59% |
| Gradient Boosting | 98.50% | 59.74% | 96.30% | 66.23% | 98.03% | 65.29% |
| XGB | 0.00% | 0.00% | 70.95% | 64.36% | 73.73% | 63.91% |
| MLP | 58.10% | 57.40% | 97.34% | 60.68% | 58.57% | 57.42% |
| Logistic Regression | 83.22% | 62.53% | 98.27% | 60.68% | 86.00% | 60.21% |
| SVC | 71.42% | 59.75% | 96.53% | 59.27% | 58.33% | 58.33% |

Tabela 29 – Acurárias obtidas para a combinação de Desgosto e Confiança no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|--------|--------|--------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 99.72% | 73.19% | 95.67% | 70.40% | 95.95% | 79.89% |
| kNN | 100.00% | 73.19% | 80.58% | 79.33% | 79.89% | 77.65% |
| Passive Aggressive | 96.10% | 76.51% | 99.86% | 75.38% | 99.86% | 77.06% |
| Naive Bayes | 0.00% | 0.00% | 96.65% | 75.43% | 81.28% | 79.33% |
| Gradient Boosting | 93.72% | 73.78% | 98.46% | 80.46% | 99.44% | 80.44% |
| XGB | 0.00% | 0.00% | 82.40% | 77.67% | 84.92% | 76.56% |
| MLP | 79.33% | 79.33% | 95.53% | 78.78% | 79.33% | 79.33% |
| Logistic Regression | 85.19% | 82.68% | 98.18% | 77.68% | 79% | 79.33% |
| SVC | 80.59% | 79.33% | 97.07% | 79.35% | 79.33% | 79.33% |

Tabela 30 – Acurárias obtidas para a combinação de Medo e Alegria no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 98.17% | 64.33% | 93.51% | 60.97% | 95.33% | 66.70% |
| kNN | 100.00% | 63.50% | 69.71% | 62.57% | 70.53% | 61.00% |
| Passive Aggressive | 84.77% | 61.03% | 100.00% | 52.83% | 100.00% | 66.67% |
| Naive Bayes | 0.00% | 0.00% | 96.54% | 60.93% | 75.20% | 69.93% |
| Gradient Boosting | 96.54% | 57.70% | 97.56% | 69.17% | 99.39% | 66.73% |
| XGB | 0.00% | 0.00% | 75.41% | 65.87% | 78.86% | 60.20% |
| MLP | 69.92% | 69.93% | 98.37% | 66.67% | 69.92% | 69.93% |
| Logistic Regression | 80.28% | 69.10% | 98.17% | 68.27% | 70.53% | 69.93% |
| SVC | 71.95% | 69.10% | 97.15% | 67.43% | 69.92% | 69.93% |

Tabela 31 – Acurárias obtidas para a combinação de Medo e Tristeza no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 97.57% | 59.85% | 92.97% 51.25% 94.49% 46.10% |
| kNN | 100.00% | 59.85% | 65.86% 53.53% 59.16% 51.18% |
| Passive Aggressive | 97.27% | 58.53% | 100.00% 41.47% 100.00% 41.40% |
| Naive Bayes | 0.00% | 0.00% | 98.17% 45.07% 98.17% 51.25% |
| Gradient Boosting | 100.00% | 59.63% | 98.78% 53.82% 100.00% 51.10% |
| XGB | 0.00% | 0.00% | 70.42% 54.85% 75.91% 52.35% |
| MLP | 52.15% | 50.07% | 98.78% 55.81% 50.61% 55.96% |
| Logistic Regression | 87.80% | 59.78% | 98.78% 50.00% 97.25% 52.43% |
| SVC | 74.99% | 59.78% | 98.47% 52.35% 54.88% 54.93% |

Tabela 32 – Acurárias obtidas para a combinação de Medo e Surpresa no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 99.60% | 59.23% | 94.48% 58.34% 94.10% 64.65% |
| kNN | 100.00% | 61.57% | 74.61% 67.78% 73.03% 67.02% |
| Passive Aggressive | 92.14% | 63.85% | 100.00% 53.69% 100.00% 65.35% |
| Naive Bayes | 0.00% | 0.00% | 95.86% 55.32% 77.17% 70.89% |
| Gradient Boosting | 97.44% | 58.37% | 97.64% 70.92% 99.21% 71.72% |
| XGB | 0.00% | 0.00% | 76.97% 68.58% 80.12% 67.78% |
| MLP | 70.87% | 70.89% | 96.85% 70.89% 70.87% 70.89% |
| Logistic Regression | 77.95% | 70.86% | 97.64% 68.58% 71.06% 70.89% |
| SVC | 71.66% | 70.09% | 95.67% 69.35% 70.87% 70.89% |

