# DESENVOLVIMENTO DE UMA PLATAFORMA DE VISUALIZAÇÃO E ANÁLISE DE DADOS DE SOCIAL CRM

Ismael Freitas de Oliveira[1]

Antonio Fernando Lavareda Jacob Junior[2]

## INTRODUÇÃO

A inovação tecnológica, definida pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE) como a implementação de novos produtos e processos no mercado (Erdin; Mehmet, 2023), é um pilar fundamental para o avanço competitivo. Na atual era digital, essa inovação é cada vez mais impulsionada pela capacidade de extrair informações úteis de grandes volumes de dados. Dentre as diversas fontes disponíveis, as mídias sociais se destacam como um ambiente rico em dados em tempo real, onde a análise de Big Data se tornou uma ferramenta indispensável para a tomada de decisões estratégicas no contexto da transformação digital corporativa (Rathore, 2023; Rahman; Reza, 2022).

Neste contexto dinâmico, a gestão de relacionamento com o cliente tradicional evoluiu para o Social CRM, que utiliza a Análise de Mídias Sociais para compreender profundamente a percepção do público. Contudo, diversas ferramentas de mercado atuais ainda se concentram de forma isolada em análises descritivas, carecendo de capacidades preditivas e prescritivas mais robustas que poderiam guiar decisões gerenciais de forma muito mais eficaz (Wittwer; Reinhold; Alt, 2016).

Para preencher essa lacuna, este projeto de pesquisa propôs a criação de uma plataforma funcional que atue como prova de utilidade para a aplicação avançada de aprendizado de máquina (*Machine Learning*) em ecossistemas de Social CRM. O foco reside em desenvolver e validar um artefato de software que integre métodos contínuos de extração de informações, classificação de conteúdo em massa e predição de engajamento a partir de dados coletados em redes sociais.

Ao automatizar processos complexos de análise de texto e aprimorar a confiabilidade dos insights gerados, a plataforma tem o potencial de trazer vantagens competitivas reais às corporações. Os benefícios envolvem a redução significativa de custos operacionais e o refinamento da gestão de relacionamento com o cliente. A solução é direcionada a um amplo espectro de atores do mercado, englobando empresas de marketing digital, grandes varejistas, micro e pequenos empreendedores e provedores de tecnologia que buscam otimizar suas estratégias de comunicação em plataformas de mídias sociais.

Portanto, o objetivo geral deste estudo é desenvolver uma plataforma de visualização e análise de dados de Social CRM que integre a captura de informações e algoritmos de aprendizado de máquina, facilitando a interpretação desses dados pelo usuário final. De forma específica, o projeto busca: desenvolver módulos para a integração de dados de redes como Facebook, Twitter e Instagram; selecionar e adaptar algoritmos de aprendizado de máquina para análise preditiva e extração de tópicos; e, por fim, criar uma interface interativa que democratize o acesso aos resultados obtidos.

## METODOLOGIA

As atividades para a construção desta plataforma fundamentaram-se na metodologia Design Science Research – DSR (Peffers et al., 2007). Esta abordagem mostra-se altamente

1 Discente em Engenharia de Computação, Universidade Estadual do Maranhão. E-mail: ismael.20210003095@aluno.uema.br
2 Doutor em Engenharia Elétrica, Docente do Departamento de Engenharia de Computação da Universidade Estadual do Maranhão. E-mail: antoniojunior@professor.uema.br

alinhada ao desenvolvimento de sistemas de software orientados para a resolução de problemas práticos, como a categorização e o monitoramento de dados de marketing nas redes sociais. Em relação às tecnologias utilizadas, os critérios adotados exigiram que as ferramentas fossem rigorosamente de código aberto (open-source), assegurando a viabilidade econômica do projeto; que oferecessem integração nativa com as bibliotecas essenciais para a modelagem dos algoritmos; e que possuíssem uma curva de aprendizado rápida, garantindo a entrega do produto dentro do cronograma.

## Conjunto de Dados

Para a realização dos experimentos e testes do sistema, foi necessário construir um conjunto de dados (*dataset*) capaz de representar o comportamento dos consumidores de forma realista. Inicialmente, o projeto planejava utilizar técnicas de captura de dados diretamente das APIs oficiais de redes como Twitter, Facebook e Instagram. Contudo, desafios técnicos e legais forçaram uma adaptação metodológica.

As recentes restrições de acesso às APIs oficiais, os altos custos cobrados pelas plataformas para a extração acadêmica de dados e as severas exigências da Lei Geral de Proteção de Dados - LGPD (Brasil, 2018) tornaram a utilização de informações orgânicas um processo instável. O uso de dados reais de usuários exigiria um trabalho longo e complexo de anonimização, o que fugiria do foco principal da pesquisa. Diante desse cenário, optou-se pela geração de um *dataset* sintético meticulosamente controlado.

