Classificação de Reclamações com LLM Open-source Mistral 7B

Rildo Demarqui Pereira

Abstract. *The exponential growth of the user base, coupled with product diversification, has led mobile operators to grapple with a significant increase in the number of complaints. Managing this growing demand has become challenging, necessitating efficient solutions to extract valuable information from large volumes of unstructured data. The use of AI in this analytical task can be a valuable tool. This article focuses on exploring the Mistral 7B LLM, a pre-trained open-source model based on the transformer architecture, for multi-label classification of complaints extracted from the Reclame Aqui website.*

Resumo. *O crescimento exponencial da base de usuários, aliado à diversificação de produtos, levou as operadoras móveis a lidarem com um aumento significativo no número de reclamações. Lidar com essa crescente demanda tornou-se desafiador, gerando a necessidade de soluções eficientes para extrair informações valiosas de grandes volumes de dados não estruturados. A utilização de IA nessa tarefa de análise pode ser uma ferramenta de grande valia. O presente artigo se concentra na exploração do LLM Mistral 7B, um modelo pré-treinado de código aberto baseado na arquitetura transformer, para a classificação multi-rótulo de reclamações extraídas do site Reclame Aqui.*

1. Introdução

Desde a privatização ocorrida no final da década de 1990, o mercado de telecomunicações brasileiro tem testemunhado uma profunda transformação, não apenas em termos tecnológicos, mas também na dinâmica competitiva entre as operadoras de telefonia móvel. Essas mudanças obrigaram as empresas a se adaptarem e a buscar estratégias que as diferenciassem de seus concorrentes. Enquanto no início dos anos 2000, as operadoras concentravam-se principalmente nos serviços de voz, atualmente, são oferecidas uma ampla gama de serviços, como dados, banda larga, TV a cabo, streaming, jogos, entre outros.

A diversificação dos produtos oferecidos, aliado ao crescimento exponencial da base de usuários, levou as operadoras a lidarem com um aumento significativo no número de reclamações. Além disso, surgiram diversos meios para os usuários registrarem suas insatisfações tais como os canais internos corporativos, órgãos governamentais como a ANATEL e PROCON, além de plataformas de redes sociais e websites especializados.

Diante desse cenário, as operadoras móveis intensificaram seus esforços no tratamento e redução das reclamações. Embora os resultados dessas iniciativas tenham surtido certo

efeito na diminuição das queixas, conforme indicado pelo relatório da ANATEL[1], é importante notar que algumas operadoras não dão a devida atenção à alguns canais, como por exemplo o Reclame Aqui, onde figuram como as piores empresas no índice de solução do site[2].

A compreensão e o efetivo atendimento às demandas dos clientes não apenas representam desafios, mas também se revelam como oportunidades valiosas para as empresas. As reclamações, quando analisadas de maneira apropriada, constituem em uma fonte rica de informações, subsidiando a tomada de decisões e proporcionando insights cruciais para o aprimoramento contínuo dos serviços. Diante desses desafios, os modelos de linguagem de grande escala (LLM) podem ser úteis como ferramentas de análise.

2. Modelos de Linguagem de Grande Escala (LLM)

Os Modelos de Linguagem de Grande Escala (*Large Language Models* ou LLMs) representam uma categoria avançada de ferramentas de inteligência artificial baseadas em aprendizado profundo. Esses modelos generativos, compostos por redes neurais extensas contendo de milhões a trilhões de parâmetros, como o notável GPT-4 da OpenAI, são capazes de criar textos autonomamente, se destacando por sua habilidade em compreender nuances sintáticas e semânticas da linguagem humana. Essa capacidade é viabilizada pela arquitetura *transformer*, que se tornou fundamental para a eficiente captura e manipulação de padrões linguísticos complexos.

2.1 Arquitetura Transformer

As redes *transformers* representam uma inovação significativa no campo da inteligência artificial e processamento de linguagem natural. Proposta no artigo *Attention Is All You Need* por [Vaswani et al. 2017], essa arquitetura revolucionou a forma como os modelos de aprendizado de máquina lidam com sequências de dados, sendo amplamente adotada em uma variedade de aplicações, desde tradução automática até reconhecimento de fala e geração de texto.

O conceito central por trás dessa arquitetura é o mecanismo de atenção, uma abordagem que permite que o modelo atribua diferentes pesos a diferentes partes de uma sequência de entrada. Isso permite que o modelo se concentre em partes específicas do texto, capturando relações de longo alcance de maneira mais eficaz do que as arquiteturas anteriores.

2.2 Mistral 7B

O Mistral 7B é um modelo de código aberto, com licença Apache 2.0[3], oficialmente lançado em setembro de 2023 pela Mistral AI, uma startup francesa cofundada por profissionais oriundos da Meta e Google DeepMind[4].

