

APP LINGUAGEM NATURAL

BTG PACTUAL

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

APP LINGUAGEM NATURAL

BTG PACTUAL

Autores: Dayllan de Souza Alho

Eric Tachdjian

Gabriela de Moraes da Silva

Giovanna Furlan Torres

Lucas de Britto Vieira

Michel Mansur

Data de criação: 17 de Abril de 2023

SÃO PAULO – SP

2023

Sumário

Controle de Documento.....	8
Histórico de Revisões.....	8
1. Introdução.....	9
1.1 Parceiro de Negócios.....	9
1.2 Definição do Problema.....	10
1.2.1 Problema.....	10
2. Objetivos.....	11
2.1 Objetivos Gerais.....	11
2.2 Objetivos Específicos.....	11
2.3 Justificativa.....	12
2.3.1 Como o MVP suporta o problema do cliente.....	12
3. Compreensão do Problema.....	13
3.1 Análise de cenário: Matriz SWOT.....	13
3.2 Proposta de Valor.....	15
3.3 Matriz de Risco.....	15
3.4 Matriz Oceano Azul.....	16
3.6 Análise Financeira.....	19
4. Lean Inception.....	21
4.1 Visão do Produto.....	21
4.2 O Produto (É – Não É – Faz – Não Faz).....	21
4.3 Brainstorming de Funcionalidades.....	23
4.5 Revisão Técnica, de Negócios e de UX.....	24
4.6 Sequenciador.....	25
4.7 Canvas MVP.....	26
5. Análise de Experiência do Usuário.....	27
5.1 Personas.....	27
5.2 Jornada do Usuário.....	29
5.3 User Stories.....	31
6. Metodologia.....	40
6.1 CRISP-DM.....	40
6.1.1 Entendimento do negócio.....	40
6.1.2 Entendimento dos dados.....	41
6.1.3 Preparação dos dados.....	41
6.1.4 Modelagem.....	42
6.1.5 Avaliação.....	42
6.1.6 Implementação.....	42
6.1.7 Ferramentas.....	42
6.2 Compreensão dos dados.....	43
6.2.1 Descrição dos dados utilizados.....	43
6.3 Preparação dos dados.....	44
6.3.1 Exclusão de Colunas não utilizadas.....	44
6.3.2 Formatação de datas.....	45
6.3.3 Remoção Comentário BTG Pactual.....	46
6.3.4 Remoção de Aspas Duplas.....	46
6.3.5 Anonimização dos Autores.....	46
6.4 Pré-Processamento dos dados.....	46
6.4.1 Pipeline.....	47
7. Análise Descritiva.....	49
7.1 Dados Analisados.....	49
7.1.1 Quantidade de linhas na tabela.....	49
7.1.2 Quantidade de palavras na coluna texto.....	49

7.1.3 A distribuição de sentimentos positivos, negativos e neutros.....	50
7.1.4 Quantidade de autores na base de dados.....	52
7.1.5 Os usuários que mais realizaram comentários.....	52
7.1.6 Quantidade de comentários por Tipo de Interação.....	53
7.2 Análise de dados.....	55
7.2.1 Gráfico de Tendência Temporal dos Sentimentos.....	55
7.2.2 Gráfico de Dispersão para Sentimento e Autor.....	56
7.2.3 Gráfico de Nuvem de Palavras.....	57
8. Arquitetura Macro da Solução.....	59
8.1 Diagrama UML.....	61
9. Algoritmos.....	63
9.1 Bag Of Words.....	63
9.2 Word2Vec.....	63
9.2 TF-IDF.....	64
10. Modelagem.....	66
10.1 Naive Bayes.....	66
10.2 Embedding Layer.....	67
10.3 Rede Neural Recorrente (RNN).....	68
10.4 SVM (Support Vector Machine).....	68
10.5 Random Forest.....	70
10.6 Regressão Logística.....	70
11. Avaliação do Modelo.....	72
11.1 Divisão dos dados.....	72
11.2 Estratégia de Avaliação do modelo.....	72
11.2.1 Matriz de Confusão.....	73
11.2.2 Acurácia.....	74
11.2.3 Recall.....	74
12. Desenvolvimento e Resultados.....	76
12.1 BOW e Word2Vec.....	76
12.2 Naive Bayes.....	77
12.3 Embedding Layer.....	77
12.4 Rede Neural Recorrente (RNN).....	78
12.4 Regressão Logística (Vetorização TF-IDF).....	79
13. Novos Resultados.....	80
13.1 Novas Features.....	80
13.1.1 Aplicação das novas features no modelo RNN.....	81
13.1.2 Aplicação das novas features no modelo Naive Bayes.....	81
13.1.3 Aplicação das novas features no modelo SVM.....	82
13.1.4 Aplicação das novas features no modelo Random Forest.....	83
13.1.5 Comparação dos modelos utilizando as novas features.....	83
13.2 Hiperparâmetros.....	85
13.2.1 Random Search.....	85
13.2.1.1 Naive Bayes.....	85
13.2.1.2 RNN.....	86
13.2.1.3 Embedding Layer.....	86
13.2.1.4 Random Forest.....	86
13.2.1.5 Regressão Logística.....	86
13.2.2 Grid Search.....	87
13.2.2.1 SVM.....	87
13.3 Novo Subconjunto.....	87
13.3.1 Subconjunto com três categorias balanceadas.....	88
13.3.2 Subconjunto com três categorias não balanceadas.....	88

13.3.4 Escolha do subconjunto.....	89
13.4 Comparativo modelos.....	89
14. Funções.....	91
14.1 Pré-Processamento.....	91
14.2 Implementação.....	92
15. Conclusões e Recomendações.....	94
16. Referências.....	95
17. Anexos.....	98
17.1. Matriz de risco.....	98
17.2. Plano de gerenciamento de riscos.....	100
17.3. Arquitetura Macro.....	102

Índice de figuras

Figura 1: Representação SOWT.....	13
Figura 2: Proposta de valor.....	15
Figura 3: Matriz de Risco.....	15
Figura 4: Gráfico - Oceano Azul.....	18
Figura 5: O produto É.....	22
Figura 6: O produto NÃO É.....	22
Figura 7: O produto FAZ.....	22
Figura 8: O produto NÃO FAZ.....	23
Figura 9: Revisão Técnica, de Negócios e de Ux.....	24
Figura 10: Sequenciador.....	25
Figura 11: Canvas MVP.....	26
Figura 12: Persona - Analista de Automação.....	27
Figura 13: Persona - Analista de Marketing.....	28
Figura 14: Persona - Analista de Produto.....	28
Figura 15: Jornada de Usuário - Analista de Automação.....	29
Figura 16: Jornada do Usuário - Analista de Marketing.....	30
Figura 17: Jornada de Usuário - Analista de Produto.....	31
Figura 18: Metodologia CRISP-DM.....	40
Figura 19: Formatação das datas.....	45
Figura 20: Pipeline Ilustrativa.....	48
Figura 21: Quant Linhas.....	49
Figura 22: Quant Palavras.....	50
Figura 23: Distribuição sentimento.....	50
Figura 24: Quant Sentimento - Antes Processamento.....	51
Figura 25: Quant Sentimento - Depois Processamento.....	51
Figura 26: Quant Autores.....	52
Figura 27: Usuário x Comentário.....	53
Figura 28: Comentários X Interação.....	53
Figura 29: Comentário X Interação - Antes Processamento.....	54
Figura 30: Comentário X Interação - Depois Processamento.....	54
Figura 31: Tendência Temporal X Sentimento.....	55
Figura 32: Comentário X Sentimento.....	56
Figura 33: Nuvem de Palavras 1.....	57
Figura 34: Nuvem de palavras 2.....	58
Figura 35: Arquitetura Macro da Solução.....	59
Figura 36: Diagrama de implantação UML.....	61
Figura 37: Fórmula Naive Bayes.....	66
Figura 38: SVM - Exemplo de aplicação.....	69
Figura 39: Regressão Logística - Exemplo de aplicação.....	70
Figura 40: Matriz de confusão.....	73
Figura 41: Matriz de Confusão - Naive Bayes.....	77
Figura 42: Matriz de Confusão - Embedding Layer.....	78
Figura 43: Matriz de Confusão - RNN.....	78
Figura 44: Matriz de Confusão - Regressão Logística.....	79
Figura 45: Sentimento X Interação - Novas Features.....	80
Figura 46: Matriz de Confusão - RNN Novas Features.....	81
Figura 47: Matriz de Confusão - Naive Bayes Novas Features.....	82
Figura 48: Matriz de Confusão - SVM novas Features.....	82
Figura 49: Matriz de Confusão - Random Forest Novas Features.....	83
Figura 50: Gráfico - Comparação dos modelos – Sem as Novas Features.....	84

Figura 51: Comparação dos Modelos - Métrica Recall.....	90
Figura 52: Matriz de Confusão - Modelo Escolhido - Embedding Layer.....	90

Índice de tabelas

Table 1: Controle de documento.....	8
Tabela 2: Matriz de oceano azul.....	16
Tabela 3: Primeira - User Story.....	31
Tabela 4: Segunda - User Story.....	33
Tabela 5: Terceira - User Story.....	34
Tabela 6: Quarta - User Story.....	36
Tabela 7: Quinta - User Story.....	37
Tabela 8: Sexta - User Story.....	38
Tabela 9: Entendimento dos dados.....	43
Tabela 10: Comparação dos modelos - Novas Features.....	84
Tabela 11: Hiperparâmetros - Naive Bayes.....	85
Tabela 12: Hiperparâmetros - RNN.....	86
Tabela 13: Hiperparâmetros - Embedding Layer.....	86
Tabela 14: Hiperparâmetros - Random Forest.....	86
Tabela 15: Hiperparâmetros - Regressão Logística.....	86
Tabela 16: Hiperparâmetros - SVM.....	87

Controle de Documento

Histórico de Revisões

Table 1: Controle de documento

Data	Autor	Versão	Resumo da Atividade
17.04.2023	Giovanna Furlan Eric Tachdjian Dayllan Alho	1	Criação do documento; Estrutura e formatação; Matriz de Risco;
18.04.2023	Gabriela Silva Lucas Britto Giovanna Furlan Eric Tachdjian Dayllan Alho Gabriela Silva	1.1	Brainstormin de Funcionalidades; SWOT; Visão do Produto;
19.04.2023	Lucas Britto Giovanna Furlan Michel Mansur Eric Tachdjian Dayllan Alho Gabriela Silva	1.2	Objetivos (Geral, específicos, justificativas); Introdução, Parceiro de Negócios, Definição do Problema User Story;
24.04.2023	Lucas Britto Giovanna Furlan Michel Mansur Eric Tachdjian Dayllan Alho Gabriela Silva	1.2	Jornada do Usuário; Análise Financeira; Objetivos do Negócio; Personas;
26.02.2023	Lucas Britto Giovanna Furlan Michel Mansur	1.3	Matriz de avaliação – Oceano Azul; Revisão Técnica, de Negócios e UX; Sequenciador; Canvas Proposta de Valor; MVP Canvas; Arquitetura Macro da Solução;

1. Introdução

As redes sociais estão cada dia mais presentes no cotidiano dos indivíduos, e com essa crescente popularidade as empresas precisam se adequar a esse ambiente. Tais plataformas tornam possível a conexão entre as organizações e seus clientes, sendo um ponto alto para receber feedbacks sobre seus produtos e serviços. Antes, as organizações precisavam se limitar a pesquisas de mercado e estudos de opinião para entender as necessidades e desejos de seus clientes. Hoje, mais de 4,7 bilhões de pessoas têm acesso a essas redes, o que representa 59% da população mundial. Dito isso, o banco BTG Pactual, percebeu a importância dessa tendência e decidiu adotar uma abordagem inovadora para coletar e analisar os comentários dos usuários sobre suas campanhas.

Com mais de 1.642 milhões de clientes, o BTG Pactual, um dos maiores bancos de investimento da América Latina, reconheceu o impacto que as redes sociais podem ter sobre seus negócios, reconhecendo que uma das formas mais eficazes de melhorar seus produtos e serviços, é ouvindo seus clientes. Para isso, cada campanha de marketing passa por um processo criativo extenso para garantir que os resultados esperados pelas postagens sejam atingidos. A fim de conseguir tais insights, o BTG firmou uma parceria com a Inteli para utilizar o Processamento de Linguagem Natural (PLN) e criar um modelo de análise de sentimentos e detecção de palavras-chave em comentários nas redes sociais.

Com o modelo PLN em operação, o BTG Pactual poderá entender as opiniões dos usuários sobre determinadas campanhas de marketing e obter feedbacks valiosos. Essas informações permitirão que o banco amplifique os resultados de suas campanhas e melhore a qualidade de seus serviços para atender às necessidades dos clientes de forma mais eficiente. Além disso, será possível identificar tendências e padrões nos comentários dos usuários e, assim, ajustar suas campanhas em tempo real.

1.1 Parceiro de Negócios

O BTG Pactual é o maior Banco de investimentos da América Latina e atua nos mercados de Investment Banking, Corporate Lending, Sales & Trading, Wealth Management e Asset Management. Desde sua criação, em 1983, o BTG Pactual tem sido

administrado com base na cultura meritocrática de partnership, com foco no cliente, excelência e visão de longo prazo. O Banco se consolidou como uma das empresas mais inovadoras do setor, tendo conquistado diversos prêmios nacionais e internacionais. Atualmente, conta com quase 3 mil colaboradores em escritórios espalhados pelo Brasil, Chile, Argentina, Colômbia, Peru, México, Estados Unidos, Portugal e Inglaterra.

Os principais critérios para o desenvolvimento do projeto é a necessidade de analisar o desenvolvimento das campanhas de marketing da empresa, visando atrair maiores insights sobre padrões e tendências dos seus clientes.

1.2 Definição do Problema

Segue a definição do problema, com uma descrição clara e objetiva da questão ou desafio que precisa ser resolvido. Incluindo informações sobre o contexto, a natureza do problema e o impacto esperado da solução. Tal definição é necessária para colaborar na eficiência e eficácia, pois ajuda a direcionar esforços, recursos e tempo para solucioná-lo.

1.2.1 Problema

O BTG Pactual enfrenta o desafio de otimizar suas estratégias de marketing digital e entender melhor o comportamento e preferências dos consumidores nas redes sociais. Com o aumento do investimento em marketing digital e a crescente utilização das redes sociais, a análise de dados de mídia social é fundamental para obter informações relevantes e tomar decisões de negócios mais eficazes. O objetivo é utilizar PLN para rastrear dados e analisar a receptividade dos usuários às campanhas em redes sociais, identificar palavras-chave nos comentários e direcionar novas campanhas baseadas nos interesses dos consumidores.

2. Objetivos

Nesta seção, apresenta-se os objetivos do projeto que são as metas e resultados esperados a serem alcançados com a execução do mesmo. Servindo como uma referência para orientar as ações do projeto e ajudar a equipe a entender o que precisa ser feito e como avaliar o sucesso do projeto.

2.1 Objetivos Gerais

Sabendo que mais de 50% da população mundial que usa redes sociais por mais de 2 horas por dia e a crescente importância do marketing nas empresas (TAPI, 2023), o BTG Pactual em parceria com o Inteli está desenvolvendo o projeto de “Análise de Sentimento das Campanhas de Marketing em Redes Sociais”. Através da tecnologia de Processamento de Linguagem Natural (PLN), será desenvolvida uma ferramenta que ajudará a empresa a compreender a receptividade dos clientes às suas campanhas de marketing e nas tomadas de decisões das áreas de negócios, através da análise de sentimento e identificação de palavras-chave nos comentários dos usuários, permitindo uma resposta rápida a possíveis problemas ou oportunidades.

2.2 Objetivos Específicos

1. Realizar um pré-processamento dos dados, visando remover palavras irrelevantes ou duplicadas. Além da conversão dos dados não estruturados em estruturados.
2. Utilizar a técnica de análise de sentimentos, visando extrair informações dos comentários de redes sociais.
3. Utilizar técnicas de mineração de texto e processamento de linguagem natural para realizar a extração de palavras-chave.
4. Realizar a classificação ternária (positiva, negativa ou neutra) de campanhas de marketing.
5. Desenvolver uma interface de usuário para realizar o monitoramento e análise de campanhas, através dos resultados obtidos com o processamento de linguagem natural.

2.3 Justificativa

A implementação de um projeto de PLN para análise de sentimentos nos comentários de usuários em campanhas é uma estratégia essencial para aprimorar a gestão de marketing e otimizar as estratégias de negócios. Atualmente as campanhas se tornaram uma das principais formas de interação com o público-alvo, porém, gerenciá-las e analisar o feedback dos usuários manualmente é uma tarefa complexa e demorada.

Para otimizar tal atividade, se faz necessário a adoção do PLN, garantindo automatização e fornecendo insights valiosos sobre a percepção dos usuários em relação à marca dos produtos/serviços. Sendo possível entender a reação dos usuários às campanhas, o que permite ajustes e melhorias necessárias para tornar as campanhas mais efetivas e alinhadas com os interesses dos consumidores. Além de, com informações obtidas, a área de negócio pode tomar decisões mais precisas e embasadas, o que impacta diretamente nos resultados das campanhas e na percepção dos usuários sobre a marca.

2.3.1 Como o MVP suporta o problema do cliente

Utilizando técnicas avançadas de PLN, a solução analisa os textos das campanhas, como anúncios, conteúdo de mídia social, e-mails e textos promocionais, e atribui uma polaridade emocional a cada comentário. Dessa forma, possibilita a identificação e a percepção geral dos clientes em relação à marca, aos produtos ou às mensagens transmitidas nas campanhas.

Ao categorizar os comentários em positivos, negativos ou neutros, consegue-se observar uma visão clara sobre como as campanhas estão sendo recebidas pelo público-alvo. Isso possibilita uma avaliação objetiva do desempenho das estratégias de marketing e permite que sejam feitos ajustes com base nos resultados obtidos.

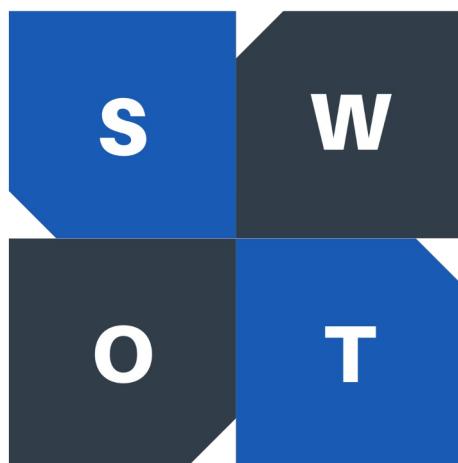
3. Compreensão do Problema

Apresenta-se nessa sessão as descrições das análises voltadas ao desenvolvimento de resultados do projeto, para a empresa BTG Pactual, a respeito da construção de um MVP (Produto mínimo viável) de um sistema de análise de sentimentos. Sendo exibido as identificações do mercado e produtos em comparação a solução prevista.

3.1 Análise de cenário: Matriz SWOT

A análise SWOT é uma ferramenta que possibilita a empresa a realizar análises de cenário ou de ambiente, sejam eles internos ou externos. Assim, é demonstrado as formas como ela atua no setor, suas fraquezas e forças (Iniciativas Internas), oportunidades e ameaças (iniciativas externas). A Figura 1, exibe uma imagem demonstrativa das quatro áreas que compõem a SWOT.

Figura 1: Representação SWOT



Fonte: Autores

I. Pontos Fortes:

- A empresa está presente em vários países, o que ajuda a obter dados de análise de texto de diferentes fontes;
- O BTG Pactual possui experiência em projetos de grande porte, o que pode auxiliar na implantação de projetos de NLP;

- Dispõe de uma equipa de colaboradores de diversas especialidades, que poderão ajudar na concretização do projeto.

II. Pontos Fracos:

- Este projeto pode ser difícil de realizar sem uma equipe de PNL dedicada;
- A empresa pode encontrar dificuldades em gerenciar a grande quantidade de dados gerados pelas redes sociais e garantir sua qualidade;
- Pode ser necessário investir em infraestrutura e tecnologia para dar suporte ao projeto.

III. Oportunidades:

- O uso do PLN pode ajudar as empresas a analisar rapidamente as respostas dos usuários nas mídias sociais. Aumentar o engajamento e a satisfação do cliente;
- As palavras-chave achadas nas avaliações dos usuários podem ajudar as empresas a criar campanhas mais eficazes;
- A análise de sentimentos pode ajudar as empresas a entender melhor as necessidades e tendências dos clientes.

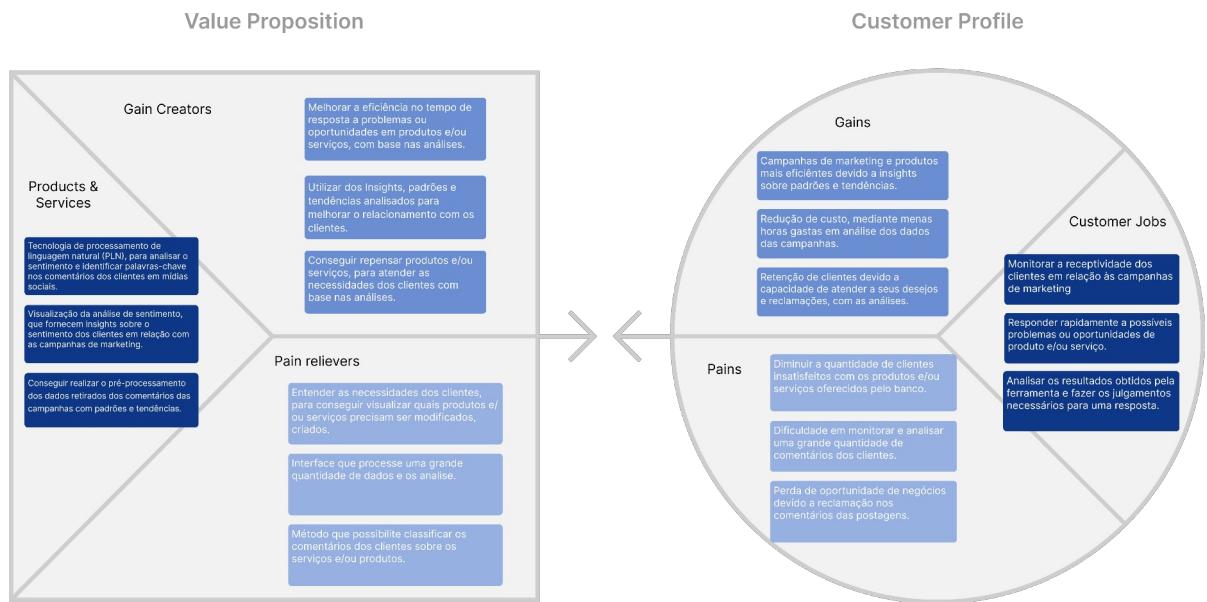
IV. Ameaças:

- Outras empresas podem usar o NLP para analisar dados de mídia social. o que aumenta a competição;
- Pode ser difícil para o PLN analisar todos os idiomas usados nas redes sociais, o que limita a quantidade de dados que podem ser analisados;
- Alterações políticas de privacidade de mídia social podem dificultar o acesso aos dados do usuário.

3.2 Proposta de Valor

A principal vantagem apresentada pela proposta de valor é conseguir auxiliar a empresa a compreender melhor os seus clientes e funcionários. Na Figura 2, é ilustrada a proposta construída para o BTG Pactual.

Figura 2: Proposta de valor



Fonte: Autores

3.3 Matriz de Risco

É uma das principais ferramentas na análise de negócios, utilizada para o gerenciamento de riscos operacionais existentes na empresa. A Figura 3, ilustra a construção da matriz de risco para o projeto e o plano de ação empregado.

Figura 3: Matriz de Risco

Matriz de Risco											
Probabilidade		Riscos					Oportunidade				
Muito Alta	1										
Alta	2					Entrega antecipada do produto final					
Médio	3										
Baixa	4										
Muito Baixa	5										
		1	2	3	4	5	5	4	3	2	1
		Muito Baixo	Baixo	Médio	Alta	Muito Alta	Muito Alta	Alta	Médio	Baixo	Muito Baixo
Impacto											

Fonte: Autores

MATRIZ	PROBABILIDADE	IMPACTO	DESCRIÇÃO DO IMPACTO	RESPONSÁVEL	MITIGAÇÃO
2-4	Alta	Alta	Com a sprint finalizando na sexta, e com a entrega antecipada para quarta feira, será necessário realizar as atividades em um menor período de tempo	Grupo inteiro	Ser responsável e finalizar todas as tarefas até terça, para validar tudo na quarta

Fonte: Autores

3.4 Matriz Oceano Azul

A Matriz de Oceano Azul é uma estratégia de negócios que ajuda a empresa a criar mercados, o “oceano azul” representa novos mercados ainda inexplorados, e a diferenciar-se da concorrência, aumentando a sua participação de mercado e, consequentemente, seu lucro.

Com base na solução proposta da Natural five, realizou-se a matriz de “oceano azul”, com base em 6 concorrentes, IBM Watson Natural Language Understanding, Google Cloud Natural Language, Google Cloud AutoML, Amazon Comprehend, Microsoft Azure Text Analytics e Python Natural Language Toolkit (NLTK). Apresenta-se na tabela 3 abaixo, os quesitos avaliados.

Tabela 2: Matriz de oceano azul

	IBM Watson Natural Language Understanding	Google Cloud Natural Language	Google Cloud AutoML	Amazon Comprehend	Microsoft Azure Text Analytics	Python Natural Language Toolkit (NLTK)	Natural Five
Maior preço	8	9	6	7	8	2	4
Maior usabilidade	8	9	9	9	9	6	10
Maior integração	9	8	8	9	9	6	10
Maior comodidade	8	10	10	10	10	3	7
Maior praticidade	7	9	9	9	9	6	9
Maior custo Benefício	6	6	6	6	6	8	8
Maior credibilidade	7	9	9	9	9	7	6
Maior Escalabilidade	7	8	8	9	7	8	9

Fonte: Autores

Abaixo se apresenta a descrição dos 8 atributos chave, sendo eles:

- Reduzir

A opção "Python Natural Language Toolkit" diminuiu o quesito "Maior preço" visto que ele é uma biblioteca gratuita, mesmo assim, também buscamos diminuir nossos custos em relação aos outros concorrentes. Assim como o aspecto de "comodidade", a opção "Python Natural Language Toolkit" diminui a praticidade do projeto pois ao usá-lo é necessário que o cliente tenha que criar por si próprio o modelo desejado, custando tempo, estrutura, planejamento e profissionais para desenvolver o modelo. Visto que todos os concorrentes são Big Tech's, diminuímos em relação a maior credibilidade, pois não possuímos a mesma estrutura de infraestrutura que elas, mesmo assim, acreditamos que não é uma perda, pois oferecemos funcionalidades personalizáveis focada no cliente.

