UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

ADICLEY DE OLIVEIRA COSTA - 202100104262
ALÍRIA DE SANTANA DE AMORIM CRUZ - 201900050813
HERBERT BARRETO FREIRE - 202200123977
IAGO SOARES DE MACEDO - 202100045644
LETÍCIA CAROLINE DA SILVA OLIVEIRA - 202000047655
LUCAS ARAGÃO DAMACENO - 202100045760
RODRIGO NUNES DE SANTANA - 202000047780

EVOLUÇÃO DE SOFTWARE

ATIVIDADE 1 - Análise de Sentimentos em Evolução de Software

SÃO CRISTÓVÃO – SE
2025

Os artefatos necessários para a realização desta atividade, bem como o vídeo tutorial dela, estão disponíveis no seguinte repositório do GitHub Evolucao_Software_2025-2_anything-llm.

Link do vídeo: https://www.youtube.com/watch?v=Mn4xuruklpE

Contribuições

Aluno	**Matrícula**	**Contribuição**
Adicley de Oliveira Costa	202100104262	Script, criação do notebook para execução dos modelos e apresentação da análise resumida
Alíria de Santana de Amorim Cruz	201900050813	Análise do impacto ao longo do tempo
Herbert Barreto Freire	202200123977	Avaliação de quais modelos foram mais efetivos
Iago Soares de Macedo	202100045644	Avaliação de quais modelos foram mais efetivos
Letícia Caroline da Silva Oliveira	202000047655	Comparação dos resultados através de tabelas
Lucas Aragão Damaceno	202100045760	Comparação dos resultados através de tabelas
Rodrigo Nunes de Santana	202000047780	Análise do impacto ao longo do tempo
Leonardo Lima Araujo	201900125781	Não entrou em contato / não contribuiu

Requisitos de Hardware

Para realizar a atividade, o projeto foi executado em um ambiente de nuvem do Google Colab com uma GPU T4 de 15GB de VRAM, 12.77GB de RAM do sistema, e disco com capacidade de 112GB

Introdução

Para realização da atividade, foi escolhido, da lista de projeto bem-sucedidos da atividade no Google Classroom, o projeto anything-llm, um projeto open source capaz de executar tarefas RAG e agentes de IA localmente na máquina sem a

necessidade de codificar ou configurar um ambiente externo ou interno para seu uso.

Nesse sentido, com o intuito de obter uma análise de sentimentos detalhada dos comentários do pull requests do projeto, foram escolhidos 3 modelos de linguagem do tipo classificação de texto (“text-classification”) da plataforma Hugging Face de acordo com o link compartilhado na atividade no Classroom para a análise de sentimento dos comentários de 100 pull requests consecutivas extraídas do projeto escolhido em questão. Para isso, foi feito um script auxiliar em Python o qual utiliza a API Rest do github para viabilizar uma extração mais efetiva dos comentários.

Em seguida, foi passado para o pipeline dos modelos selecionados cada um dos comentários extraídos das pull requests a fim de que pudesse ser feito a classificação do conteúdo de cada comentário em positivo, negativo ou neutro.

Por fim, foi criada uma tabela contendo o resumo dos dados de classificação de cada modelo selecionado com o intuito de listar quais foram os mais efetivos na análise de sentimentos proposta na atividade. Ao final, foi criada, também, uma tabela contendo as classificações dos comentários dos 3 principais colaboradores que mais deram commits no projeto na perspectiva dos modelos escolhidos.

Escolha dos modelos e resumo da análise de sentimentos para cada modelo escolhido.

Da página do hugging Face, foi escolhido pela equipe os seguintes modelos de classificação de texto:

```
models = ["clapAI/modernBERT-large-multilingual-sentiment",
          "lxyuan/distilbert-base-multilingual-cased-sentiments-student",
          "clapAI/roberta-large-multilingual-sentiment"]
```

1) Modelo: “clapIA/modernBERT-large-multilingual-sentiment”

O primeiro modelo de classificação de texto escolhido foi o “clanIA/modernBERT-large-multilingual-sentiment”, uma versão ajustada do modelo “answerdotai/ModernBERT-large”, que possui 400M de parâmetros, janela de

contexto de 8.192K de tokens e passou por um ajuste fino e foi treinado utilizando o dataset “clapAI/MultiLingualSentiment” que faz uso de um conjunto de dados de sentimentos com múltiplos idiomas.

