Desenvolvimento de *guardrails* para controle e marcação de mensagens na interação de usuários em sistemas baseados em grandes modelos de linguagem

Ricardo Augusto Franco | ricardoaugustofranco@hotmail.com
Eli Lopes da Silva | eli.lopes@ifsc.edu.br
Cristiano Mesquita Garcia | cristiano.garcia@ifsc.edu.br

RESUMO

Esta é uma pesquisa que faz parte de um trabalho de conclusão de curso do bacharelado em Sistemas de Informação do IFSC Câmpus Caçador. As soluções baseadas em inteligência artificial estão cada vez mais presentes, impactando diversos setores e gerando avanços significativos em várias áreas. A adoção crescente desses modelos requer a implementação de mecanismos que garantam a qualidade e a segurança das interações, especialmente em contextos sensíveis como o ambiente de cobrança, onde é essencial identificar conversas que não agregam valor ao negócio para evitar custos desnecessários para a empresa. Este trabalho tem como objetivo o desenvolvimento de *guardrails* para controle e marcação de mensagens enviadas por usuários durante interações em sistemas de cobrança baseados em grandes modelos de linguagem, os *Large Language Models* (LLMs). Para isso, propõe-se a criação de uma Interface de Programação de Aplicações (API), que utiliza técnicas de processamento de linguagem natural para analisar as mensagens. A API classifica e marca interações que apresentem padrões de comportamento prejudiciais, permitindo respostas rápidas das equipes de cobrança e a criação de *datasets* mais precisos. A implementação dos *guardrails*, ao focar nas mensagens dos usuários, busca filtrar interações que possam resultar em ações indevidas, como fraudes ou disputas maliciosas.

Palavras-chave: *guardrails*; aprendizado de máquina; grandes modelos de linguagem; inteligência artificial.

1 INTRODUÇÃO

A comunicação instantânea tornou-se o pilar das relações pessoais e corporativas, de tal forma que e-mails, SMS, WhatsApp e Telegram dominam o intercâmbio de informações no Brasil — país em que 85 % dos consumidores preferem tratar com empresas por mensagens (Frenay *et al*., 2024). Nesse cenário, o setor financeiro migra para canais digitais a fim de reduzir custos e aumentar a eficiência em lembretes de cobrança, renegociações e ofertas. A Monest Cobranças, fundada em 2018, ilustra essa tendência: pioneira no uso de *Large Language Models* (LLMs) em português, a empresa ampliou em 130,8% (dados coletados pelo autor) seu faturamento anual após adotar IA generativa, confirmando o impacto econômico da inovação.

Operar LLMs envolve despesas contínuas de infraestrutura, energia e mão de obra qualificada, que se agravam quando o sistema processa interações maliciosas — tentativas de manipulação, spam ou uso indevido que consomem recursos sem gerar valor. Para mitigar esses riscos, técnicas de aprendizado de máquina (*autoencoders*, SVMs e afins) podem funcionar como *guardrails*, detectando anomalias em tempo real e filtrando interações que não contribuem para a recuperação de crédito. Essa abordagem preserva a integridade do atendimento, reduz custos operacionais e reforça a conformidade ética e legal, em especial diante da presença do WhatsApp, que de acordo com Guanaes (2023) alcança 99% dos *smartphones* brasileiros.

Assim, o trabalho propõe um modelo de monitoramento inteligente que identifica e classifica interações indesejadas nos canais digitais da Monest, conciliando avanços em IA com sustentabilidade financeira. Ao otimizar o uso de recursos e proteger o sistema contra abusos, a empresa fortalece seu relacionamento com devedores legítimos, sustenta seu crescimento e ilustra como a convergência entre mensageria massificada e LLMs pode redefinir a recuperação de crédito no Brasil.

2 FUNDAMENTAÇÃO TEÓRICA

O Cross-Industry Standard Process for Data Mining (CRISP-DM) é um modelo de processo amplamente aceito oferecem uma abordagem estruturada para projetos de mineração de dados. Desenvolvido em 1996 por um consórcio de empresas, incluindo SPSS, NCR Corporation e Daimler-Benz, tornou-se o padrão na indústria de análise de dados (Chapman *et al*., 2000) É composto por seis etapas:

Compreensão do Negócio: essa etapa é dedicada a entender os objetivos e requisitos do projeto a partir de uma perspectiva de negócios;

Entendimento dos Dados: consiste na coleta inicial e análise exploratória dos dados para compreender suas características;

Preparação dos dados: inclui todas as atividades necessárias para construir o conjunto de dados final a partir dos dados brutos;

Modelagem: etapa em que técnicas específicas de modelagem são selecionadas, aplicadas e avaliadas;

Avaliação: após a construção dos modelos, é realizada uma análise para verificar se eles atendem aos objetivos do projeto e resolvem os problemas de negócios;

Implementação: na última etapa, o modelo aprovado é implementado no ambiente operacional. Isso pode incluir sua integração em sistemas de tomada de decisão, apresentação dos resultados para os stakeholders ou documentação detalhada das lições aprendidas durante o projeto.