- Eliminar

A NaturalFive optou por eliminar recursos de integração com diversas interfaces, incluindo a responsividade, para focar em uma entrega mais personalizada aos nossos clientes. Compreendemos que acrescentar essas funcionalidades poderia aumentar o preço e o tempo de entrega, o que não seria viável para atender às necessidades do mercado atual. Dessa forma, a eliminação desses recursos não foi sentida como uma perda pelos nossos clientes, pois nosso foco é oferecer soluções personalizadas e eficientes.

- Aumentar

A análise do quesito "Maior Comodidade" evidencia que a opção "Python Natural Language Toolkit" teve a menor pontuação, com uma nota 3, devido ao fato de ser uma biblioteca que requer maior conhecimento técnico para a criação de um modelo de PLN. Nesse sentido, a Natural Five buscou aumentar a comodidade do processo de utilização da solução, tornando-a mais fácil e acessível para usuários (Marketing, Produto e Automação) com diferentes níveis de conhecimento técnico. No quesito "Maior custo benefício" temos 2 principais agentes, "Python Natural Language Toolkit" e "Natural Five" visto que elevam o custo benefício por serem mais personalizáveis de acordo com as necessidades do cliente e serem mais baratas que seus concorrentes; além disso ao utilizá-los a empresa não ficará dependente de um terceiro como Google Cloud, IBM ou Amazon.

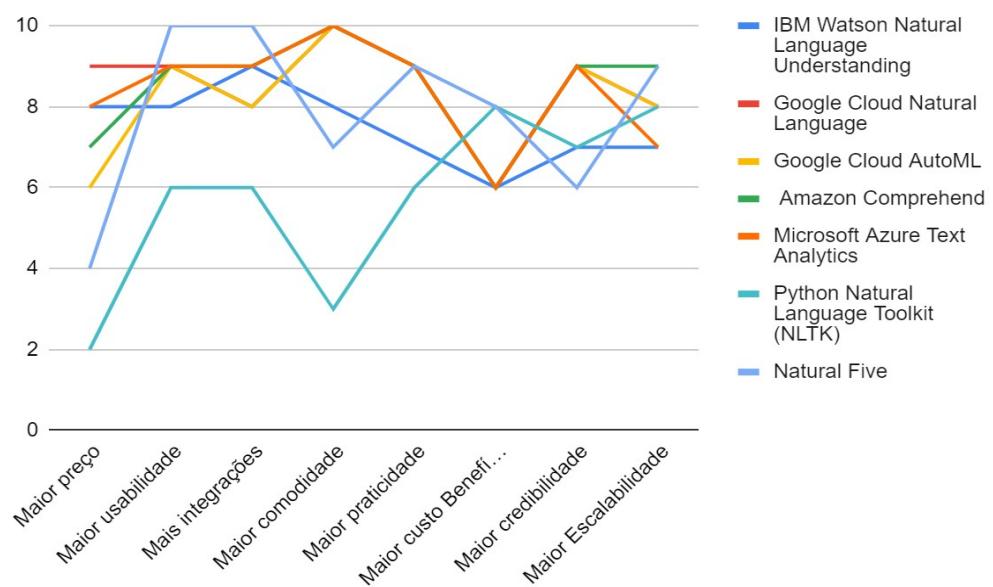
Além disso o quesito de “Maior Escalabilidade” é considerado alto comparado as outras empresas, o NaturalFive, consegue ser escalável e personalizado, podendo ser utilizado em outras redes sociais.

- Criar

Ao comparar com as grandes empresas de tecnologia que possuem soluções de PLN, podemos perceber que a Natural Five cria um nível de usabilidade personalizada muito superior para as equipes de marketing, produto e automação. Fornecendo um produto único para o cliente. Desenvolver uma plataforma pensando nas necessidades específicas de nossos clientes. Assim, fomos capazes de criar uma solução mais intuitiva, que se adapta melhor às necessidades dos usuários e é mais fácil de ser utilizada.

Esses atributos foram escolhidos com base na importância que o BTG Pactual atribui aos modelos de análise de linguagem natural como usar a tecnologia para facilitar esse processo de forma mais prática e barata, mantendo sempre uma alta qualidade. Apresenta-se abaixo um gráfico com a representação visual da tabela apresentada, na figura 4.

Figura 4: Gráfico - Oceano Azul



Fonte: Autores

Analisando o gráfico é possível perceber que a solução proposta do “Natural Five” se sobressai nas categorias de “Maior usabilidade” e “Mais integrações” ambas com nota

10/10. Por outro lado a solução se igualou com outras opções de mercado nas categorias “Maior praticidade” com nota 9/10 e “Maior custo-benefício” com nota 8/10 (ambas as notas podem ser consideradas altas). Porém a categoria “Credibilidade” é a única onde a Natural Five sai com menor pontuação "6/10".

Logo, pode-se concluir que o produto a ser desenvolvido pelo Natural Five se destaca na maioria das categorias e seu diferencial principal é por ele ser um produto mais personalizado para o cliente de forma prática e cômoda sem gerar grande gasto do cliente.

3.6 Análise Financeira

A Análise de custo das ferramentas utilizadas para a criação da solução, pode ser definida como uma estratégia adotada pelas empresas e desenvolvedores para o ponderamento do custo e benefício, visando obter maior domínio e exatidão dos gastos para a produção e implementação do serviço.

Estima-se para o desenvolvimento do MVP da solução os seguintes custos, participação de 2 desenvolvedores, hospedagem em um ambiente de cloud e tempo de desenvolvimento de 6 meses, totalizando em média de 250 à 300 mil reais. Tais dados, são apresentados, visando que o projeto fosse executado dentro da empresa, fornecidos pelo parceiro.

A projeção da receita não foi informada, portanto apresenta-se abaixo o raciocínio criado para realizar a estimativa do mesmo. Baseou-se os cálculos visando a receita, com o modelo atuando na retenção dos clientes do banco.

A retenção de clientes dentro de uma empresa, principalmente um banco, pode ser de total importância, visto que pode gerar receitas adicionais a longo prazo, uma vez que, quanto mais satisfeitos os usuários mais propensos eles estão em utilizar os produtos e/ou serviços da organização. Dito isso, espera-se que o sistema de análise de sentimentos contribua para essa retenção, fornecendo os insights necessários para que a satisfação aconteça.

Sendo assim, o BTG Pactual conta com 2,5 milhões de clientes, segundo o informativo de setembro de 2021. O valor médio das tarifas cobradas por bancos, em cada cliente que possui conta-corrente é de R\$ 23,45 por mês, de acordo com uma pesquisa realizada pelo Banco Central em 2021. Anualmente recebe-se R\$ 281,40 por

cliente, totalizando portanto uma receita de R\$ 703.500.000,00 levando em conta todos os clientes do banco.

Entretanto, em uma pesquisa da Associação Brasileira de Defesa do Consumidor, em 2020, constatou-se que cerca de 13% dos brasileiros, em média, anualmente trocam de banco por insatisfações. Com isso, o BTG Pactual perde 325 mil clientes, por ano, com insatisfações de produtos e/ou serviços. Perdendo em média R\$ 91.455.000,00 anuais com tarifas que poderiam ser cobradas de cada um desses clientes.

Ao calcular o Retorno Sobre Investimento (ROI) da solução, levando em base que a única receita vinda para o projeto é a retenção dos 13% dos clientes do banco, que agora com a aplicação do modelo, não trocariam mais de agência, por suas necessidades estarem sendo atendidas, e o custo sendo o fornecido pelos parceiros de 300 mil reais, temos:

OBS: Utilizou-se o valor de receita, para 6 meses, que é o tempo de desenvolvimento previsto. Ou seja, R\$ 91.455.000,00 dividido por 6, obtendo R\$ 15.242.500,00.

$$\text{ROI} = (\text{Receita} - \text{Custo}) / \text{Custo}$$

$$\text{ROI} = (15.242.500,00 - 300.000,00) / 300.000,00$$

$$\text{ROI} = 14.942.500,00 / 300.000,00$$

$$\text{ROI} = 49,808$$

$$\text{ROI (\%)} = 49,808 * 100$$

$$\text{ROI (\%)} = 4.980,8 \%$$

A análise financeira aponta com base no resultado apresentado pelo cálculo estimado do ROI, tal investimento gera a empresa um retorno bem-sucedido, mostrando que projeto é altamente rentável. Uma vez que, com isso à que a solução permite à empresa compreender melhor o público-alvo, identificar oportunidades de negócios e ainda atuar na retenção dos clientes que já possui.

4. Lean Inception

Nesta sessão, apresenta-se o Lean Inception, uma técnica baseada na metodologia ágil que visa definir o escopo e os requisitos do produto de forma colaborativa e eficiente, de todo o time e das partes interessadas na solução.

4.1 Visão do Produto

Responsável pela definição do objetivo principal do produto e como ele se encaixa na estratégia geral do negócio. Devendo ser clara, inspiradora e compartilhada por todas as partes interessadas. No parágrafo abaixo, segue a visão do produto, criada para o projeto de análise de sentimentos.

- Para o Banking and Trading Group (BTG), cujo à análise de dados de mídia social pode fornecer informações valiosas da criação de campanhas, produtos e/ou serviços. O projeto de Processamento de linguagem natural, para as áreas de automação, marketing e produto, é um, modelo capaz de analisar sentimentos e identificar palavras-chave nos comentários dos usuários com base nas postagens do instagram do banco, que ajuda a equipe de automação a analisar as suas campanhas realizadas durante o ano, e verificar além do seu alcance, a sensibilidade dos usuários. Diferentemente do GPT da OpenAI, do BERT do Google, da Siri da Apple e da Alexa da Amazon, o nosso produto oferece uma experiência única focada na cultura de inovação e foco no cliente, alinhada à cultura do BTG Pactual.

4.2 O Produto (É – Não É – Faz – Não Faz)

Definição das características principais do produto, especificando o que ele É e o que NÃO É, e o que ele FAZ e o que NÃO FAZ. Garantindo que todas as partes interessadas tenham uma compreensão comum do produto e evita mal-entendidos. Nas imagens 5, 6, 7 e 8 abaixo, exibe-se os critérios definidos para o produto, incluindo as condições as quais ele não atende.

Figura 5: O produto É

É	
1	Sistema de Linguagem Natural;
2	Análise de sentimentos nos posts do instagram;
3	Geração de insights para campanhas de marketing;
4	Direcionado ao banco BTG Pactual;
5	Sistema com possibilidade de filtro de palavras-chaves;
6	Ferramenta de auxílio estratégico para marketing e produto.

Fonte: Autores

Figura 6: O produto NÃO É

NÃO É	
1	Modelo preditivo;
2	Plataforma completa de gerenciamento de redes sociais;
3	Dashboard, Power BI;
4	A base de dados não é de código aberto;
5	Sistema de multi processamentos de linguagem natural;
6	Análise dos dados de todas as redes sociais;

Fonte: Autores

Figura 7: O produto FAZ

FAZ	
1	Oferece entendimento de padrões e tendências dos usuários;
2	Agrupamento de análises que coincidem.
3	Visualização dos dados em uma interface;
4	Considera na análise qualquer comentário textual;
5	Identifica palavras-chaves das campanhas e comentários;
6	Análise de sentimento dos comentários do instagram.

Fonte: Autores

Figura 8: O produto NÃO FAZ

NÃO FAZ	
1	Contempla ironia nas análises;
2	Rastreia dados de campanhas em tempo real;
3	Gera relatórios com insights das campanhas;
4	Análise de campanhas de marketing simultâneas;
5	Substitui opinião humana na tomada de decisão;
6	Contempla análise de interações pelo direct, nem GIF's.

Fonte: Autores

4.3 Brainstorming de Funcionalidades

Utilizado para gerar ideias de funcionalidades que o produto deve ter. Buscando garantir que o produto atenda às necessidades dos usuários e do negócio. Abaixo se exibe os *Clusters* criados para o mapeamento de funcionalidades.

- **Filtro:**

- Classificação de campanhas em positivas, negativa e/ou neutra;
- Visualizar palavras-chaves;
- Possibilidade de filtros para os dados.

- **Análise:**

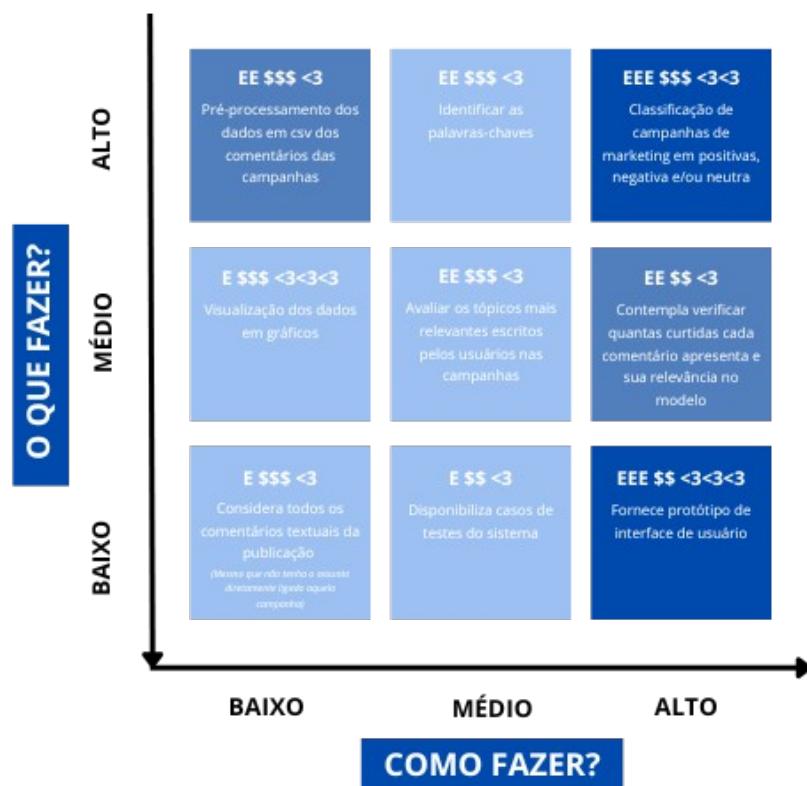
- Receptividade da campanha pelo usuário;
- Avaliar performance das campanhas;
- Analisar padrões de comportamento;
- Considera todos os comentários textuais da publicação (mesmo que não tenha o assunto diretamente ligado aquela campanha);
- Mensurar o nível de satisfação dos clientes;
- Contempla verificar quantas curtidas cada comentário apresenta e sua relevância no modelo;
- Disponibiliza casos de teste do sistema;

- Variação da polarização antes e depois da campanha;
- Visualizar o conteúdo completo dos comentários.
- **Visualização do projeto:**
 - Visualização dos dados em dashboards e/ou gráficos;
 - Fornece protótipo de interface de usuário.
- **Implementação extra:**
 - Rastrear dados em tempo real dos comentários das campanhas.

4.5 Revisão Técnica, de Negócios e de UX

Utilizado para revisar as funcionalidades geradas na etapa anterior sob as perspectivas técnicas, de negócios e de experiência do usuário (UX). Buscando garantir que as funcionalidades sejam viáveis do ponto de vista técnico, agreguem valor ao negócio e proporcionem uma boa experiência ao usuário. Na figura 9 abaixo, exibe-se a estrutura criada e categorizada entre esforço (E), negócio (\$) e UX (<3) para o projeto.

Figura 9: Revisão Técnica, de Negócios e de Ux



Fonte: Autores

4.6 Sequenciador

Técnica utilizada para priorizar as funcionalidades e requisitos do produto. Ele é uma ferramenta de classificação simples que permite que as equipes definam a ordem de execução das funcionalidades com base em critérios pré-determinado, na figura 10 abaixo, exibe-se a estrutura de sequência criada para a solução.

Figura 10: Sequenciador



Fonte: Autores

4.7 Canvas MVP

Ferramenta para estimar qual é o produto, com o menor conjunto de recursos que ainda atende às necessidades básicas dos clientes. Além de determinar como testar e validar o produto com os clientes. Na figura 11 abaixo se apresenta o Canvas MVP criado para a solução.

Figura 11: Canvas MVP

Personas Segmentadas	Proposta MVP	Resultados Esperados
<p>Personas Segmentadas</p> <ul style="list-style-type: none">• Alice Macedo Analista de Marketing;• Eduardo Júnior Analista de Automação;• Marcos Almeida Analista de produtos.	<p>Proposta MVP</p> <p>Criar um modelo de processamento de linguagem natural para análise de sentimentos em comentários em redes sociais, para encontrar padrões e tendências dos usuários.</p>	<p>Resultados Esperados</p> <p>O resultado esperado é que seja criado um modelo de processamento linguagem natural capaz de fazer classificações positivas, negativas e neutras, fornecendo dados sobre as campanhas de marketing realizadas pelo BTG.</p>
<p>Jornadas</p> <ul style="list-style-type: none">• Alice Macedo Recebimento e visualização das análises, melhoria na criação das campanhas de marketing mais personalizadas;• Eduardo Júnior Análise dos dados das campanhas e direcionamento dos insights;• Marcos Almeida Recebimento e visualização das análises, melhoria na criação e/ou modificação dos serviços e produtos oferecidos pelo banco.	<p>Funcionalidades</p> <ul style="list-style-type: none">• Avalia os tópicos mais relevantes escritos pelos usuários nas campanhas;• Classificação de campanhas em positivas, negativa e/ou neutra;• Identifica palavras-chaves	<p>Métricas validação de hipótese</p> <p>O conceito será abordado em módulos posteriores.</p>
	<p>Custo e Cronograma</p> <ul style="list-style-type: none">• Custo: Aproximadamente 300 mil reais, com 2 desenvolvedores e 1 ambiente em cloud.• Cronograma: 6 meses de desenvolvimento do projeto.	

Fonte: Autores.

5. Análise de Experiência do Usuário

Nesta sessão, apresenta-se a análise de experiência do usuário, a qual através da aplicação de estratégias, visa compreender como os usuários interagem com sistemas, produtos e serviços. O objetivo é melhorar a satisfação e a eficiência dessas interações, levando em conta aspectos subjetivos como emoções, percepções e expectativas dos usuários.

5.1 Personas

As personas do projeto são baseadas em 3 setores principais, sendo eles, 1) Analista de Automação; 2) Analista de Marketing; 3) Analista de produto. Estes representam a ideia de cliente ideal, porém fictícia, e os dados apresentados (comportamentos e características), são equivalentes ao contexto em que o BTG Pactual se encontra. As Figuras 12, 13 e 14 exibem as personas construídas.

Figura 12: Persona - Analista de Automação

Eduardo Junior
Analista de Automação

- 29 anos;
- R\$ 20 mil ao mês;
- Apaixonado por jogos FPS;
- Casado;

01 Realiza a coleta dos feedbacks (dados) nas redes sociais e plataformas do banco.

02 Realiza análises estatísticas buscando padrões e tendências nos comentários.

03 Acredita que poderia haver uma melhoria das funcionalidades (filtros, visualização, entre outras)

04 Busca fornecer relatórios e insights mais precisos para outras áreas.

Fonte: Autores

Figura 13: Persona - Analista de Marketing



Alice Macedo Analista de Marketing

- 25 anos;
- R\$ 12 mil ao mês;
- Gosta de arte e filmes;
- Solteira;



- 01 Realiza a análise dos dados das rede sociais e plataformas do banco.
- 02 Desenvolve as estratégias de marketing.
- 03 Gostaria de ter uma visualização mais clara e objetiva dos feedbacks.
- 04 Busca realizar o monitoramento contínuo dos resultados das campanhas de marketing.

Fonte: Autores

Figura 14: Persona - Analista de Produto

Marcos Almeida Analista de Produtos

- 37 anos;
- R\$ 17 mil ao mês;
- Fã do automobilismo;
- Solteiro;



- 01 Realiza a análise de mercado antes da criação de um produto.
- 02 Responsável pelo gerenciamento do ciclo de vida de um produto.
- 03 Busca entender as necessidades dos clientes para aprimorar os produtos
- 04 Gostaria de uma ferramenta que facilitasse a identificação de tendências.

Fonte: Autores

5.2 Jornada do Usuário

A jornada do usuário construída consiste na representação das etapas principais que envolvem 1) Análise dos dados das campanhas; 2) Visualização da análise e melhoria nas campanhas de marketing; e 3) Visualização da análise e melhoria na criação e/ou modificação dos produtos e serviços do banco. Divididas em 3 estruturas, exibidas nas figuras 15, 16 e 17 sendo elas respectivamente:

1. Analista de Automação;
2. Analista de Marketing;
3. Analista de produto.

Figura 15: Jornada de Usuário - Analista de Automação



Oportunidade : Conseguir otimizar recursos, além de aproveitar possíveis oportunidades na empresa, devido a eficiência de seu trabalho.

Responsabilidade : Necessidade de ter ciência de todos os novos dados que chegam à empresa, fazendo suas análises e automatizações dos processos.

Fonte: Autores

Figura 16: Jornada do Usuário - Analista de Marketing

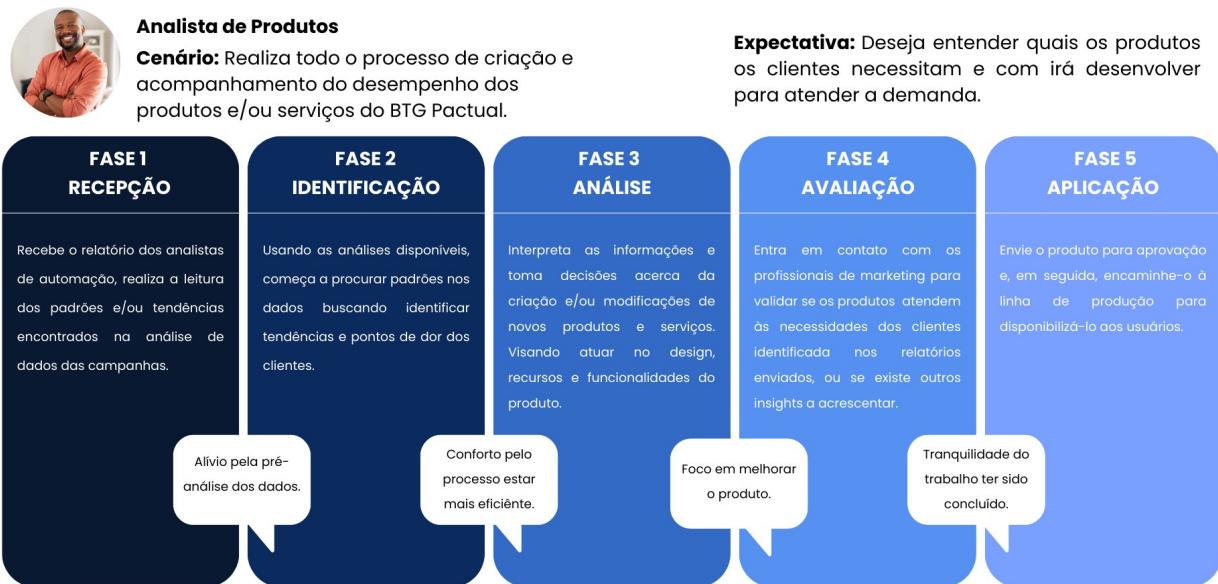


Oportunidade : Conseguir otimizar as campanhas de marketing da empresa, melhorando seu desempenho e se destacando. De forma que possa evoluir profissionalmente.

Responsabilidade : Criar campanhas de produtos e/ou serviços do banco para os usuários nas redes sociais.

Fonte: Autores

Figura 17: Jornada do Usuário - Analista de Produto



Oportunidade : Entender os desejos dos clientes melhorando a qualidade do produto e a experiência do usuário, sendo reconhecido em seu trabalho.

Responsabilidade : Garantir que as informações sobre novos produtos e/ou serviços sejam dissipadas entre os envolvidos, para realizar um melhor aprimoramento e criação.

Fonte: Autores

5.3 User Stories

Pode-se definir *User Stories* como descrições simplificadas das funcionalidades possíveis que o usuário possui e deseja dentro da aplicação, escrita com a visão dele. Além de transparecer como o sistema espera alcançar tais objetivos. As tabelas abaixo estão divididas em 6 partes: Número, Título, Personas, História, Critérios de Aceitação e Testes de Aceitação. O número e título servem para identificação, já as personas servem para associar a quem a história pertence. Os dois últimos tópicos descrevem quais são os critérios que aquele usuário deve passar no sistema para realizar a ação descrita na “história”, já o teste diz como o sistema deve agir de acordo com o critério estipulado. Nas tabelas 4,5,6,7,8 e 9 abaixo.

Tabela 3: Primeira - User Story

Título	Validação de Dados para Processamento de Modelo de PLN
Persona	Analista de Automação
História	Eu, como analista de automação, quero um ambiente que me permita validar se todas as colunas do arquivo CSV estão sendo processadas, lidas, limpas, corrigidas e utilizadas corretamente para gerar o modelo de PLN via IA.
Critério de aceitação	<p>Eu quero visualizar todas as colunas do arquivo CSV para verificar se estão corretas.</p> <ul style="list-style-type: none">• {condição: todas as colunas do arquivo CSV são exibidas corretamente}• {pós-condição: Recusa de exibição de colunas e retorno de feedback negativo} <p>Eu quero a capacidade de selecionar quais colunas devem ser processadas para gerar o modelo de PLN.</p> <ul style="list-style-type: none">• {condição: as colunas selecionadas são processadas corretamente}• {pós-condição: Recusa de processamento e retorno de feedback negativo}

	<p>Eu quero a capacidade de visualizar e corrigir dados incorretos ou faltantes nas colunas.</p> <ul style="list-style-type: none"> • {condição: os dados são exibidos corretamente e as correções são realizadas com sucesso} • {pós-condição: Recusa de exibição e correção de dados e retorno de feedback negativo}
Teste de aceitação	<p>Critério 1: Exibir todas as colunas do arquivo CSV.</p> <ul style="list-style-type: none"> • Aceitou: correto, todas as colunas são exibidas corretamente. • Recusou: errado, algumas colunas não são exibidas corretamente. <p>Critério 2: Processar as colunas selecionadas com sucesso.</p> <ul style="list-style-type: none"> • Aceitou: correto, as colunas selecionadas são processadas com sucesso. • Recusou: errado, as colunas selecionadas não são processadas corretamente. <p>Critério 3: Exibir e corrigir dados incorretos ou faltantes.</p> <ul style="list-style-type: none"> • Aceitou: correto, os dados são exibidos e as correções são realizadas com sucesso. • Recusou: errado, os dados não são exibidos e as correções não são realizadas.

