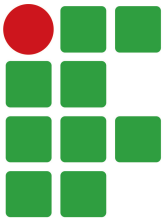

INSTITUTO FEDERAL
Goiás
Câmpus Anápolis

Nátallya Soares Costa

Aplicação de Técnicas de Inteligência Artificial para Análise de Tweets sobre Institutos Federais Brasileiros

Anápolis-GO

2023

Nátallya Soares Costa

Aplicação de Técnicas de Inteligência Artificial para Análise de Tweets sobre Institutos Federais Brasileiros

Projeto de Pesquisa apresentado ao curso de Bacharelado em Ciências da Computação, como requisito para obtenção do grau final na disciplina de Projeto Final de Curso.

Instituto Federal de Goiás - Câmpus Anápolis

Orientador: Prof. Dr. Sérgio Daniel Carvalho Canuto

Anápolis-GO
2023

Dados Internacionais de Catalogação na Publicação (CIP)

Costa, Nátallya Soares

C837a Aplicação de ténicas de inteligência artificial para análise de *tweets* sobre Institutos Federais brasileiros./ Nátallya Soares Costa. – Anápolis: IFG, 2023.

38 f. il. color.

Orientador: Prof. Dr. Sérgio Daniel Carvalho Canuto.

Trabalho de Conclusão de Curso – Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Campus Anápolis – Bacharelado em Ciência da Computação, 2023.

1. Inteligência artificial. 2. técnicas. 3. tweets. 4. Instituto Federal. 4. Brasil.
I. Canuto, Sérgio Daniel Carvalho (orient.).
II. Título.

CDD 006.3

Ficha catalográfica elaborada pela bibliotecária Claudineia Pereira de Abreu CRB1/1956
IFG Campus Anápolis

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

[ ] Tese	[ ] Artigo Científico
[ ] Dissertação	[ ] Capítulo de Livro
[ ] Monografia – Especialização	[ ] Livro
[X] TCC - Graduação	[ ] Trabalho Apresentado em Evento
[ ] Produto Técnico e Educacional - Tipo: _____	

Nome Completo do Autor:
Matrícula:
Título do Trabalho:

Autorização - Marque uma das opções

1. (X) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. ( ) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data ___/___/_____ (Embargo);
3. ( ) Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

( ) O documento está sujeito a registro de patente.
( ) O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
( ) Outra justificativa: _____

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

i. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
ii. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
iii. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Anápolis, 01/03/2024

Natallya Costa
_____
Assinatura do Autor e/ou Detentor dos Direitos Autorais

ATA DA SESSÃO PÚBLICA DE APRESENTAÇÃO E DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

Aos **dezoito** dias do mês de dezembro de **2023**, às **14h30**, em sessão pública por meio de webconferência via Google Meet, foi realizada a sessão pública de apresentação e qualificação do Trabalho de Conclusão de Curso da Graduanda **Natallya Soares Costa (matrícula 20201060140135)** do curso de Bacharelado em Ciência da Computação. A banca foi composta pelos seguintes membros: **Dr. Sérgio Daniel Carvalho Canuto, Dr. Daniel Xavier Sousa, e Dr. Thierson Couto Rosa** , sob a presidência do primeiro. O trabalho de conclusão de curso tem como título **Aplicação de Técnicas de Inteligência Artificial para Análise de Tweets sobre Institutos Federais Brasileiros**, sob orientação de **Sérgio Daniel Carvalho Canuto.** Após a apresentação, tendo sido o autor arguido pela Banca Examinadora, a nota obtida foi **9,5** e o Trabalho de Conclusão de Curso foi considerado **APROVADO** pela banca examinadora.

Encerra-se a presente sessão às **16** horas e **00** minutos. Eu, **Sérgio Daniel Carvalho Canuto**, dato e assino a presente ata que segue assinada por todos os membros da Banca e pelo graduando.

Dr. Sérgio Daniel Carvalho Canuto
(Assinado Eletronicamente)

Dr. Daniel Xavier Sousa
(Assinado Eletronicamente)

Documento assinado digitalmente
gov.br
THIERSON COUTO ROSA
Data: 22/12/2023 14:14:12-0300
Verifique em https://validar.iti.gov.br

(Assinado Eletronicamente)

Natallya Soares Costa
(Graduando)

Documento assinado eletronicamente por:

- **Natallya Soares Costa, 20201060140135 - Discente**, em 21/12/2023 15:50:17.
- **Daniel Xavier de Sousa, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 18/12/2023 16:22:11.
- **Sergio Daniel Carvalho Canuto, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 18/12/2023 15:48:31.

Este documento foi emitido pelo SUAP em 18/12/2023. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifg.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 491985
Código de Autenticação: 73178d14c5

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

GPT	Generative Pre-trained Transformer
LIWC	Linguistic Inquiry and Word Count
K-NN	K- Nearest Neighbors
MT	Modelagem de Tópicos
PLN	Processamento de Linguagem Natural
TF-IDF	Term Frequency-Inverse Document Frequency
SVM	Support Vector Machine

Sumário

1 Introdução

A conectividade entre indivíduos por meio do uso de redes sociais como ferramentas de transmissão de idéias tem se tornado cada vez mais proeminente nos últimos anos. Como resultado, há um considerável volume de dados textuais diariamente gerados por usuários de redes sociais sobre os mais diversos e efêmeros temas de interesse. Portanto, tornou-se frequente o uso de ferramentas capazes de fornecer suporte automático para análise de dados disponíveis no contexto político, social e comercial, incluindo satisfação de usuários com marcas, produtos, serviços (YUE L. CHEN, 2019).

No contexto das instituições públicas, é reconhecido um grande potencial pouco explorado por meio de técnicas baseadas em Inteligência Artificial (IA) sobre tais instituições em redes sociais (BEGANY; MARTIN, 2020). De fato, desafios para aplicação de tais técnicas no âmbito da gestão pública envolvem não somente questões de engajamento ativo da gestão pública e sociedade, mas também a necessidade de uma avaliação criteriosa da eficácia de tais técnicas para tarefas como Análise de Sentimento e Modelagem de Tópicos presentes nos textos curtos de redes sociais.

Especificamente, Análise de Sentimento visa categorizar textos de redes sociais entre positivos, negativos ou neutros, o que permite quantificar o sentimento dos usuários em relação ao contexto específico de mensagens de texto sobre os Institutos Federais. A Análise de Sentimento efetiva tem o potencial de auxiliar atividades como o acompanhamento do sentimento geral da comunidade em relação aos IFs no tempo, a antecipação de crises no âmbito da comunicação e a identificação do apoio ou reprovação da comunidade em relação a medidas adotadas pela gestão dos institutos. Nesse trabalho, foi investigada a eficácia de Análise de Sentimento com a aplicação de 11 técnicas, sendo elas baseadas em dicionários léxicos (lexicons) (RIBEIRO et al., 2016), técnicas tradicionais de aprendizado de máquina (CAMPOS et al., 2017b) e estratégias no estado-da-arte baseadas em transformers (SOUZA; FILHO, 2022a).

Adicionalmente, a aplicação de técnicas para a tarefa Modelagem de Tópicos possuem o potencial de indentificar aspectos/tópicos presentes nos discursos das redes sociais, aprofundando a compreensão e análise refinada de tópicos emergentes, preocupações sociais e temas recorrentes (RESCH; USLäNDER; HAVAS, 2023). Nesse trabalho, a Modelagem de Tópicos foi aplicada por meio da técnica recente BERTTopic (GROOTENDORST, 2022) para identificação de tópicos globais (referentes a todos Institutos Federais brasileiros) e regionais (onde os institutos foram divididos nas 5 regiões do Brasil) em tweets sobre os institutos.

Esse trabalho é pioneiro na investigação de técnicas para as tarefas de Análise de

Sentimento e Modelagem de Tópicos em redes sociais no contexto dos Institutos Federais brasileiros. Portanto foram empregadas as etapas para descoberta do conhecimento em base de dados (Knowledge Discovery in Databases (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)) incluindo, além da aplicação de técnicas de Análise de Sentimento e Modelagem de Tópicos, a coleta, pré-processamento e rotulação manual dos dados, bem como validação criteriosa dos resultados obtidos.

A partir dos resultados obtidos, foi possível observar que o métodos baseados em transformers obtiveram os melhores resultados para a tarefa de Análise de Sentimento, sendo o BERTimbau (FILHO RODRIGO WILKENS, 2018) o mais efetivo dentre os métodos avaliados, atingindo efetividade de até 86% (em Macro-F1). Tais resultados foram muito superiores aos obtidos com estratégias baseadas em dicionários léxicos e técnicas clássicas de aprendizado de máquina, que obtiveram efetividade inferior a 56%. Portanto, métodos baseados em transformers podem ser considerados satisfatórios para viabilizar futuras aplicações com potencial para análise de sentimentos no contexto dos IFs.

Dentre os tópicos mais relevantes identificados, foi observado, para os tópicos globais, um tópico referente aos avisos das equipes de comunição dos câmpus, que reflete o engajamento da gestão dos institutos para o contato com a sociedade. Outro tópico global identificado reflete ao medo e ansiedade dos alunos em relação às avaliações. Foram também observados tópicos de interesse referentes às questões regionais, como preocupações com alimentação, repercussão de projeto de extensão e organizações para votação de representantes.

Este trabalho está organizado seguindo a seguinte estrutura: a seção 2 expõe os objetivos desse trabalho, a Seção 3 apresenta a fundamentação teórica. Em seguida, na Seção 4, são descritos os passos metodológicos empregados para alcançar o objetivo deste trabalho. Na Seção 5 são expostos os resultados experimentais e na Seção 6, a conclusão e potenciais desdobramentos futuros do trabalho.

2 Objetivos

2.1 Objetivo Geral

Este trabalho visa a aplicação de técnicas de Análise de Sentimento e Modelagem de Tópicos em tweets sobre os Institutos Federais brasileiros, bem como a análise criteriosa dos resultados obtidos por meio da aplicação de tais técnicas seguindo as etapas para descoberta do conhecimento em bases de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.2 Objetivos específicos

- Coletar dados do Twitter referentes aos Institutos Federais brasileiros.
- Aplicar técnicas para o pré-processamento dos dados textuais coletados.
- Apresentar uma comparação detalhada dos modelos de análise de sentimento baseados em métodos léxicos, aprendizado de máquina e técnicas recentes baseadas em transformers na coleção de tweets sobre Institutos Federais.
- Avaliar métodos de análise de sentimento nativos em inglês por meio da tradução automática de tweets.
- Caracterizar a base de dados utilizada pelos métodos de análise de sentimento com base na evolução dos sentimentos da comunidade para com os Intitutos Federais.
- Apresentar a modelagem de tópicos em tweets sobre Institutos Federais, nos âmbitos nacional e regional.