A área do *Natural Language Processing* (NLP) ou Processamento de Linguagem Natural, investiga a interação entre computadores e a linguagem humana, visando capacitar as máquinas a compreenderem, interpretarem e gerarem texto ou fala de forma que seja natural e útil para os usuários (Goodfellow; Bengio; Courville, 2016). A NLP abrange uma variedade de tarefas, como tradução automática, análise de sentimentos, resposta automática a perguntas e geração de conteúdo textual.

A introdução de arquiteturas baseadas em *deep learning*, especialmente os modelos *Transformer*, trouxe uma revolução para o campo do NLP. Propostos por Vaswani et *al*. (2023), os Transformers são modelos de aprendizado de máquina que utilizam exclusivamente o mecanismo de atenção para processar sequências de dados. Diferentemente dos modelos anteriores, como as *Recurrent Neural Networks* (RNNs), os *Transformers* permitem o processamento paralelo dos dados e são mais eficientes ao capturar dependências de longo prazo. Esse avanço resultou em melhorias substanciais em diversas tarefas de NLP.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa adotou uma metodologia exploratória, cujo objetivo é ampliar a familiaridade com o problema e tornar suas nuances mais explícitas (Gil, 2002). Por meio de levantamento bibliográfico e análise de casos, investigaram-se as possibilidades de aplicar *autoencoders* e modelos de linguagem no monitoramento de interações mal-intencionadas em sistemas de cobrança digital. Os achados de Limna *et al*. (2018), que demonstram a eficácia dos *autoencoders* na detecção de anomalias em textos curtos, aliados às etapas do CRISP-DM, serviram de base para identificar lacunas, estruturar hipóteses e delinear uma API segura e eticamente orientada.

Paralelamente, empregou-se uma abordagem descritiva-comparativa para caracterizar e confrontar mensagens legítimas e mal-intencionadas, estabelecendo relações entre seus padrões linguísticos (Marconi; Lakatos, 2017). Essa comparação forneceu insumos para estratégias de mitigação que eliminam interações sem valor para o negócio, ao mesmo tempo que reforçam a confiabilidade do sistema.

4 RESULTADOS E DISCUSSÕES

Os resultados ainda são parciais e refletem os primeiros esforços para diferenciar, em um fluxo de cobrança, as conversas que efetivamente geram valor daquelas que apenas consomem recursos. Durante a etapa de entendimento dos dados (Chapman *et al*., 2000), a tarefa se denotou especialmente complexa, mesmo com a facilidade dos modelos que utilizam

autoencoders de entender diferentes contextos (Limna *et al.*, 2018), porque a língua portuguesa é altamente volátil: expressões coloquiais, gírias regionais e até negações, “não posso”, “não consigo”, “não agora”, compõem o diálogo legítimo de quem negocia dívidas e, portanto, não podem ser tratadas como ruído por simples ausência de positividade. Além disso, o conjunto de dados analisado é marcadamente desbalanceado: menos de 10% das interações foram rotuladas como “sem valor”, o que intensifica o risco de vieses na modelagem e exige técnicas específicas de balanceamento e avaliação para que o classificador não superestime a relevância das demais conversas.

5 CONSIDERAÇÕES FINAIS

Guiado pelo modelo CRISP-DM e por práticas ágeis, o projeto combinou etapas estruturadas de mineração de dados com ciclos iterativos e incrementais capazes de atender às rápidas mudanças do mercado de cobrança digital e de IA garantindo fluxo de trabalho consistente. A revisão bibliográfica norteou a seleção de ferramentas focadas em escalabilidade e desempenho, enquanto preocupações éticas e legais — especialmente conformidade à LGPD e mitigação de vieses — orientam decisões de *design*, reforçando a responsabilidade social da solução. Paralelamente, a análise de negócio assegurou alinhamento às necessidades reais dos usuários, integrando planejamento estratégico, pensamento crítico e resolução de problemas complexos. Esse cruzamento interdisciplinar consolida uma experiência acadêmica que desenvolve competências técnicas e de mercado, preparando o profissional para criar soluções de alto impacto positivo na sociedade.

REFERÊNCIAS

CHAPMAN, P. *et al*. CRISP-DM 1.0: Step-by-step data mining guide. **SPSS inc**, v. 9, n. 13, p. 1–73, 2000.

FRENAY, L. et al. **Messaging as a Value Driver for Brazilian Businesses**. 2024. Disponível em: https://web-assets.bcg.com/b9/32/2fdce0d9409780446961596f5aee/messaging-as-a-value-driver-for-brazilian-businesses.pdf. Acesso em: 7 dez. 2024.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo, Brasil: Atlas, 2002.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016.

GUANAES, Lucas. O discreto monopólio do meta: um estudo de caso sobre o domínio do Whatsapp no Brasil. **Seminário do LEG**, Limeira, SP, v. 14, n. 1, p. 29–41, 2024.

LIMNA, P. *et al*. The use of chatgpt in the digital era: perspectives on chatbot implementation. **Journal of Applied Learning and Teaching**, v. 6, n. 1, p. 64–74, 2023.

MARCONI, M. A.; LAKATOS, E. M. **Metodologia científica**. 7. ed. São Paulo: Atlas, 2017.

VASWANI, A. *et al*. Attention Is All You Need. **arXiv preprint arXiv:1706.03762**. 2023. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 15 out. 2025.