



Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
28/04/2023	Camila Anacleto, Eduardo França Porto , Gabriel Rocha Pinto Santos, Matheus Fidelis dos Santos Pinto, Rodrigo Campos e Sophia Dias.	0.2	Artefatos de negócios e UX
12/05/2023	Camila Anacleto, Eduardo França Porto , Gabriel Rocha Pinto Santos, Matheus Fidelis dos Santos Pinto, Rodrigo Campos e Sophia Mello Dias.	0.4	Pré processamento do corpus, aplicação do Bag of Words e análise dos resultados
26/05/2023	Camila Anacleto, Eduardo França Porto , Gabriel Rocha Pinto Santos, Matheus Fidelis dos Santos Pinto, Rodrigo Campos e Sophia Mello Dias.	0.6	Aplicação do modelo word2vec e análise dos resultados
09/06/2023	Camila Anacleto, Eduardo França Porto , Gabriel Rocha Pinto Santos, Matheus Fidelis dos Santos Pinto, Rodrigo Campos e Sophia Mello Dias.	0.8	Proposta de novas arquiteturas de modelos e modelagem de novas features.
21/06/2023	Camila Anacleto, Eduardo França Porto , Gabriel Rocha Pinto Santos, Matheus Fidelis dos Santos Pinto, Rodrigo Campos e Sophia Mello Dias.	1.0	Criação do diagrama UML, criação do frontend, deploy do melhor modelo e construção de api.

Sumário

1. Introdução	4
2. Objetivos e Justificativa	5
2.1. Objetivos	5
2.2. Justificativa	6
3. Análise do Negócio	7
3.1. Contexto da indústria	7
3.2. Ferramentas	8
4. Análise de Experiência do Usuário	21
4.1. Personas	21
4.2. User Story	23
4.3. Protótipo de interface com o usuário	30
5. Solução Proposta	32
5.1. Solução	32
5.2. Arquitetura Proposta	33
5.3. Diagrama UML	33
5.4. Pipeline de Pré processamento	34
6. Modelagem	36
6.1. Modelos aplicados	36
6.2 Comparações	46
6.3. Modelo final	49
7. Serviço	50
8. Referências	51
Anexos	

1. Introdução

1.1. Parceiro de Negócios

O BTG Pactual é o maior banco de investimentos da América Latina, abrangendo diversos mercados, como Investment Banking, Corporate Lending, Sales & Trading, Wealth Management e Asset Management. Fundado em 1983, o banco adota uma cultura meritocrática baseada em parcerias, com foco no cliente, excelência e visão de longo prazo.

Ao longo dos anos, o BTG Pactual se destacou como uma das empresas mais inovadoras do setor financeiro, conquistando inúmeros prêmios tanto nacional quanto internacionalmente. Atualmente, possui quase 3 mil colaboradores distribuídos em escritórios localizados no Brasil, Chile, Argentina, Colômbia, Peru, México, Estados Unidos, Portugal e Inglaterra.

Com sua abordagem diferenciada e compromisso com a excelência, o BTG Pactual consolidou sua posição como líder no mercado de investimentos da América Latina, oferecendo soluções financeiras sofisticadas e atendendo às necessidades de seus clientes em todo o mundo.

1.2. Definição do Problema

Com a ampla utilização das redes sociais por mais da metade da população mundial, as empresas estão investindo cada vez mais em estratégias de marketing digital para alcançar seus públicos-alvo e impulsionar o engajamento dos clientes, visto que essa área é crucial para o crescimento e desenvolvimento de todas as demais, desde vendas à tecnologia.

Considerando a quantidade de dados gerados pelas redes sociais, juntamente com o aumento do investimento em marketing digital, surge a necessidade de análise desses dados de mídia social. A análise dessas informações pode fornecer insights valiosos para ajudar as empresas a tomar decisões de negócios mais assertivas, otimizar suas estratégias de marketing e refinar suas abordagens à medida que avançam.

Em suma, o problema consiste em aproveitar a riqueza de dados disponíveis nas redes sociais e utilizar a análise de mídia social como uma ferramenta para entender o desempenho das estratégias de marketing e obter informações relevantes que permitam tomar decisões mais informadas e eficazes.

2. Objetivos e Justificativa

2.1. Objetivos

Concernem como os objetivos desse projeto:

1. Identificar o sentimento dos usuários: O objetivo é aplicar técnicas de PLN para analisar o tom emocional dos comentários dos usuários e identificar se eles expressam sentimentos positivos, neutros ou negativos em relação às campanhas. Isso permite avaliar a eficácia das estratégias de marketing e realizar ajustes necessários.
2. Responder rapidamente às interações dos usuários: A ideia é utilizar a análise de PLN para rastrear e monitorar os comentários dos usuários em tempo real. Dessa forma, a área de negócios pode responder prontamente às interações dos usuários, seja para fornecer suporte, responder a comentários negativos ou demonstrar apreço pelos comentários positivos.
3. Identificar palavras-chave e interesses dos consumidores: O objetivo é utilizar técnicas de PLN para extrair palavras-chave relevantes dos comentários dos usuários. Isso ajuda a compreender os interesses, preferências e necessidades dos consumidores em relação à marca, possibilitando direcionar futuras campanhas ou produtos com base nesses insights.
4. Refinar e otimizar as estratégias de marketing: Ao analisar a receptividade dos usuários e identificar palavras-chave, a área de negócios pode refinar e otimizar suas estratégias de marketing. Esses dados ajudam a compreender o que está funcionando e o que não está, permitindo ajustes adequados para maximizar o impacto e o sucesso das campanhas.
5. Aumentar a personalização e relevância das mensagens: Utilizando a análise de PLN, é possível personalizar as mensagens de marketing com base nas preferências e interesses específicos dos consumidores. Isso contribui para aumentar a relevância das campanhas, melhorar o engajamento dos usuários e fortalecer a conexão emocional com a marca.

2.2. Justificativa

A utilização de Processamento de Linguagem Natural (PLN) oferece uma vantagem significativa para a área de negócios ao rastrear dados em tempo real e analisar a receptividade dos usuários às campanhas realizadas nas redes sociais. Essa abordagem permite uma resposta rápida e eficaz às interações dos usuários, possibilitando ajustes e melhorias imediatas na estratégia de marketing.

Em resumo, a aplicação do PLN para análise de sentimento e identificação de palavras-chave nos comentários dos usuários nas campanhas de marketing em redes sociais permite que a área de negócios compreenda e responda rapidamente à receptividade dos consumidores, bem como a direcionar futuras campanhas com base em seus interesses e preferências. Essa abordagem contribui para aumentar a eficácia, personalização e relevância das estratégias de marketing, impulsionando o engajamento dos usuários e o sucesso das campanhas.

3. Análise do Negócio

3.1. Contexto da indústria

A indústria em que o BTG Pactual atua, o setor financeiro, é altamente competitiva e diversificada. Além do BTG Pactual, existem outros players importantes no mercado de investimentos da América Latina, como bancos de investimento globais, corretoras de valores, gestoras de ativos e instituições financeiras tradicionais.

Os bancos de investimento globais, como J.P. Morgan, Goldman Sachs e Morgan Stanley, possuem uma presença significativa na região e oferecem uma ampla gama de serviços de investment banking, asset management e wealth management. Essas instituições têm uma rede global estabelecida, com recursos e expertise para atender clientes corporativos e institucionais em todo o mundo.

No mercado local, além do BTG Pactual, há outras instituições financeiras regionais que competem em diferentes segmentos. Algumas delas têm uma abordagem mais especializada em determinadas áreas, como bancos focados em corporate lending ou asset management. Esses concorrentes locais podem ter uma presença mais forte em seus respectivos mercados domésticos.

A competição no setor financeiro também é impulsionada por corretoras de valores, que oferecem serviços de negociação de ações, renda fixa, câmbio e outros ativos financeiros. Essas corretoras podem atender tanto investidores institucionais quanto indivíduos, oferecendo acesso aos mercados financeiros e serviços de assessoria financeira.

Além disso, o cenário da indústria financeira está evoluindo com a entrada de empresas de tecnologia financeira, as chamadas fintechs. Essas empresas estão trazendo inovação e disruptão ao setor, oferecendo soluções financeiras digitais, plataformas de investimento online e serviços financeiros mais acessíveis e personalizados.

Diante desse cenário competitivo, o BTG Pactual se destaca por sua posição como o maior banco de investimentos da América Latina, oferecendo uma ampla gama de serviços financeiros e soluções sofisticadas para seus clientes. A reputação do banco, seu compromisso com a excelência e sua presença global são fatores que o diferenciam dos concorrentes e fortalecem sua posição no mercado.

Para se manter competitivo, o BTG Pactual precisa estar atento às tendências e inovações do setor, buscando constantemente aprimorar seus serviços e oferecer soluções financeiras inovadoras. A capacidade de adaptação às mudanças do mercado e a busca contínua por excelência são elementos essenciais para manter-se relevante em um setor dinâmico e competitivo como o financeiro.

3.2. Ferramentas

3.2.1. Matriz de Avaliação de Valor (Oceano Azul)

A matriz é utilizada para visualizar, a partir de demais players do mercado, possibilidades a serem exploradas e que podem criar vantagem competitiva. Dessa forma, foram analisados os seguintes aspectos: praticidade, métricas de análise de sentimento, métricas de análise de palavras chave, métricas comparativas com outras contas, custo, métricas comparativas com seus próprios posts, preço e dependência de fornecedor. Tais players foram analisados em comparação com o AdTrack: mLabs, Instagram Analytics e Iconsquare.

Matriz de avaliação de valor Oceano Azul

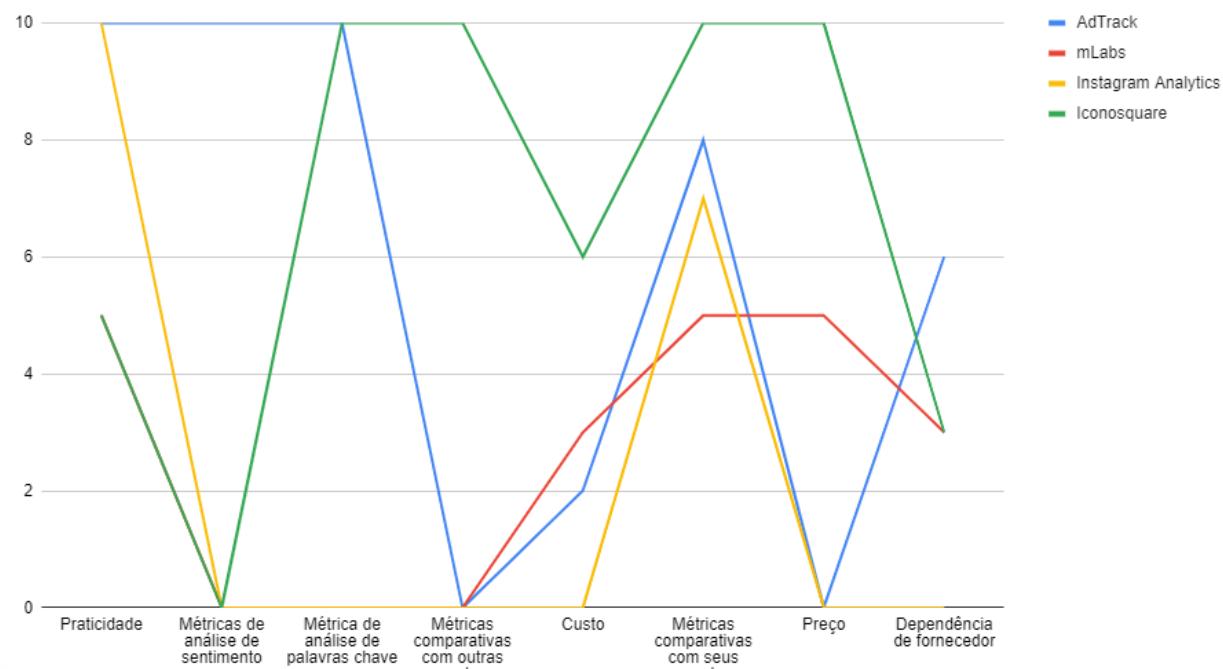


Imagen 1 - Matriz de Avaliação de Valor

Atributos	AdTrack	mLabs	Instagram Analytics	Iconosquare
Praticidade	10	5	10	5
Métricas de análise de sentimento	10	0	0	0
Métrica de análise de palavras chave	10	0	0	10
Métricas comparativas	0	0	0	10

com outras contas				
Custo	2	3	0	6
Métricas comparativas com seus posts	8	5	7	10
Preço	0	5	0	10
Dependência de fornecedor	6	3	0	3

Imagen 2- Tabela de Avaliação de Valor

Portanto, é possível concluir tais aspectos:

1. Por ser um sistema utilizado internamente pela equipe da empresa, ele é de mais fácil acesso, se tornando mais prático e, por isso, AUMENTOU tal atributo. O Instagram Analytics, por ser interno do Instagram, também é prático, diferente de serviços terceiros que requer a criação de conta, pagamento, conexão com conta do Instagram e Facebook, como é o caso do mLabs e Iconosquare.
2. Diferente dos competidores, o AdTrack criou métricas de análise de sentimentos dos comentários feitos nas postagens, mostrando quais sentimentos cada campanha gerou no seguidores. Portanto, ao CRIAR, o AdTrack tem vantagem competitiva, gerando mais valor.
3. O AdTrack CRIOU tal atributo, visto que os demais serviços não fornecem a análise de palavras chaves extraídas dos comentários feitos nas publicações.
4. O AdTrack não possui métricas comparativas das campanhas de marketing com demais contas, sendo o Iconosquare a única que possui tal serviço, por isso o AdTrack ELIMINOU esse atributo, por não ter sido requisitado pelo cliente.

