Universidade do Minho
Mestrado em Inteligência Artificial

Unidade curricular de
Engenharia e Sistemas de Dados
Ano Lectivo 2025/2026

Sistema de Modelação e Análise de Perfis de Clientes Baseada em Sentimentos no Retalho do Vinho

Carlos Bergueira	PG 11605
Diego Silva	PG 59999
Filipa Pereira	PG 42201
Rui Rodrigues	PG 7942

Dezembro, 2025

Data de Recepção	22 de dezembro
Responsável	Docente Orlando Belo
Avaliação	
Observações	

Sistema de Modelação e Análise de Perfis de Clientes Baseada em Sentimentos no Retalho do Vinho

Carlos Bergueira	PG 11605
Diego Silva	PG 59999
Filipa Pereira	PG 42201
Rui Rodrigues	PG 7942

Dezembro, 2025

Resumo

Este trabalho centra-se na conceção e implementação de um sistema de dados para suporte à decisão no setor do retalho especializado, focado no vinho. O objetivo é o estabelecimento de perfil de cliente mais abrangentes e úteis através da integração de dados estruturados e não estruturados.

A solução envolve a conciliação de dados de vendas/CRM (estruturados) com dados subjetivos provenientes de opiniões, avaliações e *feedback* (não estruturados). O núcleo do sistema é um *pipeline* de processamento que aplica Técnicas de Análise de Sentimentos ao texto para extrair emoções e níveis de satisfação dos clientes.

Os dados transformados são centralizados num Data Warehouse modelado dimensionalmente. A informação proveniente dos resultados suporta a segmentação e modelação, e alimenta *dashboards* para apoiar a decisão empresarial, permitindo a personalização de ofertas e o aumento da lealdade dos clientes.

Área de Aplicação: Engenharia e arquitetura de sistemas de dados, *data warehousing* e *business intelligence*, com foco em processamento de linguagem natural e análise preditiva no setor do retalho.

Palavras-Chave: Sistemas de suporte à decisão, data warehouse, perfil de clientes, análise de sentimentos, *big data*, *ETL*, modelagem dimensional, retalho de vinho.

Índice Geral

1. Definição do Sistema **1**
1.1 Contextualização **1**
1.2 Motivação **1**
1.3 Objectivos **2**
1.4 Viabilidade **3**
2. Desenvolvimento de um Plano de Execução e Caracterização Global **5**
2.1 Plano para a execução do projeto 5
2.2 Caracterização Global do Sistema (Arquitetura) 5
3. Levantamento de requisitos e caracterização de utilizadores **7**
3.1 Requisitos Específicos do Sistema 8
3.2 Requisitos de Transformação e Processamento 9
3.3 Requisitos de Integração e Carregamento (Load) 9
3.4 Requisitos de Análise e Visualização 9
4. Definição e Caracterização dos Sistemas de Dados **10**
4.1 Pipelines 10
4.1.1 Pipeline Carregamento de dados no chat/website (MongoDB) 10
4.1.2 Pipeline Extração de dados de feeds (MongoDB) 11
4.1.3 Pipeline Extração de dados de CSV de vendas (MongoDB) 11
4.1.4 Pipeline Extração de dados de MongoDB para Postgres 12
4.1.5 Pipeline Extração e normalização de Sentimentos e ingestão em Postgres 12
4.1.6 Pipeline Análise Sentimental 13
4.1.7 Pipeline de Ingestão em Postgres Data Warehouse 13
4.2 Base de Dados não estruturada (documento store) 14
4.3 Base de Dados Operacional 16
4.4 Análise do Diagrama de Entidade-Relacionamento 22
4.4.1 Entidades de negócio e dimensões de referência 22
4.4.2 O Fluxo Crítico de Dados e Staging 22
4.4.3 Conclusão do DER 23
4.5 Tabelas de Dados Estruturados 24
4.6 Tabelas de Staging e Processamento (Dados Não Estruturados) 24
4.7 Data Warehouse 25
4.7.1 Tabela de factos: ft_vendas (Facto de vendas) 25
4.7.2 Tabela de Factos: ft_sentimento (Facto de Sentimento) 26
4.8 Análise do Diagrama Dimensional (Modelo Estrela Dupla) 27
4.9 Dimensões Compartilhadas (Conformed Dimensions) 28
5. Definição e Caracterização do Sistema de Integração **30**
5.1 Mapeamento e Povoamento das Tabelas de Dimensão 30
5.1.1 Dimensões de baixa volatilidade (dados de suporte) 30
5.2.2 Dimensões de média/alta volatilidade (Dimensões principais) 31
5.2 Mapeamento e Povoamento das Tabelas de Fatos 32

6. Construção, Validação e Documentação do Sistema de Análise de Dados **35**
6.1. Aplicação de Modelação Preditiva e Segmentação 35
6.2. Desenvolvimento e Alimentação do Reporting (BI) 35
7. Construção, Validação e Documentação **37**
7.1 Implementação e validação da etapa de Visualização (Power BI) 37
7.1.1 KPIs (34,78 K | 2000 | 2597) 37
7.1.2 Agregação e Validação do Data Warehouse 38
7.1.3 Análise de Vendas, Produto e Fonte 38
8. Análise Crítica do Trabalho Realizado e Trabalho Futuro **40**
8.1 Pontos fortes 40
O projeto alcançou os seus objetivos centrais e demonstrou a capacidade de integração de dados complexos: 40
8.2 Pontos fracos 40
9. Conclusão Final do Trabalho **42**
Referências **43**
Anexos **45**