3 Fundamentação teórica

3.1 Definições sobre Análise de Sentimentos

A análise de sentimentos é a área de estudo que analisa as atitudes, emoções, sentimentos e as opiniões das pessoas em relação a produtos, serviços, organizações, eventos, tópicos e as características gerais deles (LIU, 2012). Segundo (CAMBRIA et al., 2013), existem 4 tipos de análise de sentimento, sendo eles:

- **Análise de sentimentos refinada:** Analisa a polaridade das opiniões, podendo ser uma avaliação simples de aprovação/desaprovação ou uma distinção entre positivo/negativo. Além disso, pode se tornar mais complexa ao incluir especificações mais detalhadas e níveis, como um sistema de classificação Likert que varia de 1 a 7, medindo o grau de acordo ou desacordo.
- **Detecção de emoções:** Se concentra em identificar e classificar as emoções expressas no texto, não se baseia em polaridades como na análise tradicional, seu foco está em compreender as emoções subjacentes nas mensagens. Este método busca identificar emoções específicas, como, raiva, tristeza, medo e arrependimento no texto e atribuir-lhes uma classificação apropriada.
- **Análise de intenção:** É um método avançado de análise de texto que busca compreender tanto o sentimento expresso em um texto quanto a intenção subjacente por trás desse sentimento. Além de determinar se uma opinião é positiva, negativa ou neutra, esse método procura identificar o propósito ou objetivo por trás das expressões de sentimento.
- **Baseada em aspectos:** Se concentra em identificar e avaliar os sentimentos expressos em relação a aspectos específicos de um produto, serviço ou tópico. Em vez de analisar o sentimento geral do texto, busca entender as opiniões e emoções associadas a aspectos particulares mencionados no texto que podem variar dependendo do domínio ou contexto. Esses aspectos podem ser recursos, características, funcionalidades ou componentes específicos do item em análise.

A análise de sentimento também é denominada mineração de opnião, pois busca descobrir pontos de vista baseados em opiniões, emoções relacionadas e informações que podem ser de interesse particular. Com o objetivo de avaliar o sentimento associado a opiniões/interesses em diferentes níveis de abstração, a análise de sentimento pode ser aplicada em diferentes níveis, sendo eles:

- **Nível de documento:** Analisa o sentimento de um texto como um todo, ignorado aspectos específicos presentes no texto, bem como divergências de sentimentos nas sentenças.
- **Nível de sentença:** Analisa o sentimento de cada sentença individualmente.
- **Nível de subsentença:** Analisa expressões presentes na sentença.

Nesse trabalho foi adotada a análise de sentimentos refinada e sua aplicação acontece em nível de documento.

3.2 Técnicas Utilizadas para Análise de Sentimento

3.2.1 Métodos Não Supervisionados baseados em Dicionários Léxicos

Métodos não supervisionados não necessitam um conjunto de dados pré-classificados para treinamento do modelo. Tais técnicas usualmente se utilizam de um dicionário de termos (léxicons), em que cada termo é associado a um grau de positividade ou negatividade. Nesse conjunto de técnicas, a ocorrência de termos positivos ou negativos na sentença é usada para indicar o sentimento geral da frase.

Uma das primeiras estratégias não supervisionadas é denominada Linguistic Inquiry and Word Count (LIWC), que utiliza um léxicon para calcular a porcentagem de termos positivos/negativos no texto (JUNIOR; GUEDES; BEZERRA, 2016). A partir dessa estratégia, métodos como SentiStrength e Vader (MELO, 2017) surgiram com aplicação em redes sociais. O SentiStrength estima a polaridade de textos curtos/informais a partir de um dicionário léxico construído por humanos e aprimorado por algoritmos de aprendizado de máquina, o que permite maior cobertura com a generalização do dicionário utilizado. A estratégia Vader se utiliza de um conjunto de regras e expressões pré-definidas, negações e intensificadores para uma análise mais detalhada do sentimento do texto a partir de termos de polaridade que incluem gírias, emoticons e acrônimos comuns em redes sociais (MELO, 2017).

Apesar dos métodos descritos anteriormente terem sido projetados para atuar na língua inglesa, há soluções recentes com foco na língua portuguesa que incluem elementos de estratégias consolidadas como o Vader e Sentistrength. O sentilex é uma estratégia baseada no uso de léxicos, é bastante robusta e conta com aproximadamente 7.000 termos classificados. O método é considerado uma das estratégias pioneiras para o português tendo obtido os melhores resultados no contexto de redes sociais em experimentos preliminares (CARVALHO; SILVA, 2015)

3.2.2 Métodos Supervisionados de Aprendizado de Máquina

Os métodos supervisionados necessitam de uma coleção de dados pré-classificados para treinamento. Dessa forma, um conjunto textos (como documentos, tweets ou review de produtos) deve ser manualmente rotulado por humanos. Por exemplo, um conjunto de tweets, cada um deles deverá ser rotulado como positivo, negativo ou neutro por um humano. O conjunto de tais dados com suas respectivas rotulações é denominado conjunto de treinamento. O objetivo de tal rotulação é a utilização de um algoritmo de aprendizado de máquina que tenta encontrar, de forma automática, padrões que associam o texto à rotulação feita. Consequentemente, o algoritmo gera um modelo preditor capaz de receber um novo texto e prever se seu sentimento é positivo, negativo ou neutro.

O método *Random Forests (RF)* é construído a partir de um conjunto de árvores de decisão individuais, que são combinadas para formar uma "floresta". Cada árvore de decisão é treinada em uma amostra aleatória do conjunto de dados e utiliza um subconjunto também aleatório de recursos (ou características) para tomar decisões de classificação. A classificação final é determinada através da combinação dos resultados das árvores individuais. Seu funcionamento é demonstrado na Figura 1. RF apresenta resultados satisfatórios em diversos contextos, incluindo cenários com coleções desbalanceadas e estratégias de pré-processamento simplistas, apresentando resultados que se destacam de forma considerável entre os classificadores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

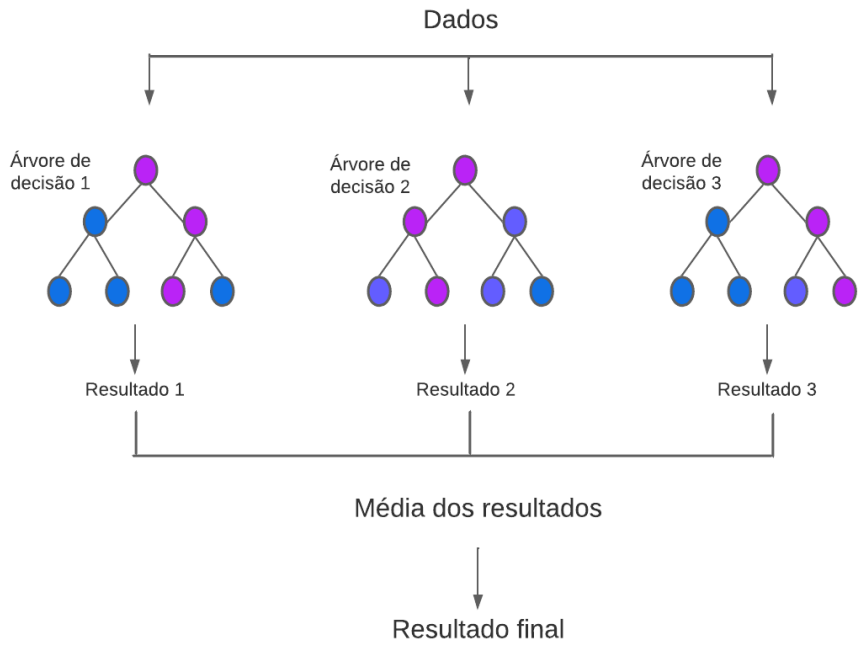

Figura 1 – Ilustração representativa do algoritmo Random Forests

O método *Support Vector Machine* (SVM) funciona mapeando os textos para um

espaço de características de alta dimensionalidade, onde é possível realizar a separação das classes. A partir dessas características, o SVM treina um modelo que cria um hiperplano que maximiza a margem entre cada par de classes, ou seja, busca encontrar a melhor separação possível (EDUARDO; CORTES, 2021). Durante o treinamento, ele identifica os vetores de suporte, que são os textos de treinamento que estão mais próximos da fronteira de decisão, isto é, aqueles que têm maior influência na definição do hiperplano de separação. Esses vetores de suporte são essenciais para a classificação de novos textos, pois fornecem informações sobre as características relevantes para a determinação do sentimento. Seu funcionamento é ilustrado na Figura 2.

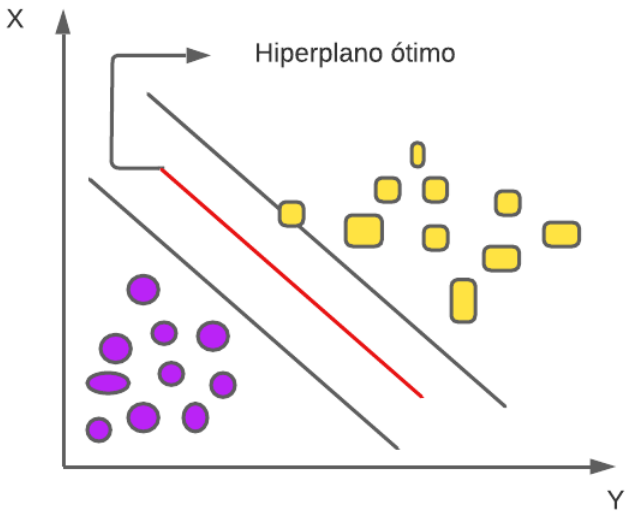

Figura 2 – Ilustração representativa do algoritmo Support Vector Machine

O método *K — Nearest Neighbors (K-NN)* funciona identificando os "k"vizinhos mais próximos de um texto não rotulado no conjunto de treinamento, onde "k"é um valor definido pelo usuário. A proximidade é medida usando métricas de similaridade, como a distância euclidiana ou distância cosseno, entre os vetores de características dos textos. Uma vez que os "k"vizinhos mais próximos são identificados, o K-NN atribui ao texto não classificado, o rótulo de sentimento mais frequente entre seus vizinhos. Por exemplo, se a maioria dos vizinhos pertencer à classe "positiva", o texto não rotulado será classificado como positivo. O método possui uma fácil implementação e um bom suporte para classificação multi-classe (MITCHELL, 1997) e seu funcionamento é demonstrado na Figura 3.