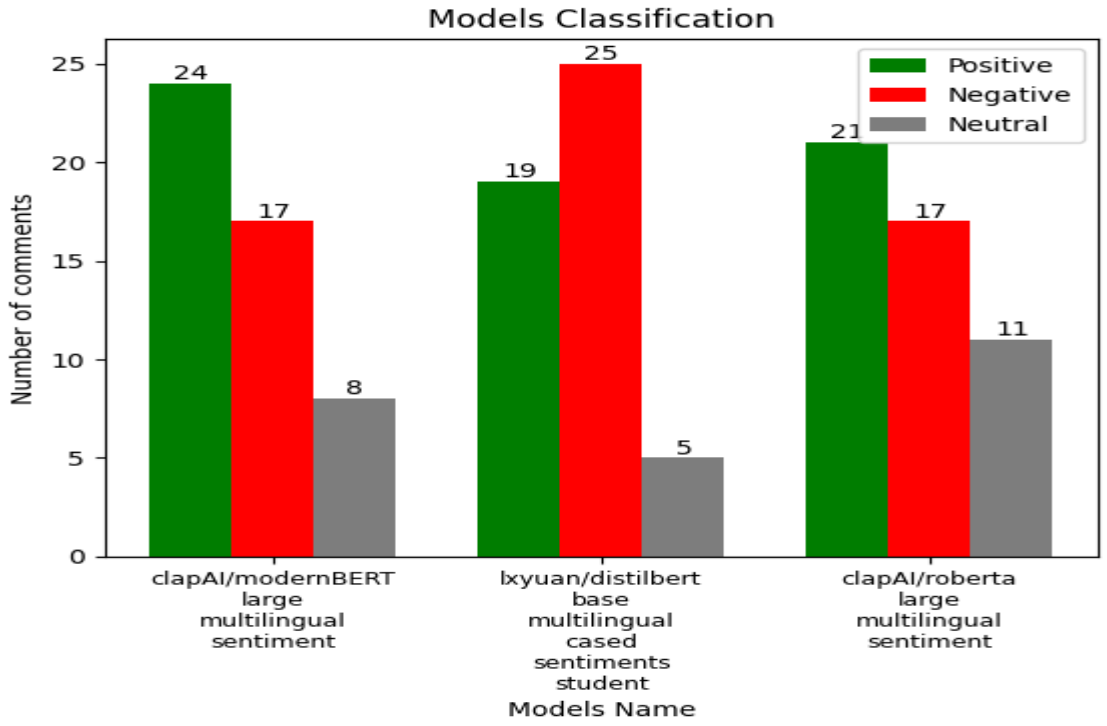
2) Modelo: “lxyuan/distilbert-base-multilingual-cased-sentiments-student”

O segundo modelo de classificação de texto escolhido foi o “lxyuan/distilbert-base-multilingual-cased-sentiments-student”, uma versão ajustada do modelo “distilbert/distilbert-base-multilingual-cased” (do tipo Fill-Mask), um modelo base com 100M de parâmetros, janela de contexto de 512 tokens que passou por um ajuste fino e foi treinado com o dataset “tyqiangz/multilingual-sentiments”.

3) Modelo: “clapAI/roberta-large-multilingual-sentiment”

Por fim, o terceiro modelo de classificação de texto foi o “clapAI/roberta-large-multilingual-sentiment", uma versão ajustada do modelo “FacebookAI/xlm-roberta-base”, um modelo roberta-large-multilingual-sentiment possui 600M de parâmetros, janela de contexto de 514 tokens o qual sofreu um ajuste fino usando o mesmo dataset do primeiro modelo mostrado anteriormente.