Fonte: Autores

A primeira User Story está direcionada para a primeira etapa a ser cumprida pelo time de desenvolvimento, focado em processamento dos dados, com sua leitura, limpeza e correção para que sejam interpretadas por uma IA, partindo assim para a segunda User Story.

Tabela 4: Segunda - User Story

Título	Análise de feedbacks
Persona	Analista de automação
História	Eu, como analista de automação, quero um processo para

	abrir um dashboard, para poder analisar o desempenho das campanhas.
Critério de aceitação	<p>Meio para acessar um dashboard com os números da campanha.</p> <ul style="list-style-type: none"> • {condição: Gerar dashboard} <p>O dashboard deve permitir a visualização clara e organizada dos feedbacks positivos, negativos e neutros.</p> <ul style="list-style-type: none"> • {condição: Seleção de feedbacks positivos, negativos e neutros} • {pós-condição: Recusa de seleção}
Teste de aceitação	<p>Critério 1: Ao gerar dashboard, renderizar o modo de visualização desejado.</p> <ul style="list-style-type: none"> • Aceitou: correto, demonstrar modo de visualização. • Recusou: errado, retornar feedback negativo de demonstração de visualização.

Fonte: Autores

A segunda User Story tem como objetivo a implementação de uma interface gráfica de usuário para exibir os resultados gerados pela IA a partir dos dados processados na primeira User Story. Essa interface deve ser intuitiva e fácil de usar, permitindo que os usuários possam visualizar e interagir com as informações de forma clara e objetiva. Para isso, o time de desenvolvimento trabalhará na construção de uma interface com recursos de seleção e visualização dos dados, a fim de atender às necessidades dos usuários finais.

Tabela 5: Terceira - User Story

Título	Análise de Sentimento e Palavras-chave Automatizada
Persona	Analista de automação
História	Como analista de automação, eu gostaria de construir um modelo de análise de sentimento e palavras-chave

	automatizado para os comentários do Instagram do BTG Pactual, a fim de fornecer insights valiosos para o analista de produto.
Critério de aceitação	<p>A interface deve possuir uma opção para visualizar a análise de sentimento e palavras-chave de cada campanha de marketing.</p> <ul style="list-style-type: none"> {condição: Ter a opção de análise de sentimento e palavras-chave} <p>A análise de sentimento deve ser realizada através de um modelo de análise de sentimento construído pelo analista de automação.</p> <ul style="list-style-type: none"> {condição: Ter um modelo de análise de sentimento} {pós-condição: Retornar o resultado do modelo para o usuário} <p>A análise de palavras-chave deve ser realizada através de uma lista de palavras-chave previamente definida pelo analista de produto.</p> <ul style="list-style-type: none"> {condição: Ter uma lista de palavras-chave definida} {pós-condição: Retornar a lista de palavras para o usuário}
Teste de aceitação	<p>Critério 1: O analista de produto seleciona a opção de visualização de análise de sentimento e palavras-chave de uma campanha específica.</p> <ul style="list-style-type: none"> Aceitou: correto, e apresenta as informações de análise de sentimento e palavras-chave da campanha. Recusou: errado, e não apresenta as informações.

	<p>Critério 2: O analista de produto verifica se a análise de sentimento foi realizada através do modelo de análise de sentimento construído.</p> <ul style="list-style-type: none"> • Aceitou: correto, e apresenta as informações realizadas através do modelo de análise de sentimento construído. • Recusou: errado, retorna erro e rever procedimento de modelagem da análise de sentimento.
--	---

Fonte: Autores

A terceira User Story tem como objetivo construir um modelo de análise de sentimento e palavras-chave automatizado para os comentários do Instagram do BTG Pactual, a fim de fornecer insights valiosos para o analista de produto. O analista de automação será o responsável por criar esse modelo. Nossa interface realizará a análise de sentimento e demonstrar quais são as palavras-chaves, indicando para quem for utilizar quais são elas. Os testes de aceitação incluem verificação da correta apresentação das informações e uso dos modelos definidos.

Tabela 6: Quarta - User Story

Título	Análise de comentários do instagram
Persona	Analista de marketing
História	Como analista de marketing, eu gostaria de analisar os comentários do Instagram do BTG Pactual através de uma plataforma, a fim de avaliar a efetividade de campanhas específicas.
Critério de Aceitação	<p>A plataforma deve permitir a seleção de campanhas.</p> <ul style="list-style-type: none"> • {condição: Ter a seleção baseada nos dados disponibilizados na base} • {pós-condição: retornar feedback de erro} <p>A plataforma deverá apresentar gráficos e estatísticas para auxiliar na visualização.</p>

	<ul style="list-style-type: none"> {condição: ter uma forma de visualização} {pós-condição: retornar feedback de que não tem forma de visualização}
Teste de Aceitação	<p>Critério 1: O analista seleciona a seleção que não existe na base dos dados.</p> <ul style="list-style-type: none"> Aceitou: errado, deve aparecer uma mensagem de erro e indicar que a seleção não existe. Recusou: Correto, informando que seleção escolhida não existe. <p>Critério 2: O analista acessa e disponibiliza o processo de visualização.</p> <ul style="list-style-type: none"> Aceitou: correto, e abre o modo de visualizar com o resultado da seleção selecionada. Recusou: errado, rever processo.

Fonte: Autores

A quarta User Story voltada para o analista de marketing, busca permitir a análise por meio de uma interface que permita a seleção de campanhas baseada nos dados disponibilizados na base. Além disso, a interface deve apresentar gráficos e estatísticas para auxiliar na visualização, garantindo uma análise mais eficiente e precisa. Os critérios de aceitação dessa User Story garante que a plataforma apresente feedback adequado ao usuário, informando caso a seleção escolhida não exista e permitindo a visualização dos resultados de maneira clara e objetiva.

Tabela 7: Quinta - User Story

Título	Realiza uma análise de mercado.
Persona	Analista de produto
História	Eu, como analista de produto, quero analisar o rendimento de cada campanha de marketing, para obter insights do produto, e realizar uma análise de sentimentos e palavras chaves.
Critério de aceitação	Quero selecionar as campanhas e os produtos na mesma pesquisa.

	<ul style="list-style-type: none"> {condição: ter a possibilidade de selecionar mais de uma opção} {pós-condição: retornar erro de seleção e rever processo.}
Teste de aceitação	<p>Critério 1: O analista vai selecionar duas ou mais opções (de campanha e de produto) que existem ao mesmo tempo.</p> <ul style="list-style-type: none"> Aceitou: correto, e apresentar a representação visual escolhida. Recusou: errado, retornar um erro.

Fonte: Autores

Na quinta User Story, o analista de produto deseja analisar o rendimento de cada campanha de marketing para obter insights do produto, além de realizar uma análise de sentimentos e palavras-chave. É importante que a interface permita a seleção de várias opções de campanha e produto ao mesmo tempo na mesma pesquisa.

O critério de aceitação estabelece que a interface deve permitir a seleção de várias opções de campanha e produto na mesma pesquisa, e caso isso não seja possível, deve-se retornar um erro e o processo deve ser revisto. Portanto, a interface deve ser capaz de lidar com seleções múltiplas de campanha e produto e apresentar resultados precisos e coerentes com os critérios de aceitação estabelecidos.

Tabela 8: Sexta - User Story

Título	Analizar resultados das campanhas de marketing.
Persona	Analista de marketing
História	Eu, como analista de marketing, quero uma plataforma que tenha como selecionar a visualização de produtos e sentimento, para saber o andamento da campanha.
Critério de aceitação	<p>Seleção de forma de visualização.</p> <ul style="list-style-type: none"> {condição: selecionar visualização de sentimento e produto.} {pós-condição: retornar feedback com erro para esta visualização, e rever processo}

Teste de aceitação	Critério 1: O analista visualiza quais são os modos de visualização disponíveis e os seleciona. <ul style="list-style-type: none"> • Aceitou: correto e demonstra as informações. • Recusou: errado, e retorna erro para o usuário.
---------------------------	--

Fonte: Autores

A sexta User Story se trata da necessidade do analista de marketing em ter uma plataforma que permita selecionar a visualização de produtos e sentimentos para analisar o andamento de uma campanha de marketing. O critério de aceitação envolve a seleção da forma de visualização e a apresentação das informações de forma adequada. O time de desenvolvimento trabalhará para atender essas necessidades e garantir a satisfação do usuário final.

A partir das definições de todas as User Stories definidas para este projeto, foca-se na hierarquização das tarefas e priorização de cada uma delas, a fim de atender aos critérios de aceitação definidos. Dessa forma, completa-se cada User Story uma a uma, garantindo que o projeto seja entregue com todos os requisitos cumpridos e dentro do prazo estabelecido.

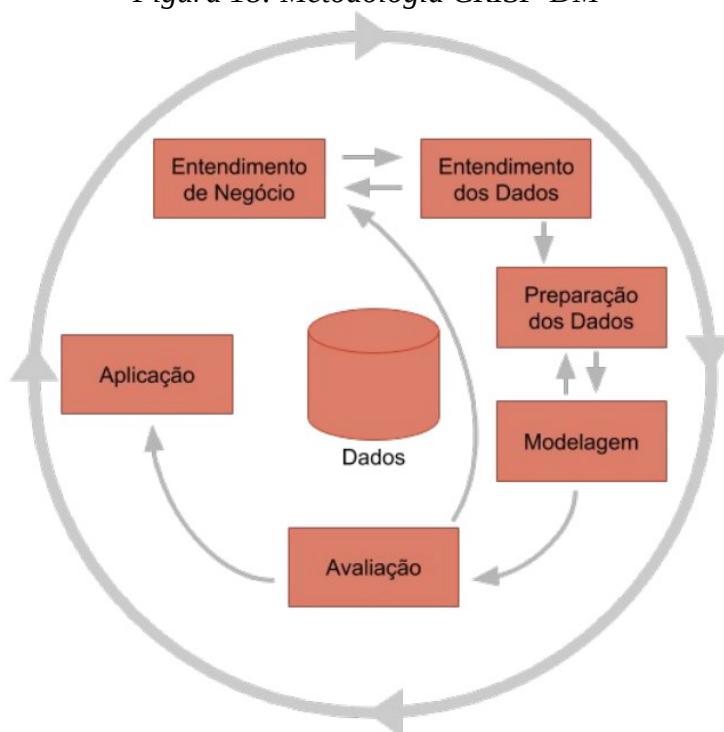
6. Metodologia

Nesta sessão é apresentado as metodologias utilizadas como base para a criação do modelo de processamento de linguagem natural como um todo.

6.1 CRISP-DM

Exibe-se as etapas que correspondem a metodologia do CRISP-DM. Na figura 18 abaixo, encontra-se uma imagem ilustrando como funciona a sequência de processos a serem exercidos quando a metodologia é implementada.

Figura 18: Metodologia CRISP-DM



Fonte: Medium 2019

6.1.1 Entendimento do negócio

Busca-se ter uma visão clara do problema que se precisa resolver, é nesta fase que se deve traçar os objetivos do negócio, buscar mais detalhes do problema, listar os recursos disponíveis e o impacto esperado. Tem como características estabelecer métricas e os critérios quantitativos para os possíveis resultados. Priorizando aqueles que

influenciam sua meta e também criar uma análise da vantagem do projeto, além do custo-benefício. Define-se os modelos, relatórios, apresentações e os dados.

6.1.2 Entendimento dos dados

Nesta segunda fase, se obtêm os dados e verifica-se se eles são adequados às suas necessidades. É importante ter feito uma boa fase 1, para que nesta fase não tenha que revisar o entendimento do negócio, nem repensar metas e planos. Os objetivos desta fase são coletar os dados, descrevê-los, explorá-los e verificar a qualidade dos mesmos. Estabelecer formato para esses dados, é possível que seja necessário reunir novos dados, enfrentar limitações de software ou hardware. E encontrar imperfeições nos dados. Na parte da documentação é importante estabelecer o feature selection, especificar os campos relevantes e criar uma descrição geral dos dados que possui, assim como os formatos, variáveis, técnicas estatísticas e qualquer informação que possa ser relevante. É o lugar para criar, testar e documentar hipóteses geradas após a exploração dos dados.

6.1.3 Preparação dos dados

Agora que a maioria dos dados usados já foram coletados, necessita de refinamento antes de ser usado na modelagem. Esta fase possui cinco principais tarefas:

1. Selecionar os dados: É o momento de justificar quais dados serão ou não utilizados, documentar a relevância desses para seu objetivo, os problemas técnicos,
2. Limpar esses dados: Corrigir alguns dados específicos, excluir ou substituir por valores padrões para uma técnica de modelagem mais sofisticada.
3. Documentar bem detalhadamente os processos utilizados nesta etapa e o possível impacto gerado por essa escolha
4. Construção dos dados: Criar campos e documentá-los explicando os motivos.
5. Integração dos dados: Diversos conjuntos de dados, para mesclá-los e prepará-los para a fase de modelagem. Formatar os dados para o formato mais conveniente para o projeto.

6.1.4 Modelagem

Nesta fase serão escolhidas as técnicas mais adequadas para modelagem, ou seja, está etapa envolve a seleção e a utilização de técnicas e algoritmos que atendam as necessidades do negócio. Geralmente os dados são divididos em duas partes: um de treino (que são gerados os modelos) e um de teste (que se refere a validação do modelo). Com base nisso, é definido se continua o desenvolvimento da modelagem (avaliação) ou se retorna para a fase de preparação de dados.

6.1.5 Avaliação

Nesta fase será avaliada a qualidade e a segurança dos resultados obtidos na etapa anterior. De modo que seja possível verificar se esse resultado corresponde às expectativas do projeto. Caso não atenda, devem ser realizadas as modificações necessárias (como correção na entrada de dados, correção no tratamento dos atributos, entre outros).

6.1.6 Implementação

Nesta fase é realizada o desenvolvimento dos modelos criados e avaliados. Durante essa etapa são realizadas tarefas, como: implantação da solução, monitoramento e manutenção, geração de relatórios e avaliação os resultados finais. Vale ressaltar que essa forma de implementação depende do tipo de modelo e projeto. Além disso, é preciso que o usuário final consiga interpretar e operar o produto com facilidade

6.1.7 Ferramentas

As ferramentas utilizadas para a construção da solução, consiste em aquelas utilizadas para o desenvolvimento, organização e compartilhamento de arquivos. Primeiramente, definiu-se uma ferramenta para a organização, tendo como base o aplicativo Notion, que permite organizar, através de cards, todas as tarefas da equipe, sendo possível visualizar o que está sendo feito pelos integrantes e gerenciar as entregas já concluídas. Em paralelo a isso, tem-se a ferramenta de desenvolvimento. Para isso, utilizou-se o Google Collaboratory, onde criou-se o notebook do projeto, o qual é utilizado para criação, organização e execução do código. As ferramentas de compartilhamento de arquivos. Para os arquivos de desenvolvimento do trabalho, é utilizado o Google Drive,

que possui integração com o Google Collaboratory. Assim, sendo possível compartilhar em tempo real os arquivos referentes ao desenvolvimento. E por fim, é utilizado o Github, que possibilita compartilhar todos os arquivos do projeto, referente a descrição, organização e desenvolvimento em um ambiente que será possível ter uma visão ampla do que foi desenvolvido

6.2 Compreensão dos dados

As sessões abaixo apresenta o conjunto de dados trabalhado, seus principais atributos, descrições e análises estatísticas.

6.2.1 Descrição dos dados utilizados

Neste tópico apresenta-se os dados disponibilizados na “2-base_10052023”. A base de dados a ser trabalhada foi disponibilizada pela empresa BTG Pactual, possuindo 12355 linhas de conteúdo. Na tabela 9 abaixo, é descrita os principais atributos, suas descrições e tipos da planilha.

Tabela 9: Entendimento dos dados

Atributo	Descrição	Tipo
dataPublicada	Data de publicação da postagem a qual o comentário foi feito.	datetime64[ns]
autor	Pessoa a qual realizou o comentário na publicação.	object
texto	Comentários realizados até 11 de Dezembro de 2022.	object
sentimento	Classificação de sentimento dos comentários (positivo, negativo e neutro).	object
tipointeracao	Categoria a qual remete-se o texto do comentário (marcação, comentário e resposta).	object
anomalia	Classificação se o comentário pode ser malicioso, um golpe, em binário 0 ou 1.	int64
probabilidadeAnomalia	Classificação em porcentagem do quanto aquele comentário pode ser malicioso.	int64
linkPost	Fornece o link do post do instagram a qual o comentário foi realizado.	object
processado	Retorna se a análise de sentimentos já ocorreu naquela linha.	int64
contemHyperlink	Retorna se o comentário possui algum hyperlink.	int64

Fonte: Autores

6.3 Preparação dos dados

Nesta seção, colocou-se em prática a análise exploratória dos dados fornecidos pela empresa, de acordo com a execução proposta pelo modelo CRISP-DM, a fim de tornar nossa base de dados mais adequada para o sistema de processamento de linguagem natural para análise de sentimentos.

6.3.1 Exclusão de Colunas não utilizadas

Nesta sessão, descrevemos as colunas relevantes selecionadas para o modelo de análise de sentimentos, como data de publicação, autor, texto, sentimento e tipo de interação. Também explicamos por que colunas como anomalia, probabilidade de anomalia, link do post, processado e contém hyperlink não são relevantes para a análise de sentimentos proposta.

Colunas Utilizadas:

1. dataPublicada: A data de publicação é relevante, pois permite identificar a cronologia dos comentários e compreender possíveis tendências ou mudanças nos sentimentos ao longo do tempo.
2. autor: A identificação do autor é relevante para analisar a inclinação dele na expressão dos sentimentos positivos, negativos ou neutros em cada publicação.
3. texto: O texto do comentário em si é o elemento central para a análise de sentimentos. É por meio do texto que podemos extrair informações linguísticas e identificar palavras, frases ou padrões.
4. sentimento: Essa coluna é a variável-alvo da análise de sentimentos. A classificação de sentimento atribui uma categoria (positivo, negativo ou neutro) a cada comentário com base na análise do seu conteúdo textual.
5. tipointeracao: A categoria de interação do texto do comentário pode fornecer informações adicionais sobre o contexto em que o comentário foi feito.

Colunas Descartadas:

1. anomalia: Essa coluna não é relevante para a análise de sentimentos, pois está relacionada à classificação de possíveis comportamentos maliciosos ou golpes nos comentários, não ao sentimento expresso neles.
2. probabilidadeAnomalia: Essa coluna também está relacionada à classificação de possíveis anomalias ou comportamentos maliciosos, não sendo relevante para a análise de sentimentos.
3. linkPost: O link da postagem do Instagram não está diretamente relacionado ao sentimento expresso nos comentários e, portanto, não é relevante para a análise de sentimentos.
4. processado: Essa coluna indica se a análise de sentimentos já foi realizada na linha. Embora seja útil para rastrear o progresso do processamento dos dados, não contribui diretamente para a análise de sentimentos em si.
5. contemHyperlink: Essa coluna indica se um comentário possui um hyperlink. Embora possa ser relevante para outras análises, como a detecção de spam, não é diretamente relevante para a análise de sentimentos dos comentários.

6.3.2 Formatação de datas

Para a manipulação correta das datas e horários na base de dados, todas precisam estar no mesmo formato, sendo o modelo escolhido dd-mm-yyyy (Exemplo: 03-04-2022). A coluna afetada pela formatação foi a “dataPublicada”. Essa Feature foi selecionada pois sem a formatação das datas resultaria em um difícil manuseio dos dados. A Figura 19, ilustra o antes e o depois da formação, sendo o lado esquerdo da imagem o antes da aplicação e o lado direito o depois.

Figura 19: Formatação das datas

Antes		Depois	
id	"dataPublicada"	id	"dataPublicada"
1	"2022-03-04 09:38:00"	1	03-04-2022

Fonte: Autores

6.3.3 Remoção Comentário BTG Pactual

A remoção dos comentários relacionados à empresa BTG Pactual da coluna “autor” da base de dados, foi realizada uma vez que o objetivo é analisar as opiniões e sentimentos expressos pelos usuários em relação à empresa, e incluir os comentários feitos pela própria empresa poderia enviesar os resultados. Ao remover os comentários da BTG Pactual, é possível obter uma análise mais imparcial e focada nas percepções dos usuários externos à empresa, o que permite uma compreensão mais precisa e representativa dos sentimentos em relação à organização.

6.3.4 Remoção de Aspas Duplas

A remoção das aspas duplas em cada título das colunas, foi realizado para formatação adequada dos títulos. A presença de aspas duplas poderia dificultar a manipulação e o processamento dos dados, especialmente ao realizar consultas ou acessar as colunas da estrutura de dados. Ao remover, os títulos ficam mais limpos e prontos para serem utilizados em análises e visualizações, contribuindo para uma melhor organização e tratamento dos dados.

6.3.5 Anonimização dos Autores

A etapa de anonimização dos autores foi necessário por se tratar de informações que relacionam diretamente comentários aos seus autores, para remover viés e/ou julgamentos. Este processo envolve a substituição das identidades dos usuários, substituindo-as por identificadores numéricos. Realizada para garantir a privacidade e a proteção dos dados pessoais dos usuários. Além disso, a anonimização dos autores permite que a análise seja focada nos padrões, tendências e informações coletivas dos dados, em vez de identificar indivíduos específicos.

6.4 Pré-Processamento dos dados

Nesta sessão, define-se o pré-processamento de dados, etapa essencial no processo de análise de dados, responsável pela transformação e limpeza dos dados brutos tornando-os adequados para análise.

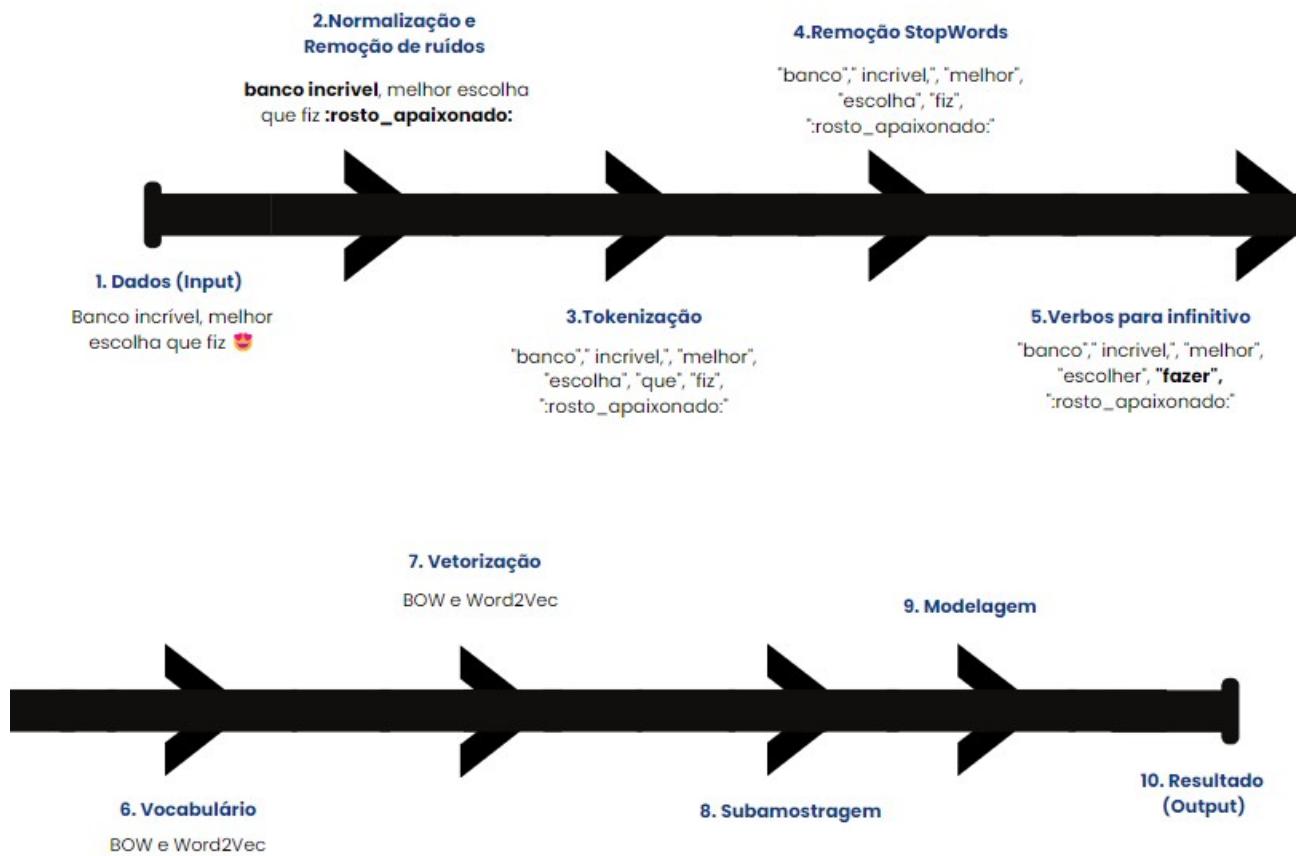
São aplicadas diversas técnicas e métodos para lidar com casos que prejudiquem a análise e objetivo proposto, o objetivo é melhorar a qualidade dos dados, torná-los mais coerentes e garantir que estejam prontos para a aplicação de modelos. As etapas realizadas foram:

1. Descrição Emojis: Identificação e análise dos emojis presentes nos textos, conversão das figuras utilizadas para palavras. A função “removeEmoji” foi utilizada para identificar todos os emojis presentes no DataFrame e realizar a conversão de símbolo para palavra.
2. Remoção de Ruídos: Eliminação de caracteres especiais, símbolos, pontuações, quaisquer elemento que possa interferir na análise dos dados. A função “removeCaracteres” e “removeAcentos”, foram criadas para remover os ruídos presentes no DataFrame, ou seja, os caracteres especiais e pontuações dos comentários.
3. Conversão de gírias e abreviações: Transformação de gírias, abreviações e termos informais em palavras completas para melhorar a compreensão e consistência dos textos. A função “normalizar_texto” é responsável por desempenhar dois papéis, sendo eles, normalizar os textos para minúsculo e converter gírias e abreviações presentes nos comentários para as palavras de onde se originaram, através de um dicionário de gírias setado.
4. Tokenização: Divisão dos textos em unidades menores, neste caso palavras, para facilitar a análise e o processamento dos dados. E, StopWords: Remoção de palavras, como artigos, preposições e pronomes, que não contribuem significativamente para a análise textual. A função “processarTexto” foi programada para realizar dois processos, tokenizar as palavras e remover as stopwords dela, visando facilitar a análise e processamento dos dados, melhorando a qualidade e relevância dos resultados obtidos.