Essa estratégia alternativa ofereceu vantagens metodológicas importantes. A principal delas foi o controle total sobre o balanceamento das amostras de treinamento, permitindo definir a proporção exata de textos com sentimentos positivos, negativos e neutros. Esse ambiente equilibrado é ideal para treinar algoritmos de aprendizado de máquina, evitando os vieses comuns em conjuntos de dados orgânicos (que costumam ser muito polarizados). Além disso, a simulação garantiu conformidade ética imediata, preservando a privacidade de usuários reais.

A base de dados foi projetada em torno de uma marca de tecnologia hipotética, denominada "OmniTech". A geração dos dados obedeceu a quatro etapas lógicas de estruturação: (1) Definição dos eixos centrais, contendo colunas de identificação, plataforma de origem, marcação de data/hora, conteúdo bruto da postagem, métricas de curtidas/comentários e o rótulo do sentimento. (2) Engenharia de Personas, mapeando perfis que variam desde o cliente entusiasta (focado em elogiar o unboxing e o ecossistema) até o usuário frustrado e irritado com ineficiências logísticas. (3) Redação controlada de 500 postagens, emulando defeitos técnicos, atrasos, dúvidas de compatibilidade ou satisfação com o design minimalista. (4) Simulação do Engajamento, onde, por meio de lógicas correlacionais, posts relatando problemas de conectividade recebiam uma tração maior de comentários colaborativos (ou queixas de outros usuários), enquanto fotos bem elaboradas de produtos recebiam um salto em curtidas, especialmente se atreladas ao cenário do Instagram.

## Adoção do Framework Streamlit

Inicialmente, planejou-se desenvolver a aplicação web utilizando o padrão Model-View-Controller (MVC), com o apoio de frameworks tradicionais como Django ou Flask. Entretanto, o desenvolvimento nesse formato exigiria um grande esforço focado na programação de interfaces visuais (HTML, CSS e JavaScript), além da configuração de rotas e integração de banco de dados, o que se mostrou improdutivo para a etapa de prototipagem rápida prevista pela DSR.

Para solucionar esse obstáculo, adotou-se o Streamlit (Streamlit, 2024; Smith; Wang; Li, 2024), uma biblioteca em Python desenvolvida especificamente para facilitar o trabalho

de cientistas de dados. Diferente de aplicações tradicionais, o Streamlit possui um modelo declarativo que converte a lógica de análise de dados diretamente em dashboards interativos na web com poucas linhas de código. A Tabela 1 resume as principais diferenças que justificaram a escolha dessa ferramenta.

**Tabela 1 -** Comparação de Frameworks de Desenvolvimento Web para Dados.

| Critério | Streamlit | Flask/Django (MVC) |
|---|---|---|
| Velocidade de Desenvolvimento | Altíssima. Converte scripts de análise em aplicações web com poucas linhas de código. | Moderada. Exige a configuração de rotas, templates e interface front-end. |
| Curva de Aprendizagem | Baixa. Requer apenas conhecimento de Python e bibliotecas de dados. | Alta. Exige o uso de Python, HTML, CSS, JavaScript e conceitos específicos do framework. |
| Customização da Interface | Limitada. A interface utiliza componentes visuais pré-definidos e limpos. | Total. Controle completo e autônomo sobre o código visual e de interação. |
| Escalabilidade | Limitada. Menos indicada para sistemas concorrentes com muitos usuários simultâneos. | Alta. Estruturada para sistemas robustos, com suporte a sessões e bancos de dados complexos. |
| Caso de Uso Ideal | Prototipagem rápida, dashboards analíticos, validação de hipóteses e modelos de ML. | Sistemas web completos, APIs REST, lojas virtuais e plataformas empresariais. |

**Fonte:** Autor, 2025.

Dessa forma, o uso do Streamlit permitiu que a maior parte do esforço do projeto fosse direcionada à implementação e ajuste dos algoritmos de inteligência artificial, à mineração de textos e à calibração das métricas, que representam o verdadeiro foco desta pesquisa.