---

[1] gov.br/anatel/pt-br/consumidor/destaques/anatel-registra-queda-de-23-em-volume-de-reclamacoes

[2] reclameaqui.com.br/ranking/

[3] xml.apache.org/xindice/license.pdf

[4] en.wikipedia.org/wiki/Mistral_AI

Atualmente, não existe uma definição oficial ou padronizada que determine exatamente a partir de qual tamanho um modelo de linguagem é considerado uma LLM (Large Language Model) ou uma SLM (Small Language Model). Com 7,3 bilhões de parâmetros, podemos definir o Mistral 7B como um LLM "leve", que foi treinado em diversas línguas para realizar múltiplas tarefas, alcançando excelente resultados e superando modelos maiores em diversos benchmarks [Jiang et al. 2023].

Sua arquitetura, fundamentada em *transformers*, emprega dois mecanismos-chave de atenção para otimizar seu desempenho:

- Atenção de consulta agrupada (GQA), proporcionando tempos de inferência mais rápidos;
- Atenção em janela deslizante (SWA), conferindo ao Mistral 7B a capacidade de lidar eficientemente com sequências de texto mais extensas a um custo computacional reduzido.

Essas abordagens permitem ao Mistral 7B direcionar seu foco computacional para segmentos mais relevantes da entrada, promovendo ganhos em desempenho, escalabilidade e eficiência computacional.

3. Estudo de Caso

3.1 Conjunto de Dados

Coletamos dados do Reclame Aqui, um site brasileiro que abrange mais de 30 milhões de consumidores e conta com o registro de 500.000 empresas, alcançando a marca de 1,5 bilhão de visualizações de páginas por ano[5]. Uma limitação da plataforma é que só é possível classificar a reclamação em uma única categoria. Contudo, na indústria, é comum que os usuários expressem insatisfação por diversos motivos simultaneamente, como, por exemplo, cancelar a assinatura devido a problemas recorrentes relacionados à cobertura e fatura incorreta.

Para esse estudo, extraímos um conjunto de dados com 202 amostras e rotulamos manualmente as reclamações com mais de uma categoria, quando aplicável. Esses rótulos servirão como comparativo de desempenho do modelo.

3.2 Implementação

Essa seção apresenta o experimento realizado para avaliação de desempenho do Mistral-7B-Instruct-v0.2, uma derivação do modelo base, porém ajustada como um modelo de instrução para conversação. Nos tópicos a seguir, apresentaremos algumas abordagens adotadas na configuração e aplicação do modelo durante os testes.

3.2.1 Temperatura

A temperatura é um hiperparâmetro importante no desempenho do modelo, onde em termos simples, controla a aleatoriedade das previsões. Um valor mais baixo, como 0.0, torna as previsões mais determinísticas e concentradas, resultando em respostas mais

[5] blog.reclameaqui.com.br/reclame-aqui-bate-recorde-de-reclamacoes-em-dezembro-de-2021/

conservadoras e com menor variabilidade. Por outro lado, um valor mais alto, como 1.0, introduz mais aleatoriedade, gerando respostas mais diversificadas e inovadoras, porém, por vezes, menos coerentes. Para esse estudo, foi utilizado o valor de 0.0, no intuito de dar mais consistência nas respostas trazidas pelo modelo.

3.2.2 Quantização

Embora o Mistral 7B seja considerado leve devido à sua arquitetura e tamanho comparativamente menores em relação a outros LLMs, optamos por utilizar seu modelo quantizado[6]. Resumidamente, a quantização é uma técnica que consiste na conversão dos pesos do modelo de uma representação de ponto flutuante de alta precisão para representações de ponto flutuante ou inteiros de menor precisão, como 16 bits ou 8 bits. Esse processo reduz significativamente o tamanho do modelo e melhora a velocidade de inferência sem comprometer muito a precisão, otimizando o consumo de recursos computacionais [Gholami et al. 2021, Wu et al. 2020].

3.2.3 Engenharia de Prompt

Ao elaborar o prompt, é importante considerar o formato que é mais eficiente para a interpretação do modelo, uma vez que diferentes modelos podem exigir estruturas específicas de entrada. No caso do Mistral, o prompt foi construído conforme as orientações fornecidas em sua documentação oficial[7].

Testamos três técnicas de prompt, sendo elas:

- *zero-shot*, onde é solicitado ao modelo a gerar uma resposta sem qualquer informação prévia ou exemplo fornecido;
- *few-shot learning*, onde alguns exemplos ou contextos são apresentados, fornecendo um leve direcionamento para a tarefa;
- *multi-turn conversation*, onde o modelo é exposto a uma sequência de interações, simulando um diálogo mais extenso.