6.4.1 Pipeline

Apresenta-se a pipeline ilustrada do projeto, sendo ela uma sequência de processos que são executados em uma determinada ordem, buscando modificar no caso, realizar o pré-processamento dos textos fornecidos. Assim, utiliza-se o resultado da etapa anterior como entrada para a próxima. Na figura 20 abaixo se pode visualiza-lá.

Figura 20: Pipeline Ilustrativa



Fonte: Autores

7. Análise Descritiva

Nesta sessão apresenta-se a etapa inicial e fundamental na análise de dados.

7.1 Dados Analisados

Apresenta-se uma descrição resumida das características e padrões presentes nos dados coletados. Permitindo uma compreensão inicial dos dados e auxilia na tomada de decisões e na formulação de estratégias com base nas informações disponíveis.

7.1.1 Quantidade de linhas na tabela

Após o tratamento dos dados, notou-se uma considerável redução no número de linhas da tabela. Isso indica que foram aplicadas etapas de limpeza e remoção de registros inválidos, duplicados ou irrelevantes. Essa redução pode ter impacto na análise subsequente, uma vez que os dados estão mais refinados e selecionados. Pode-se visualizar na figura 21 abaixo.

Figura 21: Quant Linhas

```
# Antes do tratamento
quantidadeLinhas = base.shape[0]
print("A tabela possui", quantidadeLinhas, "linhas.")

A tabela possui 12355 linhas.

# Depois do tratamento
quantidadeLinhas = dados.shape[0]
print("A tabela possui", quantidadeLinhas, "linhas.")

A tabela possui 4032 linhas.
```

Fonte: Autores

7.1.2 Quantidade de palavras na coluna texto

Após o tratamento dos textos, foi observada uma significativa redução na quantidade de palavras presentes na coluna. Isso pode indicar que foram removidas palavras irrelevantes, stop words, símbolos e outros elementos que não contribuíam para

a análise. A limpeza dos textos pode facilitar a identificação de padrões e informações relevantes posteriormente. Pode-se visualizar na figura 22 abaixo.

Figura 22: Quant Palavras

```
# Antes do tratamento
totalPalavras = base["texto"].str.split().str.len().sum()

print("A coluna 'texto' possui", totalPalavras, "palavras no total.")

A coluna 'texto' possui 412928.0 palavras no total.

# Depois do tratamento
totalPalavras = dados['texto'].str.split().str.len().sum()

print("A coluna 'texto' possui", totalPalavras, "palavras no total.")

A coluna 'texto' possui 142311.0 palavras no total.
```

Fonte: Autores

7.1.3 A distribuição de sentimentos positivos, negativos e neutros

Mesmo após o tratamento dos dados, a distribuição dos sentimentos expressos pelos usuários ainda apresenta uma grande quantidade de sentimentos neutros. Esse fato indica que deverá ser necessário um tratamento e balanceamento adequado para lidar com essa desproporção entre os sentimentos. Esse tratamento pode ser importante para garantir uma análise mais precisa e equilibrada dos dados. Pode-se visualizar na figura 23 abaixo.

Figura 23: Distribuição sentimento

```
# Antes do tratamento
contagemSentimento = base["sentimento"].value_counts()

print("Contagem de sentimentos:")
print(contagemSentimento)

Contagem de sentimentos:
NEUTRAL      5344
POSITIVE     4487
NEGATIVE     2524
Name: "sentimento", dtype: int64

# Depois do tratamento
contagemSentimento = dados['sentimento'].value_counts()

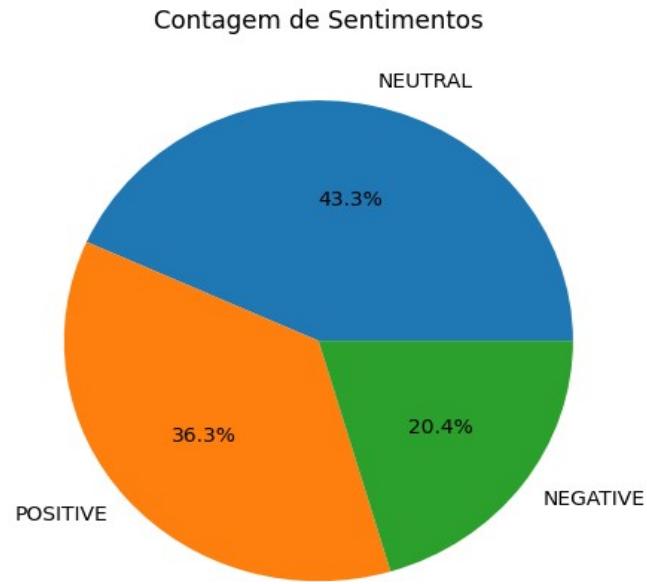
print("Contagem de sentimentos:")
print(contagemSentimento)

Contagem de sentimentos:
NEUTRAL      1642
POSITIVE     1564
NEGATIVE     826
Name: sentimento, dtype: int64
```

Fonte: Autores

Exibe-se o gráfico antes do tratamento dos dados, na figura 24 abaixo:

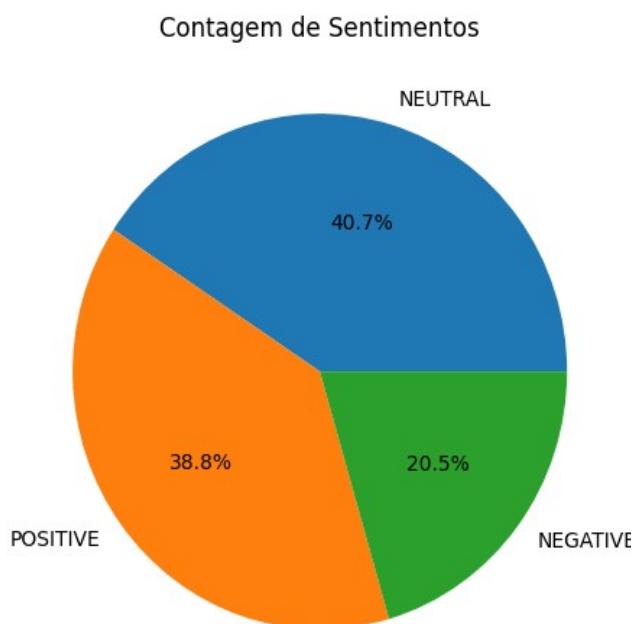
Figura 24: Quant Sentimento - Antes Processamento



Fonte: Autores

Exibe-se o gráfico depois do tratamento dos dados, na figura 25 abaixo:

Figura 25: Quant Sentimento - Depois Processamento



Fonte: Autores

7.1.4 Quantidade de autores na base de dados

Após o tratamento dos dados, a quantidade de autores na base de dados foi reduzida. Essa redução pode estar relacionada a diferentes motivos, como a remoção de registros de autores inválidos ou a anonimização dos autores por questões de privacidade. Essa informação é relevante para entender o engajamento e a participação dos diferentes autores na base de dados. Pode-se visualizar na figura 26 abaixo.

Figura 26: Quant Autores

```
# Antes do tratamento

quantidadeAutores = base['autor'].nunique()

print("A base de dados possui", quantidadeAutores, "autores únicos.")

A base de dados possui 5839 autores únicos.

# Depois do Tratamento

quantidadeAutores = dados['autor_anonimo'].nunique()

print("A base de dados possui", quantidadeAutores, "autores únicos.")

A base de dados possui 2142 autores únicos.
```

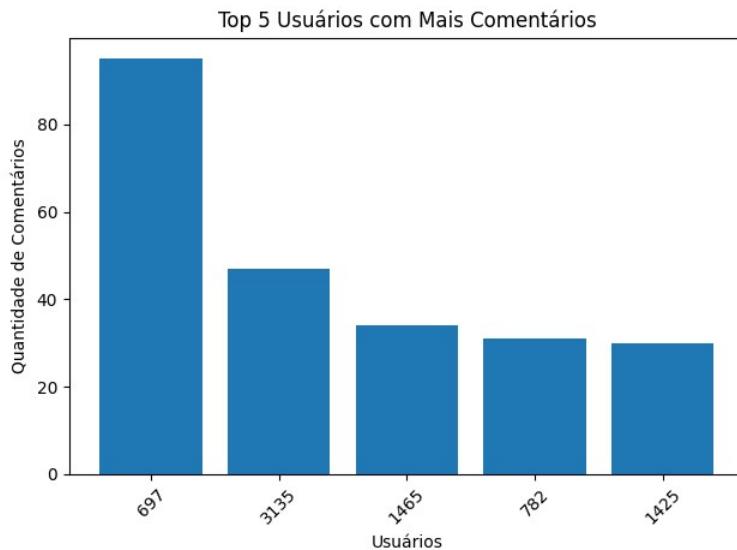
Fonte: Autores

7.1.5 Os usuários que mais realizaram comentários

A análise identificou os usuários que mais realizaram comentários na base de dados. Essa informação é útil para compreender quais usuários são mais ativos e engajados na interação, podendo indicar a presença de usuários influentes ou que possuem uma participação mais frequente nas discussões.

Porém, devido à anonimização dos autores para preservar a privacidade, não é possível exibir a identificação específica dos usuários, antes do tratamento de dados. Utilizando um gráfico de barras, podemos visualizar a distribuição dos comentários entre os usuários e identificar aqueles com maior participação. Dessa forma, garantimos a confidencialidade dos dados enquanto analisamos a contribuição dos usuários mais engajados. Pode-se visualizar na figura 27 abaixo.

Figura 27: Usuário x Comentário



Fonte: Autores

7.1.6 Quantidade de comentários por Tipo de Interação

Após o tratamento dos dados, a análise revelou a quantidade de comentários por tipo de interação. Observou-se que a marcação apresentou a maior quantidade de comentários, seguida por comentário e resposta. Essa informação permite entender como os usuários interagem e qual tipo de interação é mais comum na base de dados. Essa análise pode ajudar a identificar os padrões de comunicação e a dinâmica das interações entre os usuários. Pode-se visualizar na figura 28 abaixo.

Figura 28: Comentários X Interação

```
# Antes do tratamento
contagemSentimento = base['tipoInteracao'].value_counts()
print("A quantidade de comentários por Tipo de Interação: ")
print(contagemSentimento)

A quantidade de comentários por Tipo de Interação:
marcação      5999
comentário     5389
resposta       967
Name: "tipoInteracao", dtype: int64

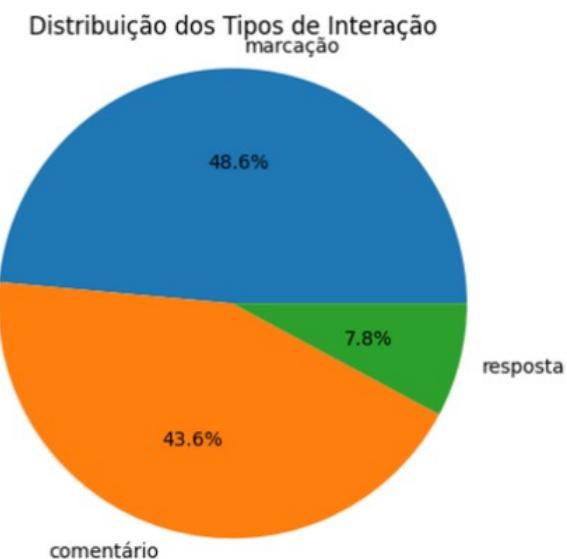
# Depois do tratamento
contagemSentimento = dados['tipoInteracao'].value_counts()
print("A quantidade de comentários por Tipo de Interação: ")
print(contagemSentimento)

A quantidade de comentários por Tipo de Interação:
marcação      1989
comentário     1739
resposta       304
Name: tipoInteracao, dtype: int64
```

Fonte: Autores

Exibe-se o gráfico antes do tratamento dos dados, na figura 29 abaixo:

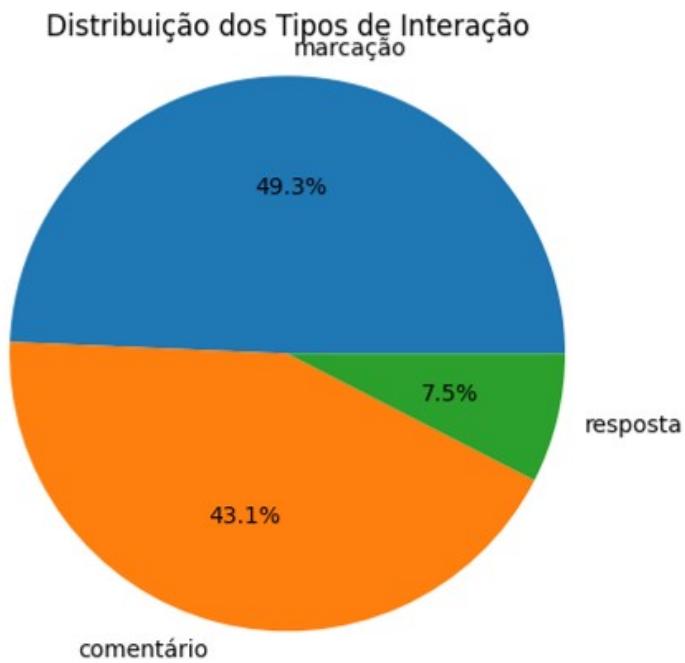
Figura 29: Comentário X Interação - Antes Processamento



Fonte: Autores

Exibe-se o gráfico após do tratamento dos dados, na figura 30 abaixo:

Figura 30: Comentário X Interação - Depois Processamento



Fonte: Autores

Essas informações são cruciais para compreender melhor a percepção dos usuários em relação ao banco e para orientar futuras estratégias de comunicação e relacionamento com o público, garantindo uma maior assertividade em futuras publicações do banco BTG.

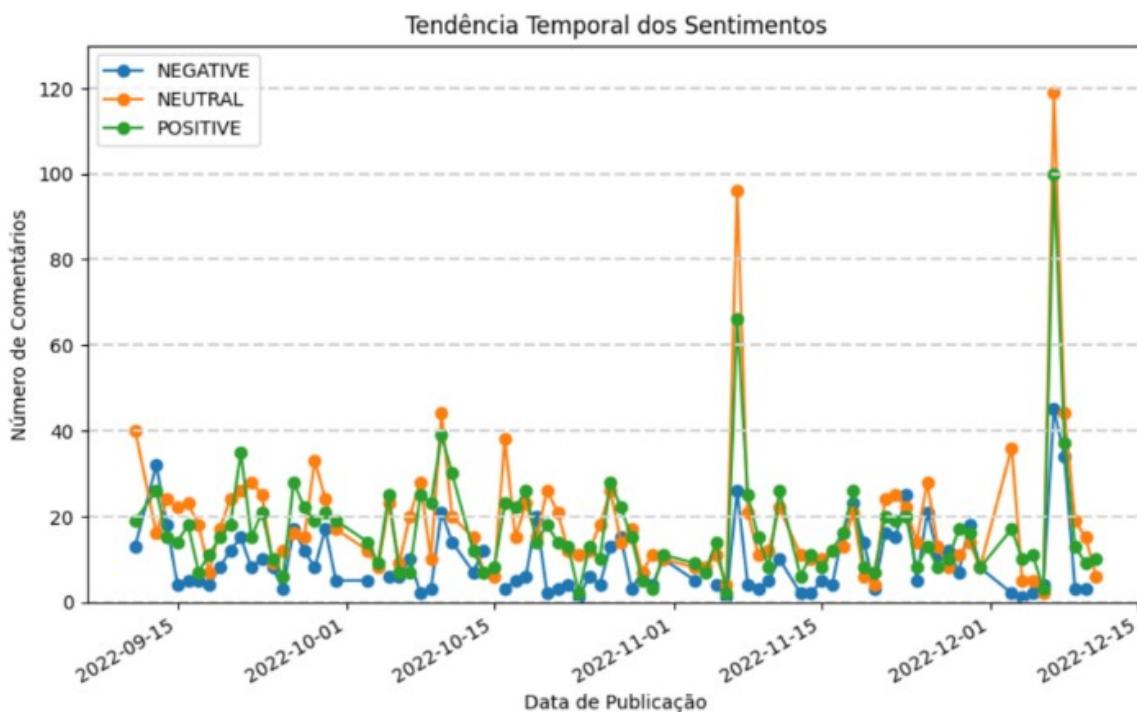
7.2 Análise de dados

Essa sessão apresenta a análise preliminar dos padrões, tendências e relações identificadas nos dados por meio de gráficos. Destaca-se os principais achados e insights obtidos a partir dos gráficos, sua análise e significado.

7.2.1 Gráfico de Tendência Temporal dos Sentimentos

O gráfico de tendência temporal dos sentimentos foi utilizado para explorar a evolução dos sentimentos (positivos, negativos e neutros) ao longo do tempo com base na coluna "dataPublicada". Com ele, identifica-se picos de comentários positivos, negativos e neutros ao longo dos três últimos meses antes da última inserção de comentários na base (11-12-2022). A figura 31 abaixo ilustra o gráfico.

Figura 31: Tendência Temporal X Sentimento



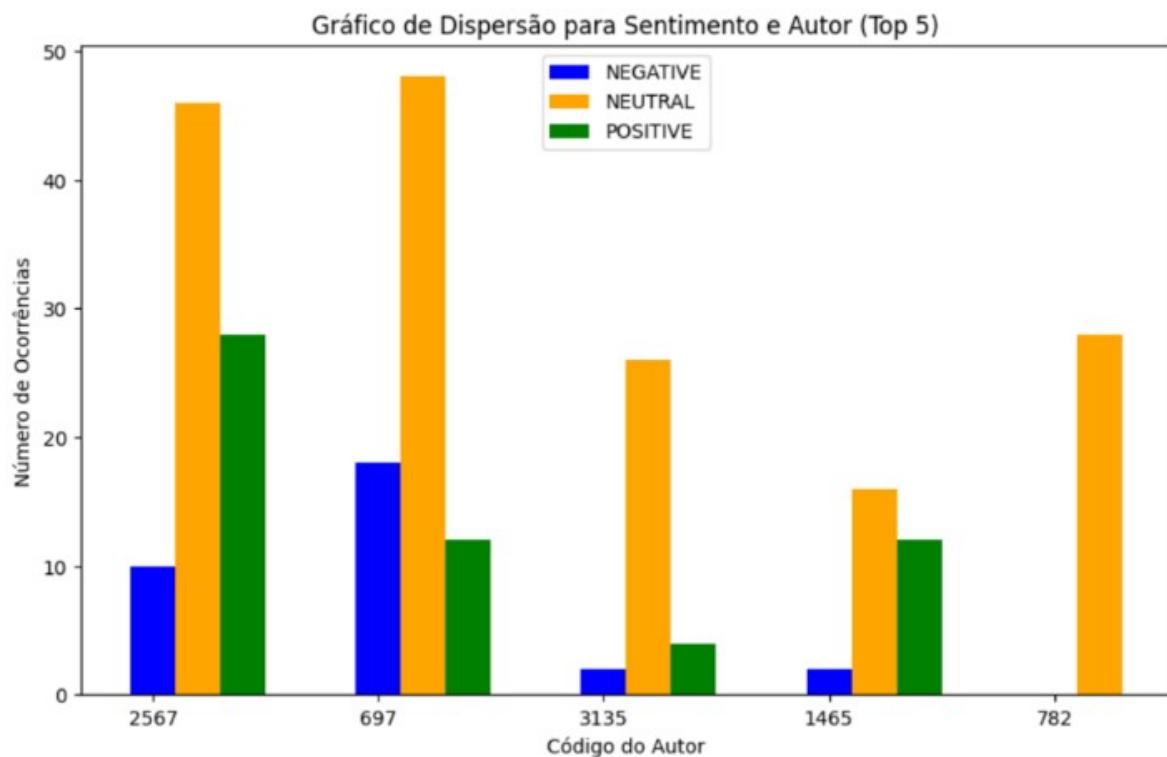
Fonte: Autores

Ao analisar o período de 02-12-2022 a 15-12-2022, observamos um padrão de comportamento nos comentários positivos. Durante esse período, houve um aumento notável na contagem de comentários positivos, atingindo um pico de 100 comentários. Esse pico pode ser um indicativo de um evento ou acontecimento especial que gerou um grande número de reações positivas dos usuários. Inclusive porque durante o mês de Dezembro o BTG realiza o evento de Natal, o que pode atrair uma maior quantidade de acessos, por conta de shows e decorações no local.

7.2.2 Gráfico de Dispersão para Sentimento e Autor

O Gráfico de Dispersão para Relação entre Sentimento e Autor é utilizado para explorar a relação entre o sentimento expresso nos comentários e os respectivos autores. Nesse gráfico, o eixo x representa os autores e o eixo y representa a polaridade do sentimento, com valores negativos, neutros e positivos. Na figura 32 abaixo se exibe o gráfico.

Figura 32: Comentário X Sentimento



Fonte: Autores

Ao observar o gráfico, podemos identificar a interação entre os autores e o sentimento expresso em seus comentários. Notamos que os comentários neutros ainda são os mais recorrentes, sugerindo uma tendência de neutralidade predominante nas

interações. No entanto, é importante ressaltar a necessidade de realizar um balanceamento dos dados, uma vez que a alta recorrência de comentários neutros pode influenciar a análise geral. A sequência de comentários positivos, é notória e merece uma análise mais aprofundada. Essa sequência de comentários positivos pode revelar insights sobre a satisfação dos autores, a eficácia de determinadas ações ou até mesmo a qualidade do conteúdo gerado.

7.2.3 Gráfico de Nuvem de Palavras

O Gráfico de Nuvem de Palavras, também conhecido como Word Cloud, é uma ferramenta para analisar e visualizar as palavras mais frequentes em um texto. Nesse contexto, o uso se mostra especialmente útil para identificar as palavras mais recorrentes nos comentários expresso pelos usuários.

DPara começar, foi criada uma nuvem de palavras considerando a base geral de dados. Essa nuvem de palavras mostra as palavras mais frequentes em todos os comentários, independentemente do período em que foram feitos. É uma forma de entender os temas e tópicos mais abordados pelos usuários de maneira geral. Palavras maiores indicam uma frequência maior na base de dados, enquanto palavras menores indicam uma frequência menor. Pode-se visualizá-lo na figura 33 abaixo.

Figura 33: Nuvem de Palavras 1



Fonte: Autores

Além disso, foi criada uma segunda nuvem de palavras focada no período específico de 02-12 a 15-12 de 2022, no qual foi identificado um pico significativo de comentários positivos. Essa nuvem de palavras permite analisar as palavras mais frequentes nesse período específico e identificar possíveis mudanças nos temas e tópicos discutidos pelos usuários. Através dessa análise, é possível compreender melhor os fatores que levaram a esse pico de comentários positivos e explorar as causas prováveis. Na imagem 34 abaixo se pode visualizá-lo.

Figura 34: Nuvem de palavras 2



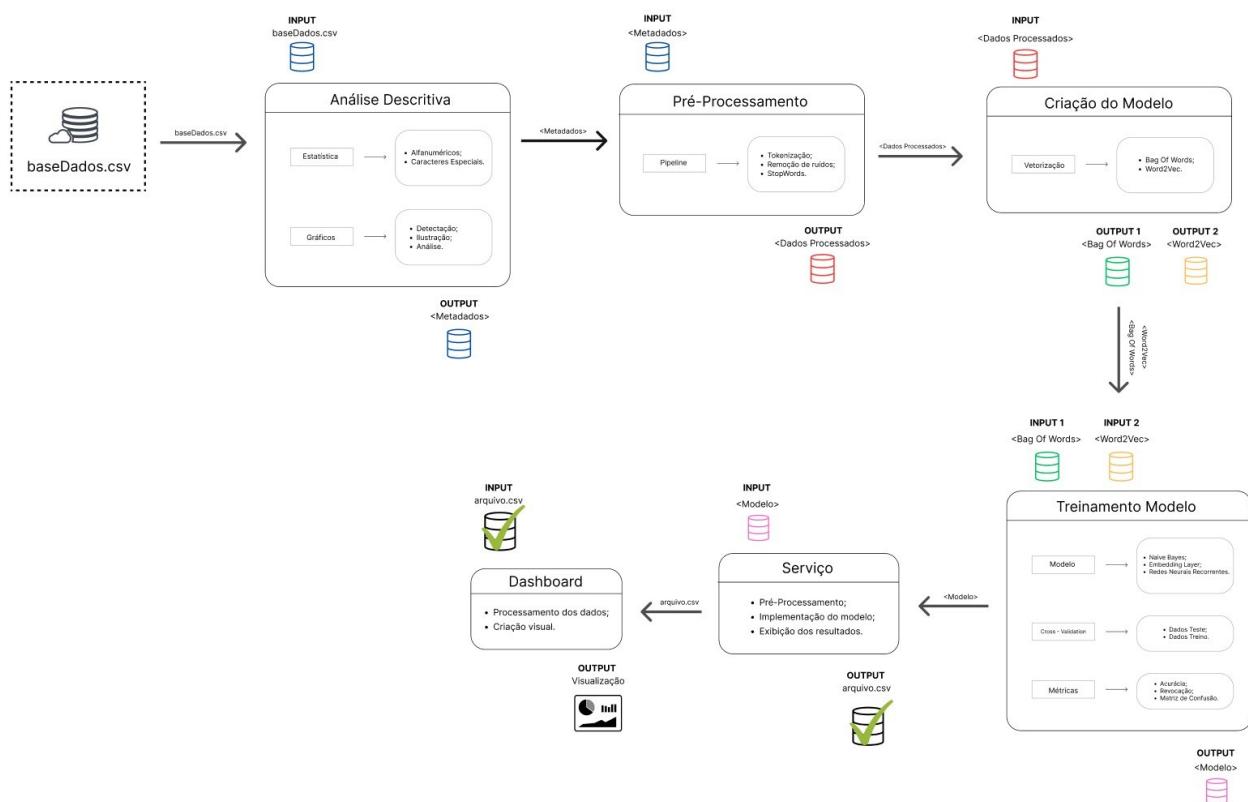
Fonte: Autores

Outro aspecto importante é que se pretende remover o nome "BTG Pactual" da nuvem de palavras, a fim de focar em palavras mais claras e específicas do assunto. Isso permite uma análise mais precisa e a identificação de termos relacionados ao sentimento dos usuários, sem a interferência do nome da empresa.