Tabela 33 – Acurárias obtidas para a combinação de Medo e Confiança no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 98.35% | 65.78% | 94.41% 65.37% 94.95% 59.22% |
| kNN | 76.97% | 64.90% | 63.81% 60.94% 66.34% 61.84% |
| Passive Aggressive | 93.20% | 69.27% | 100.00% 63.56% 100.00% 64.45% |
| Naive Bayes | 0.00% | 0.00% | 97.59% 65.32% 82.35% 62.71% |
| Gradient Boosting | 97.59% | 61.43% | 95.18% 66.23% 97.81% 63.61% |
| XGB | 0.00% | 0.00% | 72.48% 59.20% 74.45% 57.88% |
| MLP | 62.28% | 62.28% | 98.68% 64.45% 62.28% 62.28% |
| Logistic Regression | 84.87% | 69.72% | 98.90% 64.88% 79.72% 61.37% |
| SVC | 75.66% | 68.82% | 96.93% 65.34% 62.28% 62.28% |

Tabela 34 – Acurárias obtidas para a combinação de Alegria e Tristeza no primeiro dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.60% | 71.10% | 95.45% |
| kNN | 100.00% | 71.71% | 75.94% |
| Passive Aggressive | 91.72% | 69.54% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 97.33% |
| Gradient Boosting | 93.58% | 66.32% | 99.20% |
| XGB | 0.00% | 0.00% | 78.34% |
| MLP | 75.94% | 75.93% | 97.46% |
| Logistic Regression | 80.21% | 74.32% | 98.40% |
| SVC | 76.20% | 75.93% | 97.06% |
| | | | 75.93% |
| | | 75.93% | 75.94% |
| | | 75.94% | 75.93% |

Tabela 35 – Acurárias obtidas para a combinação de Alegria e Surpresa no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.52% | 59.88% | 93.96% |
| kNN | 100.00% | 59.92% | 67.99% |
| Passive Aggressive | 90.51% | 56.89% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 94.72% |
| Gradient Boosting | 97.74% | 58.61% | 94.72% |
| XGB | 0.00% | 0.00% | 71.77% |
| MLP | 61.21% | 61.20% | 96.98% |
| Logistic Regression | 81.25% | 63.35% | 97.63% |
| SVC | 66.05% | 61.64% | 94.29% |
| | | 62.53% | 97.20% |
| | | 62.11% | 62.53% |
| | | 60.32% | 76.29% |
| | | 62.53% | 61.21% |
| | | 63.83% | 82.11% |
| | | 65.12% | 61.21% |
| | | 61.20% | 61.20% |

Tabela 36 – Acurárias obtidas para a combinação de Alegria e Confiança no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 98.47% | 67.15% | 91.80% |
| kNN | 100.00% | 66.44% | 35.30% |
| Passive Aggressive | 91.40% | 58.03% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 86.26% |
| Gradient Boosting | 99.43% | 58.72% | 90.07% |
| XGB | 0.00% | 0.00% | 68.51% |
| MLP | 65.65% | 65.64% | 99.24% |
| Logistic Regression | 79.77% | 67.89% | 96.75% |
| SVC | 71.57% | 67.92% | 92.55% |
| | | 54.25% | 99.43% |
| | | 55.10% | 67.15% |
| | | 62.59% | 79.20% |
| | | 55.04% | 63.33% |
| | | 65.65% | 66.41% |
| | | 55.87% | 65.64% |
| | | 65.65% | 65.64% |

Tabela 37 – Acurárias obtidas para a combinação de Tristeza e Surpresa no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 98.44% | 59.08% | 97.82% | 49.38% | 94.03% | 48.90% |
| kNN | 100.00% | 59.63% | 60.36% | 50.02% | 61.65% | 54.57% |
| Passive Aggressive | 91.75% | 60.21% | 100.00% | 60.29% | 100.00% | 60.25% |
| Naive Bayes | 0.00% | 0.00% | 98.58% | 56.84% | 98.72% | 59.14% |
| Gradient Boosting | 100.00% | 64.19% | 99.72% | 60.24% | 100.00% | 58.57% |
| XGB | 0.00% | 0.00% | 73.01% | 55.14% | 78.27% | 52.30% |
| MLP | 52.27% | 50.00% | 99.72% | 55.17% | 49.29% | 48.29% |
| Logistic Regression | 83.66% | 60.79% | 99.86% | 61.41% | 97.73% | 56.30% |
| SVC | 73.87% | 60.75% | 99.29% | 56.33% | 51.14% | 51.14% |