Para obtermos uma maior consistência de resposta do modelo, os rótulos foram pré-definidos dentro do prompt, sendo eles: sinal/conexão de rede, cobrança indevida, consumo saldo/crédito, plano/benefício, cancelamento linha/plano, chip/sim card, spam, portabilidade, recarga/pagamento e dificuldade de contato.

3.3 Métricas de Avaliação

Para avaliar o desempenho do modelo, dada a natureza multi-rótulo do problema, empregamos as métricas de *precision*, *recall* e *f1-score* fornecidas pela função *classification_report* do scikit-learn. Existem várias maneiras de agregar essas métricas, como a *micro-average*, *macro-average*, *weighted average* e *samples average*. Optamos pela *samples average*, uma abordagem especialmente projetada para cenários multi-rótulo, que calcula as métricas individualmente para cada instância e, posteriormente, realiza a média.

---

[6] huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-AWQ

[7] docs.mistral.ai/

4. Resultados

Ao examinar os dados da Tabela 1, observamos que, em comparação com a abordagem *zero-shot*, tanto os prompts *few-shot learning* quanto os *multi-turn conversation* obtiveram pontuações superiores. Notavelmente, o uso do prompt *multi-turn conversation* resultou em uma melhoria na métrica de precisão em relação à abordagem *zero-shot*, sem comprometer o *recall*. Esses resultados corroboram com a ideia de que fornecer exemplos ao modelo contribui para aprimorar a qualidade da resposta.

Tabela 1. Comparativo de desempenho entre os prompts

	precision	recall	f1-score	support
zero_shot	0.6972	0.6738	0.6535	301.0
few_shot	0.6453	0.6563	0.6164	301.0
multi_turn	0.7195	0.6860	0.6740	301.0

Como o prompt *multi-turn conversation* apresentou os melhores resultados de *f1-score*, trouxemos abaixo, na Figura 1, os resultados obtidos para cada rótulo. Podemos notar que o modelo possui uma melhor pontuação para algumas categorias, como portabilidade, spam, sinal/conexão de rede e recarga/pagamento, isso é esperado, visto que são reclamações mais diretas quando comparadas com outras que são mais implícitas, como por exemplo, dificuldade de contato.

	precision	recall	f1-score	support
sinal/conexão de rede	0.96	0.65	0.77	34
cobrança indevida	0.78	0.71	0.74	45
consumo saldo/crédito	0.64	0.56	0.60	16
plano/benefício	0.44	0.42	0.43	40
cancelamento linha/plano	0.61	0.52	0.56	42
chip/sim card	0.88	0.45	0.60	31
spam	0.75	0.90	0.82	10
portabilidade	0.77	0.91	0.83	22
recarga/pagamento	0.68	0.85	0.76	33
dificuldade de contato	0.44	0.68	0.54	28
micro avg	0.66	0.64	0.65	301
macro avg	0.69	0.67	0.66	301
weighted avg	0.69	0.64	0.65	301
samples avg	0.72	0.69	0.67	301

Figura 1. Desempenho por rótulo

Foi criado um repositório no GitHub[8] com o código e conjunto de dados, no intuito de proporcionar a possibilidade de reprodução dos resultados deste estudo.

[8] github.com/rdemarqui/llm_complaint_management/tree/main/tim_data_academy

5. Conclusão e Discussões Finais

A utilização de modelos LLM para a classificação de textos surge como uma estratégia interessante tanto para as empresas de telecomunicações, quanto para outras indústrias, especialmente em cenários desprovidos de rótulos para treinar modelos supervisionados.

Vale ressaltar que os modelos *open source* com 7 bilhões de parâmetros apresentam a vantagem de serem leves, demandando recursos computacionais moderados. Outro ponto relevante é a possibilidade de utilização local do modelo, o que elimina a necessidade de envio de dados a plataformas externas, atendendo assim a requisitos de privacidade e conformidade com a LGPD.

Embora o Mistral 7B-instruct tenha demonstrado um desempenho satisfatório, é possível aprimorar os resultados através de treinamento adicional por meio de *fine-tuning*. Não obstante, a exploração de outros modelos, como Llama, Falcon, Zephyr, Openchat, ou até mesmo modelos maiores, como o Mixtral 8X7B, podem trazer resultados ainda melhores.

Referências

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020) “Language Models are Few-Shot Learners”, arXiv:2005.14165

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W. and Keutzer, K. (2021) “A Survey of Quantization Methods for Efficient Neural Network Inference”, arXiv:2103.13630

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M. -A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T. and Sayed, W.E. (2023) “Mistral7B”, arXiv:2310.06825

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017) “Attention Is All You Need”, arXiv:1706.03762

Wu, H., Judd, P., Zhang, X., Isaev, M. and Micikevicius, P. (2020) “Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation”, arXiv:2004.09602