8. Arquitetura Macro da Solução

A arquitetura macro da solução, apresenta os blocos responsáveis pelo funcionamento da solução, independente da tecnologia que será adotada ao desenvolvimento, comunicando-se entre si, apresentam a estrutura e ligações mínimas que a solução tem que exibir para ter o funcionamento previsto ao MVP. Abaixo na Figura 35, encontra-se a visualização inicial prevista da arquitetura macro.

Figura 35: Arquitetura Macro da Solução



Fonte: Autores.

A arquitetura se inicia recebendo os dados de entrada a partir de um arquivo "baseDados.csv", armazenado dentro de uma nuvem. Após o recebimento dos dados, o sistema passa por uma fase de análise descritiva, na qual são aplicadas técnicas estatísticas para compreender as características dos dados, além de serem gerados gráficos para facilitar a detecção, ilustração e análise dos padrões encontrados.

Na etapa seguinte, ocorre o pré-processamento dos dados. Utilizando os metadados gerados na análise descritiva, os dados são submetidos a um processo de tokenização, que consiste na separação em unidades menores, como palavras, dependendo do contexto. Além de técnicas de remoção de ruídos e stop words, que são palavras comuns e irrelevantes para a análise.

Com os dados já pré-processados, é realizado o passo de criação do modelo. Nessa etapa, os dados são convertidos em representações numéricas adequadas para o treinamento do modelo. Duas técnicas comumente utilizadas são "bag of words" e "word2vec". A representação "bag of words" atribui um vetor numérico para cada palavra presente no conjunto de dados, enquanto o "word2vec" utiliza técnicas de aprendizado de máquina para mapear as palavras em vetores de alta dimensionalidade.

Após a criação das representações dos dados, o sistema realiza o treinamento do modelo. São empregados os modelos de Naive Bayes, embedding layer e redes neurais recorrentes, diversificando os tipos de análises realizadas. Para avaliar o desempenho do modelo, é aplicada a técnica de validação cruzada, na qual os dados são divididos em conjuntos de treinamento e teste. As métricas aplicadas são acurácia, revocação e matriz de confusão, para medir a eficácia do modelo treinado.

Uma vez que o modelo é escolhido, a partir da comparação das melhores métricas e já está treinado, ele é colocado em serviço. Essa etapa engloba a automatização do pré-processamento dos novos dados recebidos e a implementação do modelo para realizar previsões ou classificações. Os resultados obtidos são então exibidos e armazenados em "arquivo.csv", que contém as saídas do sistema.

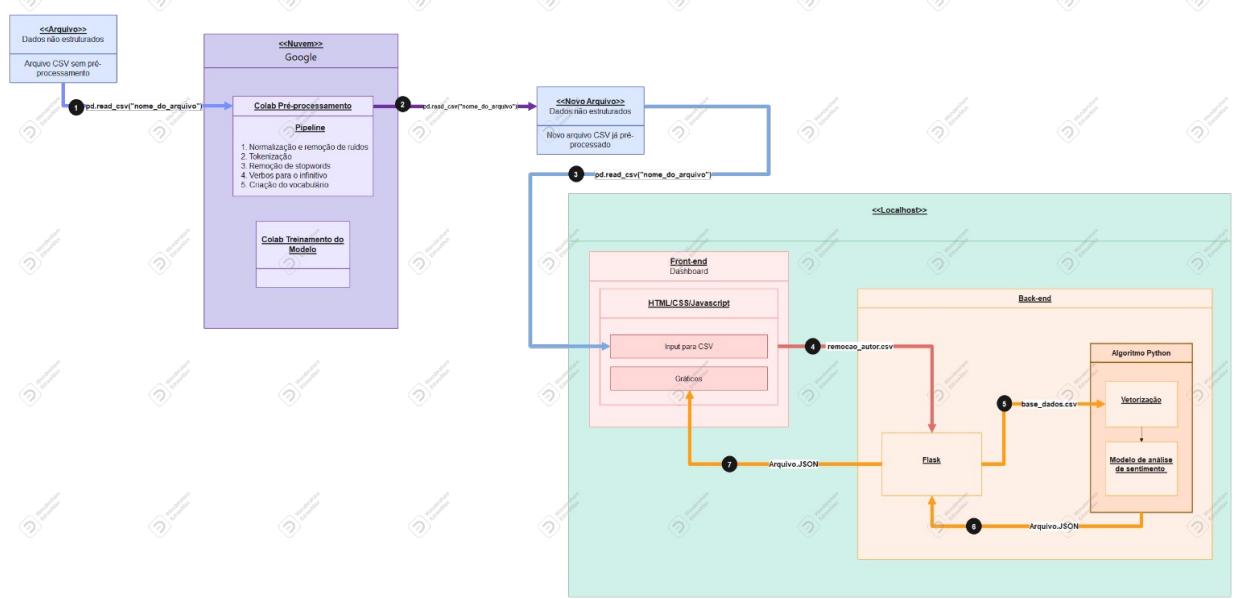
Por fim, há a etapa do dashboard, na qual o "arquivo.csv" é processado para criar as visualizações dos resultados de forma gráfica. Essas visualizações fornecem auxílio na tomada de decisões e na análise dos dados de maneira mais visual e comprehensível.

Em conclusão, essa arquitetura permite automatizar a análise de sentimentos dos comentários do Instagram do BTG Pactual, fornecendo insights valiosos sobre a percepção dos usuários. Isso pode ser útil para identificar tendências, monitorar a reputação da marca, compreender a satisfação dos clientes e tomar decisões estratégicas com base nos sentimentos expressos nas redes sociais.

8.1 Diagrama UML

O diagrama UML é uma representação visual que ilustra de forma abrangente o funcionamento do sistema, demonstrando as diversas interações entre suas partes e conexões. Esses diagramas desempenham um papel crucial na obtenção de clareza e compreensão do comportamento esperado da solução, permitindo definir de maneira precisa as entradas e respostas esperadas.

Figura 36: Diagrama de implantação UML



Fonte: Autores

O diagrama tem início no bloco superior esquerdo, denominado "arquivo". Esse arquivo contém dados não estruturados no formato CSV, os quais ainda não passaram pelo processo de pré-processamento. Em seguida, o usuário realiza a inserção desse arquivo utilizando a função "pd.read_csv("nome_do_arquivo")" (1) no ambiente Colab "Pré-processamento", uma plataforma colaborativa em nuvem fornecida pelo Google. Essa etapa é essencial para limpar e preparar os dados, a fim de torná-los adequados para a próxima fase do sistema.

Após o processo de pré-processamento, é gerado um novo arquivo CSV (2), pronto para ser inserido na aplicação local. No front-end, o usuário seleciona o arquivo que deseja classificar (3) e, por meio do Flask, um popular framework web em Python que oferece recursos básicos para o gerenciamento de rotas, renderização de templates,

tratamento de solicitações e respostas HTTP, além de suporte para sessões e cookies, os dados são encaminhados para o back-end (4).

No back-end, encontra-se o algoritmo do modelo final selecionado, o qual engloba desde a etapa de vetorização até a modelagem (5). É importante ressaltar que o treinamento do modelo não ocorre no back-end, uma vez que ele já foi previamente realizado e está armazenado na nuvem do Google.

Por fim, os resultados da análise de sentimento são retornados ao front-end por meio do Flask (6 e 7). Esses resultados são apresentados ao usuário de forma visual, por meio de gráficos e outras ilustrações, proporcionando uma visualização ampla e comprehensível das informações e insights extraídos por meio da análise dos dados.

9. Algoritmos

Nessa seção apresenta-se os dois algoritmos utilizados como base para a aplicação dos dados nos modelos escolhidos para o projeto.

9.1 Bag Of Words

O modelo "Bag of Words", ou saco de palavras, é uma técnica utilizada no processamento de linguagem natural para analisar textos. Nesse modelo, o texto é tratado como um conjunto de palavras, ignorando-se a ordem e a estrutura gramatical. O objetivo é identificar as palavras-chave presentes no texto e sua frequência.

Ao utilizar o modelo Bag of Words, é possível construir um vetor numérico que representa o texto, onde cada posição do vetor corresponde a uma palavra e o valor associado indica a frequência dessa palavra no texto. Isso permite visualizar quais palavras são mais frequentes e sua relevância dentro do contexto.

Embora o modelo Bag of Words seja útil para identificar palavras-chave e avaliar suas frequências, ele apresenta algumas limitações. Por não levar em consideração a estrutura do texto e a semântica das palavras, pode haver perda de informações importantes. A ordem das palavras e a relação entre elas não são consideradas, o que pode afetar a precisão da análise de sentimentos em textos mais complexos.

No entanto, o modelo Bag of Words oferece uma abordagem simples e eficiente para a classificação preliminar de sentimentos em textos. Ele permite identificar palavras-chave que indicam sentimentos positivos, negativos ou neutros, fornecendo uma visão geral do sentimento expresso no texto. É um ponto de partida útil para a análise de sentimentos, embora seja necessário considerar abordagens mais avançadas para uma análise mais precisa e sofisticada.

9.2 Word2Vec

Word2Vec é um dos métodos estatísticos mais utilizados no pré-processamento de dados de texto a fim de uso do PLN (processamento de linguagem natural), neste projeto sendo utilizado para análise de sentimento. Foi criado para facilitar o treinamento de embeddings baseados em redes neurais. O modelo pode ser comparado com a técnica

de incorporação de palavras, que é uma técnica em que palavras individuais são transformadas em vetores. Cada vetor captura várias características das palavras como: relação semântica da palavra, definições, contexto, etc.

O Word2Vec tem a capacidade de agrupar vetores de palavras semelhantes; o modelo pode fornecer estimativas fortes sobre os significados das palavras a partir da ocorrência delas nos textos. Para obter sucesso no resultado pode-se utilizar o CBoW (tenta prever uma palavra-alvo a partir de uma lista de palavras de contexto) e o skip-gram (rede neural simples com uma camada oculta treinada para prever a probabilidade de uma determinada palavra estar presente quando uma palavra de entrada está presente).

Neste caso o modelo é pré-treinado (construído em grandes conjuntos de dados de texto, como a Wikipedia, usando algoritmos como Skip-gram ou Continuous Bag of Words). O modelo pré-treinado captura informações linguísticas gerais e contextuais (com base no banco de texto que foi utilizado) sem a necessidade de treinamento adicional. Esses modelos pré-treinados são frequentemente disponibilizados publicamente e podem ser usados diretamente para tarefas de NLP, como classificação de texto ou análise de sentimentos.

9.2 TF-IDF

A vetorização TF-IDF (Term Frequency-Inverse Document Frequency) é uma técnica utilizada no processamento de linguagem natural para representar textos como vetores numéricos. Essa abordagem permite quantificar a importância de cada palavra em um documento, levando em consideração tanto a frequência da palavra no documento quanto sua relevância em relação a todo o corpus.

O processo de vetorização TF-IDF envolve duas etapas principais. A primeira etapa é o cálculo da frequência dos termos em cada documento. A frequência do termo (TF) mede quantas vezes um determinado termo aparece em um documento específico. Quanto mais vezes um termo aparecer em um documento, maior será o seu valor de TF.

A segunda etapa é o cálculo da frequência inversa do documento (IDF). O IDF mede a importância de um termo em um conjunto de documentos. Ele é calculado levando em consideração a frequência do termo em todos os documentos do conjunto.

Quanto menos frequente um termo for em todos os documentos, maior será o seu valor de IDF.

Combinando o TF e o IDF, obtemos o vetor TF-IDF para cada documento. Esse vetor representa a importância relativa de cada palavra em relação ao documento e ao conjunto de documentos. Palavras frequentes no documento e raras no conjunto de documentos terão valores de TF-IDF mais altos, indicando sua relevância para o documento específico.

10. Modelagem

As sessões abaixo apresentam os algoritmos escolhidos para teste do modelo preditivo, suas descrições, principais funções e exemplificações.

10.1 Naive Bayes

O algoritmo Naive Bayes utilizado neste projeto é baseado no cálculo de probabilidades condicionais, utilizando a fórmula de Bayes. A fórmula é composta por duas partes: $P(A|B)$, que representa a probabilidade do evento A ocorrer dado que o evento B já ocorreu, e $P(B|A)$, que representa a probabilidade do evento B ocorrer, dado que o evento A já ocorreu. Essa abordagem permite realizar classificações com base na probabilidade de ocorrência, levando em consideração eventos prévios. Sua fórmula pode ser visualizada na figura 37 abaixo.

Figura 37: Fórmula Naive Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fonte: Autores

No contexto do projeto, o algoritmo de Naive Bayes é aplicado através da criação de uma tabela de probabilidades. Essa tabela contém as frequências com que cada preditor se relaciona com as variáveis de saída. A partir dessas probabilidades, o cálculo é realizado e os resultados são retornados, escolhendo-se a classificação com a maior probabilidade como solução para o problema em questão.

Nesta aplicação utilizou-se o modelo Bayes Gaussiano, sendo a principal diferença entre os distintos classificadores do método Naive Bayes é que o Gaussiano considera como gaussiana a distribuição, ou seja, assume a normalidade dos dados, enquanto o classificador Naive Bayes Multinomial considera que os dados são distribuídos multinomialmente; o uso desse método é indicado para quando as variáveis independentes são contínuas e têm distribuição normal.

O modelo Naive Bayes foi escolhido pois, para classificação de textos e análise de sentimentos devido à sua simplicidade e eficiência, é capaz de lidar com grandes volumes de dados e realizar classificações rápidas, tornando-o uma opção viável para lidar com o grande número de comentários nas redes sociais. Além disso, o Naive Bayes tem um bom desempenho quando as palavras-chave relevantes estão bem definidas, o que pode ser útil na identificação de sentimentos positivos ou negativos nos comentários.

10.2 Embedding Layer

A camada de incorporação (Embedding Layer) é uma componente fundamental em modelos de processamento de linguagem natural (PLN) que envolvem o uso de redes neurais. Ela é responsável por transformar representações numéricas discretas, como palavras ou caracteres, em vetores contínuos de números reais. A técnica de incorporação é baseada na ideia de que palavras ou caracteres semelhantes têm significados semânticos ou contextuais semelhantes. Portanto, ao mapear essas unidades discretas em vetores contínuos, a camada de incorporação busca capturar as relações e as características semânticas dos dados de entrada.

A camada de incorporação é treinada em conjunto com o restante do modelo de PLN. Durante o treinamento, os pesos da camada de incorporação são ajustados de forma a minimizar a perda na tarefa específica, como classificação de sentimentos ou geração de texto. Esses vetores contínuos resultantes da camada de incorporação podem então ser usados como entrada para outras camadas da rede neural. A vantagem da camada de incorporação é que ela permite que o modelo aprenda representações mais eficientes e compactas dos dados de entrada, reduzindo a dimensionalidade do problema. Isso melhora a capacidade de generalização do modelo, pois palavras ou caracteres semelhantes serão mapeados para vetores próximos no espaço de incorporação.

A Embedding Layer é um componente essencial em modelos de PLN, pois permite a representação de palavras ou caracteres como vetores contínuos. Ao utilizar essa técnica, pode-se capturar a semântica e o contexto das palavras presentes nos comentários dos usuários. A camada de incorporação ajuda a reduzir a dimensionalidade dos dados e melhora a capacidade de generalização do modelo, permitindo uma melhor compreensão do sentimento expresso pelos usuários.

10.3 Rede Neural Recorrente (RNN)

A rede neural é um método de processamento que utiliza aprendizado profundo, no qual neurônios interconectados em camadas simulam o funcionamento do cérebro humano. Uma variante desse método é a rede neural recorrente (RNN), que possui conexões ponderadas dentro de uma camada e é capaz de armazenar informações ao processar novas entradas. Isso permite que as RNNs considerem dados anteriores, tornando-as ideais para tarefas em que a sequência de entradas é relevante.

O aprendizado em redes neurais ocorre por meio do processamento de conjuntos de dados rotulados ou não rotulados. A rede neural gradualmente adquire conhecimento desses conjuntos de dados, o que lhe permite fornecer respostas corretas previamente. Ela é capaz de analisar dados não estruturados, como documentos de texto, identificar atributos relevantes e resolver problemas complexos.

Com a capacidade de armazenar memória interna e suportar comportamento temporal, graças aos loops que incorporam. Um aspecto interessante dessas redes é que, com camadas e nós suficientes, elas podem implementar qualquer função computável. No entanto, os algoritmos de treinamento das RNNs são distintos devido à sua capacidade de incorporar informações históricas na sequência temporal. Algoritmos de descida do gradiente são comumente utilizados para otimizar os pesos da RNN, minimizando o erro por meio do ajuste proporcional à derivada do erro em relação a esses pesos.

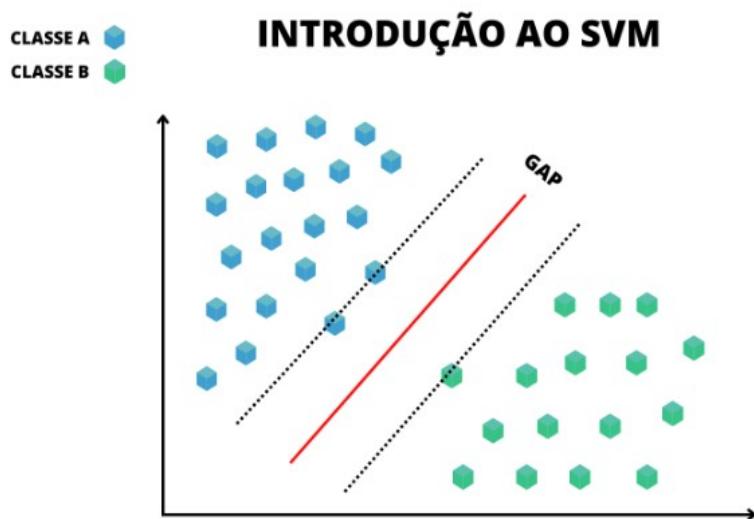
A Rede Neural Recorrente é adequada para tarefas em que a sequência de dados é relevante, como a análise de sentimentos em comentários de redes sociais. Devido à sua capacidade de armazenar informações de entradas anteriores por meio de conexões ponderadas dentro de uma camada. Podem levar em consideração o contexto das palavras nos comentários. Isso permite que o modelo comprehenda melhor a estrutura e a evolução dos sentimentos ao longo de uma sequência de texto, fornecendo resultados mais precisos à análise.

10.4 SVM (Support Vector Machine)

A definição do Support Vector Machine(SVM), pode ser dada por um algoritmo que visa encontrar o hiperplano de separação ideal para os dados propostos, sendo o seu maior objetivo a maximização das distâncias das variáveis deixando-as o mais definidas

possível. Este tende a ser mais complexo que o KNN e apresentar resultados mais estruturados, por ambos apresentarem formas de analisar as variáveis que estão mais próximas entre si e definir suas correlações, alterando apenas as métricas utilizadas. O hiperplano de separação utilizado para as análises pode ser descrito como uma linha, que passa entre os dados, tentando delimitar uma separação dos atributos selecionados, como visto na Figura 38 abaixo:

Figura 38: SVM - Exemplo de aplicação



Fonte: Autores

O hiperplano utilizado é basicamente a generalização de um plano qualquer, com mais de três dimensões. Visto isso, o objetivo primordial do SVM é conseguir traçar mediante os dados manipulados o hiperplano de separação ideal, visando a classificação de maneira correta dos atributos. O algoritmo SVM tem sua adequação a solução mediante a classificação dos dados escolhidos em categorias, sendo possível a visualização definida dos atributos referentes a variável alvo, resultando em “Sim” para probabilidade de sair e “Não” para pouca probabilidade de sair, a partir da análise do hiperplano citada anteriormente.

10.5 Random Forest

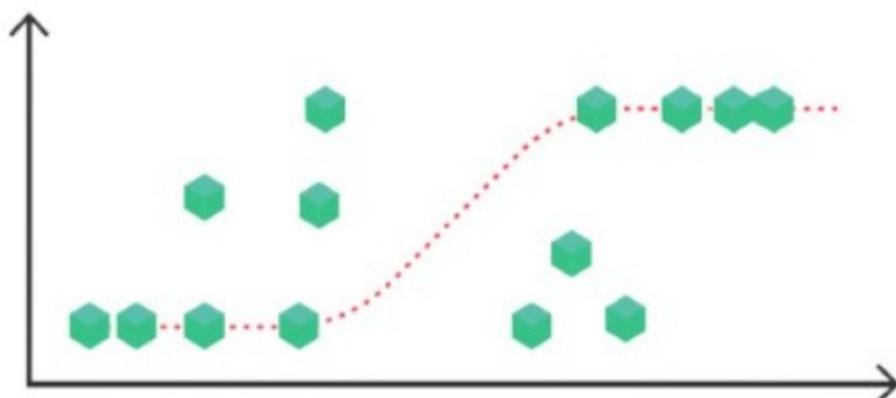
Random Forest, ou Floresta Aleatória, é um algoritmo de aprendizado de máquina que é utilizado para tarefas de classificação e regressão. É baseado na combinação de várias árvores de decisão individuais para obter uma classificação mais precisa e robusta. Uma Random Forest é construída através de um processo de treinamento onde várias árvores de decisão são criadas utilizando diferentes subconjuntos dos dados de treinamento e/ou das características. Cada árvore é treinada de forma independente, utilizando uma parcela aleatória dos dados e aplicando um critério de divisão que leva em consideração apenas um subconjunto aleatório das características.

Durante a fase de predição, as árvores individuais da floresta votam em uma classe para cada exemplo de entrada e a classe mais votada é escolhida como a predição final da floresta. A vantagem da Random Forest é que ela tende a evitar o overfitting (sobreajuste) e é menos sensível a outliers e ruídos nos dados. Além disso, a técnica é capaz de lidar com um grande número de características e oferece a possibilidade de medir a importância relativa das características para a tarefa de classificação ou regressão.

10.6 Regressão Logística

Algoritmo estatístico, de aprendizado supervisionado que é usado para a classificação e análise preditiva. Ela estima a probabilidade de um evento ocorrer se apoiando em um conjunto de dados. Na figura 39, abaixo é apresentado um exemplo gráfico de regressão logística.

Figura 39: Regressão Logística - Exemplo de aplicação



Fonte: Autores

Existem três tipos de algoritmos para regressão logística, eles são definidos com base no resultado.

- Regressão logística Binária: A variável tem apenas dois resultados possíveis, "0 e 1". Este é o algoritmo mais comumente usado.
- Regressão Logística multinomial: Nesse algoritmo a variável possui três resultados possíveis, entretanto não possui uma ordem desses valores.
- Regressão Logística Ordinal: este algoritmo é aplicado quando a variável possui 3 ou mais resultados possíveis, porém o resultado tem uma ordem já definida. Por exemplo, de "A-E" ou escalas de "1-5"

Como ela mede a relação entre uma variável alvo e outras variáveis independentes, utilizamos o algoritmo multinomial para identificar fatores importantes que impactam a nossa variável alvo: “Sentimento” nos retornando “Negativo”, “Positivo” ou “Neutro”.

11. Avaliação do Modelo

A sessão a seguir é responsável por apresentar os testes realizados no modelo de processamento de linguagem natural e seus respectivos resultados.

11.1 Divisão dos dados

Antes de modelar os algoritmos para análise de sentimento das variáveis alvo, é necessário organizar os atributos escolhidos, entre variáveis de teste e variáveis de treino, que estão descritas abaixo:

- **Dados de Treino:**

Os dados de Treino são, como o nome sugere, dados selecionados de uma base de dados que representam cerca de 70% da totalidade do conjunto da base e são levados para o treinamento do algoritmo de predição do Machine Learning;

- **Dados de Teste:**

Os dados de Teste são, como o nome sugere, dados levantados de uma base de dados que representam em torno de 30% do conjunto completo da base e servem para testar o algoritmo preditivo criado pelo aprendizado de máquina.

É importante ressaltar que haja a separação desses dados de maneira aleatória, para que não ocorra vieses nos dados por meio do aprendizado de padrões que limitam a probabilidade de predição, e a separação é necessária também para que não haja casos de overfitting, ou seja, um ajuste desproporcional aos dados apresentados

11.2 Estratégia de Avaliação do modelo

Nesta seção é apresentado todas as avaliações dos algoritmos utilizados para a construção do modelo preditivo e seus respectivos resultados obtidos.

11.2.1 Matriz de Confusão

Pode-se definir matriz de confusão como, uma tabela que representa a frequência de classificação para as variáveis declaradas no modelo. O uso dessa ferramenta de avaliação é de grande importância pois é possível realizar a análise de como o modelo se saiu nas previsões, verificando erros e acertos do modelo preditivo. Na figura 40 abaixo, ilustra-se como uma matriz de confusão funciona.

Figura 40: Matriz de confusão

		Valor Preditivo	
		SIM	NÃO
Valor Verdadeiro	SIM	VP Verdadeiro Positivo	FN Falso Negativo
	NÃO	FP Falso Positivo	VN Verdadeiro Negativo
		SIM	NÃO
		Valor Preditivo	

Fonte: Autores

Pode-se considerar para a construção do modelo, baseando-se na variável alvo, a possibilidade de um comentário ser negativo ou positivo, os quadrantes apresentados acima possuem tais significados:

1. Verdadeiro Positivo: Comentários negativos que são classificados como negativos;
2. Falso Positivo: Comentários positivos que são classificados como negativos;
3. Falso Negativo: Comentários negativos que são classificados como positivos;
4. Verdadeiro Negativo: Comentários positivos que são classificados como positivos;

Para o modelo criado, o quadrante mais importante de obter um alto índice de acerto, é o falso negativo. Ao focar no quadrante do falso negativo e buscar minimizá-lo, o objetivo é garantir que o modelo seja capaz de identificar corretamente todos os comentários negativos. Isso permite que a empresa esteja ciente dos problemas e insatisfações dos clientes.