O método *Naive Bayes* se baseia na suposição "ingênua"(naive) de independência condicional entre as características. Ele estima a probabilidade de um texto pertencer a uma determinada categoria de sentimento com base nas características presentes no texto,

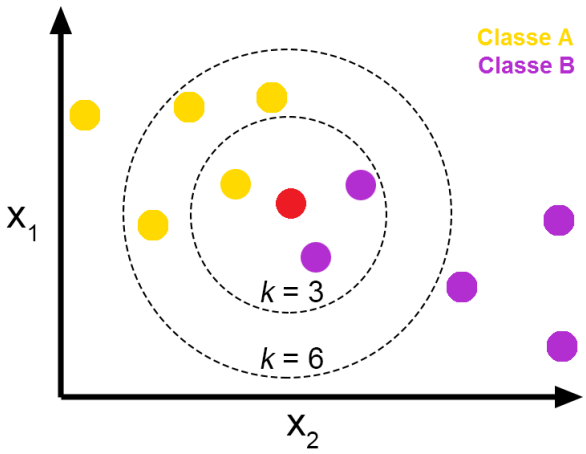

Figura 3 – Ilustração representativa do algoritmo K - Nearest Neighbors

elas podem ser palavras-chave, n-gramas, presença de determinados termos, entre outros. O Naive Bayes utiliza o Teorema de Bayes, que relaciona a probabilidade condicional de um evento, neste caso, a categoria de sentimento, dado uma evidência, as características do texto. Na prática, a suposição ingênua do Naive Bayes é que todas as características são independentes entre si, ou seja, a presença ou ausência de uma característica não afeta a presença ou ausência de outras características. Possui fácil implementação e um custo computacional baixo, além disso, apresenta bons resultados para classificações multinomiais (JIANG; WANG; LI CHAOQUN ZHANG, 2016). Seu funcionamento é exposto na figura 4

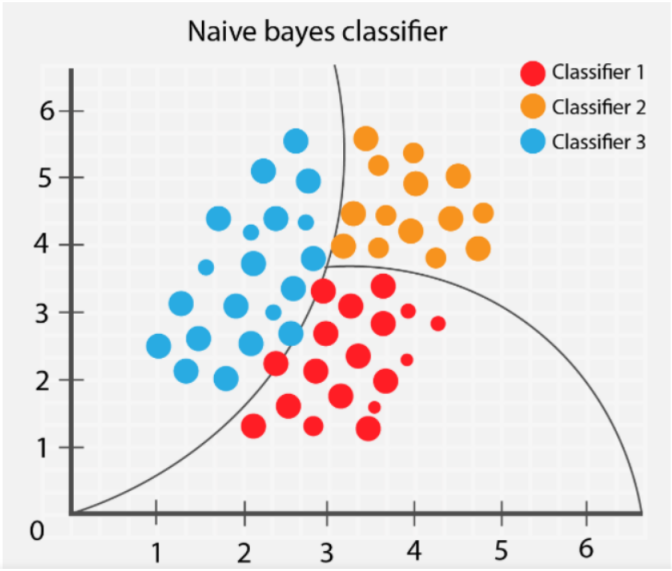

Figura 4 – Ilustração representativa do algoritmo Naive Bayes

Métodos tradicionais em classificação supervisionada de texto são comumente empregados na tarefa de análise de sentimento. Em (CAMPOS et al., 2017a), foram comparadas diversas estratégias de classificação em 10 coleções de análise de sentimento,

sendo os métodos, *Random Forests* e *Support Vector Machines* os mais eficazes na maioria das coleções avaliadas. Em particular, o método *Random Forests* foi capaz de extrair relacionamentos não lineares a partir de um comitê de árvores de decisão. Os modelos do SVM apresentaram resultados eficazes em coleções de sentimento com poucos exemplos de treino devido à simplicidade e capacidade de generalização dos hiperplanos separadores.

Recentemente, (PATIL; KOLHE, 2022) apresentou um estudo comparativo incluindo, além do SVM e RF, o classificador Multinomial Naive Bayes para análise de sentimento de tweets no contexto político indiano. Todos os métodos apresentaram alta acurácia (superior a 84%), sendo o classificador NB o mais efetivo no contexto avaliado. Os autores justificam que a simplicidade do NB torna o modelo ideal para identificação da presença de termos mais discriminativos em tweets no contexto político.

Outros métodos baseados em distância entre textos também foram largamente empregados para o problema de análise de sentimento. O método mais conhecido que compara a distância entre um texto e seus vizinhos mais próximos é o kNN (CANUTO; GONCALVES; BENEVENUTO, 2016). A utilização do kNN para avaliação do sentimento presenta na vizinhança de documentos atingiu excelentes resultados no contexto de reviews de produtos e tweets da língua inglesa.

3.2.3 Modelos de Linguagem baseados em Transformers

Os modelos de linguagem baseados em transformers são algoritmos baseados em redes neurais capazes de realizar diversas tarefas de processamento de linguagem natural (PLN), como, tradução automática, resumo de texto e geração de legendas. Eles usam uma arquitetura denominada "transformer" (VASWANI et al., 2017), popularizada pelo modelo Generative Pre-trained Transformer (GPT) e mudaram drasticamente o campo de processamento de linguagem natural (MIN et al., 2021).

O treinamento dos modelos de linguagem baseados em transformers envolve duas etapas principais: pré-treinamento e ajuste fino *(fine-tuning)*. No pré-treinamento, o modelo é alimentado com grandes quantidades de texto em uma tarefa de previsão de termos ausentes. Ele aprende a prever termos ocultos (mascarados) em uma sequência de termos. Essa etapa permite que o modelo aprenda representações ricas de palavras e contextos.

Após o pré-treinamento, o modelo é ajustado finamente em uma tarefa específica, como tradução ou geração de texto, usando um conjunto de dados menor e rotulado. Durante o ajuste fino, o modelo é refinado para se adequar melhor à tarefa em questão, aprendendo a produzir resultados mais precisos e relevantes.

Estes métodos baseados em arquiteturas de redes neurais profundas têm a propriedade de obtenção de padrões linguísticos e semânticos a partir da arquitetura de transformers (DEVLIN et al., 2019). Eles são capazes de construir modelos de linguagem

a partir do aprendizado da predição de termos em sentenças em um grande volume de dados textuais. Então, tais modelos de linguagem podem ser empregados em diversas tarefas, como a análise de sentimento, a partir do treinamento (tunagem fina) em textos previamente rotulados.

O método BERTweet (NGUYEN; V.; NGUYEN, 2020) foi recentemente desenvolvido especificamente para tarefas que envolvem processamento de dados textuais do twitter. Tal método foi pré-treinado com aproximadamente 850 milhões de tweets em inglês, os quais apresentam uma estrutura de linguagem voltada para comunicação em textos curtos. O BERTweet é considerado o estado-da-arte para tarefa de análise de sentimento no twitter e pode ser ajustado de forma facilitada para realizar diferentes tarefas e é descrito como conceitualmente simples e empiricamente poderoso (DEVLIN et al., 2019)

Os autores do BERT também criaram uma variante treinada em mais de 100 idiomas, incluindo o português, chamada m-BERT. Com base na abordagem BERT robustamente otimizada (RoBERTa) (LIU MYLE OTT, 2019), uma versão multilíngue treinada com dados de 100 idiomas foi desenvolvida e denominada XLM-RoBERTa (CONNEAU KARTIKAY KHANDELWAL, 2020). Ademais, alguns modelos de aprendizado profundo projetados e pré-treinados em português que apresentam resultados excelentes são o BERTimbau (FILHO RODRIGO WILKENS, 2018) e o GPorTuguese (SOUZA; FILHO, 2022b)

3.3 Modelagem de tópicos

Segundo (QIANG et al., 2020), a modelagem de tópicos (Modelagem de Tópicos) visa encontrar assuntos representados por grupos de palavras (tópicos) em um conjunto de textos. O propósito dessa abordagem é extrair os principais temas presentes em textos, proporcionando uma compreensão mais profunda do conteúdo. A importância da Modelagem de Tópicos está intrinsecamente ligada à capacidade de organizar, resumir e explorar grandes volumes de dados textuais de maneira eficiente.

O BERTopic é uma abordagem inovadora no campo de processamento de linguagem natural, projetada para extrair tópicos automaticamente a partir de grandes conjuntos de documentos (GROOTENDORST, 2022). Ele utiliza a representação semântica profunda fornecida pelo BERT para capturar relações contextuais entre palavras em documentos. Em vez de depender apenas de métodos estatísticos tradicionais, o BERTopic incorpora informações semânticas mais ricas para identificar tópicos latentes. A técnica envolve a extração de embeddings de documentos com BERT e, em seguida, a aplicação de técnicas de agrupamento para organizar esses embeddings em clusters representativos de tópicos.

Em comparação com métodos tradicionais de modelagem de tópicos (como LDA - Latent Dirichlet Allocation), o BERTopic geralmente se destaca na captura de relações

semânticas mais complexas. A abordagem baseada em embeddings do BERT proporciona uma representação mais rica e contextual, resultando em tópicos mais coerentes e relevantes.

4 Metodologia

Considerando que o objetivo deste trabalho está centrado na implementação e aprimoramento de métodos de análise de sentimento, bem como, a aplicação da modelagem de tópicos ao contexto dos Institutos Federais, nesta seção, serão descritos os passos metodológicos adotados para execução de tal objetivo.

Este trabalho se utiliza da pesquisa quantitativa que visa apresentar evidências empíricas sobre a capacidade preditiva dos modelos de análise de sentimento e a capacidade de geração de tópicos coerentes e representativos sobre os dados dos Institutos Federais. Neste capítulo, será descrita a estratégia de aplicação da metodologia para descoberta de conhecimento em bases de dados *(Knowledge Discovery in Databases)* (CASTRO; FERRARI, 2017) seguindo as etapas ilustradas na Figura 5 e descritas a seguir.

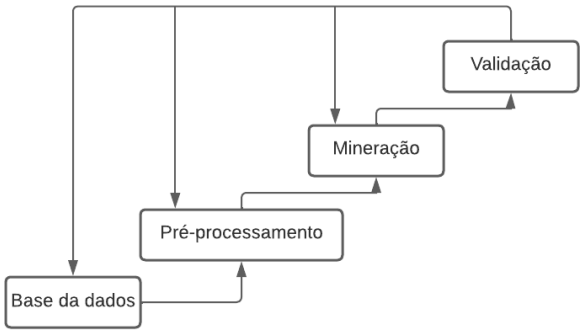

Figura 5 – Metodologia de descoberta de conhecimento em bases de dados. Fonte: (CASTRO; FERRARI, 2017)

4.1 Base de dados

A plataforma escolhida para a coleta de dados foi a rede social Twitter por meio de duas estratégias diferentes, sendo a primeira estratégia utilizada para construção de uma base de tweets para a validação de modelos de análise de sentimento e a segunda estratégia utilizada para construção da base para a modelagem de tópicos.

A primeira estratégia de coleta foi utilizada para obtenção de dados entre outubro de 2021 até fevereiro de 2022. A coleta foi realizada a partir da API V2[1] disponibilizada pelo próprio twitter aos desenvolvedores. Para tal, foi submetida e aceita uma proposta de pesquisa acadêmica na plataforma. Os tweets foram obtidos a partir do método filtered stream da API, com restrição para obtenção de tweets da língua portuguesa e remoção

[1] <https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/introduction>

de repostagens (retweets). Como restrição por palavras-chave, foram utilizados termos que necessariamente devem aparecer nos tweets coletados. A primeira palavra-chave, para obtenção de dados específicos do Instituto Federal de Goiás, foi o termo "IFG". Para a coleta de dados dos institutos federais em um âmbito nacional, foram utilizadas as expressões "Instituto Federal" e "Institutos Federais". No total, foram coletados 2500 tweets.

A segunda estratégia de coleta foi realizada por meio da contínua raspagem de dados (scrapping) de Janeiro de 2023 à Novembro de 2023. O código utilizado para raspagem de dados [2] foi continuamente executado à cada hora, durante todo o período da coleta, para obtenção dos tweets mais recentes em português. Como critério de exclusão, foram ignorandos retweets e repetições de mensagens. Nessa estratégia, foram coletados dados sobre todos os Institutos Federais do país por meio das consultas referentes às siglas correspondentes à cada instituição [3]. A Tabela 1 indica as siglas utilizadas na coleta de dados. No total, foram coletados 18 mil tweets, de forma que cada tweet contém no mínimo a sigla de um Instituto Federal brasileiro.

Região	Siglas
Centro-Oeste	IFB, IFMT, IFMS, IFG e IFGoiano.
Nordeste	IFAL, IFBA, IF Baiano, IFCE, IFMA, IFPB, IFPE, IF Sertao, IFPI,IFRN e IFS.
Norte	IFAC,IFAP,IFAM,IFPA,IFRO,IFTO e IFRR.
Sudeste	IFES,IFRJ,IFF,IFMG,IFNMG,IFSUDESTEDEMINAS,IFUSULDEMINAS,IFTM e IFSP.
Sul	IFRS,IFFarroupilha,IFSUL,IFPR,IFSC e IFC.