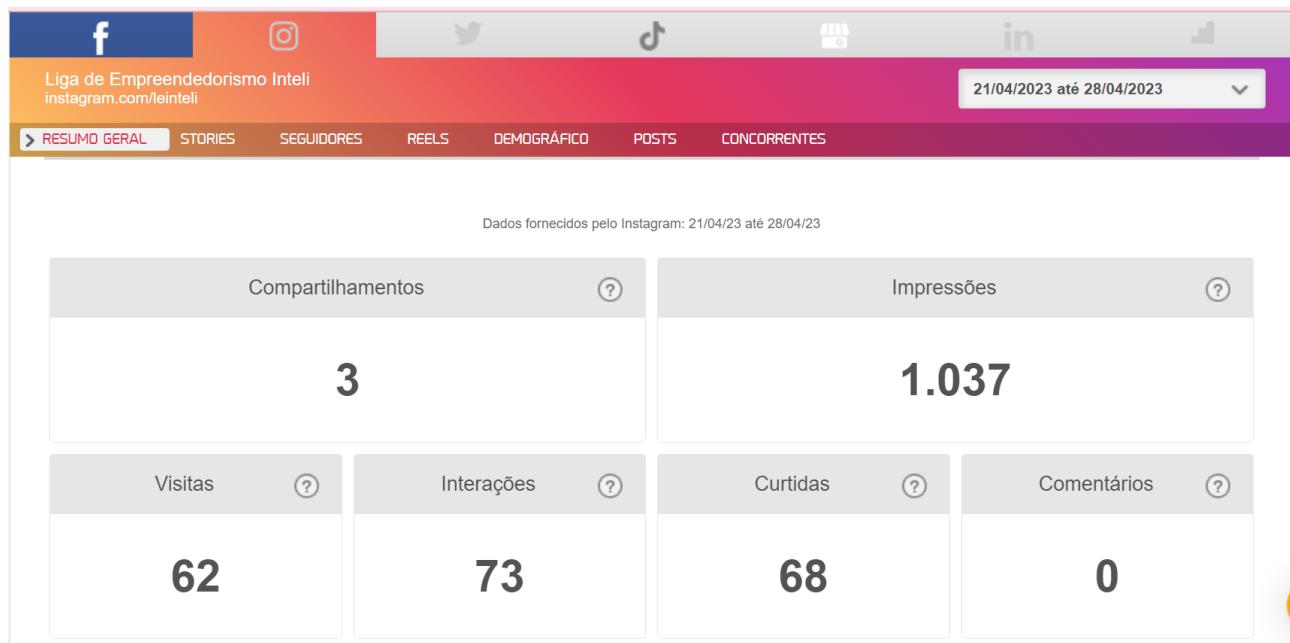
5. O Instagram Analytics é a própria ferramenta de análise do Instagram, portanto não possui custo. Já as plataformas Iconosquare e mLabs são serviços de terceiros, que possuem mensalidade de 49 euros e 29,90 reais, respectivamente. O AdTrack, por ser um serviço interno, possui custos de manutenção e utilização, o que representa ser maior que os demais, assim sendo um aspecto que deve ser REDUZIDO.

6. A ferramenta nativa do Instagram possui métricas de comparativo de posts. O mLabs fornece as mesmas ferramentas do Instagram Analytics, portanto possuem a mesma pontuação. O Iconosquare tem diferentes métricas mais detalhadas, sobre a quantidade de likes a cada 30 minutos, por exemplo. O AdTrack pode seguir algumas métricas da Iconosquare e oferecer, em tempo real, AUMENTANDO-AS, mais detalhadas sobre a postagem que o próprio Instagram não oferece.

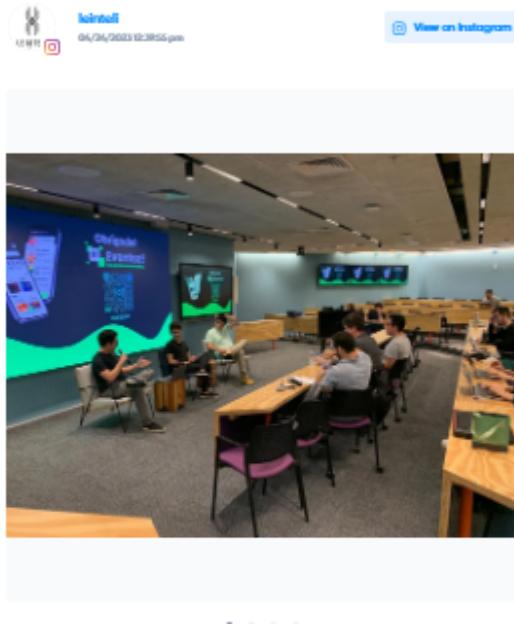
7. Visto que os serviços terceiros cobram preços mensais para o uso, o AdTrack ELIMINOU tal atributo, já que ele será nativo do sistema da empresa, não tendo que pagar um preço para uma outra empresa.

8. Para diminuir os custos e melhorar a eficiência do produto, o AdTrack pode ELIMINAR a dependência de fornecedor, ou seja, de base de dados do seu administrador, com a possibilidade de aumentar os dados e permitindo criar novas métricas comparativas.

A seguir é possível ver mais cada serviço que foi utilizado na matriz para comparar com a solução a ser desenvolvida. Foram criadas contas no mLabs e Iconosquare para realizar a pesquisa:

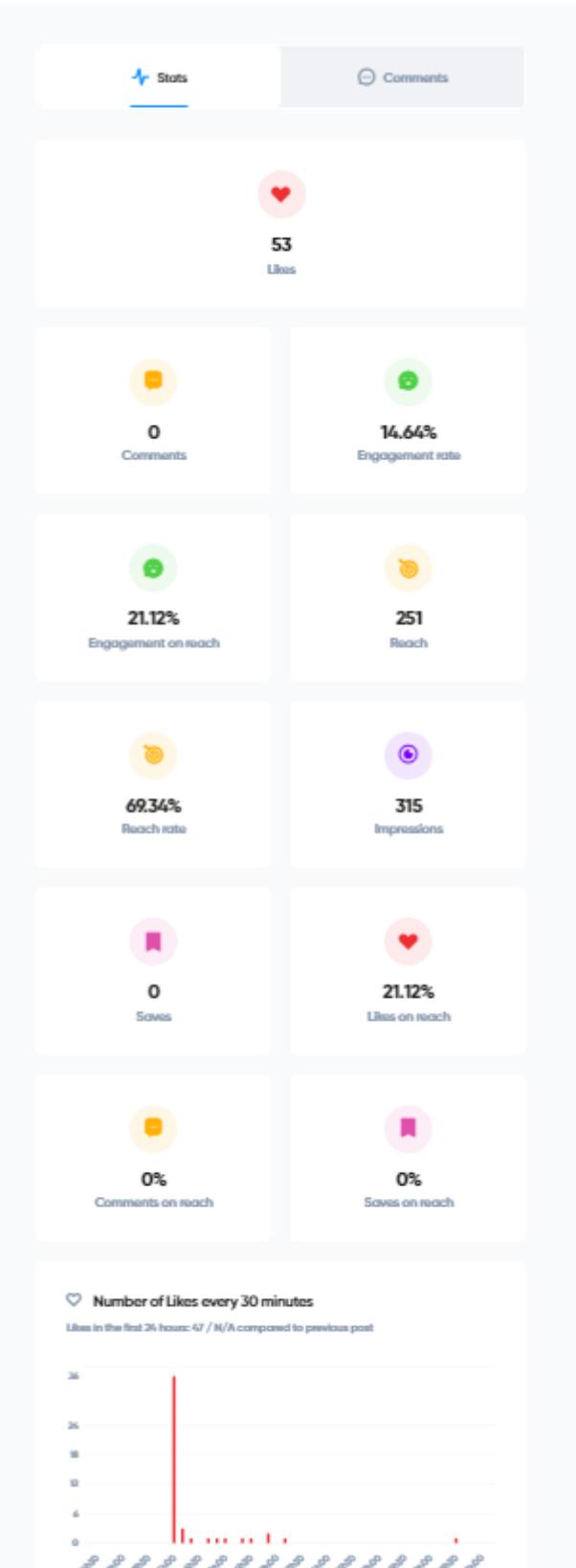


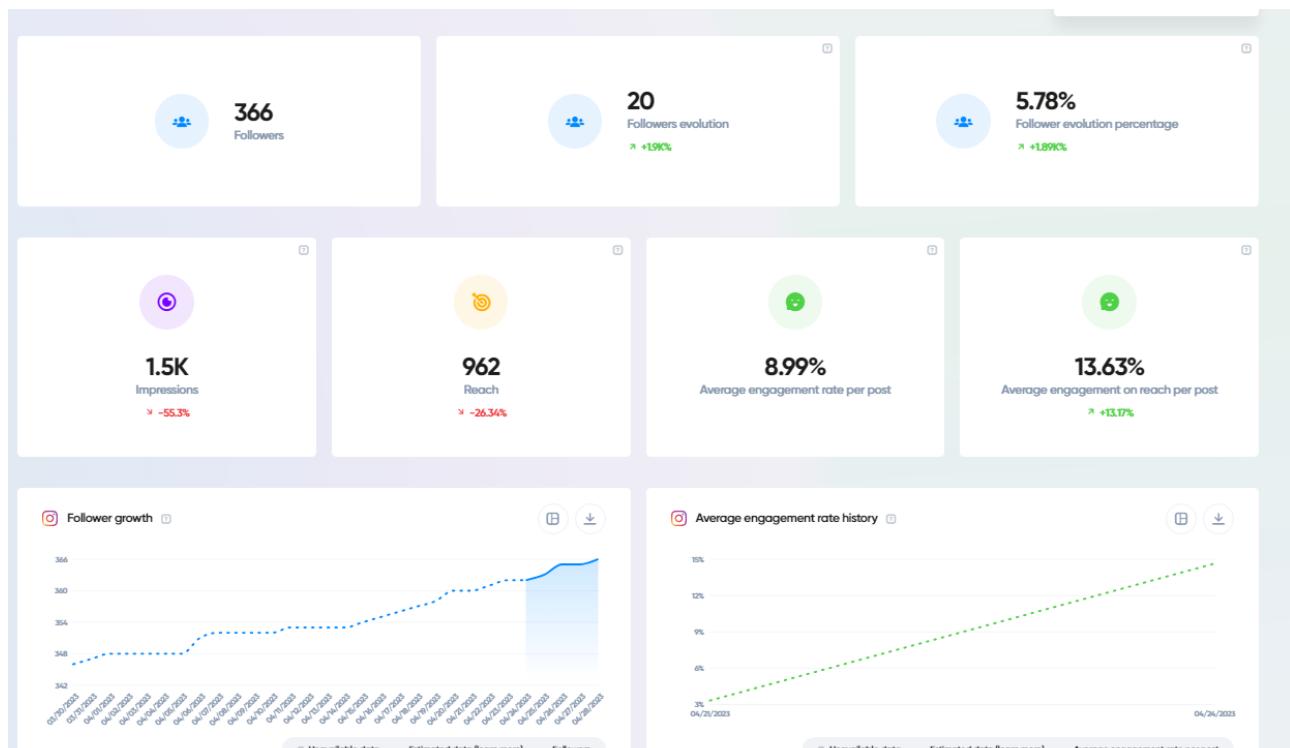
Imagens 3 e 4 - Relatórios mLabs



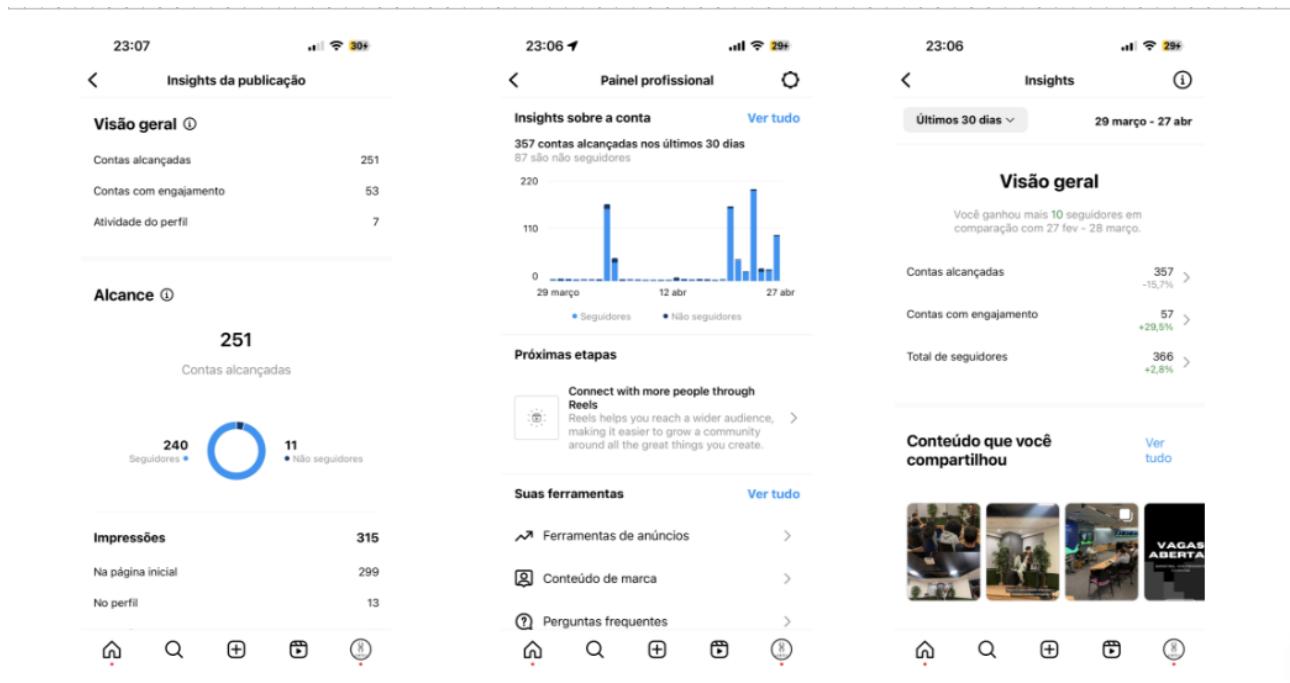
No mês passado, o LEI realizou um pitch deck com duas das startups fundadas durante os steps de 2022, a Eventse e a Zerol. As empresas apresentaram seus – [view more](#) [translate](#)

Albums & Labels





Imagens 5 e 6 - Relatórios Iconosquare



Imagens 7 e 8 - Relatórios Instagram Analytics

3.2.2. Matriz de Risco

A matriz de riscos é uma ferramenta que permite uma análise ampla das ameaças e oportunidades de um projeto. Com ela, podemos identificar as ameaças com maiores probabilidades e impactos em nosso projeto, além de identificar as oportunidades existentes no desenvolvimento. Sua principal função é auxiliar a empresa a tomar decisões com base nos impactos e na probabilidade desses riscos ocorrerem. A seguir, apresentamos a lista de todos os riscos (ameaças e oportunidades) validados pela equipe:

		Ameaças			Oportunidades			
		Ameaça 01: Acurácia pouco satisfatória, levando em consideração a subjetividade dos sentimentos humanos nos comentários	Ameaça 02: Tempo de resposta do modelo oscilado: considerando que o modelo utilizará dados em tempo real, o tempo de resposta do modelo pode variar e causar espera e insatisfação do usuário	Ameaça 03: Há um risco de modelo preditivo de classificação de sentimentos falhar e colocar os comentários dos clientes do BTG Pactual em uma categoria diferente daquela que realmente pertencem.	Oportunidade 01: Mercado em crescimento: A análise de sentimento é uma área em crescimento, à medida que as empresas buscam entender melhor as opiniões e sentimentos de seus clientes nas redes sociais	Oportunidade 02: Melhoria contínua: A análise de sentimento é uma área em constante evolução, com novas técnicas e tecnologias surgindo regularmente, o que possibilita a maior precisão dos resultados e a eficiência do produto	Oportunidade 03: Criação de Novos Produtos: Com base nos dados coletados pela sua solução de análise de sentimento, é possível identificar necessidades não atendidas pelos produtos existentes no mercado	
100%	Ameaça 03: Falta de precisão: A precisão da análise de sentimento pode ser afetada pela qualidade dos dados, bem como pela eficácia dos algoritmos de análise. Se a precisão da solução for baixa, a confiança dos clientes pode ser afetada.	Ameaça 04: Expectativas mal definidas: Se as expectativas forem claramente definidas, pode haver mal-entendidos sobre o escopo do projeto, o tempo de entrega e o que é esperado do produto final.	Ameaça 05: Interfaces mal projetadas: O design da interface do usuário é crucial para garantir que os usuários possam interagir com o projeto de maneira intuitiva e eficaz. Interfaces mal projetadas podem dificultar o uso do projeto.	Oportunidade 03: Criar novas formas de utilizar o produto, dando seguimento ao projeto e criando um produto mais robusto, pode até funcionar como um SaaS ou uma vantagem competitiva de mercado	Oportunidade 04: Utilização de tecnologias de ponta: O seu projeto pode utilizar tecnologias de ponta, como inteligência artificial e aprendizado de máquina, para fornecer resultados precisos de análise de sentimento.	Oportunidade 05: Criação de Novos Produtos: Com base nos dados coletados pela sua solução de análise de sentimento, é possível identificar necessidades não atendidas pelos produtos existentes no mercado		
75%	Ameaça 06: Dependência de bibliotecas de terceiros: O projeto de análise de sentimento pode depender de bibliotecas de terceiros para funcionar. Se essas bibliotecas estiverem desatualizadas ou forem descontinuadas, isso pode afetar a funcionalidade do projeto, entretanto o problema pode ser conformatado com outras		Ameaça 07: Mudanças nos requisitos: Se houver muitas mudanças nos requisitos durante o projeto, pode ser difícil finalizá-lo dentro de prazo estabelecido.	Oportunidade 06: Obter insights dos produtos: é possível obter insights sobre os produtos e serviços existentes e entender os desejos do consumidor, disponibilizando novas ofertas.	Oportunidade 07: Possibilidade de colaboração: O projeto pode oferecer a possibilidade de colaboração com outras empresas e organizações que possam estar interessadas em análise de sentimento.	Oportunidade 08: Gerenciamento de Crises: A análise de sentimento pode ser utilizada para gerenciar crises de imagem de empresas, entendendo o que faz gerar sentimentos ruins e contornar a situação.	Oportunidade 09: Personalização: A sua solução de análise de sentimento pode ser personalizada para atender às necessidades específicas de diferentes setores, clientes e empresas do grupo parceiro.	
50%					Oportunidade 01b: Análise de concorrência: A análise de sentimento pode ser utilizada para entender como a concorrência está sendo percebida pelos consumidores.			
25%	Baixo	Moderado	Alto	Muito alto	Muito alto	Alto	Moderado	Baixo

Imagen 9 - Matriz de Riscos

Riscos:

1. Acurácia pouco satisfatória, levando em consideração a subjetividade dos sentimentos humanos nos comentários - o risco de acontecer é alto, pois é algo que acontece frequentemente nos algoritmos de processamento de linguagem natural, já que ele é extremamente influenciado pela língua (português) e moderado, já que existem outros riscos considerados que são mais altos;
2. Tempo de resposta de modelo oscilado: considerando que o modelo utilizará dados em tempo real, o tempo de resposta do modelo pode variar e causar espera e insatisfação do usuário: o risco é alto, visto que o processamento em dados reais depende da quantidade deles e do quanto complexos estarão para passarem pela limpeza e tratamento, o que influencia diretamente no tempo de resposta;
3. Falta de precisão: A precisão da análise de sentimento pode ser afetada pela qualidade dos dados, bem como pela eficácia dos algoritmos de análise. Se a precisão da solução

for baixa, a confiança dos clientes pode ser afetada: o risco é considerado baixo, pois os dados serão tratados, o que diminui a possibilidade de interferência da precisão pela qualidade dos dados.

4. Expectativas mal definidas: Se as expectativas forem claramente definidas, pode haver mal-entendidos sobre o escopo do projeto, o tempo de entrega e o que é esperado do produto final: por ainda estarmos na primeira sprint e pela falta de recebimento dos dados, a probabilidade de as expectativas do que pretendemos entregar como solução e o que é esperado pelo parceiro podem ter sido mal compreendidas, mas com o seguimento do projeto, o recebimento dos dados e o maior contato com os parceiros, o risco tende a diminuir.
5. Interfaces mal projetadas: O design da interface do usuário é crucial para garantir que os usuários possam interagir com o projeto de maneira intuitiva e eficaz. Interfaces mal projetadas podem dificultar o uso do projeto: por se tratar de uso de gráficos na interface e seu tempo de produção ser menor do que as demais partes do projeto, foi considerado um risco muito alto de acontecer, mas que pode ser contornado com uma comunicação efetiva com o parceiro e a prototipação do design antes de sua produção de fato.
6. Dependência de bibliotecas de terceiros: O projeto de análise de sentimento pode depender de bibliotecas de terceiros para funcionar. Se essas bibliotecas estiverem desatualizadas ou forem descontinuadas, isso pode afetar a funcionalidade do projeto. Não é comum as bibliotecas serem descontinuadas, mas sim ter a adição de novas ferramentas ou melhorias, portanto apesar de ser uma ameaça, não enxergamos como alta.
7. Mudanças nos requisitos: Se houver muitas mudanças nos requisitos durante o projeto, pode ser difícil finalizá-lo dentro do prazo estabelecido. Por ainda não termos a base de dados para analisar e somente 1 Sprint Review com o parceiro, os requisitos podem ter sido mal compreendidos, mas de fácil resolução dessa ameaça no decorrer dos encontros com o parceiro e a demonstração do que compreendemos e produzimos.

Também foram identificadas Oportunidades:

1. Mercado em crescimento: A análise de sentimento é uma área em crescimento, à medida que as empresas buscam entender melhor as opiniões e sentimentos de seus clientes nas redes sociais
2. Melhoria contínua: A análise de sentimento é uma área em constante evolução, com novas técnicas e tecnologias surgindo regularmente, o que possibilita a maior precisão dos resultados e a eficiência do produto
3. Criar novas formas de utilizar o projeto: dando seguimento ao projeto e criando um produto mais robusto, pode até funcionar como um SaaS ou uma vantagem competitiva de mercado
4. Utilização de tecnologias de ponta: O seu projeto pode utilizar tecnologias de ponta, como inteligência artificial e aprendizado de máquina, para fornecer resultados precisos de análise de sentimento.
5. Criação de Novos Produtos: Com base nos dados coletados pela sua solução de análise de sentimento, é possível identificar necessidades não atendidas pelos produtos existentes no mercado
6. Obter insights dos produtos: é possível obter insights sobre os produtos e serviços existentes e entender os desejos do consumidor, disponibilizando novas ofertas.
7. Possibilidade de colaboração: O projeto pode oferecer a possibilidade de colaboração com outras empresas e organizações que possam estar interessadas em análise de sentimento.
8. Gerenciamento de Crises: A análise de sentimento pode ser utilizada para gerenciar crises de imagem de empresas, entendendo o que faz gerar sentimentos ruins e contornar a situação.
9. Personalização: A sua solução de análise de sentimento pode ser personalizada para atender às necessidades específicas de diferentes setores, clientes e empresas do grupo parceiro.
10. Análise de concorrência: A análise de sentimento pode ser utilizada para entender como a concorrência está sendo percebida pelos consumidores.

3.2.3. Canvas Value Proposition

O Canvas de Proposta de Valor é uma ferramenta visual que será útil para o projeto, pois cria e comunica de forma clara e concisa o valor que será gerado por ele. Ele consiste em um modelo de negócio dividido em seis blocos, que descrevem as principais características da proposta de valor, incluindo os problemas que motivaram o projeto, os benefícios que serão oferecidos, as soluções que propõe e a forma como se diferencia da concorrência. O Canvas de Proposta de Valor é uma ferramenta útil para entender o mercado, testar ideias e ajustar a estratégia de negócio.