## RESULTADOS E DISCUSSÃO

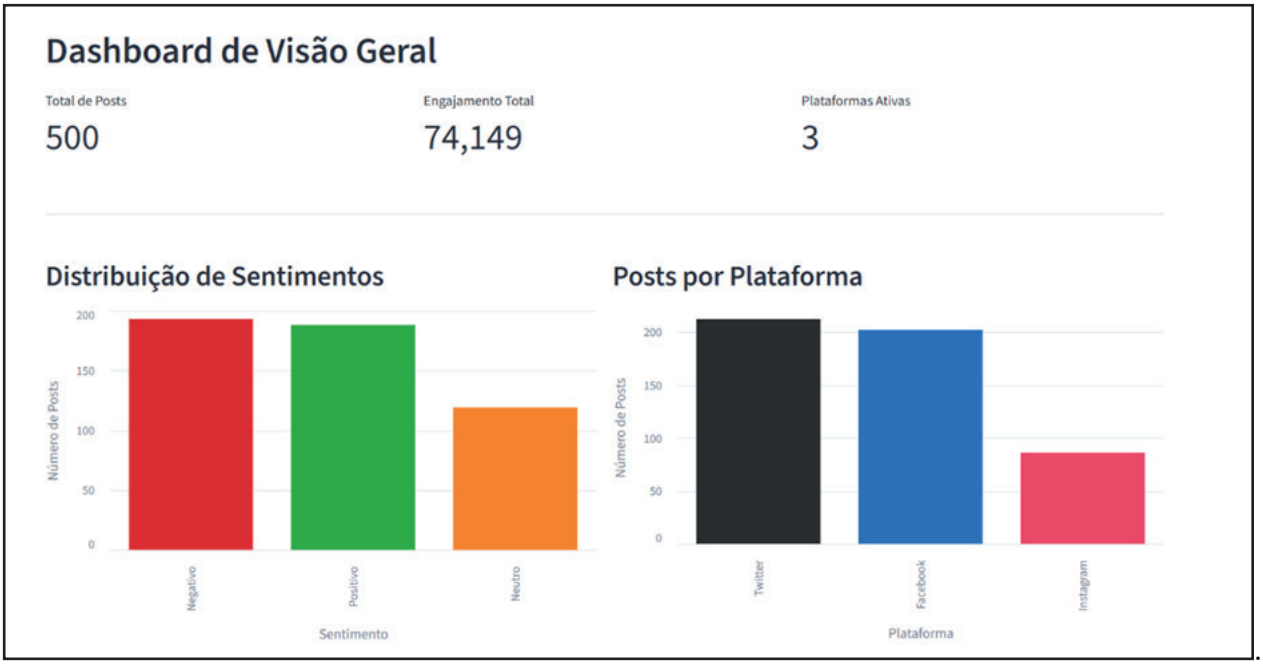
A plataforma foi estruturada por meio de módulos integrados, responsáveis desde a limpeza inicial dos dados até a geração automática de relatórios textuais. Os resultados obtidos em cada uma dessas etapas operacionais são discutidos a seguir.

### Dashboard de Visão Geral e Análise Temporal

O fluxo inicial do sistema requer o carregamento de um arquivo CSV pelo usuário. Em segundo plano, uma etapa de pré-processamento converte os campos de data para o formato reconhecido pelo sistema, padroniza as informações em texto minúsculo e cria a coluna de "Engajamento", que resulta da soma direta de curtidas e comentários de cada postagem.

O Dashboard de Visão Geral (Figura 1) apresenta a síntese inicial das interações. Os dados simulados apontam que as 500 postagens da empresa OmniTech geraram 74.149 interações no total. Os gráficos atestam que a base de treinamento possui uma divisão equilibrada de classes, essencial para a precisão dos algoritmos: 193 registros negativos, 188 registros positivos e 119 neutros. Observa-se, também, que o Twitter (212) e o Facebook (202) concentraram a maior parte das discussões escritas, enquanto o Instagram (86) agrupou interações de um perfil mais focado no engajamento visual das fotos.