Tabela 1 – Siglas utilizadas para coleta de dados na rede social Twitter.

4.2 Pré-processamento

O pré-processamento consiste na preparação dos dados para a aplicação dos modelos, isto é, são realizadas diversas etapas que incluem a preparação, organização e estruturação dos dados. Esta fase exige uma quantidade considerável de tempo e esforço e impacta diretamente a qualidade da aplicação do modelo. Entre as diversas abordagens passíveis de utilização para o pré-processamento dos dados, o presente trabalho utilizou as seguintes:

- **Rotulação manual dos dados:** Para todas as mensagens relacionadas ao IFG e ao Instituto Federal coletadas, o sentimento de cada mensagem foi avaliado manualmente entre positivo, negativo ou neutro por três usuários de redes sociais, de forma que mensagens sem concordância de avaliações foram descartadas. Tal esforço é essencial para a etapa de geração de uma coleção de treino para modelos supervisionados e para validação dos resultados na tarefa de Análise de Sentimento.

[2] <https://github.com/n0madic/twitter-scraper>
[3] <https://www.pebsp.com/lista-de-institutos-federais-do-brasil-por-estado-2020>

- **Eliminação de tweets institucionais:** Mensagens provenientes do setor de comunicação dos institutos foram removidas do conjunto, com o objetivo avaliar a tarefa de análise de sentimento com mensagens provenientes da comunidade e alunos.

- **Tokenização:** Técnica que consiste na segmentação das palavras, isto é, ocorre a quebra de sequência dos caracteres e consequentemente, a delimitação de cada palavra presente no texto.

- **Representação de texto com TF-IDF:** Cálculo de relevância de uma palavra no texto, ou seja, estatística numérica simples que representa o texto a partir do número de aparições de cada termo. A partir do método, tem-se as palavras centrais da sequência, isto é, as que mais aparecem. O TF-IDF (Term Frequency-Inverse Document Frequency) considera cada tweet como o conjunto (desordenado) de suas palavras, onde cada palavra possui um peso no tweet. O peso de cada termo t é calculado pela fórmula $TF * IDF$, onde:

$$TF = Ocorrências\ de\ t\ no\ tweet \tag{4.1}$$

$$IDF = log\left(\frac{Total\ de\ tweets}{Total\ de\ tweets\ com\ o\ termo\ t}\right) \tag{4.2}$$

- **Representação de texto com GloVe:** Técnica que utiliza a informação global de coocorrência de palavras em todos os tweets para gerar representações vetoriais. Essas representações refletem não apenas a proximidade semântica entre palavras, mas também a magnitude de suas relações. Assim, palavras que compartilham contextos semelhantes em tweets apresentam vetores mais próximos no espaço GloVe.

- **Representação de texto com FastText:** Técnica avançada de embedding que representa cada palavra como uma soma de seus n-gramas, permitindo capturar tanto o significado das palavras completas quanto de suas partes constituintes.

- **Normalização:** Adequação dos dados a um padrão de letras apenas minúsculas, esta técnica realiza a combinação de termos que estão em maiúscula e minúscula para a utilização de apenas um significado.

- **Tradução:** Para a entrada dos métodos que exigem a língua inglesa, como Vader, BERTweet e RoBERTa, os tweets em português foram traduzidos para a língua inglesa com a API Translation Basic[4] disponibilizada pelo Google.

Essas duas bases contam com aproximadamente 2.500 tweets, que foram rotulados manualmente, isto é, foram analisados de forma individual e tiveram um sentimento

[4] <https://cloud.google.com/translate/docs/basic/translating-text?hl=pt-br>

atribuído a eles. O processo de rotulação foi realizado por três estudantes e os casos onde não foi possível chegar a um consenso sobre o sentimento foram descartados.

Além disso, as duas coleções foram representadas usando três estratégias consolidadas na literatura, sendo elas, TF-IDF, FasText e GloVe. As três representações foram utilizadas como entrada para os métodos supervisionadas usados no desenvolvimento desse trabalho.

Para a coleção relacionada às siglas dos Institutos Federais do país, a divisão se deu baseada em regiões, ou seja, aproxidamente 18 mil tweets foram divididos em cinco bases de dados distintas, cada uma representando uma região do Brasil (Centro-oeste, Nordeste, Norte, Sudeste e Sul).

4.3 Mineração

Este trabalho estuda duas tarefas referentes à etapa de *Mineração de Dados*, sendo elas, a análise de sentimento e a modelagem de tópicos. A seguir, a etapa de mineração é descrita em detalhes para cada uma destas tarefas.

4.3.1 Análise de sentimento

Atualmente, existem diversos métodos para predição de sentimentos de forma automática. Considerando tal diversidade, no desenvolvimento deste trabalho foram aplicados os modelos que entregam os melhores resultados e que melhor se adaptam às características da coleção utilizada.

Desse modo, foram avaliados os métodos não supervisionados Sentilex e Vader, métodos clássicos de aprendizado de máquina que apresentaram bons resultados no contexto de tweets, como Random Forests, Naive Bayes, SVM e kNN, e métodos estado-da-arte baseados em transformers, como BERTweet, RoBERTa, Bertimbau, m-Bert e XLM-RoBERTa com a seguinte parametrização:

- **Sentilex - PT:** Como o método não supervisionado se utiliza de um dicionário e regras fixas, não há parametrização necessária (CARVALHO; SILVA, 2017).
- **Vader:** Tweets foram traduzidos para entrada desse método originalmente desenvolvido para língua inglesa. Por ser baseado em dicionários e regras fixas, não há parametrização (RIBEIRO et al., 2016).
- **Random Forests:** Segundo (CAMPOS et al., 2017a), sendo RF pouco sensível à parametrização, seguimos, como sugerido pelos autores, com 200 árvores de decisão sem poda, e com o número de termos escolhidos a cada divisão sendo $\sqrt{n}$, para os n termos do dicionário.

- **Multinomial Naive Bayes:** Por obter os melhores resultados em (PATIL; KOLHE, 2022), foi escolhida a versão multinomial do naive bayes, que considera o número de ocorrências do termo em cada tweet.
- **Support Vector Machine (SVM):** Foi utilizado o kernel linear (PATIL; KOLHE, 2022) e parâmetro de regularização C escolhido, por validação cruzada no treinamento, entre 6 valores de 10^{-3} à 10^{3}.
- **K-Nearest Neighbor:** A quantidade de vizinhos k foi escolhida entre o conjunto de valores {2, 5, 10, 20, 30} por meio da tradicional validação cruzada em 5 partes no treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2001).
- **BERTweet e RoBERTa:** A utilização do BERTweet e RoBERTa é realizada em duas etapas: a primeira consiste na tradução dos tweets em potuguês para língua inglesa, e a segunda no treinamento (tunagem-fina) do modelo para torná-lo específico ao contexto do Instituto Federal a partir dos dados coletados e traduzidos. Tal abordagem se deve ao fato de que tanto o BERTweet quanto o RoBERTa foram pré-treinados com textos da língua inglesa. A coleção de dados utilizada para pré-treinamento do BERTweet consiste na coleção de centenas de milhares de tweets em inglês, utilizados para o aprendizado de padrões de linguagem empregados no Twitter (DEVLIN et al., 2018). No caso do RoBERTa, o pré-treinamento foi realizado com uma ampla gama de textos, como artigos da web, livros e artigos acadêmicos. Essa diversidade ajuda o modelo a capturar uma representação abrangente da linguagem (LIU MYLE OTT, 2019). Para a tunagem fina, foi utilizada a parametrização original descrita em (NGUYEN; V.; NGUYEN, 2020) (i.e., 30 épocas, learning_rate=10^{-5} e batch_size=30).
- **Bertimbau, m-Bert e XLM-RoBERTa:** os três métodos utilizam textos em português em sua base de dados utilizada para o pré-treinamento, portanto, ao contrário do BERTweet e RoBERTa anteriormente mencionados, não é necessária a etapa de tradução dos tweets. Para a tunagem fina, foi utilizada a parametrização original descrita em (NGUYEN; V.; NGUYEN, 2020) (i.e., 30 épocas, learning_rate=10^{-5} e batch_size=30).

4.3.2 Modelagem de tópicos

A tarefa de modelagem de tópicos visa expressar e sumarizar aspectos-chave presentes nos tweets. Essa tarefa facilita a compreensão do conteúdo de tweets curtos, fornecendo insights valiosos para o acompanhamento de tendências e compreensão das percepções do público sobre tweets relacionados aos Insitutos Federais.

O método BERTopic foi aplicado neste trabalho para a tarefa de modelagem de tópicos, uma vez que ele explora a capacidade de modelos de linguagem baseados

em transformers para considerar o contexto e as relações semânticas entre palavras, permitindo assim, uma identificação mais refinada e detalhada dos tópicos presentes nos tweets (GROOTENDORST, 2022). Além disso, o BERTopic destaca-se em ambientes nos quais as nuances e a complexidade dos dados textuais demandam uma abordagem mais sofisticada.

A parametrização padrão do BERTopic foi adotada nesse trabalho, isso é, uso de unigramas, bigramas e trigramas em sua versão multilíngue, com tópicos representados com os 30 termos mais relevantes de cada tópico[5]. Quanto à determinação do número ideal de tópicos, a escolha foi orientada pelo escore de coerência (GROOTENDORST, 2022), de forma que a quantidade de tópicos associada à maior coerência em um intervalo de 2 a 20 tópicos foi escolhida.

4.4 Validação

Para medir a eficácia de uma tarefa de *Machine Learning*, métricas específicas são utilizadas. A seguir, são expostas as estratégias utilizadas para avaliar o desempenho dos modelos nas tarefas de análise de sentimento e modelagem de tópicos.

4.4.1 Análise de sentimentos

Para obtenção da eficácia média de cada método de análise de sentimento, foi utilizada a técnica de validação cruzada em 5 partes, também conhecida como 5-fold cross validation (HASTIE; TIBSHIRANI; FRIEDMAN, 2001). Nesse método, o conjunto de dados é dividido aleatoriamente em 5 partes, o modelo é treinado em 4 partes e testado na parte restante, repetindo esse processo 5 vezes. Portanto, há 5 pares de treino/teste.

Essa abordagem permite uma avaliação mais robusta, evitando vieses resultantes de uma única divisão dos dados. A validação cruzada em 5 partes aproveita ao máximo o conjunto de dados disponível, fornecendo uma estimativa confiável do desempenho do modelo em dados não vistos.

Para validação e análise dos resultados obtidos até o momento, utilizou-se as métricas referentes à matriz confusão que expõe a quantidade de exemplos classificados correta e erroneamente e a Macro e Micro F1. A matriz confusão permite a visualização facilitada das classificações, expondo os exemplos classificados de maneira acertada e não acertada. O método facilita na identificação de favorecimento de uma classe em detrimento de outra, bem como, auxilia na visualização geral da eficácia do modelo utilizado.