11.2.2 Acurácia

A acurácia diz respeito à proximidade entre o valor obtido experimentalmente e o valor verdadeiro. A importância dessa estratégia de avaliação se dá pelo fato de determinar a confiabilidade e grau de exatidão do modelo. É calculado como: $TP+TN/(TP+TN+FP+FN)$.

Se a acurácia for alta: Isso indica que o modelo está classificando a maioria dos exemplos corretamente, o que é um bom sinal. Uma alta acurácia pode sugerir que o modelo é capaz de fazer previsões precisas e confiáveis.

Se a acurácia for baixa: Isso pode indicar que o modelo está tendo dificuldades em classificar corretamente os exemplos. Nesse caso, é necessário investigar e analisar outras métricas, como o recall e a precisão, para entender melhor o desempenho do modelo e identificar possíveis áreas de melhoria.

Se a acurácia estiver em um meio termo: Nesse caso, é importante considerar o contexto do problema e o equilíbrio entre as classes de sentimentos. Se houver um desequilíbrio significativo nas classes, a acurácia pode ser enganosa. Portanto, é recomendado avaliar outras métricas, como o recall e a precisão, para obter uma visão mais abrangente do desempenho do modelo.

11.2.3 Recall

O recall, também conhecido como taxa de recuperação, é uma métrica que avalia a capacidade do modelo em identificar corretamente os exemplos positivos. Ele é calculado utilizando a fórmula: $recall = TP / (TP + FN)$. O recall fornece uma medida específica do desempenho do modelo na identificação dos verdadeiros positivos em relação ao total de exemplos positivos.

Se o recall for alto: Isso indica que o modelo está identificando a maioria dos exemplos positivos corretamente. Um alto recall é desejável, especialmente quando a detecção correta dos positivos é crucial, como na identificação de comentários negativos em uma análise de sentimentos.

Se o recall for baixo: Isso sugere que o modelo está deixando passar muitos exemplos positivos. Um baixo recall pode indicar falhas na capacidade do modelo em detectar corretamente os casos positivos, o que pode levar a perdas de informações importantes.

Se o recall estiver em um meio termo: Nesse caso, é importante avaliar o contexto e a importância da detecção correta dos exemplos positivos. Dependendo da aplicação, pode ser necessário ajustar o modelo para melhorar o recall e garantir uma detecção mais abrangente dos casos positivos.

12. Desenvolvimento e Resultados

Nesta sessão apresenta-se a etapa do projeto em que são revisados os progressos feitos até o momento e são apresentados os resultados alcançados. Observação: Neste momento os resultados apresentados são com o modelo Word2Vec, que após análise resultou em uma maior performance.

12.1 BOW e Word2Vec

Ao comparar o modelo "Bag of Words" (BOW) com o modelo "Word2Vec", é possível identificar diferenças significativas em relação ao consumo de memória e à quantidade de colunas no arquivo.csv gerado.

O modelo BOW tende a ocupar mais memória em comparação ao Word2Vec, 571,61MB e 1,59MB, respectivamente. Isso ocorre porque o BOW cria um vetor numérico para cada palavra única presente no texto, resultando em uma alta dimensionalidade dos dados. Cada palavra é representada por uma coluna separada no arquivo.csv gerado, o que resulta em um consumo maior de recursos computacionais.

Por outro lado, o Word2Vec consome menos memória em comparação ao BOW. Ele representa as palavras como vetores de tamanho fixo, independentemente do tamanho da frase ou do texto. A representação vetorial mais compacta permite uma redução na quantidade de memória necessária para armazenar as informações.

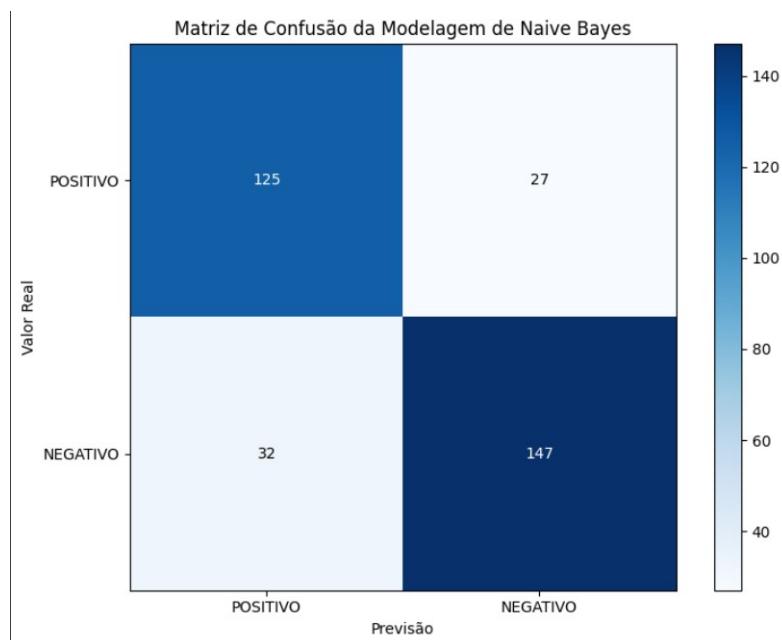
Além disso, o Word2Vec mantém uma dimensão consistente para todas as frases (52 colunas, já com acréscimo da variável-alvo e das frases utilizadas), independentemente do tamanho delas. Isso significa que o arquivo.csv gerado pelo Word2Vec terá uma estrutura mais padronizada, facilitando o processamento e a análise dos dados. Em contraste, o BOW pode resultar uma quantidade variável de colunas (Atualmente conta com 18582, já com acréscimo da variável-alvo e das frases utilizadas), tornando o processamento mais complexo e exigindo etapas adicionais de pré-processamento dos dados.

Portanto, considerando os fatores de consumo de memória e estrutura do arquivo.csv gerado, o modelo Word2Vec mostra-se mais vantajoso em relação ao BOW. Ele consome menos memória e mantém uma dimensão consistente, facilitando o processamento e a análise dos dados, até então.

12.2 Naive Bayes

Na figura 41 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Naive Bayes. Com os resultados obtidos em relação à variável alvo, “Sentimento”, que pode ser definida em sim ou não, gerou-se a matriz com 147 verdadeiros negativos, 27 falsos negativos, 32 falsos positivos e 125 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. Os resultados da acurácia para o algoritmo de Naive Bayes, para os dados de treino foram de 82%.

Figura 41: Matriz de Confusão - Naive Bayes

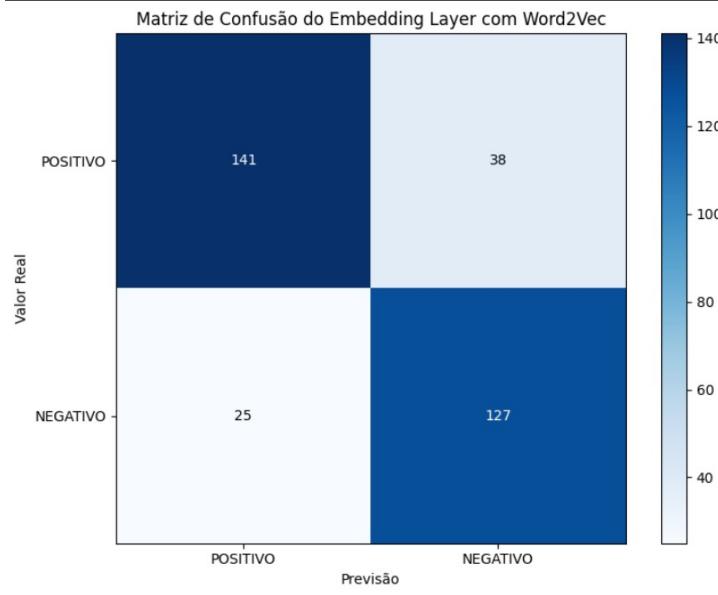


Fonte: Autores

12.3 Embedding Layer

Na figura 42 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Embedding Layer. Com os resultados obtidos em relação à variável alvo, “Sentimento”, que pode ser definida em sim ou não, gerou-se a matriz com 127 verdadeiros negativos, 25 falsos negativos, 38 falsos positivos e 141 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. Os resultados da acurácia para o algoritmo de Embedding Layer, para os dados de teste foram de 80%.

Figura 42: Matriz de Confusão - Embedding Layer

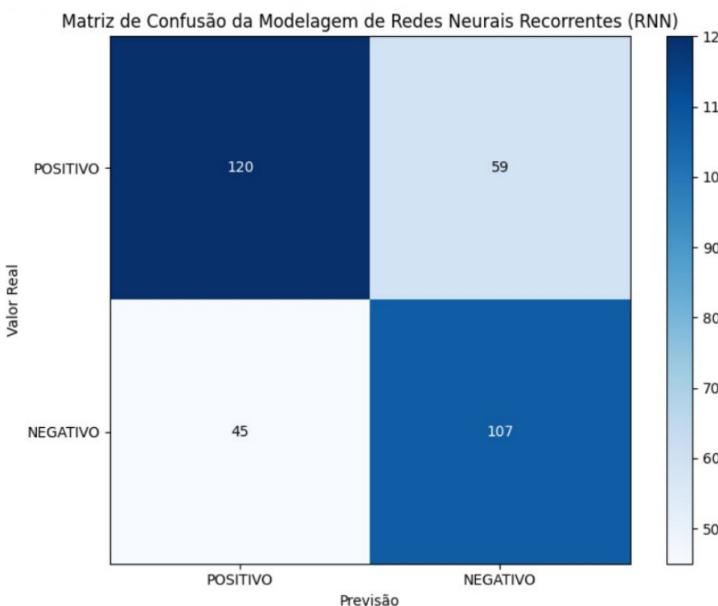


Fonte: Autores

12.4 Rede Neural Recorrente (RNN)

Na figura 43 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Rede Neural Recorrente. Com os resultados obtidos em relação à variável alvo, “Sentimento”, que pode ser definida em sim ou não, gerou-se a matriz com 107 verdadeiros negativos, 59 falsos negativos, 45 falsos positivos e 120 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. Os resultados da acurácia para o algoritmo de Rede Neural Recorrente, para os dados de teste foram de 68%.

Figura 43: Matriz de Confusão - RNN



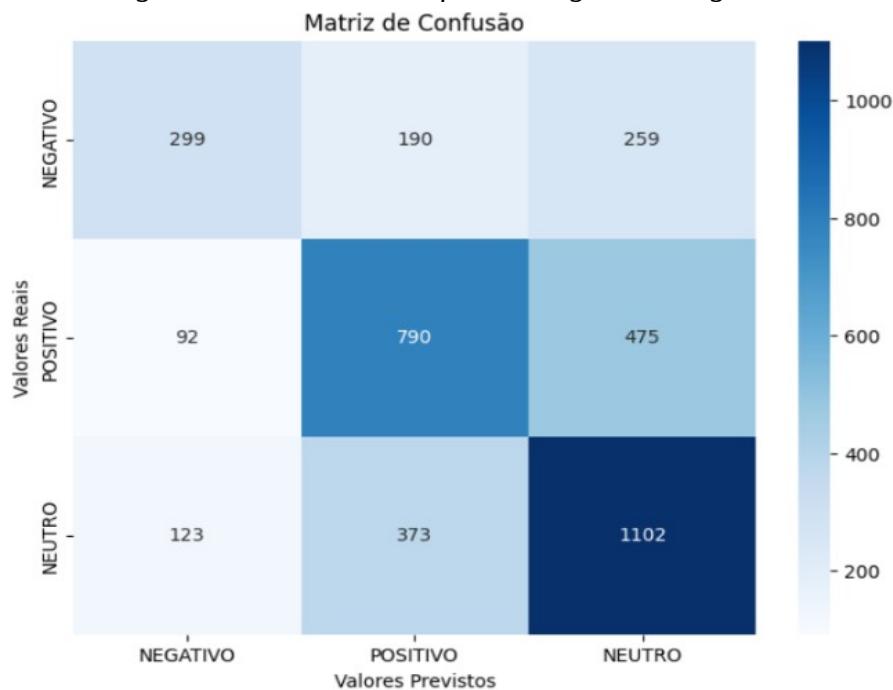
Fonte: Autores

12.4 Regressão Logística (Vetorização TF-IDF)

Na matriz de confusão do modelo de análise de sentimentos baseado em regressão logística e TF-IDF, podemos observar diferentes resultados. O modelo apresentou um comportamento de evitar classificar erroneamente as mensagens como negativas, com apenas 514 frases classificadas dessa maneira. Em contraste, o modelo fez 1355 classificações corretas de mensagens positivas e 1835 classificações corretas de mensagens neutras, indicando uma preferência por essas duas categorias.

A precisão geral do modelo, medida pela acurácia, foi de 59%. Isso significa que 59% das classificações feitas pelo modelo estavam corretas em relação ao sentimento real das mensagens. Além disso, a revocação, que é a proporção de casos positivos corretamente identificados em relação ao total de casos positivos, também foi de 59%. Apresenta-se a imagem abaixo, na figura 44.

Figura 44: Matriz de Confusão - Regressão Logística



Fonte: Autores

13. Novos Resultados

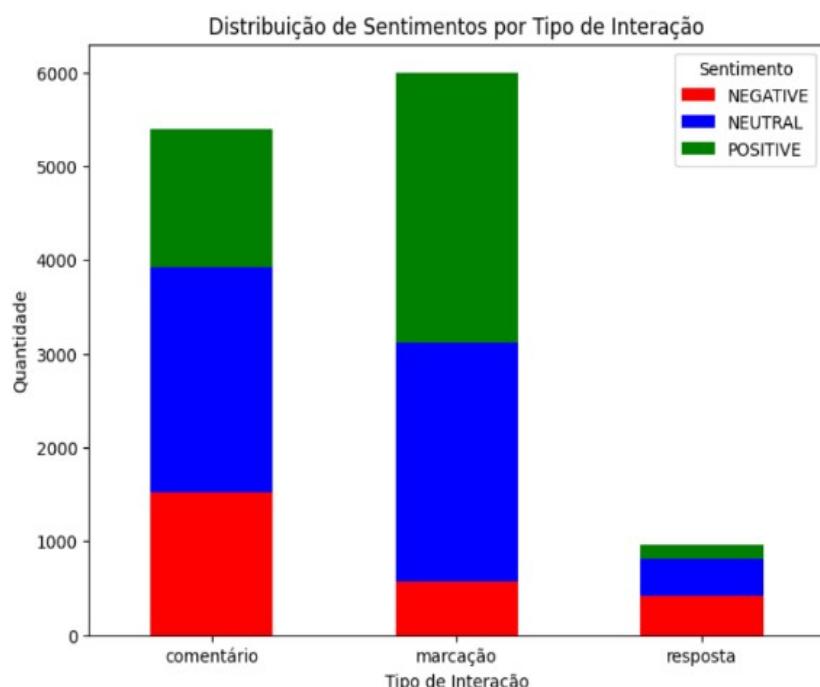
Nesta sessão apresenta-se a etapa do projeto em que são revisados os progressos feitos até o momento e são apresentados os resultados alcançados. Neste caso, mostra-se os novos resultados, após a implementação de novas features, subconjuntos e hiperparâmetros nos modelos.

13.1 Novas Features

No projeto foram utilizadas as seguintes features: “sentimento” e “frase”. Porém neste momento são aplicados novas features ao projeto, sendo elas as colunas “comentário”, “marcação” e “resposta”, todas elas são provenientes da coluna “tipointeração” que após a aplicação do método one hot encoding formou as 3 novas features. O método one-hot é uma técnica para representar variáveis categóricas como vetores binários, onde 1 representa a presença da categoria e 0 representa a ausência. A biblioteca pandas em Python oferece a função `get_dummies()` para realizar essa codificação.

Além disso, os modelos foram todos aplicados com a base de dados separado 30/70 de treino e teste com 3 categorias (positivo, negativo e neutro) sem balanceamento. Por fim o gráfico apresentado na figura 45 abaixo apresenta a distribuição dos sentimentos por tipo de interação, visto que acreditávamos que esse suposto “padrão” melhoraria nossos modelos.

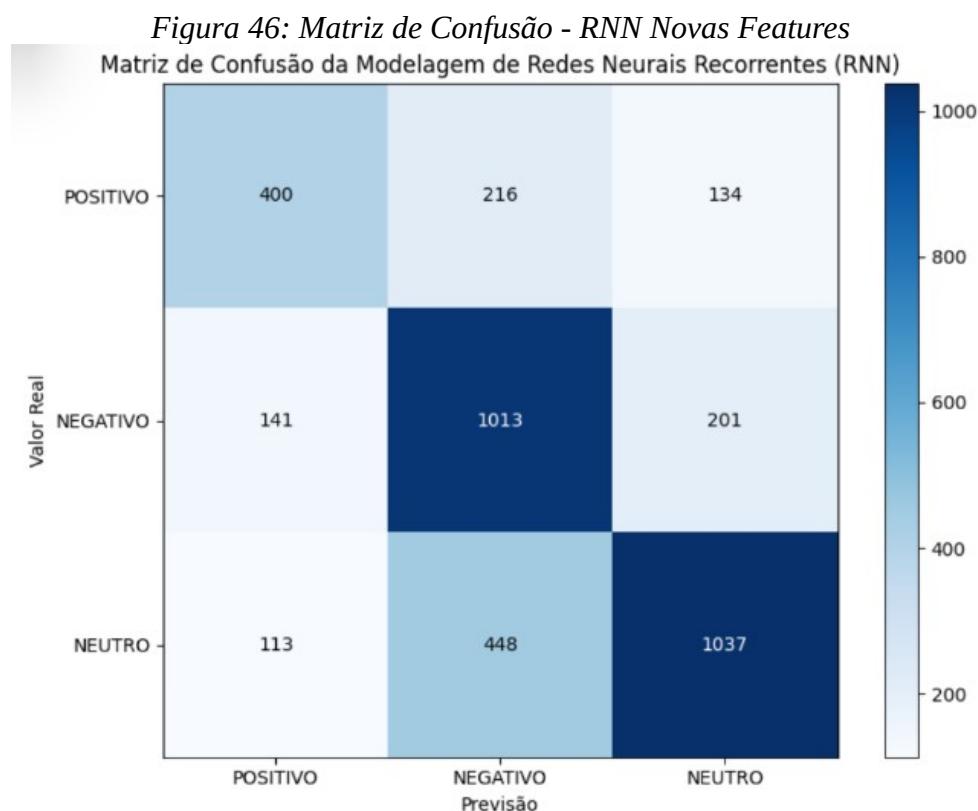
Figura 45: Sentimento X Interação - Novas Features



Fonte: Autores

13.1.1 Aplicação das novas features no modelo RNN

Na figura 46 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Rede Neural Recorrente. Com os resultados obtidos em relação à variável alvo, “Sentimento”, foi gerada a matriz com 1013 verdadeiros negativos, 342 falsos negativos, 350 falsos positivos e 400 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. O resultado do recall para o algoritmo de Rede Neural Recorrente foi de 37%.

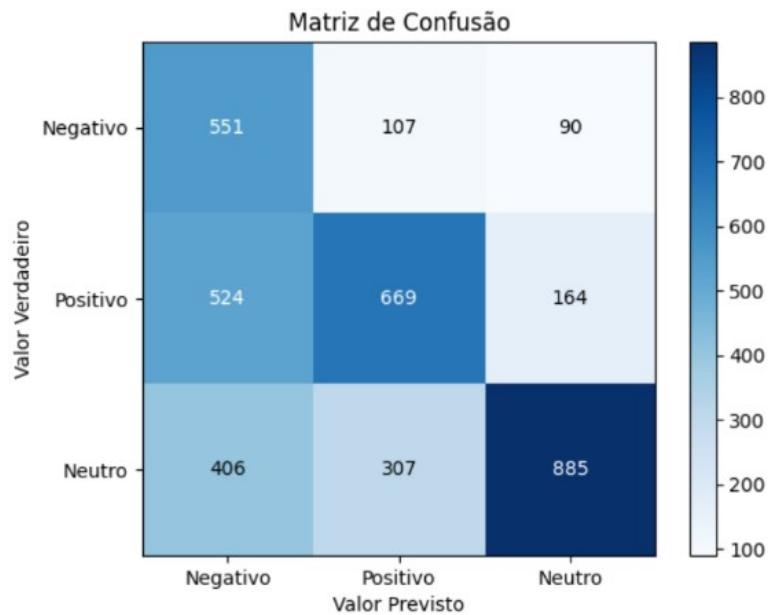


Fonte: Autores

13.1.2 Aplicação das novas features no modelo Naive Bayes

Na figura 47 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Naive Bayes. Com os resultados obtidos em relação à variável alvo, “Sentimento”, foi gerada a matriz com 551 verdadeiros negativos, 197 falsos negativos, 688 falsos positivos e 699 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. O resultado do recall para o algoritmo de Naive Bayes foi de 56%.

Figura 47: Matriz de Confusão - Naive Bayes Novas Features

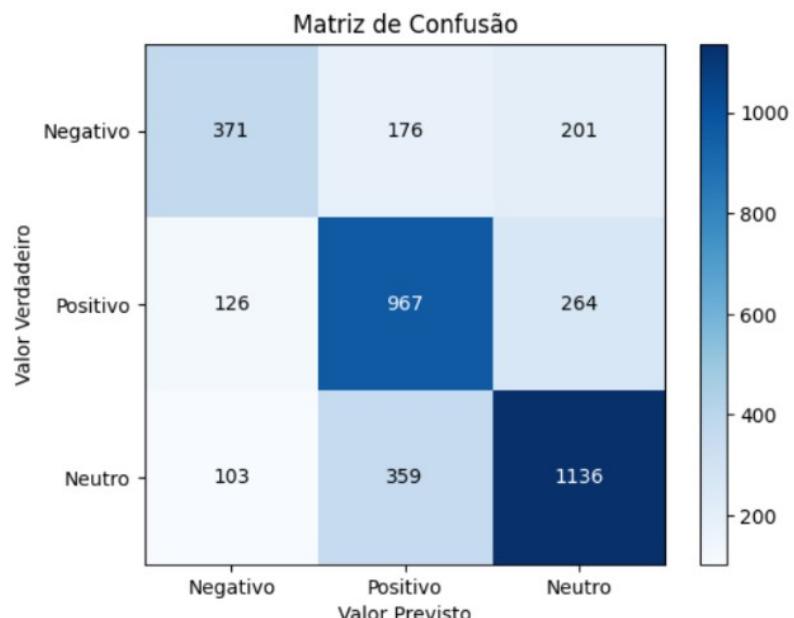


Fonte: Autores

13.1.3 Aplicação das novas features no modelo SVM

Na figura 48 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Support Vector Machine(SVM). Com os resultados obtidos em relação à variável alvo, “Sentimento”, foi gerada a matriz com 371 verdadeiros negativos, 377 falsos negativos, 390 falsos positivos e 967 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. O resultado do recall para o algoritmo de Support Vector Machine(SVM) de 71.26%.

Figura 48: Matriz de Confusão - SVM novas Features

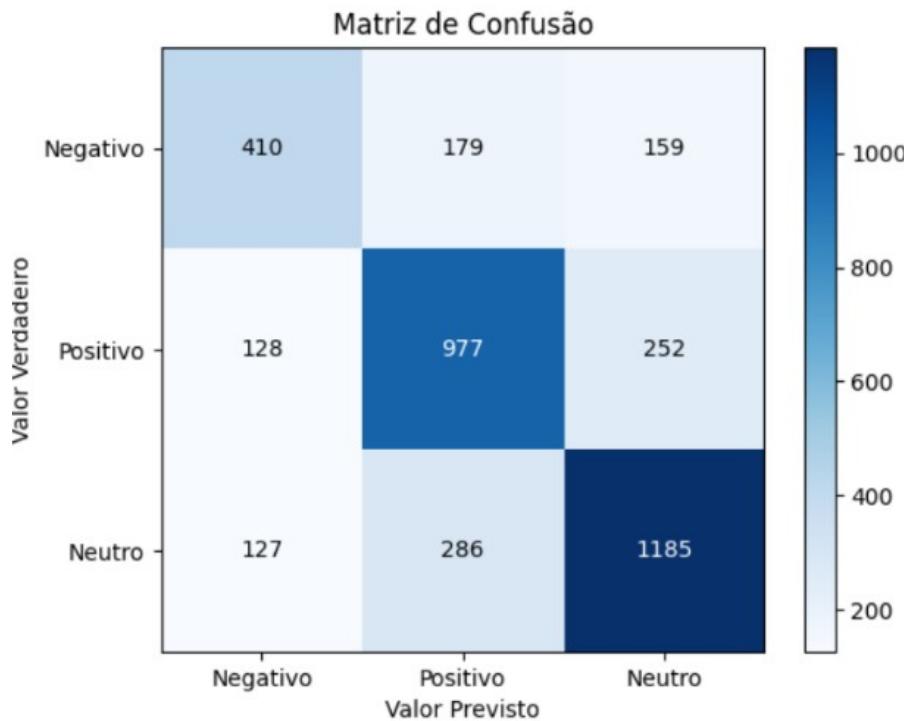


Fonte: Autores

13.1.4 Aplicação das novas features no modelo Random Forest

Na figura 49 abaixo é exibido uma visualização gráfica da matriz de confusão para o algoritmo de Random Forest. Com os resultados obtidos em relação à variável alvo, “Sentimento”, foi gerada a matriz com 410 verdadeiros negativos, 338 falsos negativos, 380 falsos positivos e 977 verdadeiros positivos. Assim sendo possível visualizar o resultado obtido pelo modelo. O resultado do recall para o algoritmo de Random forest para os dados de teste foram de 66%.