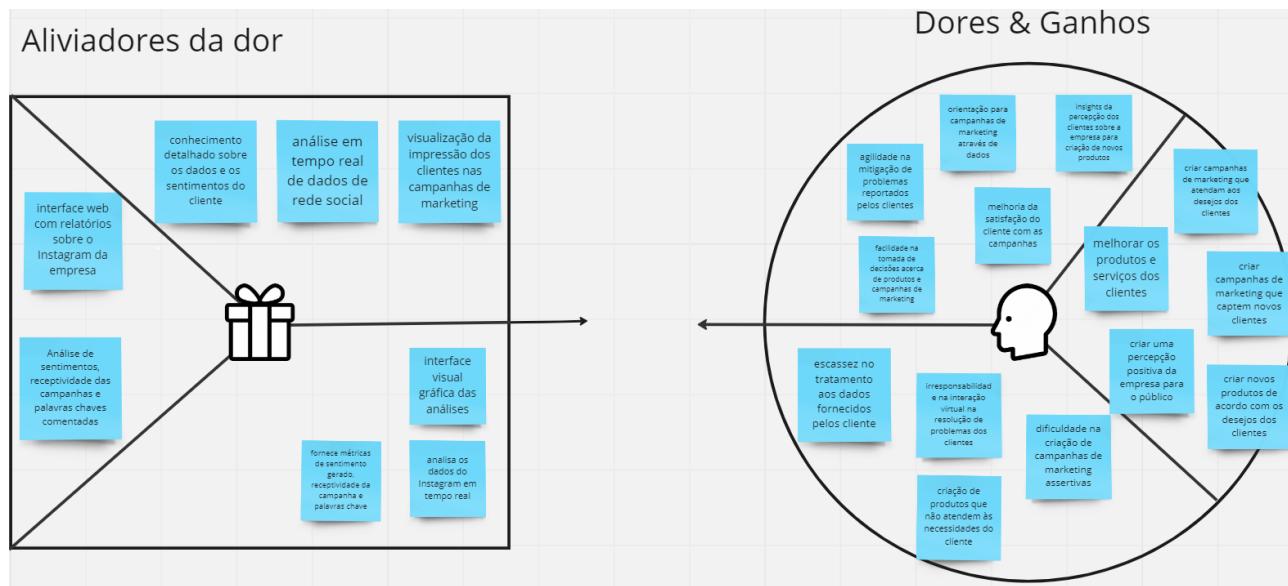


Imagen 10 - Canvas Value Proposition

Para a confecção do Canvas Value Proposition foram utilizadas as Personas (que podem ser vistas na seção abaixo), pensando nas suas necessidades core e dores. Dessa forma, tais pontos foram levantados:

Value proposition:

A) Produtos e serviços:

- Interface web com relatórios sobre o Instagram da empresa e
- Análise de sentimentos, receptividade das campanhas e palavras-chave comentadas.

B) Criadores de ganho:

- Análise em tempo real de dados de rede social;
- Visualização da impressão dos clientes nas campanhas de marketing e
- Conhecimento detalhado sobre os dados e os sentimentos do cliente.

C) Aliviadores de dor:

- Interface visual gráfica das análises;
- Fornece métricas de sentimento gerado, receptividade da campanha e palavras chave e
- Analisa os dados do Instagram em tempo real.

Perfil do consumidor:

A) Atividades do cliente:

- Criar campanhas de marketing que atendam aos desejos dos clientes;
- Melhorar os produtos e serviços dos clientes;
- Criar campanhas de marketing que captem novos clientes;
- Criar uma percepção positiva da empresa para o público e
- Criar novos produtos de acordo com os desejos dos clientes.

B) Ganhos:

- insights da percepção dos clientes sobre a empresa para criação de novos produtos;
- melhoria da satisfação do cliente com as campanhas;
- facilidade na tomada de decisões acerca de produtos e campanhas de marketing;
- agilidade na mitigação de problemas reportados pelos clientes e
- orientação para campanhas de marketing através de dados

C) Dores:

- Dificuldade na criação de campanhas de marketing assertivas;
- Criação de produtos que não atendem às necessidades do cliente;
- Escassez no tratamento aos dados fornecidos pelos cliente e
- Irresponsabilidade na interação virtual na resolução de problemas dos clientes.

3.2.4. Análise Financeira do Projeto

Com base em entrevista com o parceiro do projeto, o banco BTG Pactual previu que, caso o projeto fosse desenvolvido internamente, o investimento médio seria de cerca de R\$ 250 mil a R\$ 300 mil. Essa cifra de investimento foi fornecida pelos responsáveis técnicos da empresa, os quais estimaram que seriam necessários cerca de 3 desenvolvedores plenos, em média 6 meses de desenvolvimento e toda a infraestrutura de cloud necessária para colocar o projeto em produção. Devido ao fato de ser um projeto interno que visa entender melhor as reações do público em relação às campanhas de ativação de marketing, não haverá geração de caixa direta.

4. Análise de Experiência do Usuário

4.1. Personas

4.1.1. Maria



Maria
Idade: 30 anos
Analista de marketing do BTG

Maria é curiosa, sempre em busca de novidades e **tendências do mercado**. Com 10 anos na área, vê o **marketing digital** e as redes sociais a nova tendência de **posicionamento de marca**, o que requer novas maneiras de produzir, **trafegar** e **analisar** as campanhas.



Coleta de dados para análise e criação de campanhas



Interpretação de dados e tomada de decisões de marketing



Criar campanhas baseadas nos interesses dos consumidores



Aumentar a atração e retenção de clientes

Solução macro

Processos de negócio

Personas

User stories

Próximos passos

Imagen 11 - Persona 1

Maria é uma analista de marketing de 30 anos que trabalha no BTG Pactual. Ela possui formação em marketing e 10 anos de experiência no mercado. Maria é uma profissional dedicada e está sempre em busca de novas informações e tendências de mercado para tomar decisões informadas. Ela busca constantemente inovações e soluções tecnológicas para melhorar suas estratégias de marketing.

Seus principais objetivos são gerar campanhas personalizadas com base nos interesses dos consumidores do BTG e melhorar a atração e retenção de clientes. Maria acredita que o

sucesso de uma campanha de marketing está em sua capacidade de se conectar com os interesses e necessidades dos clientes.

No entanto, Maria enfrenta algumas frustrações em seu trabalho, principalmente na coleta de dados relevantes para suas análises de marketing, o que pode prejudicar a eficácia de suas campanhas. Além disso, ela tem dificuldade em interpretar os dados e informações de mercado, especialmente devido à subjetividade de alguns comentários, o que pode afetar a capacidade de tomar decisões de marketing bem fundamentadas.

4.1.2. Gabriel



Gabriel
Idade: 31 anos
Gerente de Produtos do BTG

Apaixonado por produtos financeiros e **satisfação do cliente**, Gabriel sempre está em busca de **melhorias e inovações** na área. Baseia suas **tomadas de decisão** a partir dos **feedbacks** dos consumidores, portanto sempre está atento as **análises** de satisfação.

 Métricas não são claras e confiáveis para avaliar a receptividade

 Informações imprecisas para gerar novos produtos baseados no interesse

 Compreender como os clientes interagem com os serviços/produtos

 Entender comportamentos e preferências do consumidor

Solução macro

Processos de negócio

Personas

User stories

Próximos passos

Imagen 12 - Persona 2

Gabriel é um gerente de produto de 31 anos que trabalha no BTG Pactual. Ele possui formação em administração e possui 11 anos de experiência no mercado. Gabriel é apaixonado por produtos financeiros e está constantemente buscando maneiras de melhorar a oferta do banco para seus clientes.

Sua personalidade é voltada para os clientes e está sempre atento às reações e feedbacks dos clientes em relação aos produtos e serviços do BTG. Gabriel busca compreender

o que os clientes valorizam e como eles interagem com os produtos oferecidos pelo banco, visando melhorar a atração e retenção de clientes.

No entanto, Gabriel enfrenta algumas frustrações em seu trabalho, principalmente em avaliar a performance de campanhas de produto para tomar decisões rápidas e eficazes, especialmente quando as métricas não são claras ou confiáveis. Além disso, gerar campanhas baseadas nos interesses dos consumidores pode ser desafiador se as informações sobre os interesses e tendências dos clientes forem imprecisas ou desatualizadas para ele.

4.2. User Story

User stories são descrições curtas e simples de funcionalidades que um usuário precisa para alcançar um objetivo específico em relação a um produto ou sistema. Elas nos ajudam fornecendo uma descrição clara e concisa dos requisitos que o sistema deve ter, mantendo o foco na solução das dores do usuário.

Número	User Story 1
Épico	Análise de sentimento dos comentários em postagens do BTG Pactual no Instagram.
Persona	Analista de Marketing do BTG Pactual (Maria)
História	Eu, como usuário do sistema, quero uma ferramenta de análise de sentimento que possa identificar automaticamente se os comentários em postagens do BTG Pactual são positivos, negativos ou neutros, para que eu possa entender como os clientes estão reagindo às postagens do banco.
Critérios de Aceitação	Critério 1: A ferramenta deve ser capaz de identificar automaticamente se um comentário é positivo, negativo ou neutro.
Testes de Aceitação	Teste 1 para o critério 1: A ferramenta identifica corretamente um comentário positivo.

	<ul style="list-style-type: none"> Conseguiu: correto Não conseguiu: essa funcionalidade precisa ser corrigida. <p>Teste 2 para o critério 1: A ferramenta precisa ter uma acurácia de, no mínimo, 80%</p> <ul style="list-style-type: none"> Conseguiu: correto, o modelo tem uma acurácia satisfatória. Não conseguiu: errado, essa funcionalidade precisa ser corrigida
--	--

Número	User story 2
Épico	Análise de sentimento dos comentários em postagens do BTG Pactual no Instagram.
Persona	Analista de Marketing do BTG Pactual (Maria)
História	Eu, como usuário do sistema, quero ter acesso a uma interface de fácil utilização que me permita visualizar os resultados da análise de sentimento de forma clara e comprehensível, como gráficos ou relatórios, para que eu possa interpretar os dados facilmente e tomar decisões informadas de marketing.
Critérios de Aceitação	Critério 1: A interface deve ser fácil de usar e permitir visualizar os resultados da análise de sentimento de forma clara e comprehensível, como gráficos ou relatórios.
Testes de Aceitação	<p>Teste 1 para o critério 1: A interface deve ter uma navegação intuitiva e fácil de usar</p> <ul style="list-style-type: none"> Conseguiu: correto, a interface é fácil de usar e permite visualizar os resultados de forma clara. Não conseguiu: errado, a interface precisa ser melhorada para facilitar a navegação e visualização dos resultados. <p>Teste 2 para o critério 1: Os gráficos e</p>

	<p>relatórios devem ser compreensíveis e mostrar os resultados da análise de sentimento de forma clara.</p> <ul style="list-style-type: none"> • Conseguiu: correto, os gráficos e relatórios mostram os resultados da análise de sentimento de forma clara e compreensível. • Não conseguiu: errado, os gráficos e relatórios precisam ser melhorados para mostrar os resultados de forma clara e compreensível.
--	---

Número	User story 3
Épico	Análise de Sentimento em Tempo Real para Postagens do BTG Pactual
Persona	Analista de Marketing do BTG Pactual (Maria)
História	Eu, como usuário do sistema, quero que a ferramenta seja capaz de processar grandes volumes de comentários em tempo real, para que eu possa obter insights em tempo hábil sobre a percepção dos clientes em relação às postagens do BTG Pactual.
Critérios de Aceitação	Critério 1: A ferramenta deve ser capaz de processar pelo menos 200 comentários por hora em tempo real, sem atrasos significativos.
Testes de Aceitação	<p>Teste 1 para o critério 1: Inserir 200 comentários fictícios em um teste da ferramenta e avaliar se ela é capaz de processá-los em um período de 2 horas, sem atrasos significativos.</p> <ul style="list-style-type: none"> • Conseguiu: correto, a ferramenta é capaz de processar, pelo menos, 100 comentários por hora em tempo real. • Não conseguiu: errado, a ferramenta não é capaz de processar e será preciso otimizar.

Número	User story 4
Épico	Melhorar a análise de dados do BTG Pactual
Persona	Gerente de Produtos do BTG Pactual (Gabriel)
História	<p>Eu, como usuário do sistema, quero que a ferramenta seja capaz de identificar palavras-chave relevantes nos comentários dos clientes, para que eu possa entender os principais tópicos de discussão e identificar tendências emergentes.</p>
Critérios de Aceitação	<p>Critério 1: A ferramenta deve ser capaz de identificar palavras-chave relevantes em cada comentário.</p> <p>Critério 2: As palavras-chave identificadas pela ferramenta devem ser precisas e relevantes para os tópicos de discussão.</p>
Testes de aceitação	<p>Teste 1 para o critério 1: A ferramenta é testada com um conjunto de 10 comentários e a identificação de pelos menos 5 palavras-chave relevantes é verificada</p> <ul style="list-style-type: none"> • Conseguiu: correto, a ferramenta identificou pelo menos 5 palavras-chave em todos os 10 comentários. • Não conseguiu: errado, a ferramenta não identificou e será preciso otimizar. <p>Teste 1 para o critério 2: A ferramenta é testada com um conjunto de 10 comentários e há precisão e relevância das palavras-chave identificadas.</p> <ul style="list-style-type: none"> • Conseguiu: correto, todas as palavras-chave identificadas são precisas e relevantes. • Não conseguiu: errado, pelo menos uma das palavras-chave identificadas

é imprecisa ou irrelevante.

Número	User story 5 (Nice to have)
Épico	Melhorar a acessibilidade e usabilidade do BTG Pactual
Persona	Gerente de Produtos do BTG Pactual (Gabriel)
História	Eu, como usuário do sistema, quero que a interface seja responsiva e amigável para uso em dispositivos móveis, para que eu possa acessar e utilizar a ferramenta em qualquer lugar.
Critérios de Aceitação	<p>Critério 1: Todos os recursos e funcionalidades do BTG Pactual devem ser acessíveis a pessoas com deficiência visual, incluindo suporte para leitores de tela e compatibilidade com tecnologias assistivas.</p> <p>Critério 2: A interface do usuário deve ser intuitiva e de fácil compreensão, com uma navegação clara e lógica, garantindo que os usuários possam encontrar facilmente as informações e recursos que estão procurando.</p>
Testes de aceitação	<p>Teste 1 para o critério 1: A interface é testada em diferentes dispositivos, por diferentes pessoas (por exemplo, smartphones, notebooks e tablets).</p> <ul style="list-style-type: none"> ● Conseguiu: correto, a interface se adapta de forma responsiva a todos os tamanhos de tela e dispositivos testados e todos os componentes são compreendidos pelos usuários. ● Não conseguiu: errado, a interface não se adapta de forma responsiva ou não são comprehensíveis a certos usuários e será preciso otimizar.