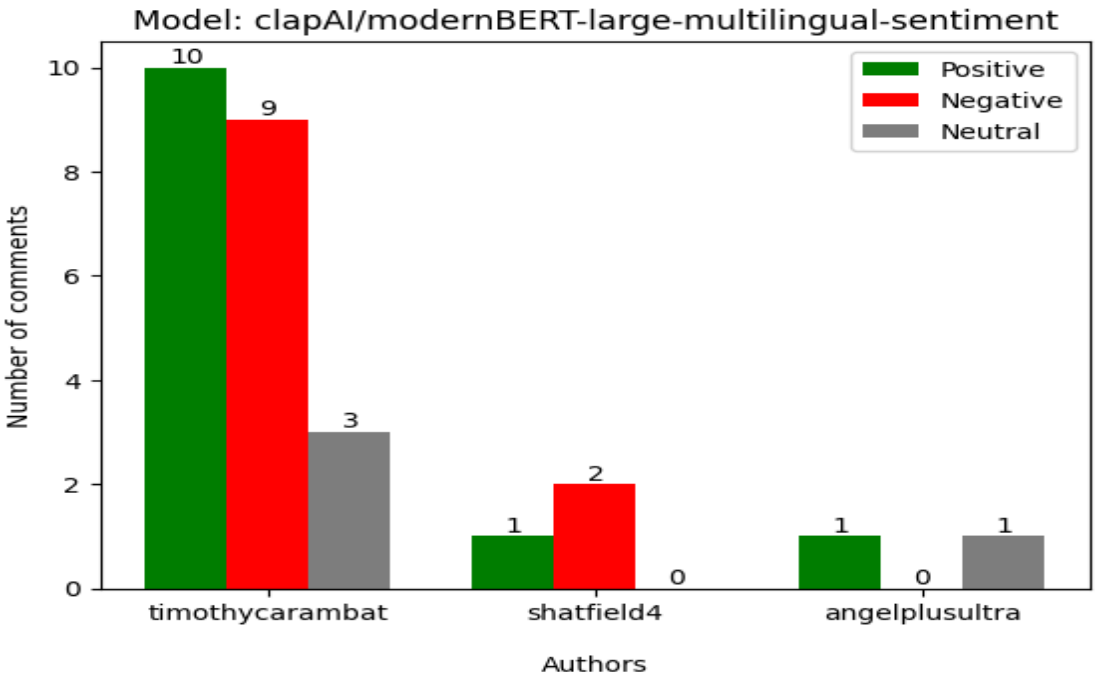
Dos resultados resumidos dos modelos para todos os comentários das PR’s, foi montado o seguinte gráfico de barras para uma visualização resumidas das classificações:



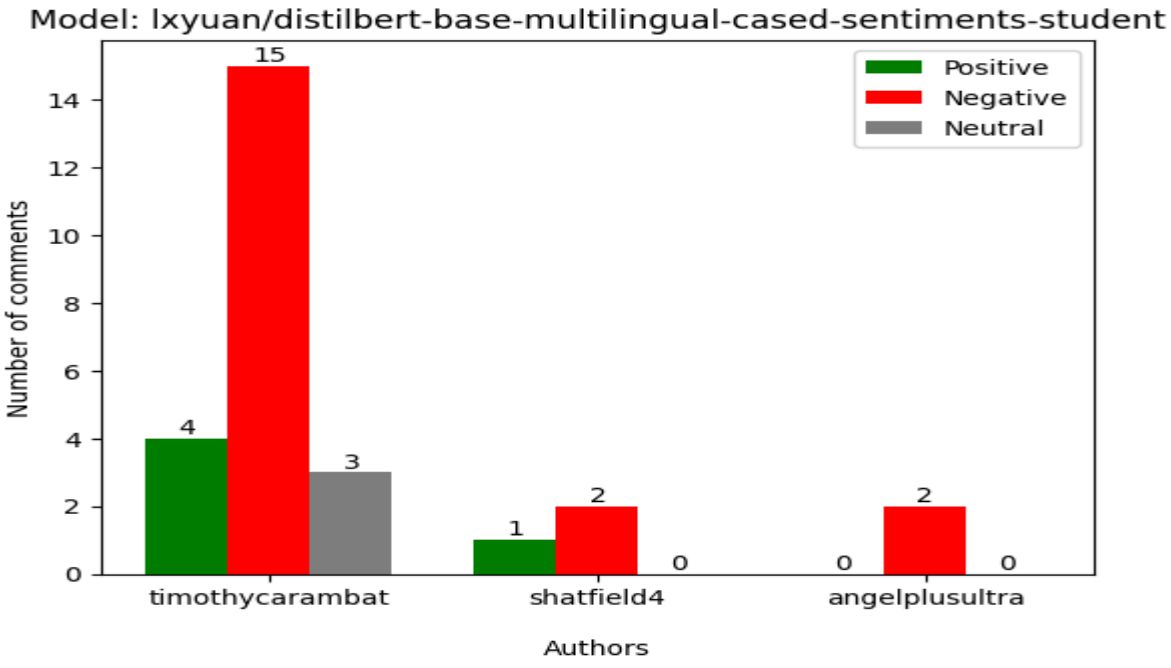
Dos 3 modelos escolhidos, todos apresentaram bastante similaridade na classificação dos comentários das pull requests, onde, após ser feita uma análise manual das classificações pela equipe, todos eles apresentaram, na maior parte dos casos, classificações corretas para cada comentário, com poucas inconsistências encontradas, onde, da análise dos dados de classificação, percebeu-se que os comentários classificados como positivos se referem, majoritariamente, a agradecimentos e correções de falhas e a inclusão de novas features no código fonte do projeto. Já a maior parte dos os comentários classificados como negativos se tratavam de duplicatas indevidas nas pull requests e bugs de software, e, por fim, o neutro foi aplicado, em maior parte, a discussões produtivas acerca de contribuições no projeto.

A partir dos dados coletados das classificações pelos modelos, também foi montado pela equipe os gráficos contendo, de forma resumida, a classificação dos 3 colaboradores mais ativos no projeto com o objetivo de observar, na perspectiva dos comentários feitos por esses colaboradores ao longo das diferentes pull requests, a análise de sentimentos realizadas pelas diferentes LLM's escolhidas no projeto:

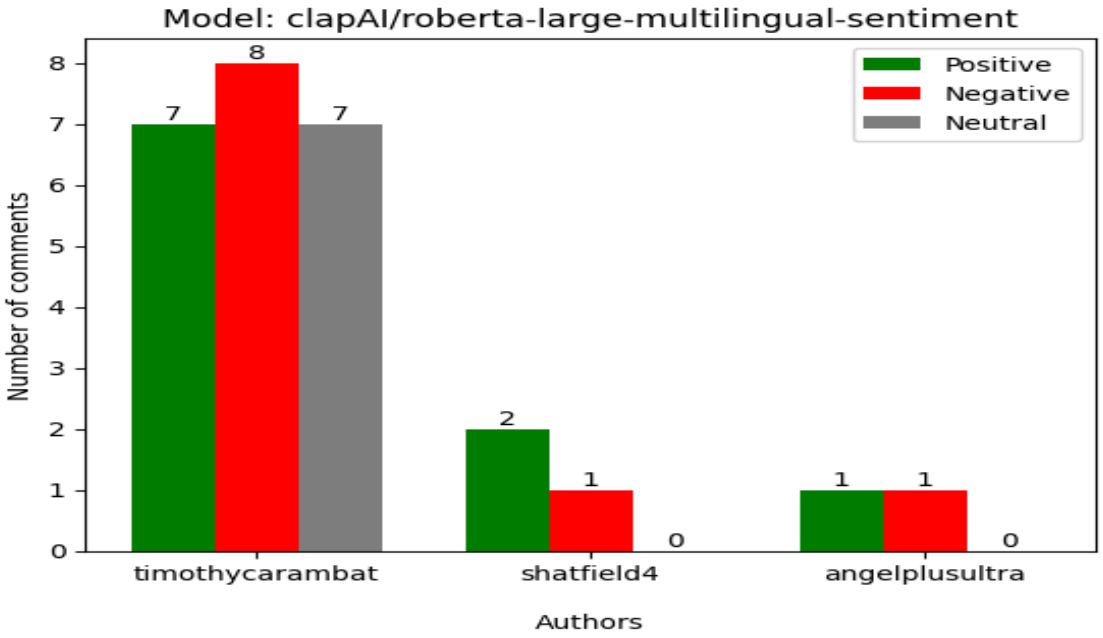
1) Modelo:



2) Modelo:



3) Modelo:



Da análise dos gráficos e também dos dados do arquivo de classificações dos modelos feita pela equipe, como também será discutido posteriormente nos demais tópicos da atividade, é que o tipo do modelo e também os dados ao qual o mesmo foi submetido para treinamento influenciam muito no resultado final, como foi observado no segundo modelo escolhido, sendo menos preciso que os demais modelos selecionados para a análise de sentimento feita pela equipe.

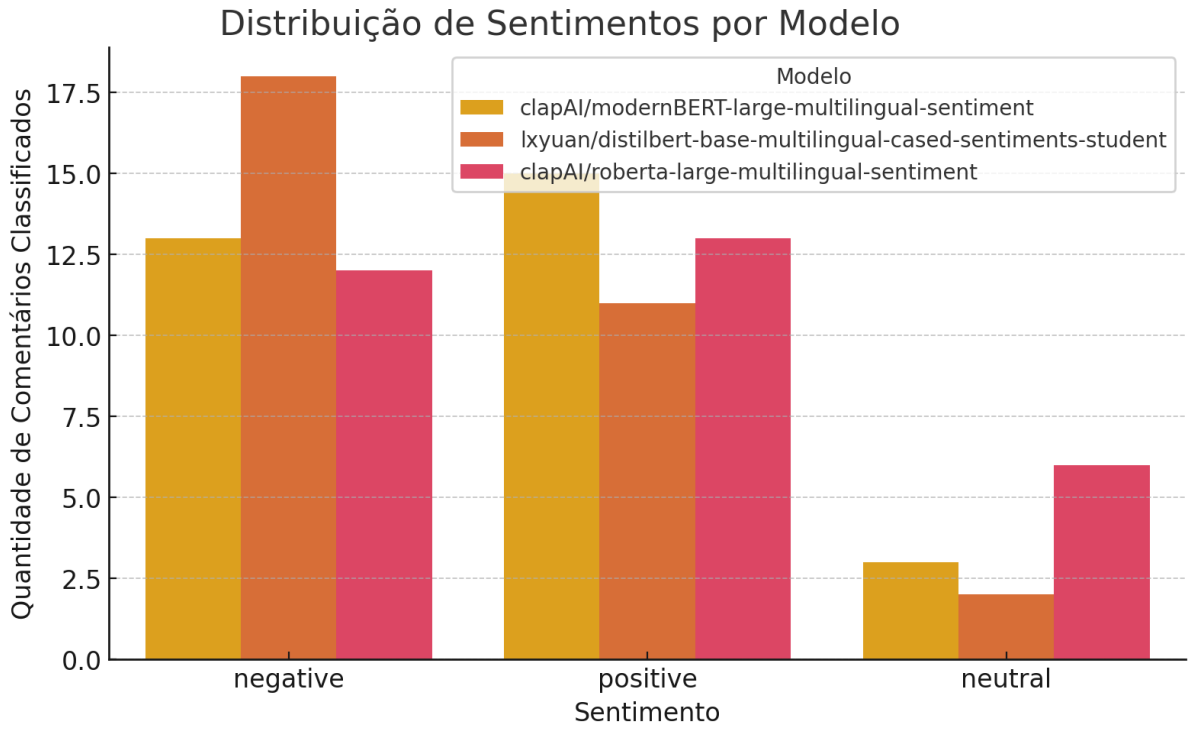
2 - Compare, através de uma tabela ou algo similar, os resultados da análise de sentimentos obtidos entre os modelos para cada um dos pull requests.