Figura 49: Matriz de Confusão - Random Forest Novas Features



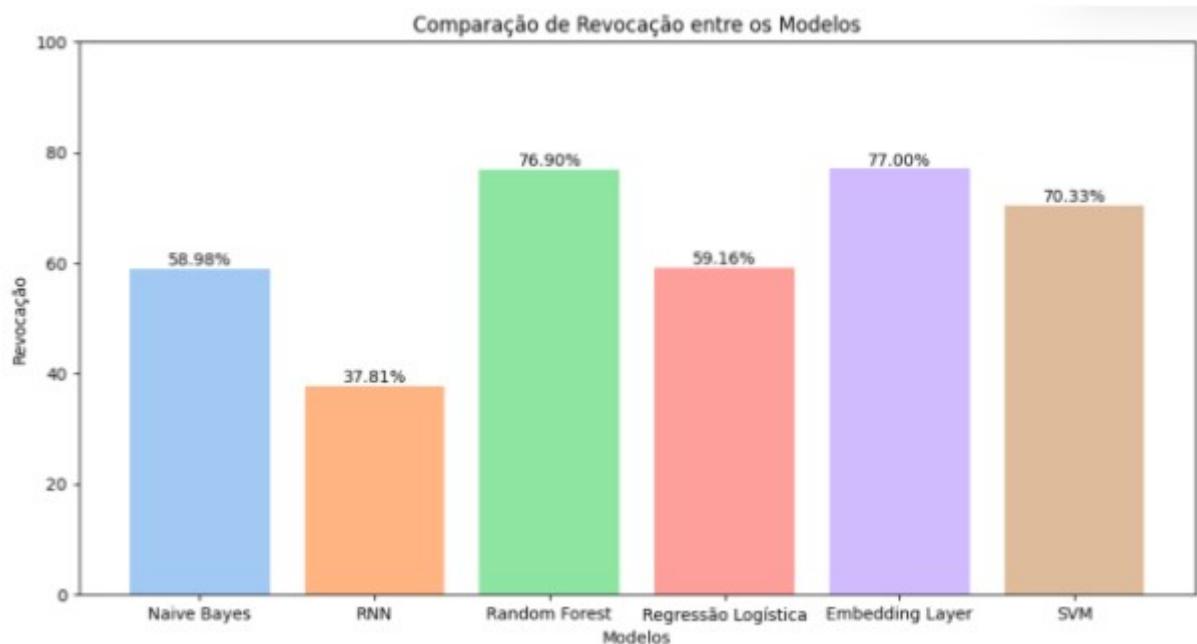
Fonte: Autores

13.1.5 Comparação dos modelos utilizando as novas features

Nesta sessão, são comparados os modelos utilizando as novas features implementadas. Os modelos são avaliados com base na maneira como lidam com essas novas features, buscando identificar qual deles apresenta um desempenho mais satisfatório ou uma melhor capacidade de aproveitamento, em comparação com as features testadas anteriormente. A comparação considera para a análise dos resultados obtidos o recall de cada modelo.

No gráfico apresentado na figura 50 abaixo, é possível observar a comparação do recall entre os modelos sem a aplicação das novas features.

Figura 50: Gráfico - Comparação dos modelos – Sem as Novas Features



Fonte: Autores

O objetivo é identificar quais modelos se destacam no contexto das novas features e, tomar decisões se será utilizado o modelo com as features antigas ou substitui-los pelas novas. A comparação é apresentada na tabela 10 abaixo.

Tabela 10: Comparação dos modelos - Novas Features

Modelos	Antes (recall)	Depois (recall)
RNN	0,56	0,37
Naive Bayes	0,72	0,56
SVM	0,78	0,71
Random forest	0,75	0,66

Fonte: Autores

Após a análise da aplicação das novas features nos modelos, é possível verificar que não houve uma melhoria significativa no desempenho dos mesmos, em comparação

com as features antigas. Pelo contrário, observamos uma piora na métrica do recall que avalia sua eficácia e precisão.

Diante dessa constatação, leva-se em consideração a relevância e o impacto das features, "frase" e "sentimento2", mantendo o foco nas features principais, pode-se direcionar melhor os recursos e esforços para aprimorar e otimizar os modelos existentes, em vez de explorar novas features que não trazem benefícios significativos no contexto atual.

13.2 Hiperparâmetros

Os hiperparâmetros são parâmetros cujos valores são usados para controlar o processo de aprendizado de máquina, ou seja, são atributos que controlam o treinamento do modelo. Nesse sentido, eles previnem o modelo de aprender apenas com os dados mostrados (overfitting e underfitting), tornando-o capaz de generalizar para outras situações possíveis. Os hiperparâmetros foram encontrados a partir da busca aleatória (random search) e por grade (grid search). A escolha do modelo de busca seguiu o tempo de processamento, assim a maioria dos modelos foram encontrados a partir do random search.

13.2.1 Random Search

O Random Search é um método que busca encontrar o melhor hiperparâmetro de forma programática. Com isso, o algoritmo encontrará a melhor solução. Essa técnica testa combinações aleatórias e os melhores resultados funcionam como um guia para a escolha dos próximos hiperparâmetros. Os hiperparâmetros dos modelos Naive Bayes, RNN, Embedding Layer foram definidos utilizando essa técnica.

13.2.1.1 Naive Bayes

Na tabela 11 abaixo é definido o hiperparâmetro do modelo Naive Bayes.

Tabela 11: Hiperparâmetros - Naive Bayes

Hiperparâmetro	Definição
var_smoothing	Utilizado para suavizar a estimativa de variância das características em Naive Bayes.

Fonte: Autores

13.2.1.2 RNN

Na tabela 12 abaixo são definidos os hiperparâmetros do modelo RNN.

Tabela 12: Hiperparâmetros - RNN

Hiperparâmetro	Definição
batch_size	Exemplos de treinamento em cada etapa
learning_rate	Tamanho do “passo” dado em cada iteração de otimização.
num_epochs	Número de vezes que o modelo passa por todo o conjunto.
num_hidden_units	Neurônios presentes em uma camada oculta

Fonte: Autores

13.2.1.3 Embedding Layer

Na tabela 13 abaixo são definidos os hiperparâmetros do modelo Embedding Layer.

Tabela 13: Hiperparâmetros - Embedding Layer

Hiperparâmetro	Definição
embedding_dim	Dimensão do espaço de incorporação
batch_size	Exemplos de treinamento em cada etapa

Fonte: Autores

13.2.1.4 Random Forest

Na tabela 14 abaixo são definidos os hiperparâmetros do modelo Random Forest.

Tabela 14: Hiperparâmetros - Random Forest

Hiperparâmetro	Definição
max_depth: 15	Profundidade máxima das árvores na floresta.
min_samples_split: 4	Número mínimo de amostras para que um nó da árvore possa ser dividido em dois nós filhos.
n_estimators: 437	Número de árvores de decisão na floresta.

Fonte: Autores

13.2.1.5 Regressão Logística

Na tabela 15 abaixo são definidos os hiperparâmetros do modelo de Regressão Logística.

Tabela 15: Hiperparâmetros - Regressão Logística

Hiperparâmetro	Definição

C	Inversa da força de regularização na regressão logística
solver	Utilizado para otimizar a função de custo na regressão logística

Fonte: Autores

13.2.2 Grid Search

O Grid Search testa todas as combinações possíveis dos hiperparâmetros, exaustivamente. Então, ela fornece alguns valores de input e testa todas as combinações. Em seguida, selecionará os hiperparâmetros que obtiveram o menor erro. Os hiperparâmetros do modelo SVM foram definidos utilizando essa técnica.

13.2.2.1 SVM

Na tabela 16 abaixo são definidos os hiperparâmetros do modelo SVM.

Tabela 16: Hiperparâmetros - SVM

Hiperparâmetro	Definição
C: 10	Parâmetro de regularização que controla a penalidade por erros de classificação.
Gamma: 'scale'	Parâmetro que afeta a influência de treinamento no ajuste dos hiperplanos.
Kernel: 'linear'	Função para transformar os dados em um espaço de maior dimensão.

Fonte: Autores

13.3 Novo Subconjunto

Nesta sessão, exploramos um novo subconjunto de dados composto por dois conjuntos distintos, cada um deles caracterizado por diferentes quantidades de categorias e com variações em relação ao balanceamento dos dados.

O objetivo principal foi avaliar o desempenho dos modelos em cenários com características diversas, a fim de compreender como eles se comportam diante de diferentes distribuições de categorias e a influência do balanceamento dos dados nos resultados. Para isso, realizamos testes comparativos nos dois conjuntos de dados, analisando a métrica recall. Observamos como cada modelo respondeu às variações e quais foram os impactos das diferentes configurações nos resultados obtidos.

13.3.1 Subconjunto com três categorias balanceadas

O primeiro subconjunto, é composto por três categorias principais: Negativos, Neutro e Positivos. O conjunto de dados possui um total de 7.572 entradas. Para garantir uma análise confiável e robusta, optamos por dividir o conjunto em duas partes: uma para treinamento e outra para teste. A divisão foi realizada da seguinte forma: 70% dos dados, ou seja, 5.300 entradas, foram destinados ao treinamento do modelo, enquanto 30% dos dados, correspondendo a 2.272 entradas, foram reservados para o teste.

Uma etapa importante nesse processo foi o balanceamento dos dados. Levando em consideração que os dados negativos apresentavam uma quantidade menor em relação às outras categorias, utilizamos essa categoria como base para o balanceamento. Dessa forma, ajustamos a distribuição dos dados para que cada categoria tivesse uma representatividade equilibrada no conjunto de treinamento e teste. Esse balanceamento é fundamental para evitar um viés em relação às categorias majoritárias e permitir que o modelo aprenda de forma equilibrada com exemplos de cada categoria. Assim, garantimos que o desempenho do modelo seja consistente e capaz de lidar de maneira adequada com cada uma das categorias.

13.3.2 Subconjunto com três categorias não balanceadas

O segundo subconjunto, também composto por três categorias: Negativos, Neutro e Positivos, possui um total de 12.343 entradas. Assim como no caso anterior, divide-se o conjunto em duas partes para realizar a avaliação dos modelos: 70% dos dados, correspondendo a 8.640 entradas, foram destinados ao treinamento, enquanto os 30% restantes, totalizando 3.703 entradas, foram reservados para o teste.

Ao contrário do primeiro subconjunto, neste caso não foi realizado um balanceamento específico das categorias. Isso significa que a distribuição dos dados nas categorias Negativos, Neutro e Positivos não foi igualmente representativa. Cada categoria pode ter uma quantidade de dados diferente, refletindo a proporção original do conjunto de dados coletado. Essa falta de balanceamento pode influenciar no desempenho dos modelos, pois as categorias majoritárias tendem a ter mais exemplos para aprender, enquanto as categorias minoritárias podem receber menos atenção do modelo.

13.3.4 Escolha do subconjunto

Nesta sessão, foi realizada a escolha do subconjunto de dados a ser utilizado nos modelos. O selecionado é o conjunto de dados composto por três categorias sem balanceamento, mantendo os hiperparâmetros definidos anteriormente. Esse subconjunto específico apresenta as categorias Negativos, Neutro e Positivos, permitindo uma classificação abrangente dos sentimentos. Ele possui um total de 12.343 entradas, sendo dividido em 70% para treinamento (8.640 entradas) e 30% para teste (3.703 entradas).

É importante ressaltar que o conjunto de dados com três categorias balanceadas será aplicado apenas no modelo que apresentar o melhor desempenho com o processamento de dados que estamos utilizando atualmente. A escolha desse subconjunto sem balanceamento nos permite avaliar como os modelos se comportam em um cenário em que as categorias podem ter diferentes proporções de dados.

Além disso, caso o modelo escolhido apresente um bom desempenho no subconjunto sem balanceamento, poderemos aplicar o conjunto de dados平衡ado com três categorias no mesmo modelo. Isso nos permitirá explorar a capacidade desse modelo em lidar com distribuições equilibradas de dados e fornecer uma classificação precisa e confiável dos sentimentos.

13.4 Comparativo modelos

Nesta sessão, apresentamos uma análise comparativa dos modelos submetidos ao conjunto de dados composto por três categorias sem balanceamento. Utilizamos o recall como métrica principal para avaliar o desempenho de cada modelo. Na figura 50 abaixo se apresenta os resultados obtidos por modelo testado. O recall foi escolhido como métrica de comparação devido à sua relevância na avaliação da capacidade dos modelos em identificar corretamente os casos de falsos negativos em relação ao total de casos reais. Pode-se verificar na figura 51 abaixo.

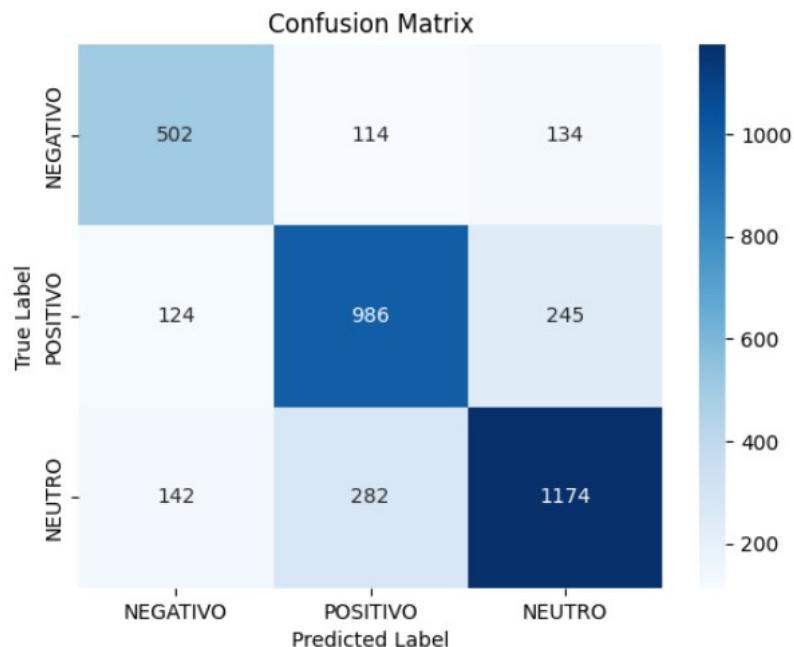
Figura 51: Comparação dos Modelos - Métrica Recall

Modelo	Resultado
0 Redes Neurais Recorrentes (RNN)	0.374139
1 Embedding Layer	0.998167
2 Naive Bayes	0.591925
3 SVM (Support Vector Machine)	0.715329
4 Random Forest	0.709425
5 Regressão Logística	0.591682

Fonte: Autores

Ao analisar os resultados apresentados na tabela, fica evidente que o modelo Embedding Layer se destacou ao obter o melhor desempenho em termos de recall. Esse resultado direciona a concentrar os esforços de aprimoramento e aplicação em interface nesse modelo específico. Apresenta-se na figura 52 abaixo sua matriz de confusão, utilizando o subconjunto com três categorias e sem平衡amento, além dos hiperparâmetros definidos para o Embedding Layer descritos na sessão “Hiperparâmetros” acima.

Figura 52: Matriz de Confusão - Modelo Escolhido - Embedding Layer



Fonte: Autores

14. Funções

Nesta sessão apresenta-se as descrições das funções desenvolvidas.

14.1 Pré-Processamento

Remover Emojis:

- A função removeEmoji() remove emojis de uma string de texto, retornando a string sem os emojis. Ela realiza a biblioteca emoji para fazer a remoção.
- A função removeEmoji_() aplica a função “removeEmoji” no dataframe.

Remover caracteres especiais:

- A função removeCaracteres(texto) remove caracteres especiais de uma string de texto fornecida.

Remover acentos:

- A função removeAcentos(x) remove os acentos do texto utilizando a biblioteca Unicode.

Normalizar texto:

- A função normalizar_texto(texto) utiliza um dicionário criado pelo grupo para transformar abreviações em palavras e também desconsiderar algumas palavras específicas.
- A função normalizar_texto_(removeAcento_) aplica a função normalizar_texto no dataframe.

Processar texto:

- A função processarTexto(texto) recebe uma string de texto como entrada e converte o texto para minúsculas, divide-o em tokens (palavras), cria uma lista de stopwords para o idioma português e filtra os tokens removendo as stopwords. Ela realiza essas etapas através da biblioteca “NLTK”.
- processarTexto_(normTexto) aplica a função processarTexto no dataframe.

Bag of Words:

- A função `bag_of_words(frases)` realiza o bow utilizando a biblioteca “`sklearn`”, montando o dicionário e criando a matriz BOW

Gráficos:

- A função `color_func()` recebe informações sobre uma palavra (como o tamanho da fonte, posição e orientação) e retorna uma cor com base no sentimento associado a essa palavra.

Codificar colunas:

- A função `codificar_coluna(coluna)` recebe uma coluna de um DataFrame como entrada e realiza a codificação dessa coluna usando a técnica "one-hot encoding". Ela utiliza a função `pd.get_dummies()` do pandas para criar colunas binárias (0s e 1s) correspondentes aos diferentes valores presentes na coluna original.
- A função `normalizar_coluna(dados)` chama a função `codificar_coluna(dados)`. Em seguida, aplica a normalização z-score na coluna codificada usando `StandardScaler()` da biblioteca `sklearn.preprocessing`. A função retorna um novo DataFrame contendo as colunas normalizadas. São utilizadas as bibliotecas “`pandas`” e “`sklearn`”

14.2 Implementação

Tratamento de texto

- A função `preprocess_sentence(sentence)` realiza o pré-processamento de um texto. Ela divide a frase em palavras individuais, remove as stopwords específicas para o idioma português e retorna a lista de palavras resultante. Ela utiliza a biblioteca “`sklearn`”
- A função `process_objects(objects)` recebe uma lista de objetos como entrada e processa cada objeto individualmente. Para cada objeto, a função chama a função `preprocess_sentence(obj)` para realizar o pré-processamento.

Modelo Word2Vec

- A função `vetorizar_word2vec()` vetoriza o conjunto de frases usando o modelo Word2Vec, cria um DataFrame com os vetores resultantes e os rótulos das colunas, e retorna o mesmo.

Modelo RNN (rede neural recorrente)

- A função `classification_rnn()` realiza a classificação usando uma RNN, treinando o modelo, avaliando sua precisão e retornando o modelo e os dados de teste.

Modelo Embedding Layer

- A função `classification_embedding_layer()` realiza a classificação usando uma camada de embedding em um modelo de rede neural. Faz a tokenização, o treinamento do modelo, e retorna a perda, precisão, previsões, rótulos reais e dados de teste.

Modelo Naive Bayes

- A função `classification_naive_bayes()` realiza a classificação usando o método Naive Bayes Gaussiano, treina o modelo, faz previsões e retorna a precisão, rótulos reais e previsões.

Modelo SVM (Support Vector Machine)

A função `classification_svm()` realiza o treinamento de um modelo SVM para classificação, faz previsões no conjunto de teste, calcula acurácia, F1-score e recall, realiza validação cruzada e retorna o modelo, o conjunto de teste e as previsões.

Modelo Random Forest

A função `classification_random_forest()` realiza o treinamento de um modelo Random Forest para classificação, faz previsões no conjunto de teste, calcula acurácia, F1-score e recall, realiza validação cruzada e retorna o modelo, o conjunto de teste e as previsões.

Modelo Regressão Logística

A função `classification_logistic_regression()` realiza o treinamento de um modelo de Regressão Logística com vetorização TF-IDF, faz previsões no conjunto de teste, calcula acurácia, F1-score e recall, realiza validação cruzada e retorna o modelo, o conjunto de teste e as previsões.

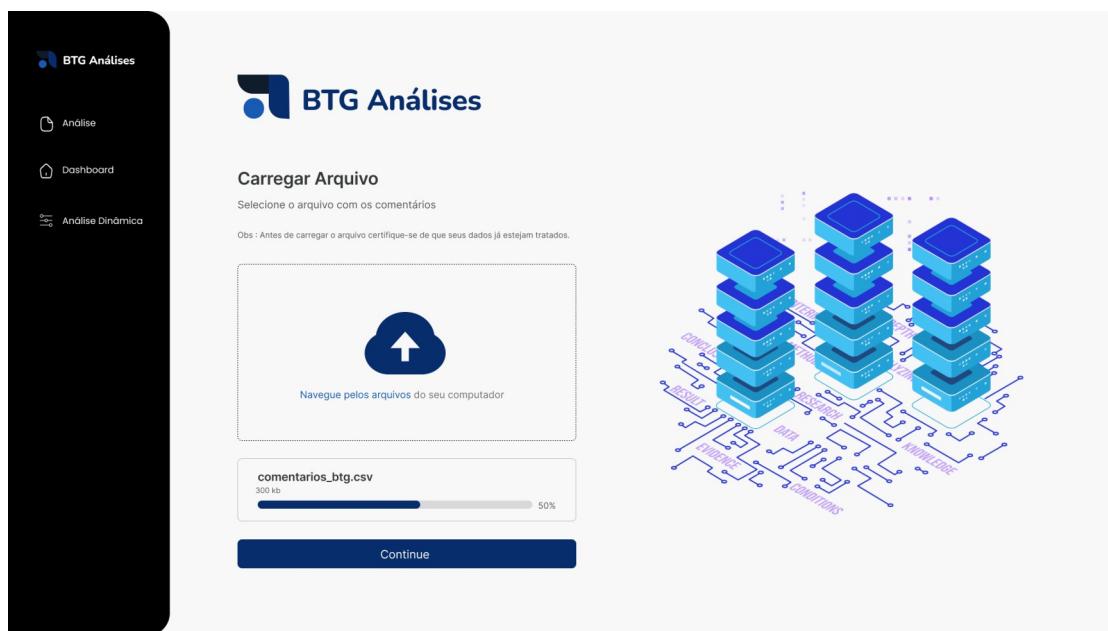
15. Prototipação

O protótipo de interface para o usuário, foi construído como um modelo que representa as telas necessárias para o aplicativo, baseando-se nas user stories planejadas anteriormente. Utilizou-se três princípios para a elaboração das telas, sendo eles: 1) Familiaridade com o usuário, priorizando termos, ícones e conceitos que o público-alvo possua base de experiência; 2) Consistência, onde as funcionalidades previstas para a solução, podem ser realizadas de forma similares; e 3) Confirmação de execuções, alarmes significativos para o usuário confirmar suas ações e evitar possíveis erros.

Tela 1: Carregamento de Arquivo CSV

A primeira tela prototipada é responsável pelo carregamento de um arquivo CSV contendo dados de comentários do Instagram. A interface apresenta uma área central para selecionar o arquivo a ser carregado. A parte superior da tela inclui uma seção de informações, explicando o formato esperado do arquivo CSV. Abaixo se pode visualizar seu resultado na figura 53 abaixo.

Figura 53: Tela 1 : Carregamento CSV



Fonte: Autores

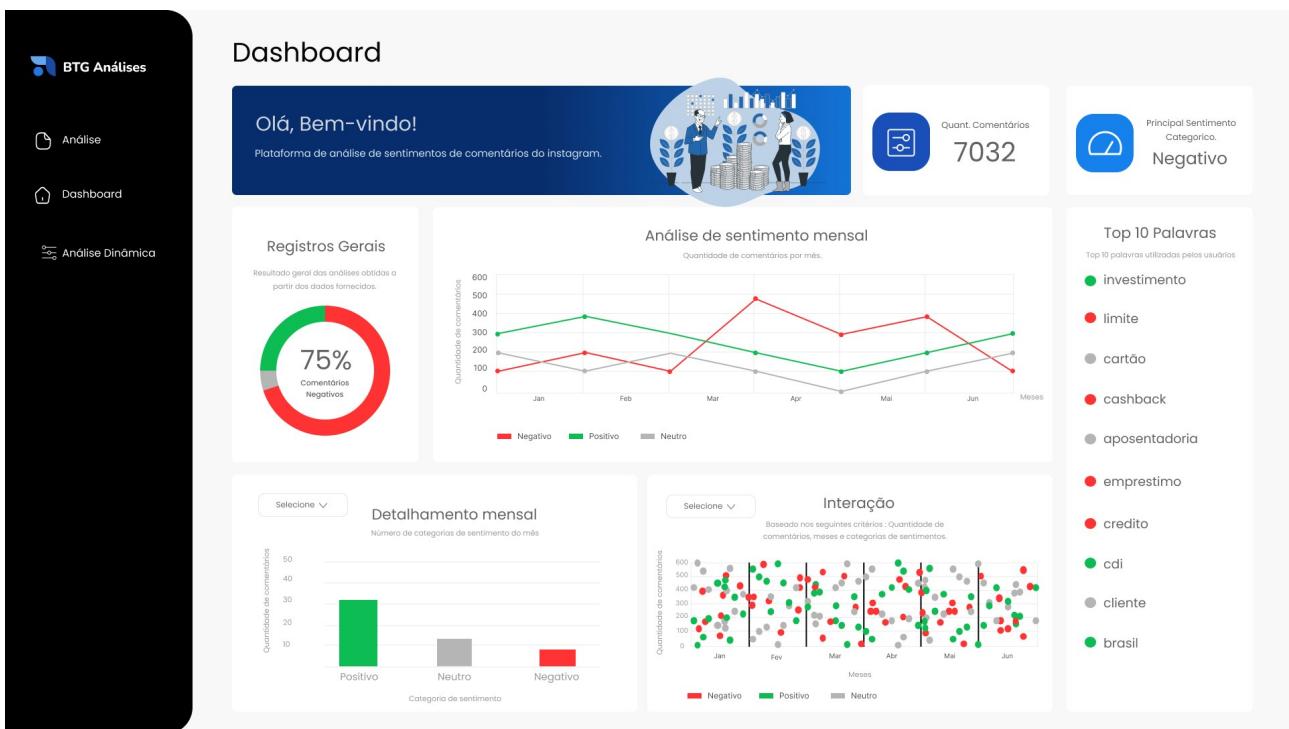
Tela 2: Dashboard de Análise Descritiva

A segunda tela prototipada é um dashboard que apresenta a análise descritiva dos dados carregados a partir do arquivo CSV. A interface é composta por vários gráficos e métricas visuais que fornecem insights sobre os sentimentos presentes nos comentários. Alguns elementos típicos incluem:

- Gráfico de barras: exibe a distribuição dos sentimentos (positivo, neutro, negativo) encontrados nos comentários. Cada barra representa a quantidade ou a proporção de comentários em cada categoria.
- Gráfico de dispersão: mostra a relação entre comentários, a data que foi publicada e seu sentimento. Os pontos são plotados em um gráfico bidimensional, onde o eixo y representa comentário e o eixo x representa os meses.

Pode-se verificar seu resultado na figura 54 abaixo.