	<p>Teste 1 para o critério 2: Vários usuários diferentes testam a interface e fornecem feedback sobre sua amigabilidade e intuitividade.</p> <ul style="list-style-type: none"> • Conseguiu: correto, a interface é considerada amigável e intuitiva pela maioria dos usuários • Não conseguiu: errado, pelo menos três usuários consideram a interface pouco amigável ou não intuitiva.
--	---

Número	User story 6 (<i>Nice to Have</i>)
Épico	Melhorar a análise de dados do BTG Pactual
Persona	Gerente de Produtos do BTG Pactual
História	Eu, como usuário do sistema, quero poder comparar as análises de diferentes campanhas ao mesmo tempo, para que consiga tomar melhores decisões.
Critérios de Aceitação	<p>Critério 1: A ferramenta deve permitir a seleção e comparação de pelo menos duas campanhas diferentes.</p> <p>Critério 2: A ferramenta deve apresentar a análise de sentimento de cada campanha de forma clara e fácil de entender.</p> <p>Critério 3: A ferramenta deve permitir a visualização da comparação entre as análises de sentimento de cada campanha.</p>
Testes de Aceitação	<p>Teste 1 para o critério 1: O usuário tenta selecionar e comparar pelo menos duas campanhas diferentes.</p> <ul style="list-style-type: none"> • Conseguiu: correto, a ferramenta permite comparar pelo menos duas campanhas diferentes. • Não conseguiu: errado, a ferramenta

não permite a comparação e será necessário aprimorar.

Teste 1 para o critério 2: O usuário examina as análises de sentimento de cada campanha selecionada.

- Conseguiu: correto, a ferramenta apresenta as análises de sentimento da campanha de forma clara e fácil de entender.
- Não conseguiu: errado, a ferramenta não apresenta as análises de sentimento de forma clara ou fácil de entender e será preciso otimizar.

Teste 1 para o critério 3: O usuário examina a visualização da comparação entre as análises de sentimento de cada campanha.

- Conseguiu: correto, a ferramenta permite visualizar a comparação entre as análises de sentimento de cada campanha de forma clara e fácil de entender.
- Não conseguiu: errado, a ferramenta não permite visualizar a comparação entre as duas análises de sentimento de forma clara ou fácil de entender e será preciso otimizar.

4.4. Protótipo de interface com o usuário

Para garantir uma melhor experiência aos usuários da solução proposta, foi desenvolvido um protótipo no Figma que pode ser acessado no link abaixo:

<https://www.figma.com/file/LDHmzOUuwnzw0vU3GwquaQ/Untitled?type=design&node-id=0%3A1&t=NR6Dhc0ubyPg7zAn-1>



O protótipo é uma interface web com o seguinte layout:

- Cabeçalho:** Logo BTG e Instagram.
- Barra superior:** Título "Analise os posts do Instagram" e placeholder "Insira aqui o link do post".
- Secção Top 10 palavras:** Placeholder para visualização de resultados.
- Secção Nuvem de palavras:** Nuvem de palavras com termos como "palmas", "azul", "cliente", "dinheiro", "foguete", "bolsonaro", entre outros.
- Secção Top perfis engajados:** Placeholder para visualização de resultados.
- Secção Sentimentos:** Gráfico de donut mostrando 35% negativos e 65% positivos, com o principal sentimento definido como "POSITIVO" e a quantidade de comentários em 117.
- Botão Sair:** Localizado no lado esquerdo da interface.

Imagen 13 - Protótipo Tela 1



Imagen 14 - Protótipo Tela 2

A interface foi projetada pensando nos seguintes pontos:

- Analisar um único post: assim, colocando o link do post (webscrapping), será possível ter uma análise detalhada dos sentimentos daquela campanha;
- Top 10 palavras: foi um ponto positivo de feedback da review com o stakeholder, que optou-se por implementar também na interface como mais um insight a ser obtido;
- Nuvem de palavras: para uma visão mais ampla do post, é possível ver todas as palavras comentadas, além das Top 10;
- Top perfis engajados: seguindo a própria interface já existente do stakeholder, optou-se por manter essa funcionalidade com o intuito de oferecer mais insights sobre o público, quem mais engaja, positiva ou negativamente;

- Sentimento: com um gráfico que mostra a parcela de sentimentos dos comentários e dá destaque aos negativos, que devem ser tratados, mas também aponta o principal sentimento percebido naquela campanha e a quantidade de comentários do post analisado e
- Visualização por comentários: permite uma visão mais detalhada de cada comentário, em que é possível ver o texto do comentário, autor que o realizou e o sentimento identificado.

Esse design permite sua adaptação para demais redes sociais, visto que as análises não seguem um padrão específico de uma única rede, permitindo sua versatilidade.

5. Solução Proposta

5.1. Solução

A solução proposta consiste em um sistema local que utiliza aprendizado de máquina supervisionado e processamento de linguagem natural para realizar a análise de sentimentos dos comentários feitos nos posts do Instagram @btgpactual.

O sistema é capaz de identificar e classificar os sentimentos expressos nos comentários, como positivos, neutros ou negativos. Ao aplicar técnicas avançadas de processamento de linguagem natural, o sistema é capaz de compreender o contexto e a semântica das mensagens, além de lidar com desafios como a interpretação de emojis e a compreensão de expressões idiomáticas.

Essa solução visa fornecer insights valiosos para o Instagram @btgpactual, permitindo uma compreensão mais aprofundada da percepção e opiniões dos seguidores em relação aos posts. Dessa forma, a equipe responsável pode tomar decisões mais informadas e estratégicas para melhorar a interação e o relacionamento com o público.

5.2. Arquitetura Proposta

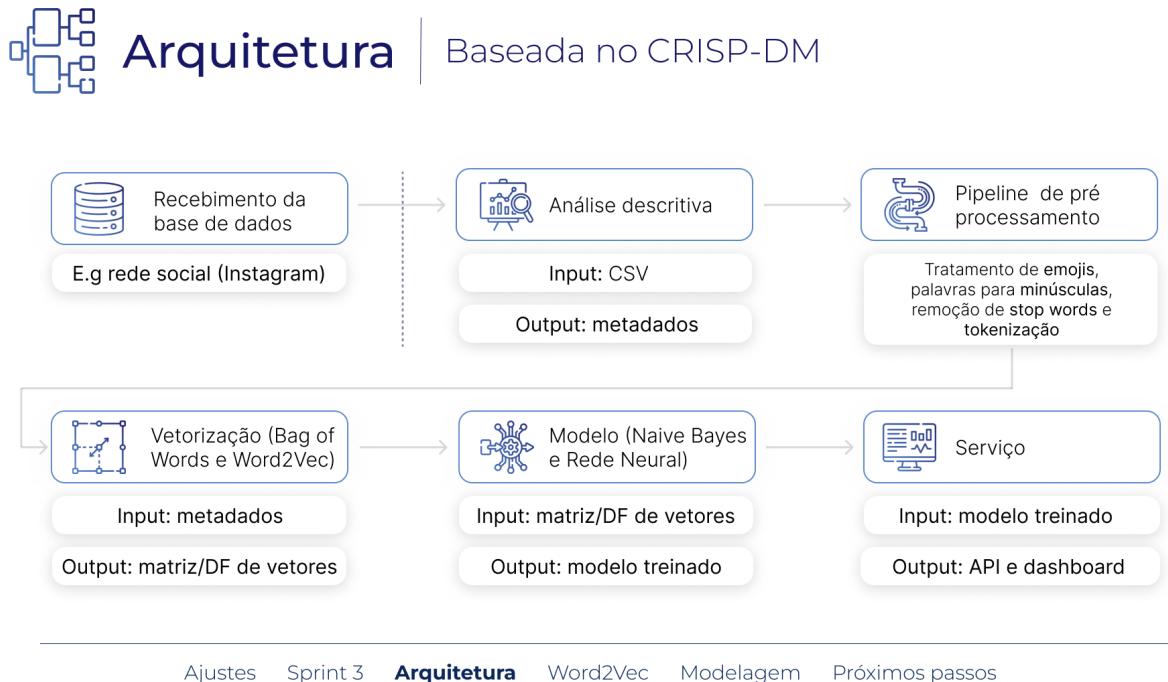


Imagen 15 - Arquitetura

A arquitetura proposta foi desenvolvida com base na metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) e segue um fluxo sequencial de etapas.

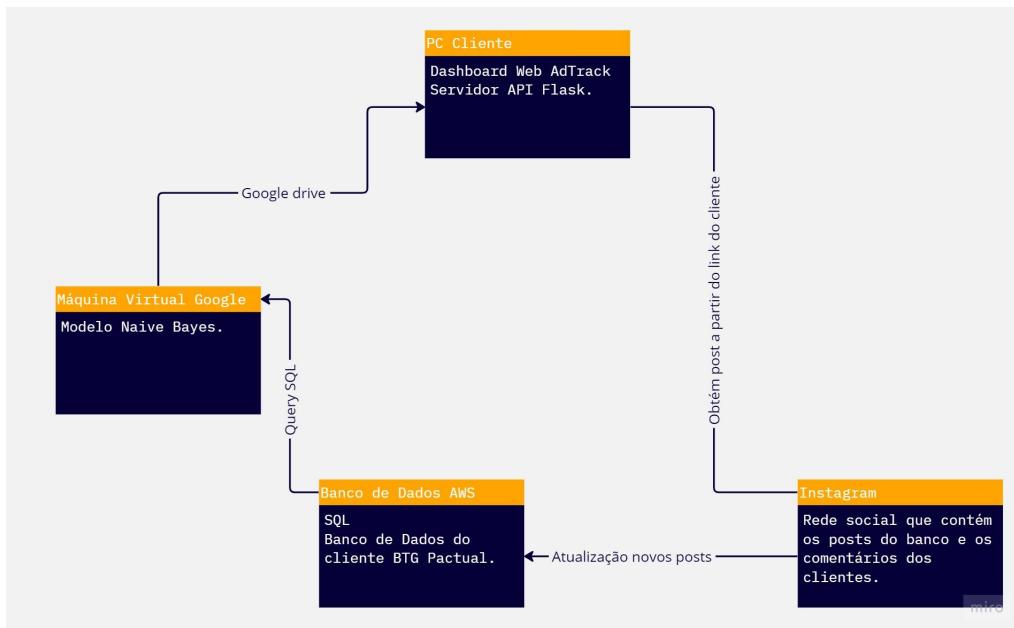
Inicialmente, os dados são coletados a partir de uma rede social, utilizando técnicas como webscraping ou a importação de um arquivo CSV. No contexto deste projeto, o Instagram é a plataforma escolhida para a coleta dos dados. Após a coleta, é realizada uma análise descritiva dos dados, empregando gráficos, comparações e outras técnicas para obter uma compreensão mais aprofundada. Essa análise resulta em metadados que serão utilizados na etapa de pré-processamento, que será detalhada posteriormente.

O pré-processamento dos dados inclui várias etapas, como o tratamento de emojis, a conversão de todas as palavras em minúsculas, a tokenização dos textos e a remoção de stopwords. Essas técnicas visam preparar os dados para a etapa seguinte. Em seguida, é realizada a vetorização dos dados, utilizando métodos como Bag of Words, Word2Vec, GloVe, TF-IDF, entre outros. O objetivo é obter uma matriz de vetores que representam os dados de forma numérica e adequada para aplicação de modelos de machine learning.

Com base nesses vetores, um modelo de machine learning é aplicado e treinado. Esse modelo servirá como base para o desenvolvimento de serviços, como uma API local ou na nuvem, bem como um dashboard para visualização dos resultados obtidos. Dessa forma, a arquitetura proposta segue a metodologia CRISP-DM, permitindo uma abordagem estruturada e sequencial na análise dos dados do Instagram, desde a coleta até a disponibilização dos resultados por meio de serviços e visualização em um dashboard.

5.3. Diagrama UML

O diagrama de implantação UML é uma representação visual da arquitetura de implantação de um sistema, mostrando a disposição física e interações entre os componentes de hardware e software. Ele descreve como os componentes do sistema são implantados em diferentes nós de hardware, como servidores, máquinas virtuais ou dispositivos físicos, e como esses nós se comunicam uns com os outros por meio de conexões de rede. Abaixo está uma imagem do diagrama que representa como o projeto do grupo Ad Track compõe suas funcionalidades, interações e conexões:



1. PC Cliente (Dashboard Web Ad Track com Servidor API Flask):

- Responsável por fornecer a interface do dashboard para o analista de marketing do BTG Pactual.