Das 100 pull requests selecionadas, apenas 31 continham comentários, e para facilitar a visualização e análise, foram consideradas apenas essas.

A primeira análise, feita a partir do gráfico abaixo de Distribuição de Sentimentos por Modelo, que representa a quantidade de comentários classificados em cada tipo de sentimento, mostra de um modo geral que a classificação de comentários negativos é predominante, o que é coerente com o contexto de pull requests em que os comentários “negativos” costumam estar associados a correções de bugs ou melhorias técnicas de código, como dito anteriormente. Os comentários positivos aparecem com uma frequência intermediária, e os neutros em menor proporção.

A comparação entre os modelos sugere que aqueles da clapAI apresentam padrões bastante próximos em suas análises, ou seja, têm sensibilidades semelhantes quanto a expressões utilizadas nos comentários. Já o modelo da lxyuan apresenta variações mais perceptíveis, principalmente na classificação negativa, que pode indicar uma menor precisão no tratamento de comentários com linguagem ambígua ou tecnicamente densa.

Apesar de todos os modelos apresentarem certa coerência entre si, o gráfico deixa claro que há diferenças entre eles na forma como o tom dos comentários é interpretado, principalmente nos casos que contêm ambiguidade. Essas diferenças decorrem do tipo de treinamento e grau de robustez dos modelos, e evidenciam a importância de considerar o contexto e o modelo a ser utilizado para análises automatizadas de sentimento em ambientes colaborativos.



As 100 pull requests também foram divididas ao meio para uma segunda análise dos dados. E das 50 primeiras, apenas 13 continham comentários a serem analisados. Vale ressaltar que as tabelas apresentadas a seguir aplicam um processo de agregação baseado no sentimento predominante em cada pull request, portanto, por mais que a PR tenha mais de um comentário, é o sentimento predominante entre eles que está apresentado na tabela para cada modelo.

pr_id	clapAI/modernBERT	lxyuan/distilbert	clapAI/roberta
4569	negative	negative	neutral
4545	positive	negative	neutral
4540	negative	neutral	negative
4526	negative	negative	neutral
4507	positive	positive	positive
4501	negative	negative	negative
4495	positive	negative	negative
4484	positive	positive	positive
4467	positive	negative	positive
4451	negative	positive	negative
4442	positive	positive	positive
4441	negative	negative	negative
4434	neutral	negative	negative

A tabela acima confirma o que o gráfico de Distribuição de Sentimentos já mostra, que de maneira geral, as 13 primeiras PRs com comentários analisados têm uma predominância de classificações negativas. E que a concordância entre os três modelos ocorre em aproximadamente ⅓ dos casos analisados neste primeiro conjunto, enquanto os demais casos apresentam leves divergências.

Dando seguimento à análise, a segunda tabela, que representa as 50 pull requests finais (das quais 18 continham comentários), demonstra um comportamento distinto do primeiro lote.

pr_id	clapAI/modernBERT	lxyuan/distilbert	clapAI/roberta
4404	negative	negative	negative
4402	positive	negative	positive
4349	positive	positive	positive
4347	neutral	neutral	neutral
4344	positive	positive	positive
4331	positive	positive	positive
4322	negative	negative	negative
4317	positive	negative	positive
4307	negative	negative	negative
4281	positive	positive	positive
4279	positive	negative	positive
4278	positive	negative	neutral
4274	positive	negative	negative
4273	negative	positive	positive
4271	negative	negative	negative
4266	negative	positive	negative
4258	negative	negative	positive
4247	neutral	positive	neutral

Neste segundo conjunto, nota-se uma predominância de classificações positivas. A concordância entre os três modelos também é significativamente maior do que no primeiro grupo (que foi de aproximadamente 1/3 dos casos), ocorrendo em 9 dos 18 casos (exatamente 50%). Destes casos de concordância, 4 foram classificados como "positive", 4 como "negative" e 1 como "neutral".