Figura 54: Tela 2: Dashboard



Fonte: Autores

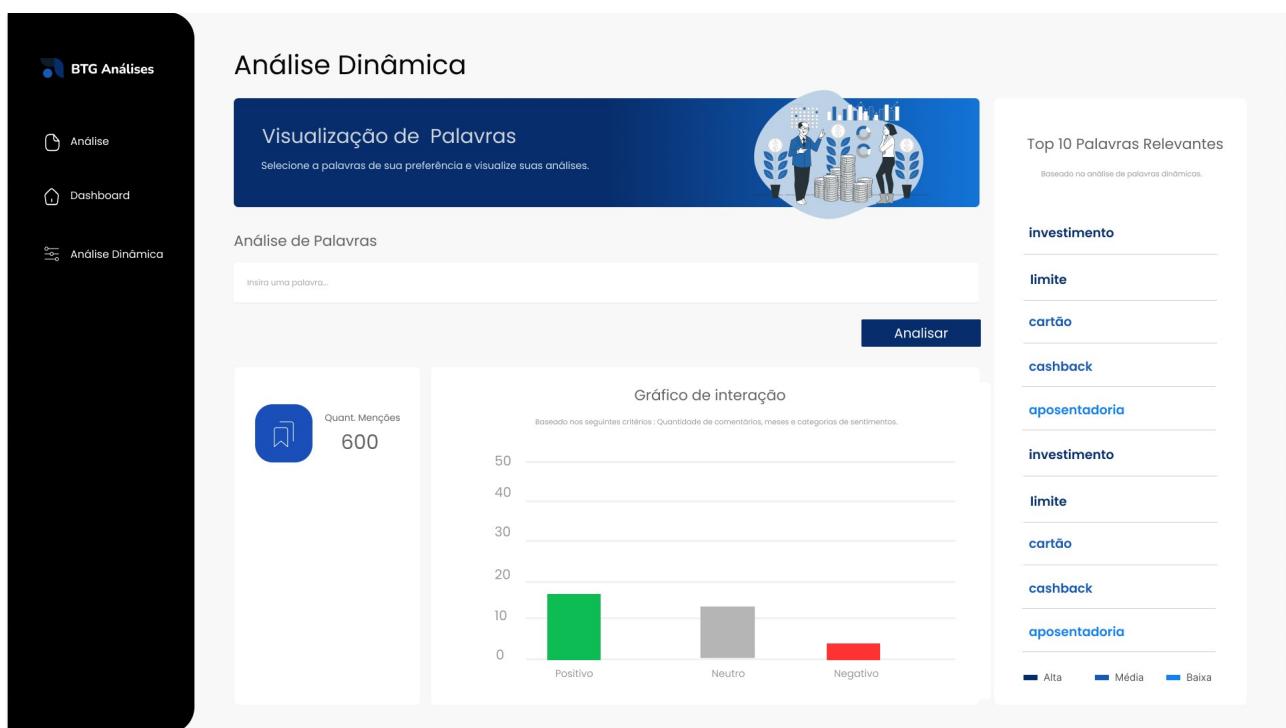
Tela 3: Análise Dinâmica

A terceira tela prototipada permite uma análise dinâmica com base em uma palavra-chave fornecida pelo usuário. A interface possui um campo de entrada de texto onde o usuário pode digitar uma palavra específica. Após a inserção da palavra, a tela exibe duas seções principais:

- Palavras Similares: apresenta uma lista de palavras semelhantes à palavra-chave fornecida pelo usuário. Essa funcionalidade permite ao usuário explorar termos relacionados e expandir sua análise.
- Sentimentos Associados: exibe os sentimentos mais frequentes associados à palavra-chave. Apresentado em forma de gráfico de barra, onde cada coluna representa uma categoria de sentimento (positivo, neutro, negativo) e sua proporção relativa. Essa visualização permite ao usuário compreender rapidamente como diferentes sentimentos estão relacionados à palavra-chave específica.

Pode ser visualizado na figura 55 abaixo.

Figura 55: Tela 3: Análise Dinâmica



Fonte: Autores

As telas prototipadas foram desenvolvidas utilizando a ferramenta Figma e podem ser acessadas [clicando aqui](#).

Além disso, as telas foram implementadas tanto no frontend quanto no backend, e os códigos correspondentes podem ser encontrados dentro da pasta "código fonte". Os arquivos estão organizados de forma a facilitar a compreensão da estrutura e lógica por trás da interface prototipada.

Para auxiliar no entendimento e utilização do sistema, foi elaborada uma documentação, disponível no arquivo "[t4_G5_V5_PLN_Instalação.pdf](#)" dentro da pasta "docs". Essa documentação contém informações detalhadas sobre a instalação do sistema, requisitos de software, configurações necessárias e instruções de uso.

Ao combinar a prototipação no Figma com a implementação completa no frontend e backend, a solução oferece uma experiência interativa e funcional para análise de sentimentos de comentários, proporcionando aos usuários uma interface intuitiva e recursos poderosos para explorar e compreender os dados.

16. Conclusões e Recomendações

Diante dos algoritmos implementados, a escolha pelo Word2Vec como modelo de representação vetorial se deu principalmente pelo menor consumo de memória, sua capacidade de compreender o contexto das frases para a classificação de sentimentos e a manutenção de dimensões fixas nas colunas do arquivo.csv gerado. Essas características favorecem a eficiência e a facilidade de processamento dos dados.

No que diz respeito à performance dos modelos, com base na análise dos resultados de recall, foi constatado que o modelo Embedding Layer obteve um desempenho superior em relação aos demais modelos considerados. Esse resultado foi alcançado após a aplicação de hiperparâmetros e utilizando o conjunto de dados contendo três categorias, sem balanceamento.

Portanto, o modelo Embedding Layer alcançou um recall de 99.81%, se mostrando eficiente em identificar e classificar corretamente os casos negativos. Essa precisão na detecção de casos negativos é de extrema importância já que visamos a análise de sentimentos de comentários para insights e melhorias dentro do contexto empresarial.

17. Referências

BANCO CENTRAL DO BRASIL (Brasil). Tarifas Bancárias. [S. I.], 2023. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/tarifas_bancarias. Acesso em: 25 abr. 2023.

BTG PACTUAL (São Paulo - Brasil). Relatório Anual 2021: Negócios, estratégia e desempenho. Com efetiva integração ESG.. [S. I.], 2021. Disponível em: <https://static.btgpactual.com/media/rs2021-btgpactual-vf1.pdf>. Acesso em: 25 abr. 2023.

IBM SPSS Modeler CRISP-DM Guide. recurso online. Disponível em: https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf. Acesso em: 22 maio 2023.

INTELIGÊNCIA artificial: uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2021. 1 recurso online. ISBN 9788521637509. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788521637509>. Acesso em: 22 maio 2023.

JUROS%BAIXO (Brasil). Compare as tarifas de 6 bancos: conta corrente. [S. I.], 17 dez. 2018. Disponível em: <https://jurosbaixos.com.br/conteudo/compare-as-tarifas-de-6-bancos-para-abrir-uma-conta-corrente/>. Acesso em: 24 abr. 2023.

Problema, Esboço do problema, TAPI, Inteli, Página 4, 2023;

SIQUEIRA, Andressa. Conta corrente: quais são os melhores bancos para abrir conta? Descubra aqui!. [S. I.], 20 abr. 2021. Disponível em: <https://blog.magnetis.com.br/conta-corrente/>. Acesso em: 24 abr. 2023.

TORRES, Vitor. O que é ROI?: como calcular retorno sobre o investimento?. Contabilizei.blog, 11 out. 2022. Disponível em: <https://www.contabilizei.com.br/contabilidade-online/o-que-e-roi-como-calcular-retorno-sobre-o-investimento/>. Acesso em: 24 abr. 2023.

"Conheça e saiba como aplicar a acurácia" - ClearSale. Disponível em: <https://blogbr.clear.sale/conheca-e-saiba-como-aplicar-a-acuracia>. Acesso em 22 maio 2023.

"Entendendo o que é matriz de confusão com Python" - Data Hackers. Disponível em:
<https://medium.com/data-hackers/entendendo-o-que-%C3%A9-matriz-de-confus%C3%A3o-com-python-114e683ec509>. Acesso em 22 maio 2023.

"O que é Word2Vec?" - edrone. Disponível em: <https://edrone.me/pt/blog/o-que-e-word2vec>. Acesso em 22 maio 2023.

"Introdução a Bag of Words e TF-IDF" - Turing Talks. Disponível em: <https://medium.com/turing-talks/introdu%C3%A7%C3%A3o-a-bag-of-words-e-tf-idf-43a128151ce9>. Acesso em 22 maio 2023

."Word2Vec Explained" - Towards Data Science. Disponível em: <https://towardsdatascience.com/word2vec-explained-49c52b4ccb71>. Acesso em 22 maio 2023.

"Word2Vec e sua importância na etapa de pré-processamento" - Medium. Disponível em:
<https://medium.com/@everton.tomalok/word2vec-e-sua-import%C3%A2ncia-na-etapa-de-pr%C3%A9-processamento-d0813acf8ab>. Acesso em 22 maio 2023.

"O que é Word2Vec?" - edrone. Disponível em: <https://edrone.me/pt/blog/o-que-e-word2vec>. Acesso em 22 maio 2023.

"Pretrained Word Embeddings in NLP" - Analytics Vidhya. Disponível em:
<https://www.analyticsvidhya.com/blog/2020/03/pretrained-word-embeddings-nlp/>. Acesso em 22 maio 2023.

"Word2Vec – Gensim" - Gensim. Disponível em: <https://radimrehurek.com/gensim/models/word2vec.html>. Acesso em 22 maio 2023.

"Pre-trained Word Embeddings or Embedding Layer: A Dilemma" - Towards Data Science. Disponível em:
<https://towardsdatascience.com/pre-trained-word-embeddings-or-embedding-layer-a-dilemma-8406959fd76c>. Acesso em 22 maio 2023.

"Como fazer embedding de frases para NLP no TensorFlow?" - Stack Overflow em Português. Disponível em: <https://pt.stackoverflow.com/questions/496584/como-fazer-embedding-de-frases-para-nlp-no-tensorflow>. Acesso em 22 maio 2023.

"What is Neural Network?" - AWS. Disponível em:
<https://aws.amazon.com/pt/what-is/neural-network/#:~:text=Uma%20rede%20neural>

%20%C3%A9%20um,camadas%2C%20semelhante%20ao%20c%C3%A9rebro%20humano. Acesso em 22 maio 2023.

"Um mergulho profundo nas Redes Neurais Recorrentes" - iMasters. Disponível em: <https://imasters.com.br/data/um-mergulho-profundo-nas-redes-neurais-recorrentes>. Acesso em 22 maio 2023.

"Tipos de Métodos - Naive Bayes" - IA Expert. Disponível em: https://iaexpert.academy/2019/04/24/tipos-de-metodos-naive-bayes/?doing_wp_cron=1684863123.7367320060729980468750. Acesso em 22 maio 2023.

"Modelo de Predição Naive Bayes" - Turing Talks. Disponível em: <https://medium.com/turing-talks/turing-talks-16-modelo-de-pred%C3%A7%C3%A3o-naive-bayes-6a3e744e7986>. Acesso em 22 maio 2023.

"Um Estudo Experimental Comparativo do Classificador Naive Bayes em Tarefas de Classificação de Textos" - Anais do CONAPESC. Disponível em: https://editorarealize.com.br/editora/anais/conapesc/2021/TRABALHO_EV161_MD1_SA105_ID2111_22092021170519.pdf. Acesso em 22 maio 2023.

"Bag of Words: One of the Simplest Techniques in Natural Language Processing" - MyGreatLearning. Disponível em: <https://www.mygreatlearning.com/blog/bag-of-words/>. Acesso em 22 maio 2023.

"Comparison between Bag-of-Words and Word2Vec" - PyImageSearch. Disponível em: <https://pyimagesearch.com/2022/07/18/comparison-between-bagofwords-and-word2vec/>. Acesso em 22 maio 2023.

"All You Need to Know about Bag of Words and Word2Vec: Text Feature Extraction" - Towards Data Science. Disponível em: <https://towardsdatascience.com/all-you-need-to-know-about-bag-of-words-and-word2vec-text-feature-extraction-e386d9ed84aa>. Acesso em 22 maio 2023.

NILC (Núcleo Interinstitucional de Linguística Computacional) : Repositório de Word Embeddings do NILC. Disponível em: <http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc> .Acesso em 22 maio 2023.

18. Anexos

Nesta seção apresenta-se o espaço destinado a informações complementares e relevantes ao conteúdo principal do projeto, utilizado para reforçar a argumentação do documento e contribuir para o entendimento completo.

18.1. Matriz de risco

A matriz de risco é uma ferramenta para identificar e avaliar potenciais riscos que possam impactar negativamente no desenvolvimento do projeto. Neste tópico do anexo apresenta-se o histórico da matriz de risco utilizada em cada sprint do projeto, desde o seu início até o momento atual.

Cada sprint do projeto é acompanhada de uma matriz de risco específica, que é atualizada de acordo com as mudanças e imprevistos que surgem durante o planejamento do projeto. O objetivo desta seção de anexo é fornecer uma visão geral das matrizes de risco, permitindo uma análise comparativa do nível de risco enfrentado em cada momento.

1. Matriz de Risco – Sprint 1

Matriz de Risco										
Probabilidade		Riscos					Oportunidade			
Muito Alta	1									
Alta	2	Conflito de ideias no grupo	Sobrecarga no time com as atividades não observadas na planning	Má organização para as entregas em semanas com feriados e pontos facultativos	Dificuldades de entender o modelo DevOps durante o desenvolvimento do projeto	Aprendizagem do modelo DevOps e aprender boas práticas	Uma maior acurácia nos dados	Diminuir os riscos de erro humano (medição e tomada de decisão), evitando perdas para o negócio		
Médio	3		Faltas não avisadas antecipadamente.	Sugestão de novos requisitos que fogem do escopo inicial do projeto	Conversas paralelas sobre assuntos diversos sem contexto com o projeto		Melhora da performance do negócio			
Baixa	4	Ausência prolongada devido a doença ou afastamento		Atraso na entrega	Falta de organização do Git-lab	Maior automação dos processos	A base de dados gerados a partir das medidas ser utilizada para análises outras análises			
Muito Baixa	5									
		1	2	3	4	5	5	4	3	2
		Muito Baixo	Baixo	Médio	Alta	Muito Alta	Muito Alta	Alta	Médio	Baixo
		Impacto								

2. Matriz de Risco – Sprint 2

3. Matriz de Risco – Sprint 3

4. Matriz de Risco – Sprint 4

18.2. Plano de gerenciamento de riscos

Um documento que descreve as estratégias, processos e ações a serem adotados para identificar, avaliar, monitorar e mitigar os riscos em um projeto, atividade ou organização. É uma ferramenta essencial para o sucesso e a segurança de qualquer empreendimento, pois permite antecipar e lidar proativamente com os possíveis eventos adversos que possam afetar os objetivos e resultados esperados.

Cada sprint do projeto é acompanhada de um plano de gerenciamento de risco específico, que é atualizado de acordo com as mudanças na matriz de risco. O objetivo desta seção de anexo é fornecer uma visão geral dos planos traçados, permitindo uma análise comparativa em cada momento.

1. Plano de Gerenciamento de Risco - Sprint 1

MATRIZ RESPONSÁVEL	MITIGAÇÃO
2-2 Dayllan	Quando houver ideias divergentes, será necessário que os idealizadores escrevam em um post-it o essencial da ideia e faça um pitch de 1 min defendendo suas respectivas ideias. Após a defesa das ideias, os integrantes do grupo votam na ideia. Confirmar que a saúde dos integrantes está sempre em dia, evitando sobrecargas que podem resultar em afastamento do projeto. Verificar pela daily se existem impedimentos relacionados a saúde
4-2 Eric	Depois da planning mensurar o peso de cada atividade balancear entre os membros
2-3 Gabriela	Toda planning realizar as anotações de quais dias serão necessários que os integrantes faltem, já prevendo a realização das atividades em outros dias, horários e/ou locais.
3-3 Giovanna	Durante a planning identificar os feriados e pontos facultativos, de modo que a distribuição das tarefas seja realizada levando em consideração o volume de trabalho e a disponibilidade de tempo
2-4 Lucas	Recomendar a todos que leiam o TAPI de maneira concentrada, para evitar dar sugestões que não são compatíveis com o escopo do projeto solicitado Questionar durante a daily o status quo de cada atividade de cada membro do grupo, e indagando quais suas atividades para o dia letivo, além de, antes do final do dev, questionar qual o status das tarefas desempenhadas.
3-4 Michel	Em caso de dúvidas para o modelo DevOps, me disponho para tentar auxiliar os outros integrantes do grupo. E caso não seja capaz de ajudar, marcarei uma reunião com o professor responsável para que tudo seja explicado de maneira
4-4 Dayllan	
2-5 Eric	

			mais clara
3-5	Gabriela		Durante o dev fechar abas que não convém com o projeto e caso ainda tenha conversa paralela alguém deve intervir de forma educada e voltar a atenção para o projeto
4-5	Giovanna		Deixar o PO da Sprint responsável por subir os conteúdos da branch de dev para main, no final de toda semana.

Fonte: Autores

2. Plano de Gerenciamento de Risco – Sprint 2

Cód.	Descrição do risco	Probabilidade	Impacto	Descrição do Impacto	Ação	Descrição da ação	Responsável	Previsão
2-3	Dificuldade na utilização do Jupyter Notebook	4 - Alta	3 - Médio	A falta de competência técnica da ferramenta entre os integrantes do time, pode ocasionar em atraso de atividades do projeto.	Mitigar	1. Trocar a plataforma de construção do código fonte para o Google Colaboratory e quando realizar alguma mudança significativa, baixar a extensão do jupyter, para continuar com os versionamentos de código. 2. Procurar o professor Hayashi para retrair dúvidas e adquirir conhecimento da ferramenta.	Lucas Britto	5/5/2023
2-4	Dificuldade da implementação do modelo Bag Of Words	4 - Alta	4 - Alto	A falta de conhecimento e uso do modelo Bag Of Words, interfere na transformação das palavras retidas dos comentários em numeral, para a análise de sentimentos.	Mitigar	Buscar meios de entendimento do problema com o professor, vídeos e leituras sobre o assunto. Além disso, é importante entender o conceito, visualizar exemplos de código e usar bibliotecas de linguagem natural.	Dayllan Alho	9/5/2023
2-5	Dificuldades de entender o modelo DevOps durante o desenvolvimento do projeto	4 - Alta	5 - Muito Alto	A falta de capacidade do time para que todos os arquivos e desenvolvimento do projeto, serem exibidos diretamente na main principal, impactando questões de versão/consulta e desempenho dos artefatos.	Mitigar	Realizar os procedimentos com commits na main, guardando um backup anterior do projeto, para ter as versões, mas enquanto isso, estudar como realizar o processo de forma adequada.	Giovanna Furlan	09/05/2023
3-2	Redundância no uso de bibliotecas no modelo	3 - Média	2 - Baixo	Aumento desnecessário de códigos, prejudicando o desempenho do sistema e afetar a eficiência do projeto.	Mitigar	Realizar uma pesquisa de modos que cortam coisas desnecessárias no código, como por exemplo, juntar tudo em uma função, tirar partes que não fazem nada, entre outros	Eric Tachdjian	9/5/2023
3-3	Dificuldade em criar dicionário de categorização de emojis e Stop Words	3 - Média	3 - Médio	Afectar a precisão do modelo, prejudicando a qualidade dos resultados. Levava a decisões erradas ou ações incorretas .	Mitigar	Utilizar as bibliotecas prontas disponíveis na programação, como a biblioteca emoji para categorização de emojis e as bibliotecas NLTK ou spacy para Stop Words.	Gabriela Silva	10/5/2023
3-4	Sugestão de novos requisitos que fogem do escopo inicial do projeto	3 - Média	4 - Alto	Dificultar o gerenciamento do escopo e das expectativas do cliente.	Prevenir	Definir o escopo do projeto, evitando sugestões de novos requisitos que estejam fora do escopo do projeto.	Michel Mansur	10/5/2023
3-5	Pré-processamento da base de dados ser realizado de forma inadequada	3 - Média	5 - Muito Alto	Interfere nos resultados deixando-os imprecisos ou inconsistentes.	Prevenir	Criar um plano de pré-processamento de dados, ou seja, incluir as etapas de limpeza, normalização e categorização de dados.	Dayllan Alho	9/5/2023
4-3	Imprevistos técnicos na entrega de atividades do projeto	2 - Baixa	3 - Médio	Leva a atrasos no cronograma do projeto e no cumprimento das metas estabelecidas.	Prevenir	Criar gerenciamento de projetos minimizando riscos de imprevistos, incluindo a definição clara de responsabilidades, prazos e marcos do projeto.	Michel Mansur	10/5/2023
4-4	Falta de organização do GitHub	2 - Baixa	4 - Alto	Tornar mais difícil o controle de versões e o gerenciamento do projeto, além dos resultados entregues.	Prevenir	Criar um guia para a organização do repositório do GitHub, incluindo padrões para nomeação de arquivos, estrutura de pastas, controle de versões e permissões de acesso.	Giovanna Furlan	9/5/2023

3. Plano de Gerenciamento de Risco – Sprint 3

MATRIZ	PROBABILIDADE	IMPACTO	DESCRIÇÃO DO IMPACTO	RESPONSÁVEL	MITIGAÇÃO
2-3	Alta	Médio	Desbalanceamento do grupo, integrantes mais atarefados que outros desorganizando o planejamento inicial	Dayllan	Fazer o balanceamento e deixar algumas pessoas (2) com 3 pontos a menos pois caso entre algo inesperado elas poderão atribuir essas tarefas sem serem prejudicadas
3-3	Média	Médio	Menos um dia de desenvolvimento do projeto e consequentemente pode acabar prejudicando outro colega pois muitas vezes uma tarefa depende da outra	Eric	Ter duas pessoas por tarefa pois caso uma da dupla falte sem avisar a tarefa sera cumprida na data correta
1-4	Muito alta	Alta	Desorganiza o planejamento e pode causar acúmulo de tarefas pendentes do entendimento do entregável	Gabriela	Sempre falar com os professores que estão criando os critérios de avaliação para estarmos sempre atualizados sobre o andamento das novidades do projeto
3-5	Média	Muito alta	Pode causar perda de partes do código, desorganização da pasta do grupo, nomes duplicados, etc...	Lucas	Durante a plannig organizar a ordem de cada um subir no github e nas dailys reforçar quem vai subir o que e quando vai fazer isso.

4. Plano de Gerenciamento de Risco – Sprint 4

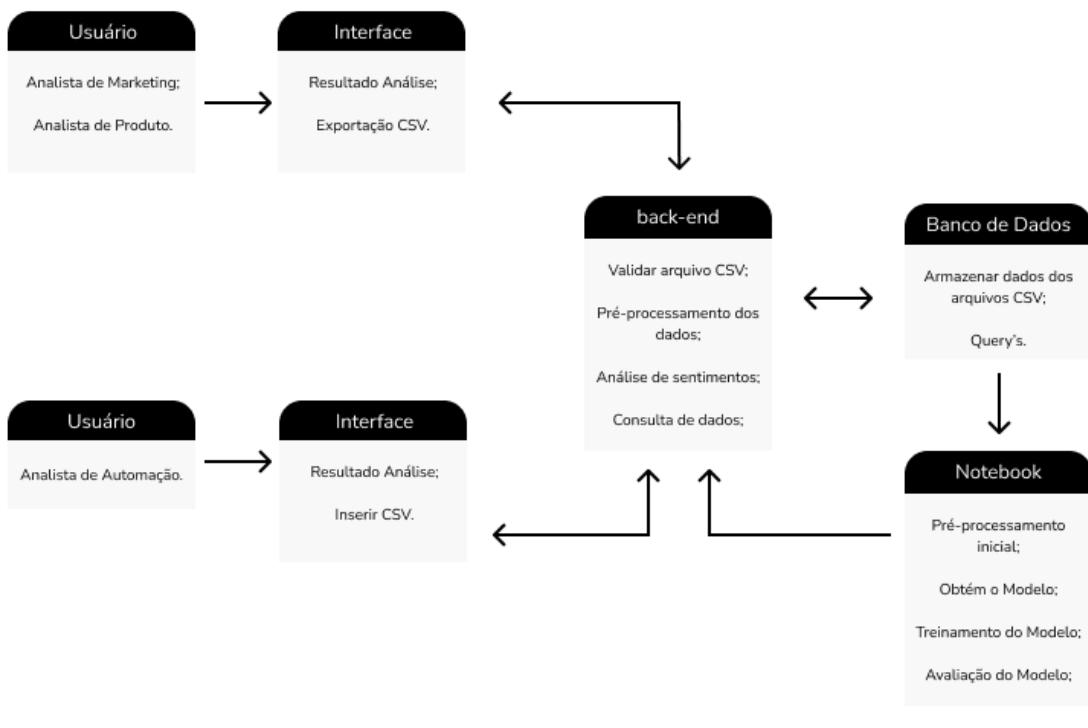
MATRIZ	PROBABILIDADE	IMPACTO	DESCRIÇÃO DO IMPACTO	RESPONSÁVEL	MITIGAÇÃO
2-4	2	4	Menos dias de desenvolvimento do projeto e consequentemente pode acabar prejudicando as entregas	Grupo inteiro	Todos os integrantes se comprometerem a finalizar as tarefas atribuidas ate a data da reunião com o cliente.
4-4	4	4	Pode causar perda de nota, além de dar erro nos códigos.	Michel Mansur	Durante a plannig organizar a ordem de cada um subir no github e nas dailys reforçar quem vai subir o que e quando vai fazer isso.

18.3. Arquitetura Macro

A arquitetura macro da solução, apresenta os blocos responsáveis pelo funcionamento da solução, independente da tecnologia que será adotada ao desenvolvimento, comunicando-se entre si, apresentam a estrutura e ligações mínimas que a solução tem que exibir para ter o funcionamento previsto ao MVP. Abaixo se apresenta as versões das arquiteturas ao longo das sprint's para manter um versionamento e desenvolvimento.

1. Arquitetura Macro – Sprint 1 e 2

ARQUITETURA MACRO DA SOLUÇÃO



A arquitetura macro proposta, conta com 5 módulos, sendo eles 1) Usuário; 2) Interface; 3) back-end; 4) Banco de Dados; e 5) Notebook. Nos módulos Usuário e Interface, apresenta-se as personas do projeto em conjunto com as interfaces as quais

terão acesso no projeto. O back-end é o módulo responsável por realizar o gerenciamento das rotas de consulta da aplicação, além das que são necessárias para acessar o modelo e suas análises. No módulo Banco de Dados, será localizado todos os dados dos arquivos CSV recebido da interface, utilizado para guardar novos dados e fornecê-los para o treinamento do modelo, quando requisitado. Por fim, o módulo do Notebook é responsável pelo processamento, implementação e avaliação do modelo de análise dos comentários e sentimentos da aplicação, obtendo os dados do banco, transformando em CSV, realizando os procedimentos necessários e exportando com uma biblioteca para o back-end que em conjunto com a interface, exibe para os usuários os resultados, em um dashboard.