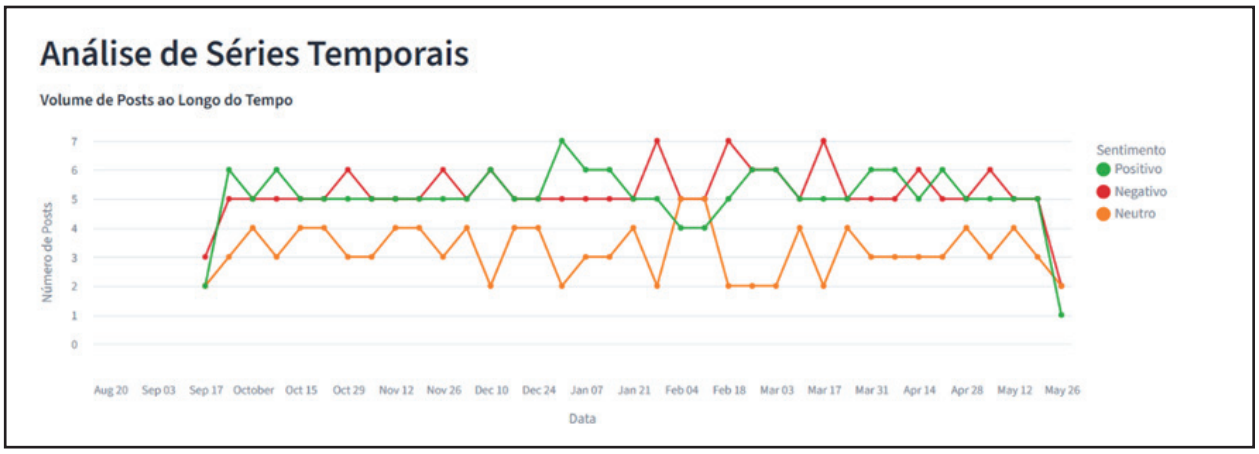
**Figura 1 -** Dashboard de Visão Geral dos Dados




**Fonte:** Autor, 2025.

Outra perspectiva importante do painel é a Análise de Séries Temporais (Figura 2), a qual permite observar o comportamento do consumidor de forma cronológica. O gráfico evidencia que as postagens de tom positivo e neutro se mantiveram relativamente lineares e contínuas ao longo do tempo. Por outro lado, a curva de menções negativas registrou picos repentinos em datas específicas. Para um gestor de marca, identificar essas variações agudas é estratégico, pois elas geralmente sinalizam crises pontuais, como problemas no lote de um novo lançamento ou instabilidades no sistema de compras da loja virtual.

**Figura 2 -** Análise de Séries Temporais




**Fonte:** Autor, 2025.

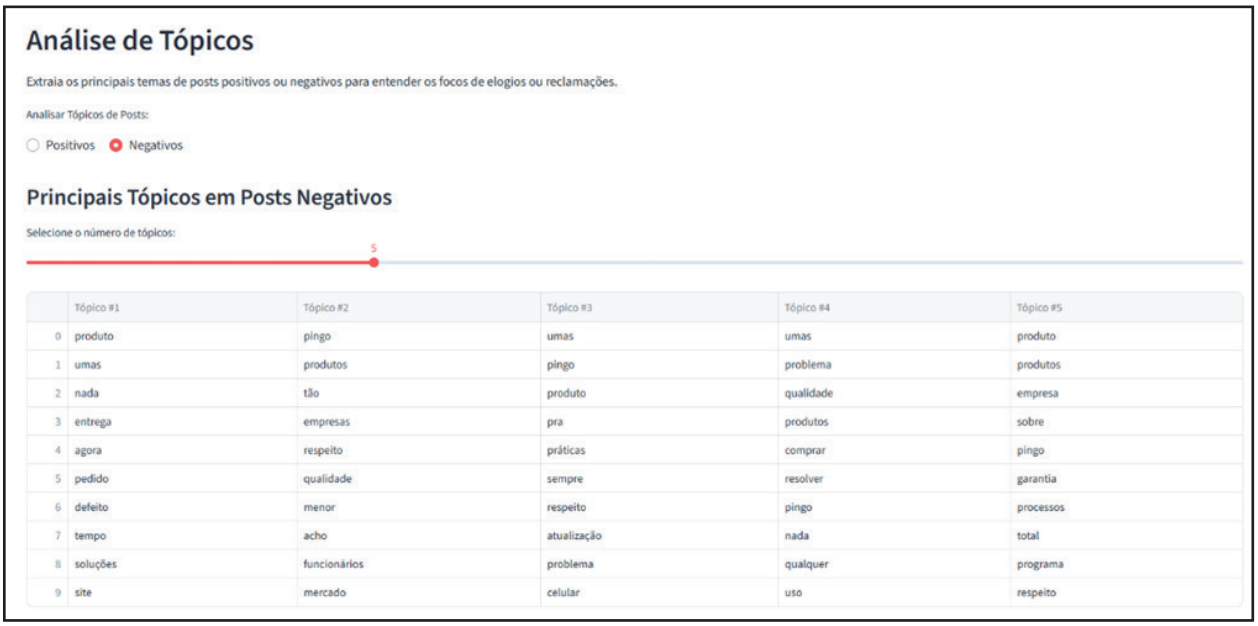
## Processamento de Linguagem: Tópicos e Bigramas

Para compreender melhor os motivos que geram as publicações, a plataforma conta com uma etapa de mineração textual que agrupa os assuntos em temas principais. A técnica escolhida para isso foi a Modelagem de Tópicos, baseada no algoritmo *Latent Dirichlet Allocation*

(LDA). Em linhas gerais, este método estatístico identifica palavras que costumam aparecer juntas nos mesmos contextos e agrupa as postagens em categorias latentes.