O Macro-F1 mede a efetividade da classificação para cada classe de sentimento individualmente e depois computa a média das avaliações das classes individuais, sendo

[5] <https://pypi.org/project/bertopic/>

capaz de avaliar a efetividade média dos acertos em cada classe, o que desfavorece métodos que possuem o viés de acertar mais exemplos da classe majoritária. Sua fórmula é exposta a seguir.

$$Macro = \frac{Sum(F1\ Scores)}{Number\ of\ classes} \tag{4.3}$$

Já a Micro-F1 mede a efetividade da classificação em todas as predições, ou seja, as tabelas de contingência combinadas de todas as classes, isto é, avalia a efetividade de acertos/erros em geral. A fórmula usada para calcular esta métrica é demonstrada a seguir.

$$Micro = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.4}$$

4.4.2 Modelagem de tópicos

A tarefa de modelagem de tópicos visa a viabilização da análise qualitativa de tópicos em tweets, isso é, a obtenção de visualizações capazes sumarizar os assuntos discutidos nas redes sociais. Portanto, foram adotadas visuações em WordCloud [6] que visam destacar os termos de maior importância nos tópicos. Visualizações WordCloud fornecem uma sumarização das informações presentes nos tópicos gerados pelo BERTopic, adotado para tarefa de modelagem de tópicos nesse trabalho.

Para avaliação quantitativa dos tópicos gerados pelo BERTopic, foi adotada a medida denominada Coerência (GROOTENDORST, 2022). Essa medida visa avaliar o quão semanticamente coesos são os tópicos, ou seja, se as palavras atribuídas a um tópico têm uma relação forte e consistente entre si. A Equação 4.5 expressa o cálculo da medida de coerência adotada:

$$Coerência = \frac{\sum_{w_i, w_j \in T} p(w_i, w_j)}{\sum_{w_i \in T} p(w_i)} \tag{4.5}$$

Onde:

- **T:** representa o conjunto de tópicos.
- **Wi e Wj:** representa o conjunto de tópicos.
- **p(Wi e Wj):** probabilidade conjunta de Wi e Wj ocorrerem juntas.
- **p(Wi):** probabilidade da ocorrência de Wi no corpus.

[6] <https://github.com/amueller/word_cloud>

5 Resultados experimentais

5.1 Caracterização das coleções

Conforme descrito na Seção 4.1, foram obtidos 2500 tweets que contêm os termos "IFG" ou "Instituto Federal", no intervalo de outubro de 2021 até fevereiro de 2022, para avaliação das técnicas empregadas para análise de sentimento. Os tweets foram separados em dois conjuntos de dados referentes ao IFG e Institutos Federais. Na Tabela 2 é apresentada a distribuição entre positivos/negativos/neutros para os dados obtidos nos dois cenários. Em ambos, é possível observar a pequena quantidade de termos utilizados nos tweets, bem como a proximidade entre a proporcionalidade de neutros e negativos, mas uma grande disparidade entre positivos e negativos.

A disparidade (de cerca de 2x) entre a quantidade de tweets positivos e negativos pode ser explicada pela tendência geral de uso da rede social twitter, onde usuários tendem a fazer mais comentários negativos e de cunho crítico (GHOSH et al., 2019). Ainda, nota-se o grande número de tweets neutros, uma vez que a rede favorece a comunicação entre o aluno e a instituição por meio da utilização das redes por parte do setor de comunicação dos institutos para o atendimento de dúvidas da comunidade.

	Caracterização das coleções para Análise de Sentimento			
	Positivos	**Negativos**	**Neutros**	**Média de palavras por tweet**
IFG	128	271	269	18
Instituto Federal	263	613	580	20

Tabela 2 – Quantidade de tweets positivos, negativos e neutros nas coleções sobre o IFG e Institutos Federais.

Ademais, a partir da análise das Figuras 6 e 7, é perceptível a ocorrência de picos negativos na distribuição temporal de tweets na base dos Institutos Federais. Tais picos ocorrem simultaneamente a eventos importantes da instituição. Como, por exemplo, entre dezembro e janeiro houve o início da discussão para o retorno presencial das atividades e o fim do semestre letivo, períodos considerados de grande agitação para os alunos. Por outro lado, as altas nos tweets positivos acontecem em períodos de resultado do vestibular, no final de dezembro e início de janeiro, e expõe a felicidade dos estudantes pela aprovação. Além disso, eventos artísticos e culturais também geram publicações positivas no início de fevereiro. O aumento dos tweets neutros no início do semestre reflete muitas das dúvidas respondidas à comunidade pelos setores de comunicação dos institutos.

Conforme descrito na Seção 4.1, foi coletado um segundo conjunto de dados (não rotulados), obtido por meio das siglas dos Institutos Federais de todo o país, com o objetivo

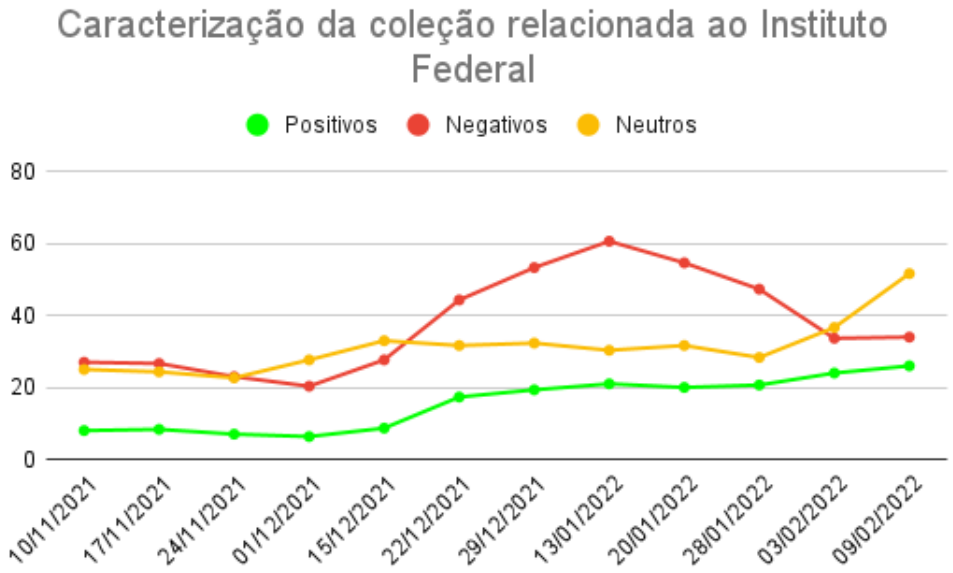

Figura 6 – Evolução temporal dos tweets relacionados ao IFG

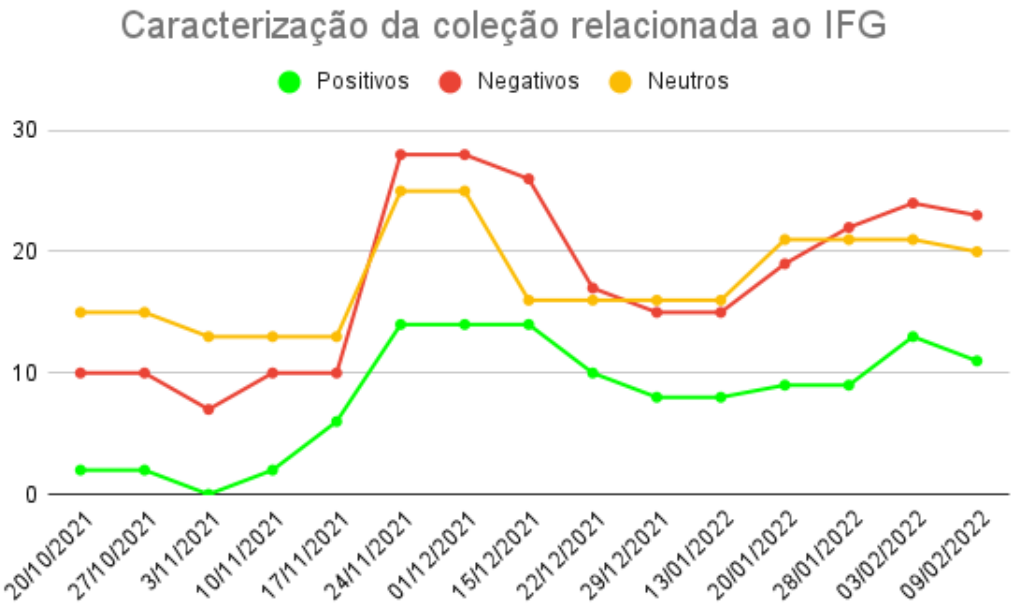

Figura 7 – Evolução temporal dos tweets relacionados ao Instituto Federal

de avaliar a modelagem de tópicos em dados regionais. O conjunto de dados possui 18.000 tweets com ao menos uma siga de instituto federal, coletados de janeiro à novembro de 2023. Nesse conjunto, a proporção de tweets por região é altamente correlacionada com a proporção de unidades dos institutos no Brasil, isto é, as regiões Nordeste e Sudeste possuem uma maior quantidade de Institutos e consequentemente, mais tweets. Ademais, a região Centro-oeste possui o menor número de tweets e Institutos, como apresentado na Figura 8.

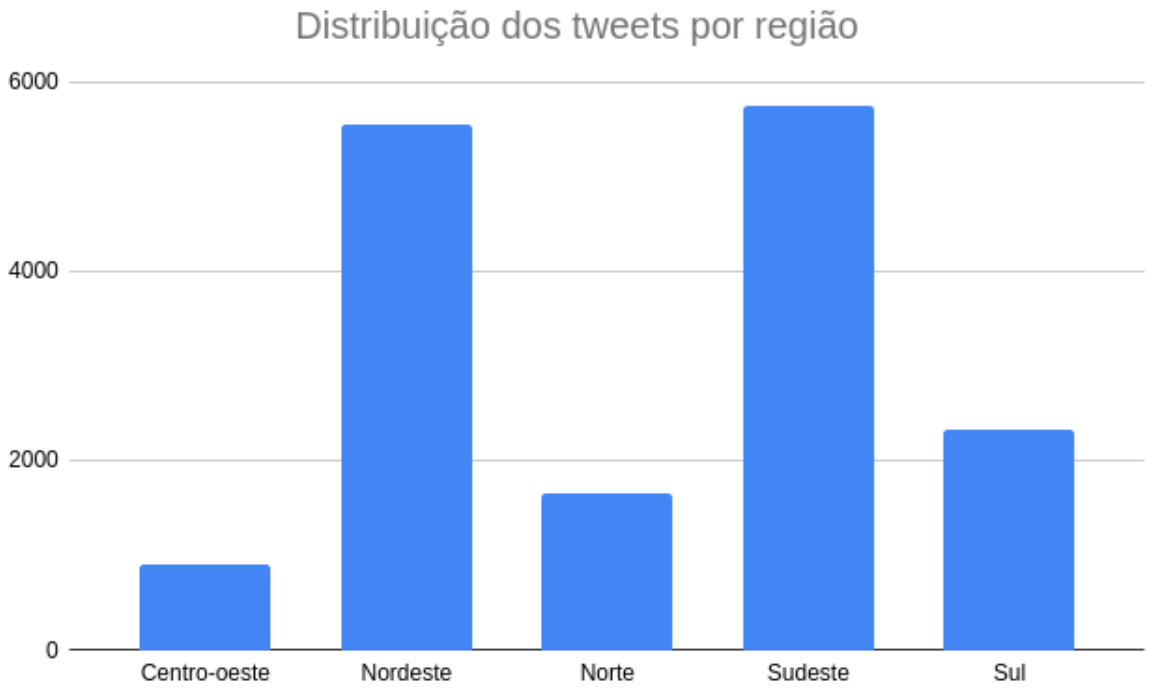

Figura 8 – Distribuição da coleção em regiões

5.2 Efetividade das técnicas para análise de sentimentos

Esta seção apresenta uma análise detalhada do desempenho dos métodos não supervisionados, supervisionados tradicionais e baseados em transformers na avaliação de sentimentos expressos em tweets relacionados aos Institutos Federais Brasileiros.