Tabela 38 – Acurárias obtidas para a combinação de Tristeza e Confiança no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 99.44% | 57.78% | 93.89% | 66.67% | 93.52% | 64.44% |
| kNN | 100.00% | 69.63% | 67.04% | 66.67% | 68.15% | 65.93% |
| Passive Aggressive | 95.74% | 60.74% | 100.00% | 60.74% | 100.00% | 59.26% |
| Naive Bayes | 0.00% | 0.00% | 98.33% | 58.52% | 78.52% | 66.67% |
| Gradient Boosting | 99.26% | 59.26% | 100.00% | 62.96% | 100.00% | 62.96% |
| XGB | 0.00% | 0.00% | 71.11% | 62.22% | 74.44% | 60.74% |
| MLP | 66.67% | 66.67% | 99.26% | 63.70% | 66.67% | 66.67% |
| Logistic Regression | 81.48% | 68.15% | 99.44% | 61.48% | 68.33% | 66.67% |
| SVC | 71.67% | 67.41% | 97.96% | 59.26% | 66.67% | 66.67% |

Tabela 39 – Acurárias obtidas para a combinação de Surpresa e Confiança no primeiro *dataset*.

| Combinação de emoções | Acurácia de teste |
|-----------------------|-------------------|
| Raiva/Ansiedade | 71.61% |
| Raiva/Alegria | 81.57% |
| Raiva/Tristeza | 69.60% |
| Ansiedade/Alegria | 65.75% |
| Desgosto/Alegria | 72.24% |
| Desgosto/Confiança | 67.78% |
| Alegria/Tristeza | 70.17% |

Tabela 40 – Acurárias de teste obtidas com o BERT no primeiro *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.87% | 61.11% | 95.75% |
| kNN | 99.60% | 62.79% | 53.33% |
| Passive Aggressive | 89.21% | 68.56% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 96.94% |
| Gradient Boosting | 99.47% | 66.49% | 97.87% |
| XGB | 0.00% | 0.00% | 77.26% |
| MLP | 50.52% | 58.48% | 98.01% |
| Logistic Regression | 87.37% | 70.18% | 98.67% |
| SVC | 74.20% | 64.88% | 98.27% |
| | | | 67.03% |
| | | | 95.88% |
| | | | 66.46% |
| | | | 99.73% |
| | | | 68.09% |
| | | | 78.99% |
| | | | 67.03% |
| | | | 56.65% |
| | | | 53.21% |
| | | | 85.11% |
| | | | 70.74% |
| | | | 75.26% |
| | | | 63.29% |

Tabela 41 – Acurárias obtidas para a combinação de Raiva e Ansiedade no segundo dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|--------|
| | treino | teste | treino |
| Random Forest | 85.91% | 40.91% | 84.20% |
| kNN | 87.50% | 39.55% | 65.11% |
| Passive Aggressive | 71.14% | 46.36% | 87.50% |
| Naive Bayes | 0.00% | 0.00% | 46.36% |
| Gradient Boosting | 87.50% | 47.27% | 86.59% |
| XGB | 0.00% | 0.00% | 45.00% |
| MLP | 50.11% | 50.91% | 68.52% |
| Logistic Regression | 76.36% | 46.82% | 49.09% |
| SVC | 63.86% | 50.00% | 48.18% |
| | | | 73.52% |
| | | | 84.89% |
| | | | 87.05% |
| | | | 47.50% |
| | | | 87.50% |
| | | | 48.18% |
| | | | 50.91% |
| | | | 49.55% |
| | | | 48.64% |
| | | | 85.68% |
| | | | 46.82% |
| | | | 52.84% |
| | | | 52.73% |

Tabela 42 – Acurárias obtidas para a combinação de Raiva e Desgosto no segundo dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.55% | 69.26% | 94.58% |
| kNN | 100.00% | 72.73% | 55.25% |
| Passive Aggressive | 96.68% | 69.85% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 66.40% |
| Gradient Boosting | 97.20% | 68.47% | 93.88% |
| XGB | 0.00% | 0.00% | 56.75% |
| MLP | 72.73% | 72.73% | 72.02% |
| Logistic Regression | 82.17% | 72.00% | 72.76% |
| SVC | 75.00% | 72.73% | 76.57% |
| | | | 72.04% |
| | | | 99.13% |
| | | | 74.12% |
| | | | 77.27% |
| | | | 72.73% |
| | | | 72.73% |
| | | | 72.73% |
| | | | 72.73% |
| | | | 72.73% |
| | | | 72.73% |
| | | | 72.73% |
| | | | 72.73% |

Tabela 43 – Acurárias obtidas para a combinação de Raiva e Medo no segundo dataset.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 99.26% | 69.40% | 94.95% 59.10% 95.57% 66.04% |
| kNN | 100.00% | 66.04% | 73.04% 61.50% 70.69% 67.46% |
| Passive Aggressive | 94.20% | 77.79% | 100.00% 69.41% 100.00% 67.49% |
| Naive Bayes | 0.00% | 0.00% | 98.65% 70.93% 97.17% 69.96% |
| Gradient Boosting | 99.63% | 77.80% | 99.51% 65.52% 99.75% 67.50% |
| XGB | 0.00% | 0.00% | 75.86% 63.11% 78.94% 66.52% |
| MLP | 51.73% | 52.73% | 99.14% 69.43% 49.63% 49.77% |
| Logistic Regression | 90.02% | 78.27% | 99.88% 69.93% 91.87% 68.00% |
| SVC | 83.62% | 76.34% | 98.77% 67.98% 76.11% 62.09% |