Quando o algoritmo avaliou exclusivamente as mensagens com sentimento negativo (Figura 3), os grupos formados trouxeram termos centrais como “produto”, “entrega”, “pedido”, “defeito” e “site”. A união dessas palavras indica claramente que a maior fonte de insatisfação está vinculada a falhas na logística de entrega e instabilidades no site oficial da empresa. Em contrapartida, ao rodar a análise sobre as mensagens positivas, o algoritmo agrupou termos como “marca”, “qualidade”, “design”, “tecnologia” e “experiência”. Isso sugere que os consumidores reconhecem e valorizam a identidade da marca e o diferencial estético e tecnológico dos produtos, fatores que atuam como principais forças de retenção do cliente.

**Figura 3 -** Análise de Tópicos Latentes para Posts Negativos

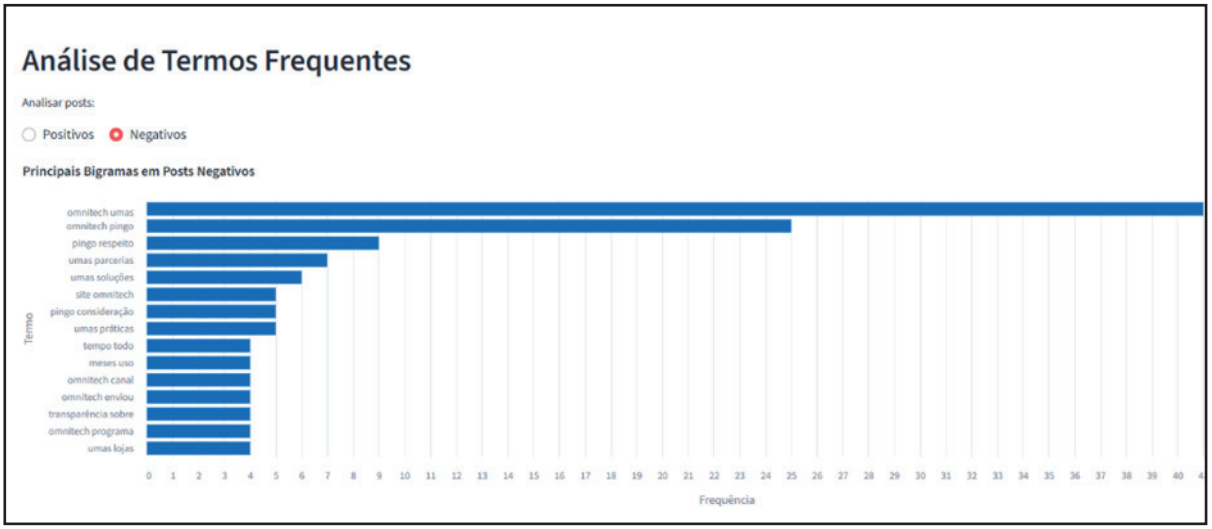



| | Tópico #1 | Tópico #2 | Tópico #3 | Tópico #4 | Tópico #5 |
|---|---|---|---|---|---|
| 0 | produto | pingo | umas | umas | produto |
| 1 | umas | produtos | pingo | problema | produtos |
| 2 | nada | tão | produto | qualidade | empresa |
| 3 | entrega | empresas | pra | produtos | sobre |
| 4 | agora | respeito | práticas | comprar | pingo |
| 5 | pedido | qualidade | sempre | resolver | garantia |
| 6 | defeito | menor | respeito | pingo | processos |
| 7 | tempo | acho | atualização | nada | total |
| 8 | soluções | funcionários | problema | qualquer | programa |
| 9 | site | mercado | celular | uso | respeito |

**Fonte:** Autor, 2025.

Para detalhar ainda mais essas informações, implementou-se a extração de Termos Frequentes, focada especificamente nos Bigramas (expressões formadas por duas palavras que aparecem lado a lado). Essa técnica complementa o LDA, trazendo as frases reais digitadas pelas pessoas. A Figura 4 ilustra os resultados dessa extração para os dados negativos, evidenciando a insatisfação e a fadiga do consumidor com o suporte por meio de pares recorrentes como “pingo respeito”, “tempo todo” e “meses uso”.