- Permite que o analista de marketing insira o link de um determinado post do Instagram do BTG Pactual.
 - Recebe informações do modelo preditivo Naive Bayes, como palavras mais ditas, autores que mais interagiram no post, porcentagem dos sentimentos e nuvem de palavras.
 - Interage com a Máquina Virtual Google para obter os resultados do modelo preditivo.
2. Máquina Virtual Google (Modelo Naive Bayes):
- Executa o modelo preditivo Naive Bayes para analisar os dados dos comentários do Instagram do BTG Pactual.
 - Está conectada ao Google Drive para acessar os dados necessários para o treinamento e a classificação dos comentários.
 - Fornece os resultados do modelo para o PC Cliente, que serão exibidos no dashboard.
3. Banco de Dados AWS (Banco de Dados do BTG Pactual):
- Armazena os dados dos posts e comentários do Instagram do BTG Pactual.
 - Permite a realização de consultas SQL para recuperar informações relevantes para o modelo preditivo e o dashboard.
 - Recebe atualizações do Instagram em relação a novos posts, mantendo o banco de dados atualizado.
4. Instagram:
- Fornece os posts do BTG Pactual para análise e interação com os usuários.
 - Atualiza o Banco de Dados AWS com novos posts e comentários à medida que são publicados.

5.4. Pipeline de pré processamento

As etapas do pipeline de pré processamento estão exemplificadas na imagem abaixo:

Pipeline



Sprint 2 Análise dos dados Processos **Bag of Words** Próximos passos

Imagen 16 - Pipeline de pré processamento

Após o recebimento da base de dados, inicia-se o processo de pré-processamento utilizando técnicas específicas para o processamento de linguagem natural. Essas etapas são realizadas em uma única função, sequencialmente:

1. Tratamento de emojis: É aplicado um dicionário para converter os emojis presentes nos textos em palavras correspondentes.
2. Minuscularização das palavras: Todas as palavras são convertidas para minúsculas, garantindo uma padronização e evitando diferenças por capitalização.
3. Tokenização e remoção de stop words e caracteres alfanuméricos: As frases são divididas em tokens, ou seja, unidades individuais, removendo-se também stop words (palavras comuns que não agregam muito significado) e caracteres alfanuméricos, como números e pontuações.
4. Vetorização utilizando o método Bag of Words: Nessa etapa, as palavras são transformadas em vetores numéricos com base em sua frequência de ocorrência na base de

dados. Essa técnica de vetorização permite a representação das palavras de forma adequada para o processamento posterior.

Portanto, as etapas de tratamento de emojis, minuscularização, tokenização, remoção de stop words e alfanuméricos, e vetorização utilizando Bag of Words, são aplicadas como parte do processo de pré-processamento para o processamento de linguagem natural.

6. Modelagem

6.1. Modelos aplicados

A partir da Sprint 2, foram aplicados diferentes modelos com distintas vetorizações e, também, com dois datasets: o original, fornecido pelo parceiro, e o com os sentimentos revistos pelo grupo. Abaixo encontram-se os métodos e conclusões obtidos por sprint:

6.1.1. Sprint 3

Em primeiro momento, foram utilizadas as vetorizações Bag of Words e Word2Vec, explicadas em detalhes na seção nos modelos, detalhados melhor abaixo:

Realizado na Sprint 2, o BoW é uma técnica de Processamento de Linguagem Natural que cria um vocabulário de palavras, referentes ao input, e retorna uma matriz de vetores dessas palavras. Dessa forma, foi produzido um CSV de vetores das frases junto com sua respectiva tag de sentimento, pois os modelos utilizam de dados numéricos para funcionarem e, dessa maneira, tal tratamento foi utilizado para transformar os dados não estruturados (texto) para estruturados (números). Dois modelos foram feitos, portanto, utilizando o Bag of Words:

1. Naive Bayes: utilizado com a variante Bernoulli do Naive Bayes os recursos são representados por variáveis discretas que podem assumir apenas dois valores (booleano), normalmente 0 ou 1, representando afirmativa ou negativa de uma condição, nesse caso do tipo de sentimento do comentário. Para melhorar a performance e robustez do modelo, foi aplicada a técnica de cross validation, em que o conjunto de dados é dividido em k partes (chamadas de folds) e o modelo é treinado k vezes, onde em cada iteração, um dos folds é utilizado como conjunto de teste e o restante como conjunto de treinamento. A métrica de avaliação é então calculada a partir das k iterações. Na aplicação desse modelo, utilizamos de 5 folds e, assim, obtendo uma acurácia de 74% e uma matriz de confusão que apresenta poucos falsos positivos e falsos negativos.
2. Rede Neural: aplicada utilizando a biblioteca Keras, a rede neural utilizada é Sequencial, em que permite a adição subsequente de camadas. O modelo aplicado utilizou duas camadas densas, sendo a primeira de entrada com 64 neurônios e função de ativação ReLU, que recebe um vetor de tamanho igual ao número de recursos (colunas) da matriz obtida com o Bag of Words. Por outro lado, a segunda camada é de saída, composta por um único neurônio e que utiliza da função de ativação sigmóide. Para melhorar a performance do modelo, foi aplicado o otimizador Adam, que ajusta os pesos da rede durante o treinamento com base na taxa de aprendizado especificada (0.001). Por fim, para treinar esse modelo, foram utilizadas 20 épocas e com tamanho de lote de 64, o que resultou em uma acurácia de 60%, mas uma tendência de falso negativo para os dados classificados como positivos para o modelo. Os resultados obtidos com os modelos podem ser vistos na imagem abaixo:

Naive Bayes

74% de acurácia

Matriz de confusão

	1719	152	81
Verdadeiro	270	2411	117
	212	154	1240
Previsto			

Rede Neural

60% de acurácia

Matriz de confusão

	1952	0	0
Verdadeiro	2798	0	0
	1606	0	0
Previsto			

Imagen 17 - Resultados Bag of Words

Por fim, comparando os dois modelos aplicados, conclui-se que o Naive Bayes, apesar de mais simples, obteve melhor desempenho, com uma acurácia maior e adequada à aplicação do Bag of Words, representando poucos erros de falsos negativos e falsos positivos aos dados, sendo a melhor escolha entre os dois, pela menor complexidade e tempo de resposta e maior acurácia aos dados de treinamento e teste.

O Word2Vec é um modelo de aprendizado de máquina utilizado para representar palavras como vetores em um espaço vetorial de muitas dimensões. Esse algoritmo foi inventado em 2013 por engenheiros do Google e com ele é possível classificar cada palavra em um espaço vetorial finito e realizar operações matemáticas simples com n palavras para buscar graus de similaridades entre as mesmas.

No presente projeto, o Word2Vec foi empregado de duas formas: em primeiro plano, utilizando como base todas as palavras obtidas nos comentários, ou seja, na coluna 'texto', já previamente tratada do DataFrame, buscando relações de similaridade entre as mesmas e, em segundo plano, foi utilizada uma base vetorizada pré treinada fornecida pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo (USP).

Para treinar o modelo com o dataset fornecido pelo stakeholder BTG Pactual, utilizou-se a biblioteca Gensim, que transforma as palavras em vetores. O modelo foi treinado utilizando as palavras contidas na base de dados disponível, e posteriormente aplicado às frases processadas. Durante esse processo, o modelo captura os vetores correspondentes a cada palavra presente na frase e os soma. Os primeiros 50 vetores resultantes são, então, armazenados em colunas específicas, o que proporciona uma representação compacta e numérica das frases, tornando-as adequadas para análises e tarefas de aprendizado de máquina.

Já para utilizar a vetorização pré treinada pelo NILC é necessário baixar o arquivo .txt, que está no site do mesmo, e carregar a base no modelo (para mais detalhes dessa etapa, consulte o arquivo .ipynb na pasta "scr" em que todo o passo a passo está documentando). Assim, após o download da base do NILC, é repetido o mesmo processo citado no parágrafo acima e são obtidas as 50 colunas que representam o vetor de cada frase.

Essa abordagem baseada no Word2Vec e na biblioteca Gensim possibilitou a obtenção de representações vetoriais para as palavras, as quais foram aplicadas às frases processadas. Ao realizar a soma dos vetores das palavras em cada frase, obtemos uma representação geral da mesma e essa representação compacta é armazenada nas colunas criadas para esse propósito, permitindo uma análise mais eficiente e possibilitando a utilização dos primeiros 50 vetores resultantes em tarefas subsequentes. Com essa abordagem, é possível explorar as relações semânticas e contextuais entre as palavras, obtendo insights valiosos a partir das frases contidas em nosso banco de dados.

Naive Bayes

59% de acurácia

Matriz de confusão

	328	12	36
Verdadeiro	132	334	107
Previsto	206	46	71

Rede Neural

39% de acurácia

Matriz de confusão

	1952	387	0
Verdadeiro	2798	447	0
Previsto	1606	300	0

Por fim, é possível perceber que, mesmo com uma acurácia mais baixa que o modelo Naive Bayes utilizando Bag of Words, a matriz de confusão do mesmo modelo com a técnica de Word2Vec demonstrou ser positiva em relação ao objetivo central do projeto: identificar os comentários negativos para tratá-los e, de tal maneira, o modelo tendeu a identificar negativos, o que ocasionou na ocorrência de falso negativo nos comentários positivos, o que não acarreta grandes problemas no projeto.

As técnicas implementadas e descritas acima têm sua análise detalhada no seguinte documento: <https://github.com/2023M6T4-Inteli/Projeto4/blob/Dev/analisesSprint3.md>

Como conclusão, pode-se entender que, em primeiro momento, a utilização da técnica Bag of Words nos modelos obteve um melhor resultado do que a Word2Vec, mas ainda há onde melhorar e, de tal maneira, os próximos passos do desenvolvimento visam a melhoria da base de dados e aplicação de novas técnicas e modelos nas duas abordagens para a definição do modelo final.

6.1.2.Sprint 4

Nessa Sprint, foram ajustados pontos do pré-processamento (detalhados melhor na seção), com o fim de otimizar os resultados obtidos com os modelos de processamento de linguagem natural obtidos na Sprint anterior, analisar e comparar diferentes métricas destes, além de testar novas modelagens, tanto aplicando Bag of Words, quanto Word2Vec.

1. CatBoost: O CatBoost é uma biblioteca de gradient boosting que lida eficientemente com dados categóricos. Ele oferece uma implementação de alta performance, eliminando a necessidade de pré-processamento adicional. Com recursos como tratamento de valores ausentes e otimização automática de hiperparâmetros, é amplamente utilizado em problemas de classificação e regressão. Sua capacidade de lidar diretamente com dados categóricos e a interpretabilidade proporcionada pela análise de características importantes o tornam uma ferramenta poderosa para análise e previsão de dados. O modelo que utilizou o CBOW obteve uma acurácia de 60% e com o corpus foi também de 60%.

2. Random Forest: O Random Forest é um algoritmo de aprendizado de máquina que faz parte da família dos métodos de ensemble. Ele combina a construção de várias árvores de decisão independentes para formar uma "floresta", em que cada árvore contribui com sua previsão individual. No caso do Random Forest, a previsão final é obtida através de uma média ou votação das previsões individuais das árvores. O Random Forest é caracterizado pela sua capacidade de lidar com conjuntos de dados complexos e realizar tanto tarefas de classificação quanto de regressão. Esse modelo é particularmente útil quando há um grande número de características (features) e algumas delas são mais importantes do que outras na tomada de decisões. Além disso, ele tem a capacidade de lidar com dados ausentes, outliers e overfitting, o que o torna uma escolha popular em diversos problemas de aprendizado de máquina.

Uso do Random Forest em PLN para compreender os sentimentos por trás dos comentários dos clientes no Instagram do BTG Pactual: O Random Forest é uma escolha interessante para compreender os sentimentos por trás dos comentários dos clientes no Instagram do BTG Pactual por várias razões. Primeiramente, o modelo é capaz de lidar com uma grande quantidade de dados, o que é essencial em plataformas de mídia social onde há uma quantidade significativa de comentários dos clientes. Além disso, o Random Forest pode ser treinado para realizar tarefas de classificação, como a análise de sentimento, em que os comentários dos clientes são classificados como positivos, negativos ou neutros. Isso permite que o BTG Pactual obtenha uma compreensão abrangente dos sentimentos dos clientes em relação aos seus produtos, serviços ou campanhas de marketing, auxiliando na tomada de decisões estratégicas. Outra vantagem do Random Forest é sua capacidade de lidar com características (features) relevantes para a análise de sentimento. Ele é capaz de identificar quais características são mais importantes na determinação dos sentimentos expressos nos comentários, permitindo que o BTG Pactual concentre seus esforços em áreas específicas para melhorar a satisfação do cliente. Por fim, o Random Forest também é robusto em relação a outliers e dados ausentes, o que é comum em comentários de mídias sociais. Isso significa que o modelo é capaz de lidar com a natureza variada e ruidosa dos dados coletados no Instagram do BTG Pactual, garantindo resultados mais confiáveis e precisos na análise de sentimentos.