Tabela 44 – Acurárias obtidas para a combinação de Raiva e Alegria no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 92.89% | 48.00% | 89.58% 54.51% 8971.00% 50.98% |
| kNN | 94.61% | 51.88% | 58.58% 54.87% 62.87% 49.98% |
| Passive Aggressive | 82.96% | 50.96% | 94.11% 48.52% 93.38% 50.01% |
| Naive Bayes | 0.00% | 0.00% | 92.89% 51.93% 92.77% 51.91% |
| Gradient Boosting | 94.24% | 55.33% | 93.75% 52.43% 94.36% 54.39% |
| XGB | 0.00% | 0.00% | 67.16% 50.96% 83.53% 49.99% |
| MLP | 49.02% | 49.51% | 93.87% 54.37% 50.12% 50.49% |
| Logistic Regression | 82.60% | 55.35% | 94.12% 53.38% 90.81% 51.46% |
| SVC | 66.42% | 53.39% | 93.26% 56.30% 53.55% 50.96% |

Tabela 45 – Acurárias obtidas para a combinação de Raiva e Tristeza no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 98.25% | 69.24% | 95.10% 72.02% 92.83% 72.02% |
| kNN | 100.00% | 69.24% | 72.20% 72.02% 76.23% 72.73% |
| Passive Aggressive | 93.69% | 69.88% | 100.00% 69.88% 100.00% 69.21% |
| Naive Bayes | 0.00% | 0.00% | 98.43% 58.00% 74.82% 72.73% |
| Gradient Boosting | 97.73% | 69.19% | 98.95% 72.71% 100.00% 71.31% |
| XGB | 0.00% | 0.00% | 75.25% 72.04% 78.50% 70.64% |
| MLP | 72.73% | 72.73% | 99.30% 69.88% 72.73% 72.73% |
| Logistic Regression | 81.12% | 69.21% | 99.30% 69.88% 72.73% 72.73% |
| SVC | 74.30% | 73.42% | 98.78% 71.31% 72.73% 72.73% |

Tabela 46 – Acurárias obtidas para a combinação de Raiva e Surpresa no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 97.30% | 62.23% | 92.97% | 57.61% | 92.97% | 62.26% |
| kNN | 100.00% | 65.85% | 70.22% | 61.76% | 73.66% | 67.74% |
| Passive Aggressive | 93.57% | 72.96% | 100.00% | 62.89% | 100.00% | 64.10% |
| Naive Bayes | 0.00% | 0.00% | 96.86% | 63.44% | 87.88% | 65.29% |
| Gradient Boosting | 99.40% | 68.81% | 97.45% | 65.29% | 99.55% | 64.71% |
| XGB | 0.00% | 0.00% | 72.01% | 62.85% | 75.00% | 57.49% |
| MLP | 62.28% | 62.28% | 98.95% | 64.05% | 62.28% | 62.28% |
| Logistic Regression | 84.58% | 74.81% | 99.25% | 65.88% | 78.45% | 62.89% |
| SVC | 73.20% | 66.45% | 97.16% | 64.06% | 62.28% | 62.28% |

Tabela 47 – Acurárias obtidas para a combinação de Raiva e Confiança no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 98.38% | 60.00% | 95.50% | 62.50% | 95.88% | 61.50% |
| kNN | 100.00% | 62.50% | 48.12% | 44.50% | 68.00% | 61.50% |
| Passive Aggressive | 93.25% | 66.00% | 100.00% | 63.50% | 100.00% | 70.50% |
| Naive Bayes | 0.00% | 0.00% | 98.25% | 54.50% | 96.37% | 72.00% |
| Gradient Boosting | 98.62% | 66.00% | 98.12% | 64.50% | 99.50% | 64.50% |
| XGB | 0.00% | 0.00% | 73.88% | 64.50% | 76.88% | 63.00% |
| MLP | 57.88% | 56.00% | 99.50% | 69.50% | 58.38% | 57.50% |
| Logistic Regression | 84.38% | 64.00% | 99.50% | 70.00% | 86.88% | 70.00% |
| SVC | 74.00% | 65.00% | 99.50% | 68.00% | 61.13% | 59.00% |