Esta maior concordância sugere que os sentimentos expressos neste conjunto de PRs podem ter sido menos ambíguos. Assim como na análise geral e no primeiro lote, o modelo 'lxyuan/distilbert' continua a ser o que mais apresenta divergências em relação aos outros dois, frequentemente classificando PRs de forma oposta aos modelos da clapAI.

3 - Dentre os modelos selecionados pela sua equipe, avalie quais foram os mais efetivos na análise de sentimentos. Justifique sua resposta.

A partir dos resultados obtidos, observou-se que o modelo ModernBERT (clapAI/modernBERT-base-multilingual-sentiment) apresentou o melhor

desempenho entre os três avaliados. Sua distribuição entre os sentimentos positivo (24), negativo (17) e neutro (8) foi a mais equilibrada, refletindo uma interpretação mais coerente com o tom técnico e colaborativo dos comentários das pull requests. Esse equilíbrio demonstra que o modelo foi capaz de captar nuances emocionais e contextuais nas interações entre os desenvolvedores, representando adequadamente comentários de agradecimento, sugestões e revisões construtivas.

O modelo DistilBERT (lxyuan/distilbert-base-multilingual-cased-sentiments-student) demonstrou comportamento mais polarizado, classificando 25 comentários como negativos, 19 como positivos e apenas 5 como neutros. Isso evidencia menor capacidade de distinguir sentimentos intermediários, resultando em um viés negativo mais acentuado em contextos ambíguos ou técnicos.

Por outro lado, o modelo RoBERTa (clapAI/roberta-large-multilingual-sentiment) apresentou bom desempenho geral, com 21 comentários positivos, 17 negativos e 11 neutros, revelando sensibilidade consistente, embora com leve tendência a classificar positivamente interações neutras. Ainda assim, manteve uma boa correlação entre o tom dos comentários e as classificações geradas.

Dessa forma, conclui-se que o ModernBERT foi o modelo mais efetivo para a análise de sentimentos no projeto anything-llm, pois apresentou maior equilíbrio e coerência nas classificações, oferecendo uma representação mais fiel do comportamento comunicativo dos colaboradores ao longo do desenvolvimento do software.

4 - Baseado na análise de sentimentos das pull requests ao longo do tempo, qual impacto pode ser constatado na evolução do projeto analisado? Justifique sua resposta.

A avaliação das LLM’s mostra que o impacto no projeto é impulsionado por dois tipos principais de interação (sentimentos) ao longo da execução: rigor técnico (muitas vezes classificado como negativo) e colaboração construtiva (classificada

como positiva). Dessa forma, é possível afirmar que existe um equilíbrio entre controle de qualidade e incentivo à colaboração ao longo do projeto.

Isso ocorre por que muitos dos comentários classificados como "negativos" ou "neutros" pelas LLMs não refletem conflito direto, mas sim um processo de revisão de código, como um pedido direto para correção de problemas; duplicatas de PRs; falta de necessidade ou ainda falta de alinhamento com a arquitetura. Esse fator é reforçado pois são raros os comentários em que há realmente algum conflito ou discrepância de ideias, prevalecendo o direcionamento ao rigor técnico anteriormente mencionado.

Em contraste, as avaliações positivas são observadas, principalmente, em contextos de reconhecimento de bom trabalho ou colaboração com a comunidade, demonstrando o forte comprometimento do projeto na visão *Open Source.* Quando há, por exemplo, contribuições de terceiros, os revisores sempre agradecem a participação, exigindo, no entanto, o seguimento dos padrões estabelecidos para o projeto.

Tal fator é perceptível pois comentários direcionados a correção de problemas ou duplicatas, por exemplo, tendem a ocorrer no início do projeto devido ao alinhamento técnico da equipe ou ainda por falta de familiaridade com o projeto. Dessa forma, conforme o desenvolvimento avança, os comentários de incentivo e colaboração tendem a aumentar.