**Figura 4 -** Análise de Termos Frequentes em Posts Negativos




**Fonte:** Autor, 2025.

De forma complementar, o processamento dos textos positivos revelou expressões de confiança e forte associação com o valor da marca. Foram consolidados bigramas que transmitem emoções claras, a exemplo de "produtos omnitech", "marca empresa" e "faz sentir". Essas métricas são importantes pois complementam os tópicos abstratos do LDA, mostrando a linguagem exata e o tom de voz do cliente, o que é um dado fundamental para treinar sistemas de atendimento e roteiros de suporte.

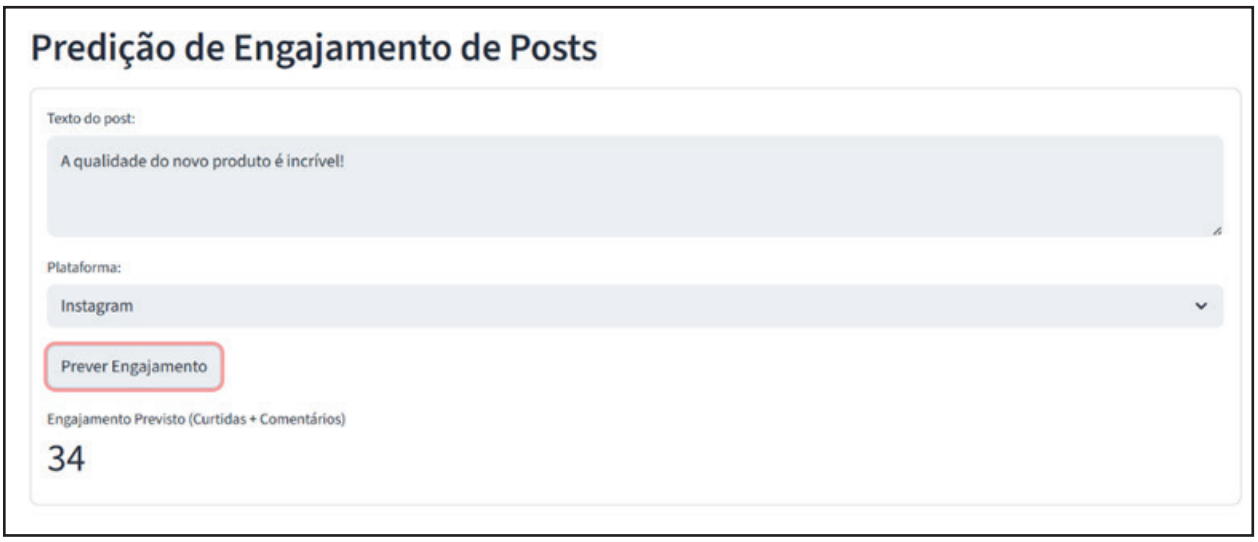
## Predição Avançada com LightGBM

Além da análise descritiva, a plataforma também conta com um módulo preditivo de engajamento, cujo papel é estimar o volume de curtidas e comentários que um texto pode receber antes mesmo de ser publicado. Para essa tarefa, utilizou-se o algoritmo LightGBM (Ke et al., 2017), um método de aprendizado de máquina focado em árvores de decisão. A escolha por essa biblioteca ocorreu devido ao seu excelente desempenho e capacidade de operar rapidamente com grandes quantidades de variáveis originadas da transformação de textos em dados matemáticos.

O LightGBM opera com alta eficiência graças a duas técnicas específicas: o método *Gradient-based One-Side Sampling* (GOSS), que foca o treinamento nos dados mais difíceis (onde o algoritmo comete mais erros) ignorando instâncias fáceis e redundantes; e a técnica *Exclusive Feature Bundling* (EFB), que agrupa variáveis semelhantes para não sobrecarregar a memória do computador. Além disso, o algoritmo constrói suas árvores de decisão expandindo as ramificações que apresentam o maior ganho de acerto de forma contínua, uma estratégia que garante melhores resultados mesmo quando o banco de textos é muito ruidoso e mal redigido.

Na prática, esse módulo demonstrou utilidade ao simular a publicação da seguinte frase teste: "A qualidade do novo produto é incrível!". Ao preencher o formulário selecionando a plataforma Twitter, o sistema utilizou seu aprendizado para estimar a marca de 26 interações. Ao alterar a seleção para o Facebook, a previsão subiu para 31 interações. Por fim, ao definir o Instagram como o canal de distribuição (Figura 5), o engajamento esperado atingiu o pico de 34 interações. Essa simulação ressalta como diferentes redes sociais exigem abordagens de linguagem e estratégias de postagem distintas.