Tabela 48 – Acurárias obtidas para a combinação de Ansiedade e Desgosto no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|--------|--------|--------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 94.10% | 60.80% | 92.67% | 64.87% | 92.88% | 70.57% |
| kNN | 95.31% | 63.27% | 78.24% | 69.70% | 76.42% | 71.40% |
| Passive Aggressive | 85.37% | 64.90% | 95.11% | 63.20% | 94.91% | 65.67% |
| Naive Bayes | 0.00% | 0.00% | 88.61% | 64.10% | 78.44% | 67.47% |
| Gradient Boosting | 94.71% | 62.40% | 91.87% | 63.33% | 94.10% | 68.13% |
| XGB | 0.00% | 0.00% | 79.46% | 68.87% | 82.31% | 68.13% |
| MLP | 68.29% | 68.30% | 93.69% | 69.67% | 68.29% | 68.30% |
| Logistic Regression | 80.48% | 69.87% | 93.89% | 67.23% | 80.27% | 69.87% |
| SVC | 76.01% | 68.30% | 92.88% | 64.80% | 68.29% | 68.30% |

Tabela 49 – Acurárias obtidas para a combinação de Ansiedade e Medo no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 98.64% | 60.56% | 94.68% |
| kNN | 100.00% | 56.89% | 59.96% |
| Passive Aggressive | 89.88% | 57.39% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 98.50% |
| Gradient Boosting | 99.73% | 59.58% | 98.09% |
| XGB | 0.00% | 0.00% | 71.45% |
| MLP | 46.44% | 51.34% | 98.77% |
| Logistic Regression | 85.38% | 61.20% | 98.91% |
| SVC | 76.91% | 62.30% | 98.22% |
| | | | 60.08% |
| | | | 99.73% |
| | | | 59.56% |
| | | | 74.86% |
| | | | 59.56% |
| | | | 51.23% |
| | | | 44.77% |
| | | | 92.76% |
| | | | 62.85% |
| | | | 54.10% |
| | | | 54.10% |

Tabela 50 – Acurárias obtidas para a combinação de Ansiedade e Alegria no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.01% | 55.39% | 95.24% |
| kNN | 100.00% | 57.64% | 64.66% |
| Passive Aggressive | 91.19% | 54.91% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 97.69% |
| Gradient Boosting | 99.73% | 58.18% | 97.15% |
| XGB | 0.00% | 0.00% | 72.01% |
| MLP | 50.26% | 47.27% | 99.18% |
| Logistic Regression | 83.56% | 58.17% | 99.32% |
| SVC | 69.83% | 54.88% | 98.78% |
| | | | 59.00% |
| | | | 99.59% |
| | | | 60.30% |
| | | | 74.46% |
| | | | 53.29% |
| | | | 47.69% |
| | | | 52.73% |
| | | | 88.99% |
| | | | 64.68% |
| | | | 56.25% |
| | | | 54.35% |

Tabela 51 – Acurárias obtidas para a combinação de Ansiedade e Tristeza no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.36% | 65.80% | 93.50% |
| kNN | 100.00% | 59.37% | 68.29% |
| Passive Aggressive | 92.47% | 61.87% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 98.17% |
| Gradient Boosting | 96.95% | 61.67% | 99.80% |
| XGB | 0.00% | 0.00% | 73.98% |
| MLP | 68.29% | 68.30% | 98.99% |
| Logistic Regression | 82.72% | 65.83% | 98.78% |
| SVC | 70.53% | 67.50% | 97.97% |
| | | | 68.30% |
| | | | 100.00% |
| | | | 67.47% |
| | | | 70.73% |
| | | | 66.60% |
| | | | 76.83% |
| | | | 68.30% |
| | | | 100.00% |
| | | | 67.43% |
| | | | 81.71% |
| | | | 63.40% |
| | | | 68.29% |
| | | | 68.30% |
| | | | 74.38% |
| | | | 68.30% |
| | | | 68.29% |
| | | | 68.30% |

Tabela 52 – Acurárias obtidas para a combinação de Ansiedade e Surpresa no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 98.81% | 61.08% | 94.90% | 62.67% | 93.03% | 56.53% |
| kNN | 100.00% | 57.06% | 59.02% | 58.53% | 68.54% | 63.95% |
| Passive Aggressive | 84.34% | 63.86% | 100.00% | 61.93% | 100.00% | 57.15% |
| Naive Bayes | 0.00% | 0.00% | 94.73% | 64.64% | 87.41% | 61.20% |
| Gradient Boosting | 99.66% | 54.34% | 98.81% | 64.62% | 100.00% | 61.20% |
| XGB | 0.00% | 0.00% | 70.24% | 61.22% | 78.56% | 63.89% |
| MLP | 55.10% | 57.15% | 97.62% | 63.24% | 58.34% | 55.08% |
| Logistic Regression | 85.38% | 64.55% | 97.62% | 63.95% | 86.73% | 61.84% |
| SVC | 72.09% | 57.15% | 95.41% | 64.60% | 57.14% | 57.15% |