A eficácia dos métodos não supervisionados na análise de sentimentos em tweets relacionados aos Institutos Federais é apresentada na Tabela 3. Observa-se que os resultados obtidos são considerados insatisfatórios. No caso do Sentilex, isso é atribuído à baixa cobertura do dicionário em relação aos termos presentes nos documentos, como gírias e erros gramaticais. Por outro lado, o Vader demonstra resultados mais promissores devido à sua abordagem, que leva em consideração gírias, pontuação e regras pré-definidas para a classificação, no entanto, a presença de ruídos de tradução impactam negativamente no resultado final.

	Coleção relacionada ao IFG		**Coleção relacionada ao Instituto Federal**	
Abordagem	**Macro-F1**	**Micro-F1**	**Macro-F1**	**Micro-F1**
Sentilex	0,37	0,43	0,39	0,45
Vader	0,47	0,48	0,54	0,55

Tabela 3 – Efetividade dos métodos não supervisionados para análise de sentimento

A Tabela 4 expõe os resultados de efetividade dos métodos supervisionados de análise de sentimento. Observa-se que os resultados foram apenas medianos, indicando

uma baixa efetividade, apesar de serem métodos amplamente utilizados em tarefas de análise de sentimento e serem independentes da linguagem. Entre esses métodos, o SVM mostrou-se mais resistente à escassez de dados disponíveis para o treinamento em ambas as coleções avaliadas. No entanto, a efetividade máxima alcançada pelo SVM foi de 0.63 em Macro-F1 e 0.59 em Micro-F1, o que dificulta a aplicação prática desse método para a predição de sentimentos em tweets no contexto estudado, devido à alta incidência de erros de predição na classe minoritária, a positiva.

As três representações de texto mencionadas anteriormente (TF-IDF, GloVe e FasText) foram aplicadas aos dados, contudo, tanto o FasText quanto o GloVe apresentaram resultados muito semelhantes e inferiores à representação utilizando TF-IDF. Assim, apenas os resultados obtidos utilizando a representação TF-IDF são expostos a seguir, na tabela 4.

	Resultados do IFG			**Resultados do Instituto Federal**		
Abordagem	**Macro-F1**	**Micro-F1**	**Desvio padrão**	**Macro-F1**	**Micro-F1**	**Desvio padrão**
Random Forests	0,56	0,49	0,04	0,58	0,52	0,03
Multinomial Naive Bayes	0,56	0,51	0,03	0,62	0,57	0,02
Support Vector Machine (SVM)	**0,58**	**0,54**	**0,03**	**0,63**	**0,59**	**0,01**
K-Nearest Neighbor	0,52	0,41	0,03	0,57	0,48	0,03

Tabela 4 – Efetividade dos métodos supervisionados para análise de sentimento

A Tabela 5 exibe os resultados de efetividade dos métodos de análise de sentimento baseados em transformers. Notavelmente, as abordagens centradas em transformers destacaram-se ao alcançar os resultados mais significativos. Essa conclusão respalda os benefícios da fase de pré-treino, na qual ocorre a transferência de conhecimento relacionado a padrões linguísticos de um extenso conjunto de dados para o modelo.

Assim, os dois desempenhos mais satisfatórios derivam do modelo BERTweet, que utiliza uma ampla base de tweets em inglês durante a fase de pré-treinamento, alinhando-se com a aprendizagem de padrões linguísticos característicos do Twitter. No entanto, é importante notar que a tradução impacta negativamente na performance desse método. O resultado mais expressivo foi alcançado pelo modelo BERTimbau pré-treinado extensivamente com um vasto conjunto de dados na língua portuguesa.

	Resultados do IFG			**Resultados do Instituto Federal**		
Abordagem	**Macro-F1**	**Micro-F1**	**Desvio padrão**	**Macro-F1**	**Micro-F1**	**Desvio padrão**
BERTweet	0,73	0,68	0,15	0,74	0,69	0,12
RoBERTa	0,54	0,44	0,21	0,76	0,73	0,09
BERTimbau	**0,78**	**0,75**	**0,18**	**0,86**	**0,84**	**0,09**
m-Bert	0,53	0,38	0,17	0,68	0,60	0,21
XML-RoBERTa	0,39	0,20	0,02	0,41	0,19	0,004

Tabela 5 – Efetividade dos métodos baseados em transformers para análise de sentimento

5.3 Modelagem de tópicos

5.3.1 Tópicos Gerais

A Figura 9 ilustra, em cada ponto, a coerência média (em 10 execuções) obtida com as execuções do BERTopic. Especificamente, cada ponto representa 10 execuções do BERTopic, com toda a base de dados coletada, no parâmetro "número de tópicos" (especificado no eixo x do gráfico). O BERTopic executado com apenas dois tópicos apresenta a menor coerência, indicando que a escolha de utilização de apenas dois tópicos corresponde a pouca representatividade dos assuntos em toda a base de dados coletada. Os resultados indicam que, de forma geral, a maior coerência pode ser obtida com a escolha de 4 tópicos principais.

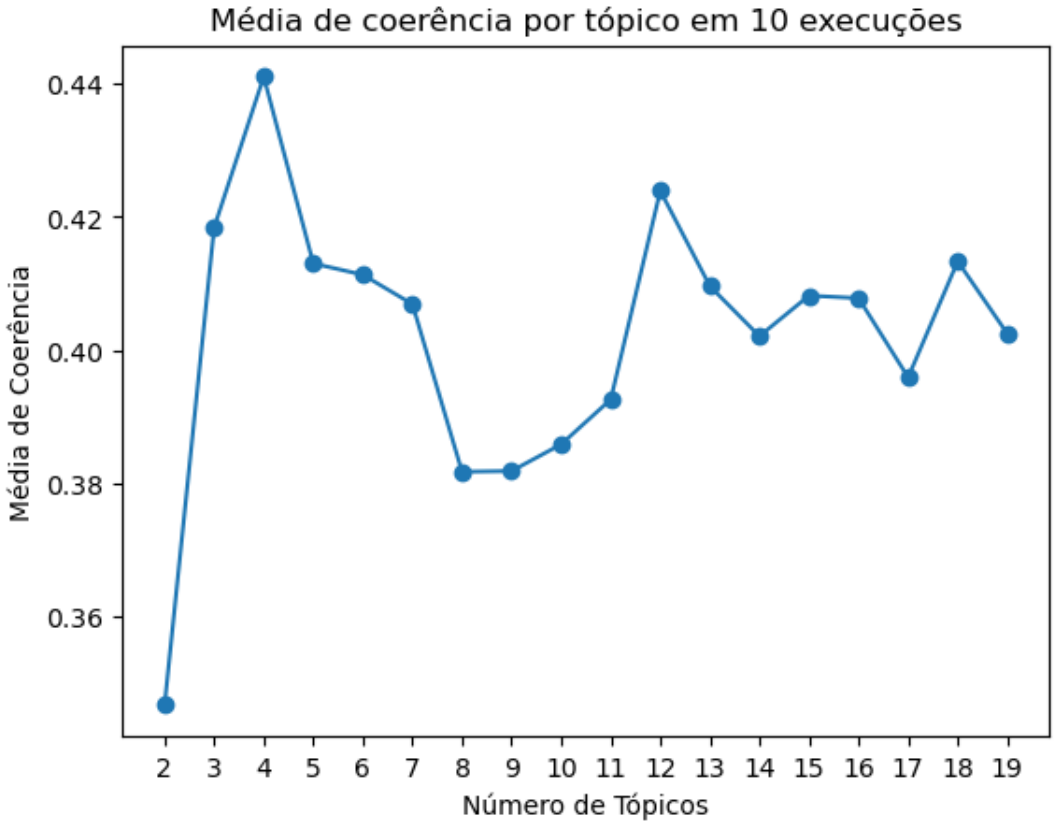

Figura 9 – Média de coerência por tópico em 10 execuções

Para realizar a caracterização dos Institutos Federais aplicou-se a modelagem de tópicos na coleção completa com o intuito de descobrir padrões gerais para o país e na coleção dividida em regiões para encontrar características específicas de partes do Brasil. A escolha dos tópicos mais representativos foi realizada analisando o *ranking* de coerências por tópico para a execução geral e para cada região.

Para a coleção completa, é possível notar, a partir da figura 10, que a comunicação entre aluno - instituição se faz presente em todas as regiões, isto é, a rede social é utilizada

como plataforma para resolução de dúvidas e envio de comunicados de uma forma agilizada e menos burocrática.

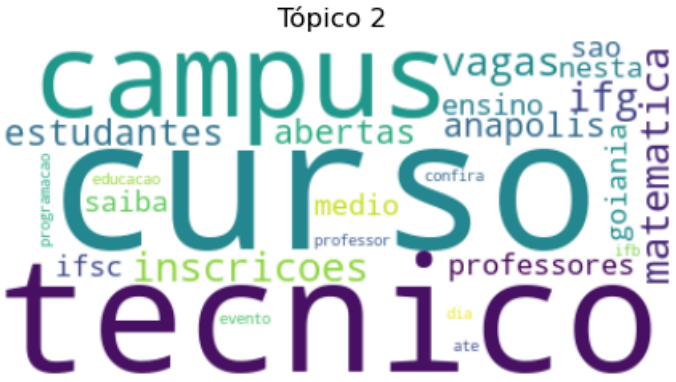

Figura 10 – Tópico global com maior coerência

Ademais, a partir da análise da figura 11, nota-se a grande ansiedade e preocupação dos alunos para com as provas em todas as regiões do país. Os estudantes demonstram grande aflição tanto para a etapa do processo seletivo para entrada nas instuições quanto para as provas cotidianas ao ser aprovado na instuição. Muitos alunos relatam a sensação de medo constante relacionado a avaliações. bem como, o surgimento de traumas e necessidade de procurar ajuda médica devido a pressão que sentem.

Figura 11 – 2º tópico global com maior coerência

5.3.2 Tópicos Específicos por Região

Região Centro-Oeste

A partir das execuções com a coleção dividida, foi possível ter a visão dos alunos sobre eventos/características dos IFs em específico. Por exemplo, a figura 12 mostra a comoção dos alunos da região Centro-oeste com uma votação do IFB sobre pesquisa e

desenvolvimento. Os alunos mostraram grande apoio à instituição e fizeram um grande esforço para divulgar e participar.

Figura 12 – Tópico com maior coerência para a região Centro-oeste

A partir da figura 13, nota-se que a análise de sentimento para a região Centro-Oeste segue o mesmo padrão observado nas coleções rotuladas relacionadas ao IFG e Institutos Federais, isto é, uma grande quantidade de tweets negativos e neutros em comparação com a pequena quantidade de positivos.