**Figura 5 -** Projeção de Engajamento para o Instagram




**Fonte:** Autor, 2025.

## Síntese Executiva Automatizada e Explorador Textual

Como grandes volumes de informações demandam muito tempo de leitura, tornou-se necessário integrar uma ferramenta capaz de resumir os textos classificados de forma automática. Em ambiente acadêmico, o uso frequente de APIs de inteligência artificial comerciais (como o modelo GPT da OpenAI) é muitas vezes inviável devido aos altos custos. Portanto, a solução adotada foi o uso da API do modelo Google Gemini 1.5 Flash (Google, 2023). Essa versão específica foi desenvolvida para tarefas rápidas que exigem processamento de grandes blocos de texto sem gerar atrasos excessivos no sistema.

Por meio de uma instrução detalhada no código (*Prompt Engineering*), o modelo de IA foi orientado a assumir o papel de um Analista de Relações Públicas, avaliando todos os comentários e criando um relatório executivo com os principais problemas enfrentados pelos usuários. A Figura 6 demonstra o resultado: o sistema estruturou um sumário claro apontando quatro áreas de crise. Foram listadas deficiências no atendimento ao cliente (esperas longas no telefone), problemas de fabricação (pixels queimados em monitores e atrasos na entrega), estratégias de publicidade pouco transparentes e falhas técnicas de acesso no aplicativo oficial da empresa.

**Figura 6 -** Resumo Analítico Automatizado para Interações Negativas

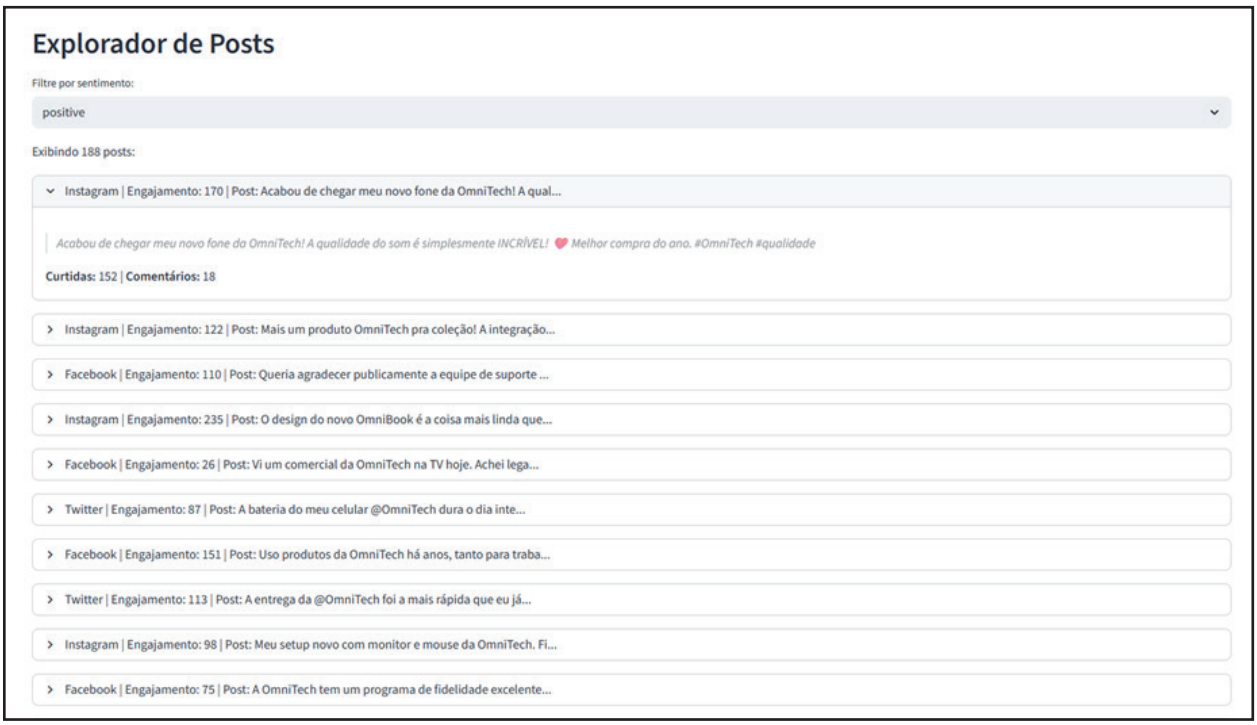



**Fonte:** Autor, 2025.

Quando aplicada à listagem de publicações elogiosas, a mesma ferramenta produziu um texto destacando os principais fatores de fidelização dos clientes. A IA identificou que o design físico dos produtos, a experiência de abertura das caixas (unboxing) e as ações voltadas para a sustentabilidade e reciclagem são os pilares que geram avaliações mais calorosas por parte do consumidor, alinhando as conclusões ao que o LDA havia identificado genericamente como o atributo de "marca" e "experiência".