Tabela 53 – Acurárias obtidas para a combinação de Ansiedade e Confiança no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|--------|--------|--------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 95.32% | 72.90% | 93.55% | 70.97% | 93.39% | 73.55% |
| kNN | 98.23% | 70.97% | 54.52% | 48.39% | 76.13% | 73.55% |
| Passive Aggressive | 86.77% | 69.03% | 98.23% | 70.97% | 98.23% | 72.90% |
| Naive Bayes | 0.00% | 0.00% | 94.35% | 61.29% | 78.06% | 74.84% |
| Gradient Boosting | 93.23% | 67.74% | 95.81% | 74.19% | 97.58% | 73.55% |
| XGB | 0.00% | 0.00% | 77.26% | 73.55% | 80.97% | 73.55% |
| MLP | 74.84% | 74.84% | 93.71% | 72.90% | 74.84% | 74.84% |
| Logistic Regression | 82.90% | 74.19% | 94.03% | 72.26% | 75.48% | 74.84% |
| SVC | 76.61% | 74.84% | 93.23% | 72.90% | 74.84% | 74.84% |

Tabela 54 – Acurárias obtidas para a combinação de Desgosto e Medo no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 99.53% | 63.72% | 96.63% | 63.26% | 95.81% | 63.26% |
| kNN | 100.00% | 67.91% | 73.72% | 55.35% | 70.81% | 61.40% |
| Passive Aggressive | 93.95% | 68.84% | 100.00% | 67.44% | 100.00% | 66.05% |
| Naive Bayes | 0.00% | 0.00% | 98.95% | 65.12% | 99.30% | 66.98% |
| Gradient Boosting | 99.65% | 71.16% | 99.42% | 65.12% | 99.65% | 67.44% |
| XGB | 0.00% | 0.00% | 71.28% | 59.07% | 74.77% | 56.74% |
| MLP | 51.74% | 49.77% | 99.77% | 64.19% | 49.19% | 50.23% |
| Logistic Regression | 88.84% | 74.88% | 99.88% | 66.05% | 93% | 63.26% |
| SVC | 81.74% | 73.95% | 99.42% | 65.58% | 57.91% | 53.02% |

Tabela 55 – Acurárias obtidas para a combinação de Desgosto e Alegria no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|--------|
| | treino | teste | treino |
| Random Forest | 85.76% | 56.93% | 83.22% |
| kNN | 87.15% | 49.50% | 54.97% |
| Passive Aggressive | 74.07% | 55.52% | 87.15% |
| Naive Bayes | 0.00% | 0.00% | 87.03% |
| Gradient Boosting | 86.57% | 53.20% | 86.34% |
| XGB | 0.00% | 0.00% | 68.86% |
| MLP | 49.42% | 51.37% | 86.80% |
| Logistic Regression | 78.47% | 50.90% | 87.03% |
| SVC | 64.81% | 50.42% | 86.69% |
| | | | 54.14% |
| | | | 53.70% |
| | | | 53.70% |

Tabela 56 – Acurárias obtidas para a combinação de Desgosto e Tristeza no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.74% | 70.32% | 94.52% |
| kNN | 100.00% | 72.26% | 74.84% |
| Passive Aggressive | 94.03% | 68.39% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 98.71% |
| Gradient Boosting | 96.94% | 69.03% | 99.35% |
| XGB | 0.00% | 0.00% | 76.13% |
| MLP | 74.84% | 74.84% | 99.19% |
| Logistic Regression | 80.65% | 73.55% | 99.52% |
| SVC | 76.29% | 74.84% | 98.55% |
| | | | 73.55% |
| | | | 100.00% |
| | | | 74.19% |
| | | | 72.90% |
| | | | 74.19% |
| | | | 74.84% |
| | | | 74.84% |
| | | | 74.84% |
| | | | 74.84% |
| | | | 74.84% |
| | | | 74.84% |

Tabela 57 – Acurárias obtidas para a combinação de Desgosto e Surpresa no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 96.65% | 62.57% | 95.11% |
| kNN | 100.00% | 57.46% | 69.70% |
| Passive Aggressive | 93.58% | 68.08% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 97.07% |
| Gradient Boosting | 97.35% | 65.35% | 97.07% |
| XGB | 0.00% | 0.00% | 73.60% |
| MLP | 64.80% | 64.81% | 98.46% |
| Logistic Regression | 81.00% | 66.46% | 98.60% |
| SVC | 69.27% | 64.24% | 97.07% |
| | | | 68.70% |
| | | | 66.51% |
| | | | 99.58% |
| | | | 64.80% |
| | | | 64.81% |
| | | | 65.94% |
| | | | 64.80% |
| | | | 64.81% |