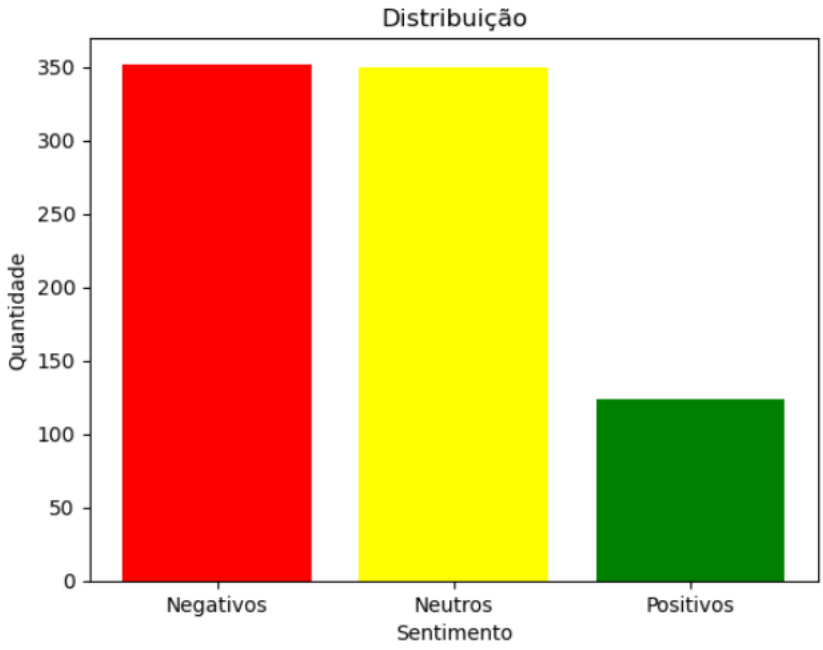

Figura 13 – Distribuição do sentimento para a região Centro-oeste

Região Nordeste

Já a partir da figura 14 que representa o Nordeste, nota-se a impolgação da comunidade e estudantes para com o acontecimento do Eclipse Anular no dia 14 de outubro.

O IFBA recebeu os interessados em acompanhar o evento e além disso, disponibilizou óculos especiais para todos.

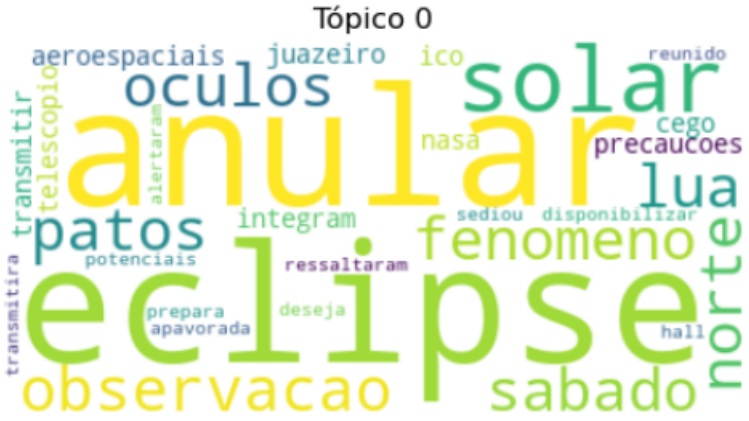

Figura 14 – Tópico com maior coerência para a região Nordeste

A figura 15 mostra a distribuição dos sentimentos previstos para a região Nordeste. É possível ver uma pequena alteração em relação ao padrão conhecido através rotolação, isto é, a maior quantidade de tweets neutros. Contudo, no geral, a distribuição se assemelha bastante com o abservado anteriormente.

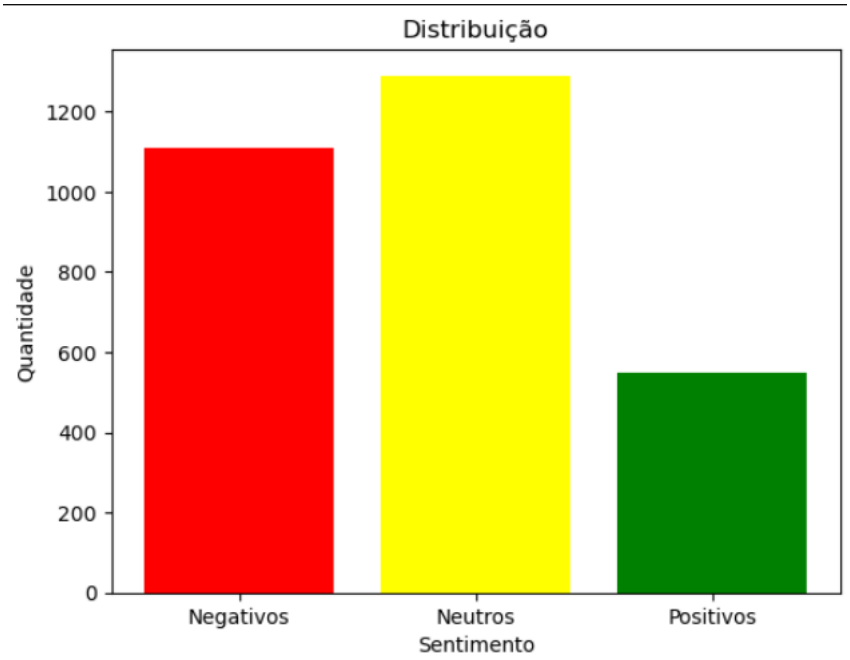

Figura 15 – Distribuição do sentimento para a região Nordeste

Região Norte

A figura 16, mostra o tópico de maior coerência para a região Norte. A figura não expõe um evento ou característica específica da região mas sim o dia a dia dos alunos que

estudam nas IFs dessa parte do Brasil. Ela mostra palavras usandas em *trends* como review, mostra a rotina dos estudantes nos câmpus e comentários sobre suas vidas particulares e rotinas.

Figura 16 – Tópico com maior coerência para a região Norte

A figura 17 mostra a distribuição dos sentimentos previstos para a região Norte. Nota-se que o padrão se assemelha bastante ao das coleções rotuladas.

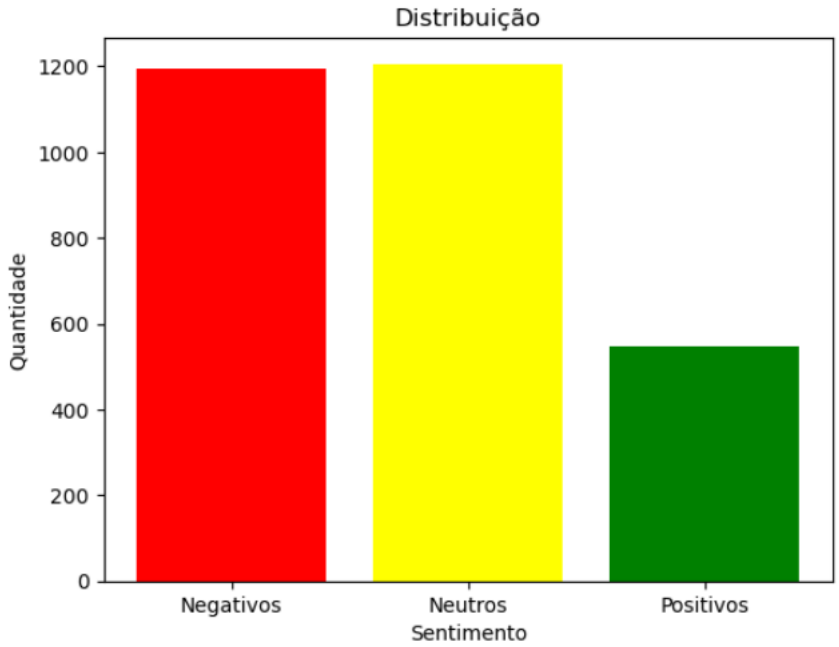

Figura 17 – Distribuição do sentimento para a região Norte

Região Sudeste

A figura 18, expõe a preocupação dos estudantes para com o processo seletivo dos Institutos Federais da região Sudeste, em especial, do Instituto Federal Do Espiríto Santo e do Instituto Federal Fluminense. O tópico de maior coerência da região Sudeste trás

a tona uma preocupação geral como exposto nos tópicos globais, contudo, com foco em instituições específicas.

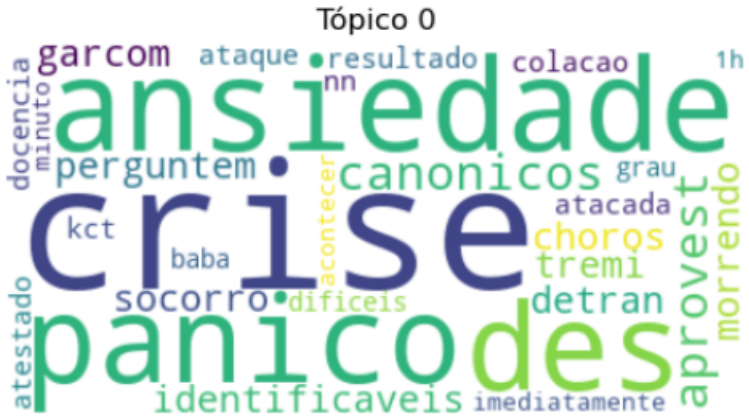

Figura 18 – Tópico com maior coerência para a região Sudeste

A figura 19 mostra a distribuição dos sentimentos previstos para a região Sudeste. Nota-se um maior balanceamento entre as classes em comparação com outras regiões uma vez que a quantidade de tweets é maior. Ademais, o padrão é semelhante ao rotulado.

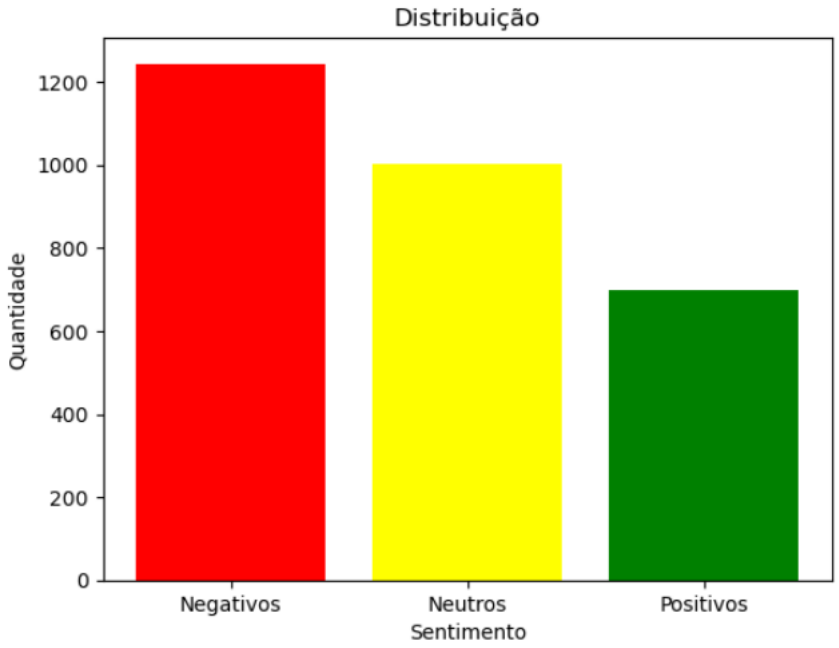

Figura 19 – Distribuição do sentimento para a região Sudeste

Região Sul

A partir da figura 20, que mostra o tópico de maior coerência da região Sul, nota-se a reação dos alunos ao cancelamento das aulas nos institutos por conta do grande volume de chuva na região. Em outubro de 2023, algumas cidades do Sul passaram por um período de grande volume de chuvas e inundações, provocando muitas perdas. Assim, as aulas

de algumas instituições foram canceladas e os alunos se manifestaram sobre o ocorrido, demostrando apreensão e insegurança com o fato.