- Bag of Words:

- Base original: Recall de 70%
 - Base de sentimentos revistos: Recall de 69%
- TF-IDF:
 - Base original: Recall de 71%
 - Base de sentimentos revistos: Recall de 70%
- 3. Naive Bayes: O modelo naive bayes é considerado o mais simples dentre os existentes para classificação. Ele possui distintas aplicações, podendo ser aplicado em resultados binomais (booleanos) com distribuição de Bernoulli, multinominais (mais de 2 resultados possíveis), ou de regressão, com a distribuição Gaussiana. Ele também permite diversos ajustes de hiperparâmetros e divisões entre treino e teste. Após a revisão do target de sentimento da base e dos ajustes finais do pré processamento, o modelo de classificação multinomial foi aplicado novamente com distintas vetorizações e obteve os seguintes resultados:
 - Bag of Words:

Modelo	Dataset	Acurácia	Recall
Naive Bayes Simples	Original	54	54
Naive Bayes Simples	Revisto	73	73
Naive Bayes com cross validation	Original	72	72
Naive Bayes com cross validation	Revisto	61	61
Naive Bayes com Grid Search e Cross Validation	Original	72	72
Naive Bayes com Grid Search e Cross Validation	Revisto	70	70

Imagen 19 - Tabela Resultados Naive Bayes com Bag of Words

- TF-IDF:

Modelo	Dataset	Acurácia	Recall
Naive Bayes Simples	Original	53	53
Naive Bayes Simples	Revisto	69	69
Naive Bayes com cross validation	Original	71	71
Naive Bayes com cross validation	Revisto	70	70
Naive Bayes com Grid Search e Cross Validation	Original	71	71
Naive Bayes com Grid Search e Cross Validation	Revisto	70	70

Imagen 20 - Tabela Resultados TF-IDF com Bag of Words

3. Rede Neural Long Short-Term Memory (LSTM):

A rede neural do tipo LSTM tem o nome em inglês de neural network long short-term memory traduzindo para o português com uma tradução livre obtém-se : de Rede Neural de memória de curto longo prazo. Assim, são redes capazes de obter insights sobre dependências entre sequências de dados ou análises de dados a partir de padrões temporais. Com isso, é uma arquitetura que tem aptidão para o processamento de linguagem natural.

Adentrando mais a fundo no funcionamento do algoritmo. Nota-se os detalhes peculiares do LSTM que é a capacidade da rede enviar dados da camada de output para os nós anteriores, o que não existe em outras arquiteturas. Assim o modelo começa recebendo um dado e o mesmo vai passando para os novos nós da rede, esse número de nós é definido como números de camadas da rede. Ao chegar na camada final o modelo considera os dados de saída com o resultado fornecido e vai reajustando os pesos nos nós passados até atingir o máximo de acerto possível.

Aplicação do modelo Com o modelo já explicado, foi aplicado o mesmo no dataset tratado com o word2vec na base [corpus](#) e os seguintes resultados foram obtidos:

- Acurácia 56 %

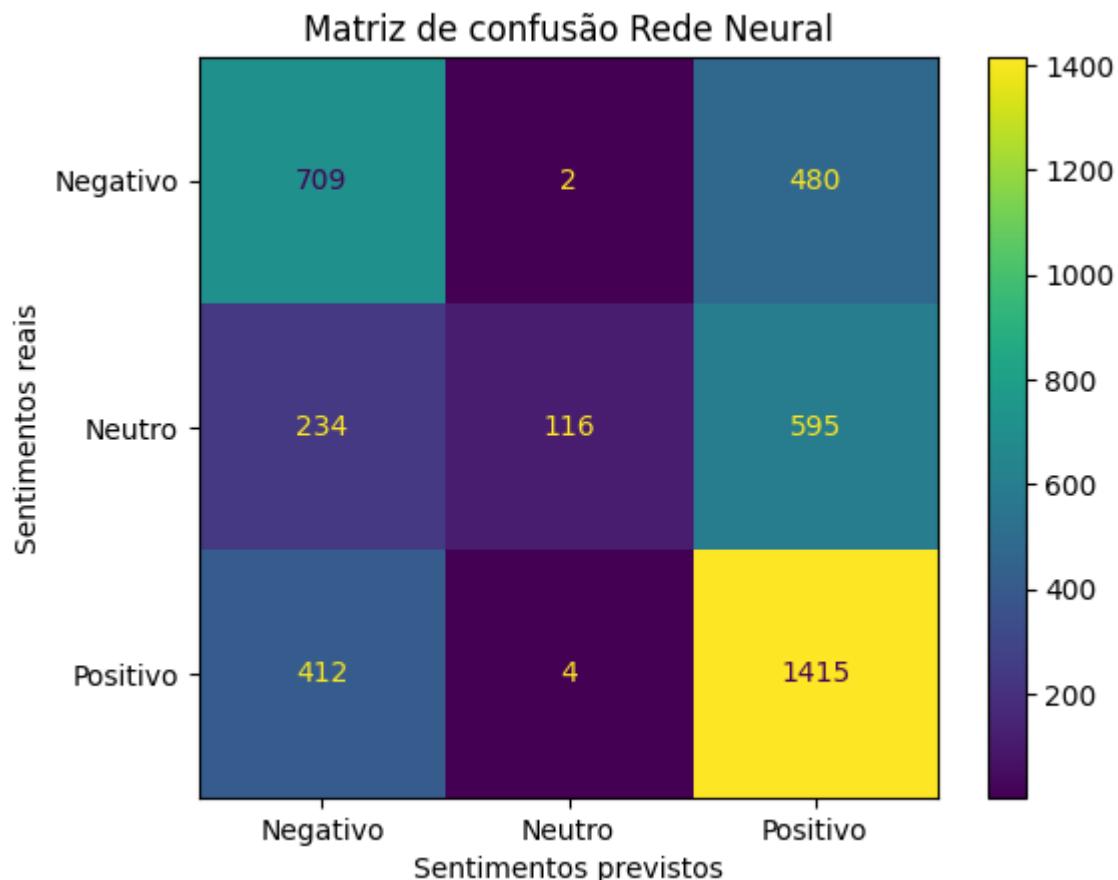


Imagen 21 - Matriz de Risco Rede Neural

Com isso, a rede neural não se mostrou tão boa para lidar com o modelo word2vec nesse caso específico, visto que a mesma tem uma tendência muito grande para comentários positivos.

Para acessar todo o notebook dessa parte acesse o seguinte link:

As métricas utilizadas tiveram sua análise baseadas no artigo da Microsoft 'How to understand automated machine learning', que diz: "Recall é a capacidade de um modelo para detectar todos os exemplos positivos e a precisão é a capacidade de um modelo evitar rotular amostras negativas como positivas. Alguns problemas empresariais podem exigir uma recall mais alta e uma precisão mais alta, dependendo da importância relativa de evitar falsos negativos versus falsos positivos." e também do livro 'Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems', que diz: "A acurácia em machine learning é uma métrica utilizada para medir a taxa de acertos de um modelo de aprendizado de máquina. Ela representa a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo."

6.2. Comparações

Seguindo o artigo da Microsoft 'Results from Machine Learning Models', que diz "A matriz de confusão fornece um meio de avaliar o êxito de um problema de classificação e onde ele comete erros (ou seja, onde ele se torna "confuso").", tal abordagem foi utilizada para comparar os resultados dos modelos.

Todas as comparações podem ser vistas no seguinte notebook:

https://github.com/2023M6T4-Inteli/Projeto04/blob/main/src/Analises_Sprint_4.ipynb

Após as análises dos resultados dos modelos, priorizando o recall, mas, principalmente, a matriz de confusão, que aponta os falsos negativos e falsos positivos, os dois melhores modelos de cada vetorização foram escolhidos para análises mais profundas e, por fim, escolhido o modelo final a ser utilizado.

1. Bag of Words:

Matriz de confusão Naive Bayes

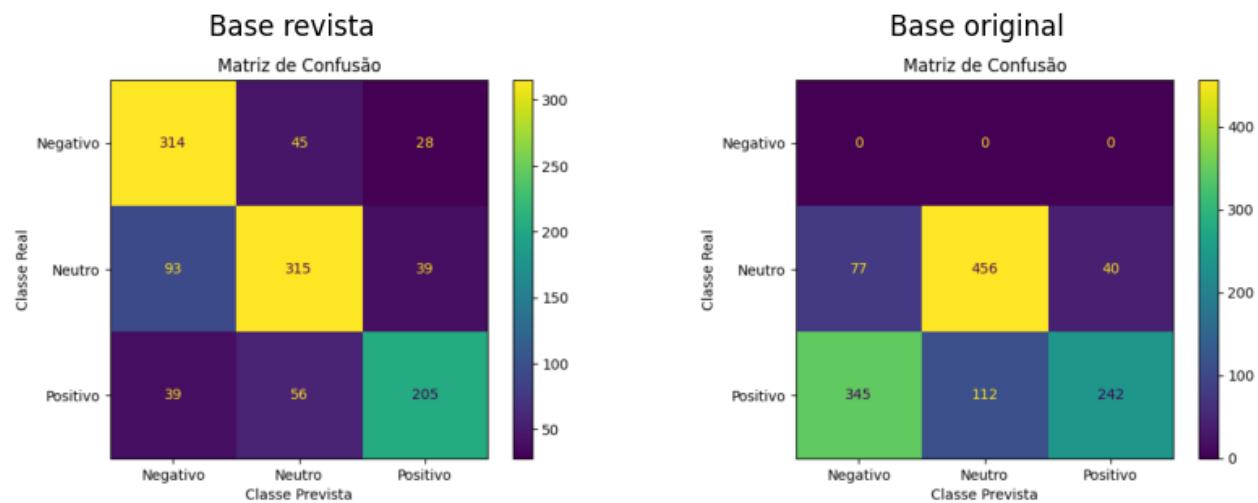


Imagen 22 - Matrizes de Confusão Naive Bayes com Bag of Words

É possível ver, pela imagem acima, que o modelo Naive Bayes simples obteve um alto acerto de comentários negativos, ou seja, do principal objetivo do parceiro com a solução.

Matriz de confusão Naive Bayes com Grid Search e cross validation

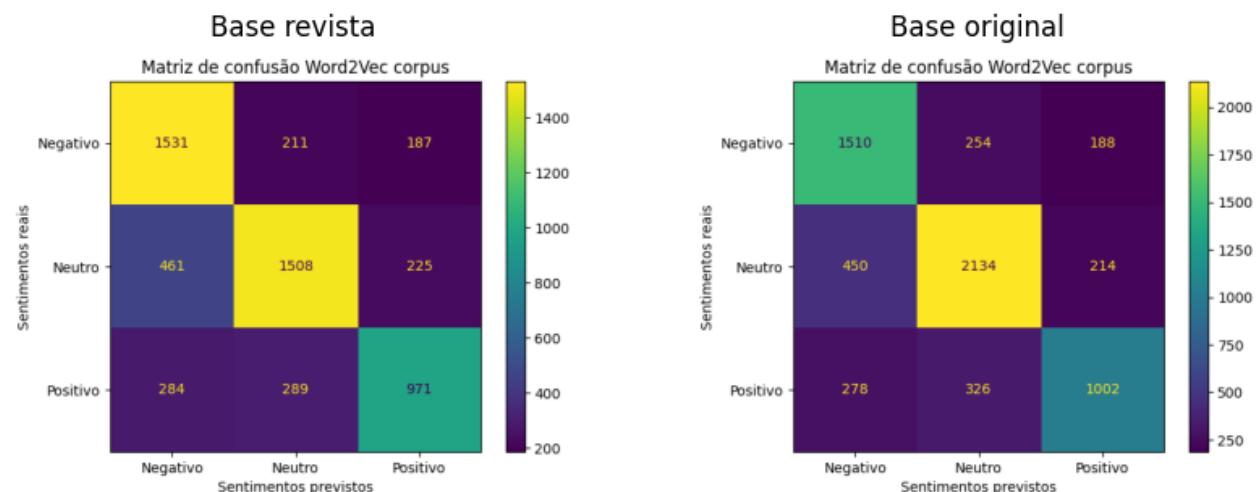


Imagen 23 - Matrizes de Confusão Naive Bayes com Grid Search e Cross Validation com Bag of Words

O mesmo ocorre quando aplicado o Naive Bayes com Grid Search e Cross Validation, entretanto, apesar da base original apresentar maior recall (71%), o acerto de comentários negativos foi menor, pois tende a acertar para neutros.