Tabela 58 – Acurárias obtidas para a combinação de Desgosto e Confiança no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 99.64% | 64.50% | 95.11% |
| kNN | 100.00% | 79.05% | 73.00% |
| Passive Aggressive | 100.00% | 81.80% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 93.29% |
| Gradient Boosting | 99.10% | 75.21% | 97.28% |
| XGB | 0.00% | 0.00% | 81.16% |
| MLP | 71.74% | 71.75% | 94.93% |
| Logistic Regression | 90.58% | 78.92% | 95.29% |
| SVC | 83.15% | 76.03% | 93.66% |
| | | 72.43% | 72.43% |
| | | 71.74% | 71.75% |

Tabela 59 – Acurárias obtidas para a combinação de Medo e Alegria no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|--------|
| | treino | teste | treino |
| Random Forest | 92.45% | 63.36% | 89.57% |
| kNN | 92.98% | 71.96% | 72.66% |
| Passive Aggressive | 87.78% | 66.93% | 92.62% |
| Naive Bayes | 0.00% | 0.00% | 87.23% |
| Gradient Boosting | 91.01% | 62.65% | 90.11% |
| XGB | 0.00% | 0.00% | 75.00% |
| MLP | 71.94% | 71.96% | 87.59% |
| Logistic Regression | 81.29% | 69.10% | 88.49% |
| SVC | 72.84% | 71.96% | 86.33% |
| | | 70.53% | 70.53% |
| | | 71.94% | 71.96% |

Tabela 60 – Acurárias obtidas para a combinação de Medo e Tristeza no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 99.36% | 58.92% | 94.54% |
| kNN | 100.00% | 51.25% | 50.00% |
| Passive Aggressive | 99.68% | 61.58% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 99.37% |
| Gradient Boosting | 100.00% | 60.33% | 100.00% |
| XGB | 0.00% | 0.00% | 73.71% |
| MLP | 51.30% | 46.17% | 100.00% |
| Logistic Regression | 92.62% | 57.50% | 100.00% |
| SVC | 83.64% | 57.67% | 99.37% |
| | | 61.42% | 61.42% |
| | | 75.80% | 75.80% |
| | | 59.92% | 59.92% |

Tabela 61 – Acurárias obtidas para a combinação de Medo e Surpresa no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 98.53% | 55.86% | 96.08% 58.81% 96.32% 64.62% |
| kNN | 100.00% | 65.52% | 53.74% 41.19% 70.34% 65.57% |
| Passive Aggressive | 97.07% | 67.52% | 100.00% 70.38% 100.00% 69.38% |
| Naive Bayes | 0.00% | 0.00% | 97.06% 67.67% 90.68% 64.67% |
| Gradient Boosting | 99.76% | 61.67% | 97.31% 59.81% 99.51% 62.71% |
| XGB | 0.00% | 0.00% | 74.75% 65.43% 81.36% 66.43% |
| MLP | 61.76% | 61.76% | 99.02% 70.38% 61.76% 61.76% |
| Logistic Regression | 85.54% | 67.57% | 99.26% 70.43% 82.36% 61.67% |
| SVC | 72.54% | 64.62% | 96.81% 65.62% 61.76% 61.76% |

Tabela 62 – Acurárias obtidas para a combinação de Medo e Confiança no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-------------------------------|
| | treino | teste | treino teste |
| Random Forest | 99.12% | 63.35% | 94.60% 51.73% 93.85% 53.26% |
| kNN | 100.00% | 63.87% | 68.46% 51.26% 66.96% 52.28% |
| Passive Aggressive | 94.10% | 65.32% | 100.00% 54.79% 100.00% 52.27% |
| Naive Bayes | 0.00% | 0.00% | 97.61% 55.29% 97.49% 52.29% |
| Gradient Boosting | 99.62% | 62.82% | 98.62% 54.27% 99.75% 57.28% |
| XGB | 0.00% | 0.00% | 70.48% 50.23% 74.62% 54.72% |
| MLP | 49.37% | 50.26% | 99.25% 51.23% 50.13% 49.74% |
| Logistic Regression | 87.19% | 65.31% | 99.75% 51.73% 94.85% 54.27% |
| SVC | 77.64% | 65.82% | 97.61% 52.27% 63.29% 51.26% |

Tabela 63 – Acurárias obtidas para a combinação de Alegria e Tristeza no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|-----------------------------|
| | treino | teste | treino teste |
| Random Forest | 94.37% | 62.25% | 91.66% 65.85% 89.49% 68.81% |
| kNN | 95.82% | 65.82% | 71.74% 71.75% 74.27% 68.78% |
| Passive Aggressive | 88.39% | 61.48% | 95.82% 65.13% 93.64% 62.33% |
| Naive Bayes | 0.00% | 0.00% | 93.47% 61.46% 75.54% 71.75% |
| Gradient Boosting | 92.92% | 60.74% | 94.56% 67.33% 95.64% 65.82% |
| XGB | 0.00% | 0.00% | 74.81% 67.35% 77.17% 65.13% |
| MLP | 71.74% | 71.75% | 94.38% 66.53% 71.74% 71.75% |
| Logistic Regression | 78.98% | 68.07% | 94.38% 66.59% 71.92% 71.75% |
| SVC | 72.46% | 70.29% | 93.65% 65.85% 71.74% 71.75% |