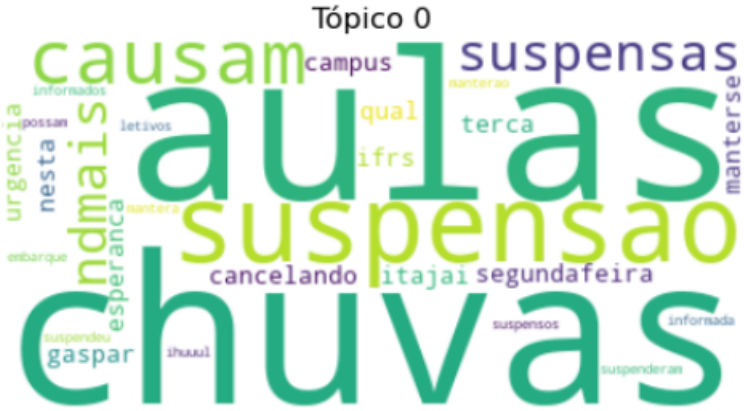

Figura 20 – Tópico com maior coerência para a região Sul

A figura 21 mostra a distribuição dos sentimentos previstos para a região Sul. Percebe-se que balanceamento é o mais próximo do visto nas coleções rotuladas, ou seja, uma maior quantidade de tweets negativos seguido de uma quantidade mediana de neutros e uma parcela bem inferior de dados positivos.

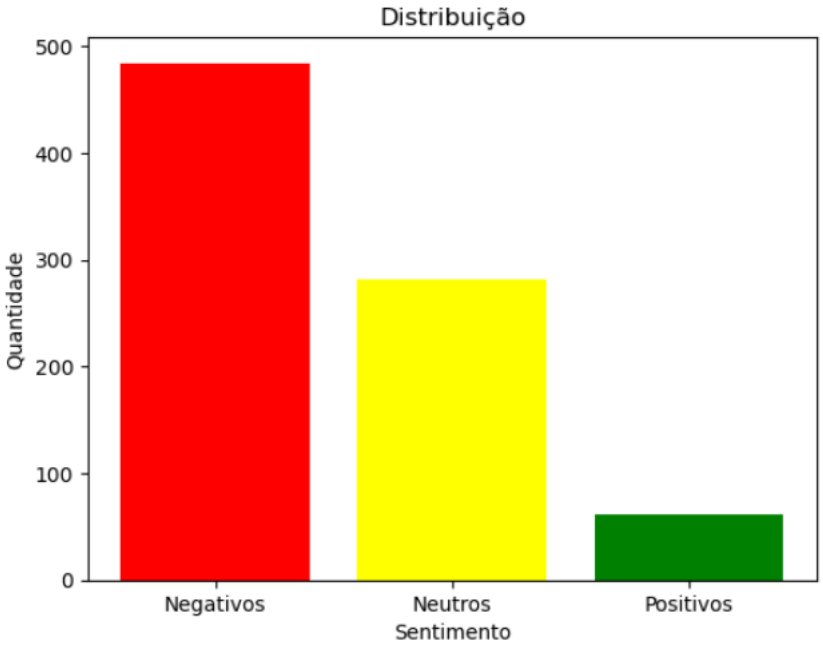

Figura 21 – Distribuição do sentimento para a região Sul

6 Conclusão

Nesse trabalho foram aplicados diferentes métodos de análise de sentimento, bem como a modelagem de tópicos, no contexto específico de textos curtos sobre Insitutos Federais brasileiros obtidos da rede social Twitter. Para tal, dados relacionados aos institutos advindos do Twitter foram coletados de forma contínua entre outubro de 2021 e novembro de 2023. Em seguida passaram por etapas de pré-processamento e aplicação de 11 técnicas de IA para a tarefa de análise de sentimento, sendo elas: Sentilex-pt, Vader, Random Forest, Support Vector Machine, Multinomial Naive Bayes, K-Nearest Neighbor, BERTweet, m-Bert, BERTimbau, XML-RoBERTa e RoBERTa. Por fim, o modelo BERTopic foi aplicado para realização da modelagem de tópicos.

A avaliação dos métodos de análise de sentimento alcançou o objetivo de identificação de estratégias efetivas ao contexto estudado. De fato, estratégias baseadas em métodos léxicos e técnicas baseadas em aprendizado de máquina tradicional se mostraram muito inferiores às técnicas recentes baseadas em transformers. Especificamente, a partir da análise dos resultados, foi possível observar a maior eficácia do BERTimbau, isto é, o método pré-treinado com uma quantidade extensiva textos genéricos em português foi a que gerou resultados mais eficazes, atingindo até 86% de efetividade (em Macro-F1).

Ademais, as limitações encontradas durante o desenvolvimento deste estudo estão voltadas à quantidade inferior de tweets considerados positivos (desbalanceamento de dados) e a linguagem utilizada pelos usuários do Twitter. Tal linguagem contém um grande número de gírias e erros gramaticais, características que podem dificultar a classificação correta.

Como trabalhos futuros, podemos citar a investigação de outras técnicas de modelagem de tópicos, bem como a evolução do sentimento em cada tópico considerando o nível mais específico (de estados da federação ou câmpus) de análise. Outras possibilidades englobam a detecção de anomalias em séries temporais, aplicada para identificação de padrões temporais discrepantes na evolução dos sentimentos em diferentes períodos.

Por fim, a identificação da modelagem adequada para análise de sentimentos no contexto estudado é um pressuposto essencial para criação de uma ferramenta efetiva para o monitoramento das opiniões expressas nas redes. Tal monitoramento tem o potencial de auxiliar, em tempo real, na identificação, prevenção e gestão de crises no âmbito da comunicação institucional e a identificação do apoio ou reprovação da comunidade em relação a medidas adotadas pela gestão dos institutos.

Referências

BEGANY, G. M.; MARTIN, E. G. Moving towards open government data 2.0 in u.s. health agencies: Engaging data users and promoting use. *Info. Pol.*, IOS Press, NLD, v. 25, n. 3, p. 301–322, jan 2020. ISSN 1570-1255. Disponível em: <https://doi.org/10.3233/IP-190169>. Citado na página 6.

CAMBRIA, E. et al. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 2013. Citado na página 9.

CAMPOS, R. et al. Stacking bagged and boosted forests for effective automated classification. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2017. (SIGIR '17), p. 105–114. ISBN 9781450350228. Disponível em: <https://doi.org/10.1145/3077136.3080815>. Citado 2 vezes nas páginas 13 e 20.

CAMPOS, R. C. et al. Stacking bagged and boosted forests for effective automated classification. *SIGIR*, 2017. Citado na página 6.

CANUTO, S.; GONCALVES, M.; BENEVENUTO, F. xploiting new sentiment-based meta-level features for effective sentiment analysis. *Proceedings of the ninth ACM WSDM international conference*, 2016. Citado na página 14.

CARVALHO, P.; SILVA, M. Sentilex-pt: principais características e potencialidades. *Oslo Studies in Language*, v. 7, n. 1, p. 425–438, 2017. Citado na página 20.

CARVALHO, P.; SILVA, M. J. Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language. 7.*, 2015. Citado na página 10.

CASTRO, L. D.; FERRARI, D. Introdução à mineração de dados: Conceitos básicos, algoritmos e aplicações. *ISBN*, 2017. Citado 2 vezes nas páginas 2 e 17.

CONNEAU KARTIKAY KHANDELWAL, N. G. V. C. G. W. F. G. E. G. M. O. L. Z. V. S. A. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Citado na página 15.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://doi.org/10.18653/v1/n19-1423>. Citado 2 vezes nas páginas 14 e 15.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. Citado na página 21.

EDUARDO, W.; CORTES, O. C. Utilizando análise de sentimentos e svm na classificação de tweets depressivos. *Computer on the Beach 2021*, 2021. Citado na página 12.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado 2 vezes nas páginas 7 e 8.

FILHO RODRIGO WILKENS, M. I. A. V. J. A. W. The brwac corpus: A new open resource for brazilian portuguese. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. Citado 2 vezes nas páginas 7 e 15.

GHOSH, K. et al. Imbalanced twitter sentiment analysis using minority oversampling. In: *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*. [S.l.: s.n.], 2019. p. 1–5. Citado na página 24.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022. Citado 4 vezes nas páginas 6, 15, 22 e 23.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001. (Springer Series in Statistics). Citado 2 vezes nas páginas 21 e 22.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random forests. in: The elements of statistical learning. *Springer Series in Statistics.*, 2009. Citado na página 11.

JIANG, L.; WANG, S.; LI CHAOQUN ZHANG, L. Structure extended multinomial naive bayes. *Information Sciences. 329*, 2016. Citado na página 13.

JUNIOR, L. A. D. P.; GUEDES, G. P.; BEZERRA, E. Inferindo o sexo de usuários de redes sociais utilizando o liwc em português do brasil. *Brazilian Workshop On Social Network Analysis And Mining (BRASNAM)*, 2016. Citado na página 10.

LIU, B. Sentiment analysis and opinion mining. *Morgan and Claypool Publishers, N. May.*, 2012. Citado na página 9.

LIU MYLE OTT, N. G. J. D. M. J. D. C. O. L. M. L. L. Z. V. S. Y. Roberta: A robustly optimized bert pretraining approach. *Arxiv*, 2019. Citado 2 vezes nas páginas 15 e 21.

MELO, R. L. Avaliando o dicionário em português do método de análise de sentimentos sentistrength. *TCC (Graduação em Engenharia de Software) - Universidade Federal do Ceará, Campus Quixadá*, 2017. Citado na página 10.

MIN, B. et al. Recent advances in natural language processing via large pre-trained language models: A survey. *Arxiv*, 2021. Citado na página 14.

MITCHELL, T. M. Machine learning. *Hill Science/Engineering/Math*, 1997. Citado na página 12.

NGUYEN, Q.; V., T.; NGUYEN, A. Bertweet: A pre-trained language model for english tweets. *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020. Citado 2 vezes nas páginas 15 e 21.

PATIL, R. S.; KOLHE, S. R. Supervised classifiers with tf-idf features for sentiment analysis of marathi tweets. *Social Network Analysis and Mining*, Springer, v. 12, n. 1, p. 51, 2022. Citado 2 vezes nas páginas 14 e 21.

QIANG, J. et al. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2020. Citado na página 15.

RESCH, B.; USLäNDER, F.; HAVAS, C. Combinação de modelos de tópicos de aprendizado de máquina e análise espaço-temporal de dados de mídia social para avaliação de pegadas de desastres e danos. *International Journal of Disaster Risk Reduction*, v. 75, p. 102760, 2023. Citado na página 6.

RIBEIRO, F. et al. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science 5*, 2016. Citado 2 vezes nas páginas 6 e 20.

SOUZA, F. D.; FILHO, J. a. B. d. O. e. S. Embedding generation for text classification of brazilian portuguese user reviews: From bag-of-words to transformers. *Neural Comput. Appl.*, Springer-Verlag, Berlin, Heidelberg, v. 35, n. 13, p. 9393–9406, dec 2022. ISSN 0941-0643. Disponível em: <https://doi.org/10.1007/s00521-022-08068-6>. Citado na página 6.

SOUZA, F. D.; FILHO, J. a. B. d. O. e. S. Embedding generation for text classification of brazilian portuguese user reviews: From bag-of-words to transformers. *Neural Comput. Appl.*, Springer-Verlag, Berlin, Heidelberg, v. 35, n. 13, p. 9393–9406, dec 2022. ISSN 0941-0643. Disponível em: <https://doi.org/10.1007/s00521-022-08068-6>. Citado na página 15.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado na página 14.

YUE L. CHEN, W. . L. X. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, v. 60, p. 617–663, 2019. Citado na página 6.