Para finalizar o ciclo de análise, a plataforma dispõe do módulo Explorador de Posts (Figura 7). Nele, o gestor da marca pode filtrar as mensagens diretamente por sentimento e fazer a leitura do texto original, sem a intervenção dos algoritmos matemáticos ou da inteligência artificial generativa. Essa etapa é crucial para a verificação manual dos dados, conectando a abstração dos relatórios gerados automaticamente aos relatos genuínos deixados pelos clientes nas plataformas digitais.

**Figura 7 -** Explorador de Posts da Interface




**Fonte:** Autor, 2025.

## CONCLUSÃO

Este trabalho demonstrou a viabilidade e a utilidade do desenvolvimento de uma plataforma prática voltada para a análise de dados em Social CRM. Adotando a metodologia de pesquisa focada na criação de soluções de design (DSR) em conjunto com a flexibilidade da linguagem Python e do framework Streamlit, confirmou-se que é plenamente possível construir ferramentas de inteligência analítica eficientes, superando as limitações técnicas e financeiras impostas por softwares comerciais restritos.

A integração de métodos como o LDA para a identificação latente de tópicos, a extração detalhada de bigramas para análise de contexto e as predições de alcance baseadas em árvores de decisão no LightGBM formam um painel de indicadores muito robusto. Quando aliada aos relatórios sintéticos gerados pela API do modelo Gemini 1.5 Flash, essa combinação de técnicas supera amplamente os painéis puramente quantitativos convencionais, transformando um mar de textos desconexos em inteligência corporativa acionável.

Cabe ressaltar uma limitação importante deste estudo experimental: a utilização de um dataset criado de forma 100% sintética. Embora essa medida tenha preservado a privacidade de usuários reais e criado um ambiente ideal (balanceado) para aferir a eficácia matemática dos algoritmos preditivos, o banco de dados controlado não contempla erros graves de ortografia, o sarcasmo implícito, as gírias voláteis e o caos generalizado típicos do ambiente orgânico das mídias sociais. Futuros testes conduzidos com fluxos reais de clientes serão fundamentais para testar a margem de erro final do sistema em ambientes de produção.

Como sugestão para trabalhos e implementações futuras, é recomendado a inclusão de técnicas baseadas em Análise de Sentimentos por Aspecto (ABSA). Em uma postagem comum de um cliente, que afirme, por exemplo, que “a tela do novo monitor tem cores belíssimas, mas o suporte plástico quebrou no primeiro dia”, a abordagem global tradicional acaba classificando o texto como “neutro” ou pendendo para um único lado, mascarando o real problema. Com a tecnologia ABSA, o algoritmo se torna capaz de isolar os elementos gramaticais, atribuindo um viés positivo unicamente à “tela” e uma crítica severa ao “suporte”, garantindo um diagnóstico muito mais preciso e direcionado para as correções operacionais e fabris da empresa.

## REFERÊNCIAS

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. Brasília, DF: Presidência da República, 2018.

ERDIN, C.; MEHMET, C. **National innovation efficiency: a DEA-based measurement of OECD countries**. *International Journal of Innovation Science*, v. 15, n. 3, p. 427-456, 2023.

GOOGLE. **Gemini: a family of highly capable multimodal models**. Mountain View: Google AI, 2023. (Technical Report).

KE, G. et al. **LightGBM: a highly efficient gradient boosting decision tree**. *Advances in Neural Information Processing Systems*, v. 30, 2017.

PEFFERS, K. et al. **A design science research methodology for information systems research**. *Journal of Management Information Systems*, v. 24, n. 3, p. 45-77, 2007.

RAHMAN, Md. S.; REZA, H. **A systematic review towards big data analytics in social media**. *Big Data Mining and Analytics*, v. 5, n. 3, p. 228-244, 2023.

RATHORE, B. **Digital transformation 4.0: integration of artificial intelligence & metaverse in marketing**. *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal*, v. 12, n. 1, p. 42-48, 2023.

SMITH, T.; WANG, H.; LI, C. **Web application development with Streamlit**. [S. l.]: Springer, 2024.

STREAMLIT. **Streamlit: facilitando a criação de aplicações web com Python**. [S. l.: s. n.], 2024.

WITTWER, M.; REINHOLD, O.; ALT, R. **Social media analytics in social CRM: towards a research agenda**. *In*: BLED eCONFERENCE, 29., 2016, Bled. *Proceedings* [...]. Bled: [s. n.], 2016.