Tabela 64 – Acurárias obtidas para a combinação de Alegria e Surpresa no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 97.99% | 59.91% | 61.72% |
| kNN | 100.00% | 54.94% | 64.67% |
| Passive Aggressive | 91.35% | 69.07% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 95.52% |
| Gradient Boosting | 99.84% | 62.99% | 59.30% |
| XGB | 0.00% | 0.00% | 76.54% |
| MLP | 60.96% | 61.12% | 97.69% |
| Logistic Regression | 87.96% | 72.16% | 62.97% |
| SVC | 75.61% | 63.56% | 96.91% |
| | | 62.92% | 61.11% |
| | | | 61.12% |

Tabela 65 – Acurárias obtidas para a combinação de Alegria e Confiança no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|--------|
| | treino | teste | treino |
| Random Forest | 93.52% | 68.36% | 88.84% |
| kNN | 94.96% | 67.65% | 71.94% |
| Passive Aggressive | 84.35% | 60.48% | 94.96% |
| Naive Bayes | 0.00% | 0.00% | 93.70% |
| Gradient Boosting | 93.34% | 58.99% | 94.78% |
| XGB | 0.00% | 0.00% | 73.38% |
| MLP | 71.94% | 71.96% | 93.52% |
| Logistic Regression | 78.78% | 66.96% | 69.81% |
| SVC | 73.02% | 71.96% | 92.45% |
| | | 68.39% | 71.94% |
| | | | 71.96% |

Tabela 66 – Acurárias obtidas para a combinação de Tristeza e Surpresa no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF |
|---------------------|---------------------------|------------------|---------|
| | treino | teste | treino |
| Random Forest | 98.01% | 50.28% | 93.71% |
| kNN | 100.00% | 56.42% | 63.50% |
| Passive Aggressive | 91.41% | 57.05% | 100.00% |
| Naive Bayes | 0.00% | 0.00% | 97.55% |
| Gradient Boosting | 98.62% | 53.43% | 97.24% |
| XGB | 0.00% | 0.00% | 72.24% |
| MLP | 61.35% | 61.36% | 98.31% |
| Logistic Regression | 79.90% | 61.36% | 98.16% |
| SVC | 65.95% | 61.99% | 97.55% |
| | | 62.61% | 78.52% |
| | | | 60.15% |
| | | 59.55% | 61.35% |
| | | | 61.36% |

Tabela 67 – Acurárias obtidas para a combinação de Tristeza e Confiança no segundo *dataset*.

| Classificador | Word Embedding (fastText) | Count Vectorizer | TF-IDF | | | |
|---------------------|---------------------------|------------------|---------|--------|---------|--------|
| | treino | teste | treino | teste | treino | teste |
| Random Forest | 99.27% | 50.90% | 94.61% | 62.71% | 93.38% | 60.81% |
| kNN | 100.00% | 58.95% | 62.98% | 61.76% | 69.11% | 61.86% |
| Passive Aggressive | 96.31% | 64.71% | 100.00% | 61.86% | 100.00% | 63.86% |
| Naive Bayes | 0.00% | 0.00% | 99.51% | 53.05% | 89.47% | 61.76% |
| Gradient Boosting | 99.51% | 61.62% | 99.27% | 60.81% | 100.00% | 57.81% |
| XGB | 0.00% | 0.00% | 69.62% | 57.95% | 77.22% | 58.81% |
| MLP | 61.76% | 61.76% | 100.00% | 59.95% | 61.76% | 61.76% |
| Logistic Regression | 80.88% | 64.71% | 99.75% | 60.90% | 77.21% | 61.76% |
| SVC | 71.81% | 66.71% | 99.02% | 59.95% | 61.76% | 61.76% |

Tabela 68 – Acurárias obtidas para a combinação de Surpresa e Confiança no segundo *dataset*.

| Combinação de emoções | Acurácia de teste |
|-----------------------|-------------------|
| Raiva/Ansiedade | 67.66% |
| Raiva/Alegria | 77.65% |
| Raiva/Confiança | 78.09% |
| Ansiedade/Desgosto | 72.40% |
| Desgosto/Alegria | 75.19% |
| Desgosto/Confiança | 70.67% |
| Alegria/Tristeza | 76.80% |
| Alegria/Confiança | 67.80% |

Tabela 69 – Acurárias de teste obtidas com o BERT no segundo *dataset*.