2. TF-IDF:

Matriz de confusão Naive Bayes com Cross Validation

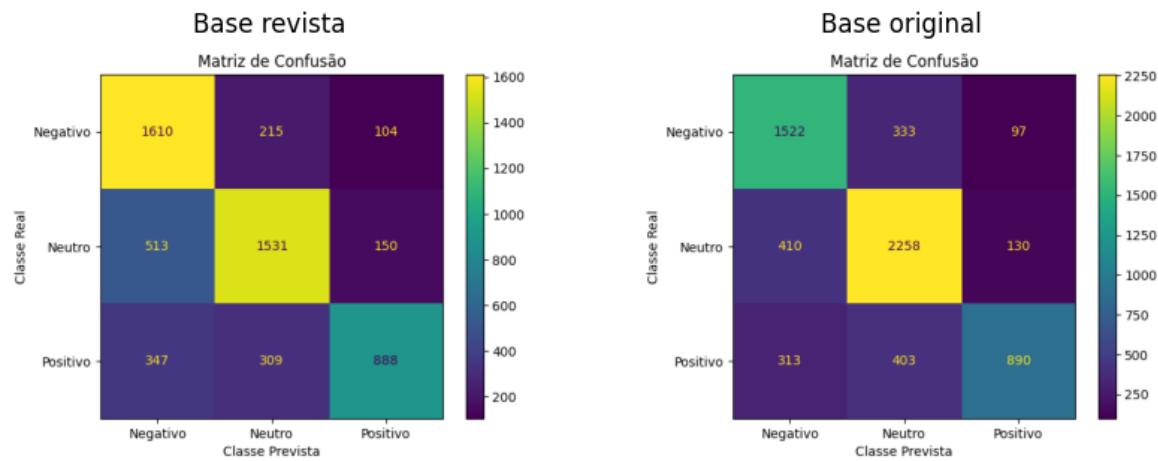


Imagen 24 - Matrizes de Confusão Naive Bayes com Cross Validation com TF-IDF

O cenário se repete com o TF-IDF, em que a base original tende para comentários neutros, mesmo apresentando um recall maior.

Matriz de confusão Random Forest

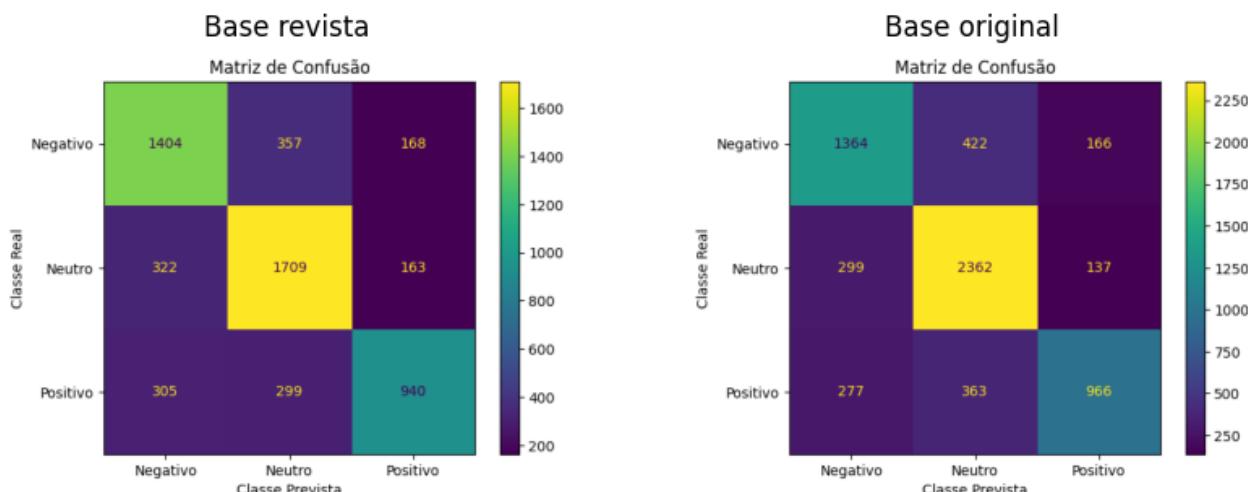


Imagen 25 - Matrizes de Confusão Random Forest com TF-IDF

E com o Random Forest o resultado foi muito semelhante ao do Naive Bayes.

6.3. Modelo final

Por fim, para a escolha do modelo final, foi levado em consideração o tempo de resposta do modelo, sendo que o modelo Naive Bayes Simples com Bag of Words, utilizando a base de sentimentos revistos, teve tempo de resposta de 0,18 segundos, com 73% de recall e o modelo Naive Bayes com Grid Search e Cross Validation com Bag of Words, utilizando a base original teve tempo de resposta de 0,35 segundos e, portanto, o primeiro foi a escolha final. O modelo foi exportado utilizando a biblioteca Pickle (<https://docs.python.org/3/library/pickle.html>) e testado com novos dados retirados dos comentários do perfil do BTG Pactual (@btgpacutal).

Os novos dados passam por um pipeline de execução, que realiza o pré processamento (tratamento de emojis, minúscularização das letras, tokenização e remoção de stop words), vetoriza a frase e, por fim, aplica o modelo. Abaixo é possível ver os resultados, sendo 0 negativo, 1, neutro e 2, positivo.

```
entrada1 = 'Melhor desempenho dentre todas as carteiras do mercado financeiro! Vamos para cima!!! 🎉📊'
```

```
pipeline(entrada1)
```

```
array([2])
```

```
entrada2 = 'não tem mais a opção atendimento via chat no app, só via email 😞 preciso de ajuda'
```

```
pipeline(entrada2)
```

```
array([0])
```

```
entrada3 = 'Vem pro Flu vem 📌📌📌'
```

```
pipeline(entrada3)
```

```
array([1])
```

Imagen 26 - Resultados do Modelo Final

7. Serviço

A etapa final do MVP consistiu na criação de uma API em Flask para ser integrada a um dashboard visual. A API foi desenvolvida para funcionar localmente, mas também pode ser implantada na nuvem. Com o objetivo de corresponder às visualizações do dashboard, diferentes respostas são devolvidas, como explicadas abaixo:

1. Classificação de comentários:

É realizado um pipeline de pré-processamento para as novas entradas de comentários. Esse pipeline envolve o tratamento de emojis, etapas de Processamento de Linguagem Natural (PLN) e vetorização. Em seguida, é aplicado um modelo escolhido, capaz de classificar cada comentário individualmente como positivo, neutro ou negativo. Por fim, a resposta corresponde ao autor do comentário, o texto e o sentimento predito.

2. Proporção dos sentimentos:

Após a classificação individual dos comentários, é realizado um cálculo para determinar as proporções de cada sentimento em relação ao total de comentários presentes na requisição. Essas proporções são utilizadas para exibir visualmente um gráfico representativo no dashboard.

3. Nuvem de palavras:

Nesta rota, os comentários passam pelo mesmo processo de pré-processamento e vetorização descrito anteriormente. Em seguida, é criada uma nuvem de palavras com base nesses dados e uma imagem é gerada e salva.

4. Top 10 palavras:

Assim como na rota anterior, as palavras são contabilizadas e ordenadas em ordem decrescente de frequência. Nessa rota, as 10 palavras mais frequentes, juntamente com o número de ocorrências, são exibidas.

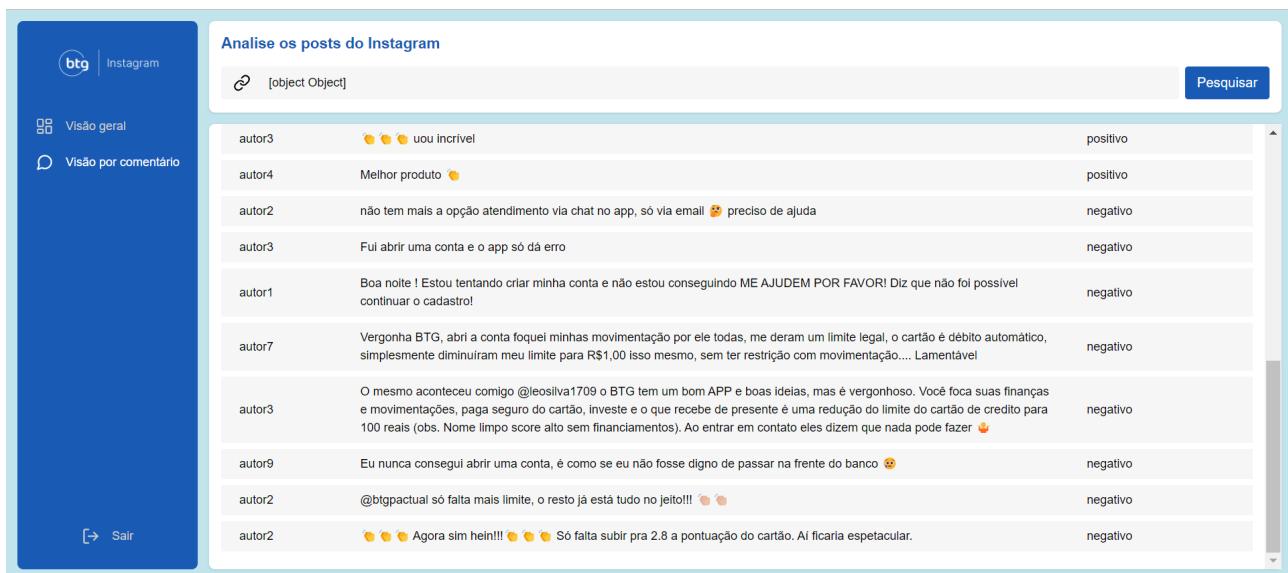
5. Top perfis

Essa rota segue a mesma lógica da rota de "Top 10 palavras". Sabendo os autores dos comentários, é contada a frequência de aparição de cada um e, então, feito um ranking decrescente.

Por meio dessas rotas, a API em Flask oferece funcionalidades de classificação de comentários, visualização de proporções de sentimentos, criação de nuvens de palavras, identificação das palavras mais frequentes e de autores mais frequentes. Essas informações foram integradas e exibidas no dashboard visual, demonstrado abaixo:



Imagen 27 - Dashboard Visão Geral



The screenshot shows a dashboard titled "Analise os posts do Instagram". On the left, there's a sidebar with a "btg" logo and "Instagram" text, followed by two buttons: "Visão geral" and "Visão por comentário". Below these buttons is a "Sair" link. The main area has a search bar with placeholder "[object Object]" and a "Pesquisar" button. The results table lists 10 comments from users "autor3", "autor4", "autor2", "autor3", "autor1", "autor7", "autor3", "autor9", "autor2", and "autor2". Each comment includes the author, some text, and a sentiment classification: positivo (positive), negativo (negative), or neutral. For example, the first comment from "autor3" is "uuu incrivel" and is classified as positivo.

Autor	Comentário	Sentimento
autor3	uuu incrivel	positivo
autor4	Melhor produto	positivo
autor2	não tem mais a opção atendimento via chat no app, só via email 🙄 preciso de ajuda	negativo
autor3	Fui abrir uma conta e o app só dá erro	negativo
autor1	Boa noite ! Estou tentando criar minha conta e não estou conseguindo ME AJUDEM POR FAVOR! Diz que não foi possível continuar o cadastro!	negativo
autor7	Vergonha BTG, abri a conta foquei minhas movimentação por ele todas, me deram um limite legal, o cartão é débito automático, simplesmente diminuíram meu limite para R\$1,00 isso mesmo, sem ter restrição com movimentação.... Lamentável	negativo
autor3	O mesmo aconteceu comigo @leosilva1709 o BTG tem um bom APP e boas ideias, mas é vergonhoso. Você foca suas finanças e movimentações, paga seguro do cartão, investe e o que recebe de presente é uma redução do limite do cartão de crédito para 100 reais (obs. Nome limpo score alto sem financiamentos). Ao entrar em contato eles dizem que nada pode fazer 🙄	negativo
autor9	Eu nunca consegui abrir uma conta, é como se eu não fosse digno de passar na frente do banco 😞	negativo
autor2	@btgpactual só falta mais limite, o resto já está tudo no jeito!!! 🌟🌟	negativo
autor2	Agora sim hein!!! 🌟🌟 Agora sim hein!!! 🌟🌟 Só falta subir pra 2.8 a pontuação do cartão. Ai ficaria espetacular.	negativo

Imagen 28 - Dashboard Visão por Comentário

Devido à uma limitação das bibliotecas que fornecem Web Scraping, os dados utilizados como entrada para test front-api foram estados manualmente, retirados do próprio perfil do Instagram @btgpactual, colocados em um arquivo JSON na pasta 'utils', dentro de 'front'. Para o correto uso, recomenda-se utilizar a API fornecida pelo próprio Instagram para o scraping, em que as credenciais de proprietário da conta são necessárias.

Para utilizar a API feita, os dados necessários são: comentários (nomeados 'dados' na rota) e autores dos comentários (nomeados 'authors' na rota). Esses dados devem estar em formato de lista, respectivamente, em JSON, para a correta interpretação e funcionamento da API.

O vídeo que demonstra todas as respostas da API pode ser consultado para maiores esclarecimentos: <https://youtu.be/MraHmdM0RPQ>

8. Referências

[1] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Proceedings of International Conference on Learning Representations Workshop (ICLR-2013).

[2] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.

[3] Ling, W., Dyer, C., Black, A., and Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

[4] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014), 12:1532–1543.

[5] MICROSOFT. Como entender o Automated ML. Disponível em:
<https://learn.microsoft.com/pt-br/azure/machine-learning/how-to-understand-automated-ml?view=azureml-api-2>. Acesso em: [19/06/2023].

[6] Géron, Aurélien."Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems". Acesso em: [19/06/2023]

[7] MICROSOFT. Resultados de Modelos de Machine Learning. Disponível em:
<https://learn.microsoft.com/pt-br/dynamics365/finance/finance-insights/confusion-matrix>. Acesso em: [19/06/2023].