Índice de Figuras

Imagem 1: workflow carregamento de dados do chat 10
Imagem 2: workflow extração de dados dos feeds 11
Imagem 3: workflow extração de dados de vendas 11
Imagem 4: workflow extração de dados do mongodb para postgres 12
Imagem 5: workflow de normalização de dados de sentimentos 13
Imagem 6: workflow do processamento da análise de sentimentos 13
Imagem 7: workflow do processamento de carga do data warehouse 14
Imagem 8: workflow de segmentação de utilizadores 14
Imagem 9: Diagrama de Entidade-Relacionamento 24
Imagem 10: ft_vendas 26
Imagem 11: ft_sentimento 27
Imagem 12: Diagrama de relacionamento das dimensões e factos 28
Imagem 13: Visualização (Power BI) 36

Índice de Tabelas

Tabela 1: Vantagens e ganhos operacionais (Viabilidade) 04
Tabela 2. Plano de execução do projeto 05
Tabela 3: Processo ETL 06
Tabela 4: Perfis de utilização 07
Tabela 5: "*feeds_rss_amarante_vinhos*", "*feeds_rss_decanter*" e "*feeds_rss_wine_inspector*" 15
Tabela 6: "file_csv_comentarios" 16
Tabela 7: Collection "vendas" 17
Tabela 8: feeds_rss 18
Tabela 9: feeds_rss_processada 18
Tabela 10: file_csv_comentarios 19
Tabela 11: file_csv_comentarios_processada 19
Tabela 12: website_feedback 19
Tabela 13: website_feedback_processada 20
Tabela 14: Loja 20
Tabela 15: metodo_pagamento 20
Tabela 16: utilizador 20
Tabela 17: produto 21
Tabela 18: utilizador_email 21
Tabela 19: sentimento 21
Tabela 20: venda 22
Tabela 21: venda_processada - 22
Tabela 22: utilizador_segmento 22
Tabela 23: feature_utilizador 23
Tabela 24: dim_metodo_pagamento 31
Tabela 25: dim_tipo_sentimento 31
Tabela 26: dim_source 32
Tabela 27: dim_date 32
Tabela 28: dim_regiao 32
Tabela 29: dim_utilizador 33
Tabela 30: dim_produto 33
Tabela 31 :dim_segmento 33
Tabela 32: ft_sentimento 34
Tabela 33: ft_vendas 35

1. Definição do Sistema

1.1 Contextualização

O presente trabalho insere-se no domínio da Engenharia de Sistemas de Dados e visa a concepção e implementação de um sistema de modelação e análise de perfis de clientes para o setor de retalho de vinho.

O mercado do vinho é caracterizado pela sua riqueza de produtos e pela forte componente subjetiva associada à experiência de consumo. Neste contexto, as estratégias de retenção e a personalização de ofertas dependem da capacidade da empresa retalhista em compreender as preferências e o nível emocional de satisfação dos seus clientes.

A conceptualização deste trabalho lida com a dificuldade de integrar e analisar diversas fontes de dados heterogéneas, nomeadamente:

1. Dados estruturados: Registos transacionais, histórico de compras, dados demográficos e dados de CRM.
2. Dados não estruturados: Opiniões, *reviews* de vinhos, comentários e *feedback* em redes sociais ou *websites* (dados cruciais para a análise de sentimento).

O sistema proposto pretende consolidar, transformar e analisar este volume de informação, utilizando Técnicas de Análise de Sentimentos como eixo central para enriquecer os perfis de clientes.

1.2 Motivação

A implementação deste sistema de modelação e análise de perfis de clientes, baseada em sentimentos, é motivada pela necessidade das empresas de retalho de vinho compreenderem os clientes para além da análise tradicional de dados.

Num mercado de elevada concorrência e valorização da experiência sensorial, a motivação reside em:

- Obter uma visão holística do cliente: Os dados sobre o que o cliente compra são incompletos. É fundamental integrar o componente emocional, obtido através da Análise de Sentimentos, para criar perfis de clientes de alta qualidade.
- Melhorar a tomada de decisão: A incapacidade de processar dados não estruturados e volumosos resulta em decisões de marketing e *stock* baseadas em informação parcial. O sistema centraliza dados num Data Warehouse (DW), permitindo que os agentes de decisão acedam a KPIs de satisfação e a modelos preditivos com maior precisão.
- Aumentar a fidelidade e personalização: O conhecimento do sentimento dos clientes permite a criação de ofertas e comunicações personalizadas.

Deste modo, a motivação é transformar grandes volumes de dados heterogêneos em conhecimento valorizado para otimizar o envolvimento do cliente e aumentar a rentabilidade do negócio de retalho de vinho.

1.3 Objectivos

O objetivo principal deste trabalho é a concepção, modelação e implementação de um sistema de dados para suporte à decisão, focado no retalho de vinho.

Para concretizar este objetivo, o projeto cumpre os seguintes objetivos específicos:

Objetivos de Engenharia e Arquitetura:

- Consolidar dados heterogéneos: Projetar e implementar um *pipeline* de integração capaz de recolher, limpar e integrar dados estruturados e dados não estruturados provenientes de variadas fontes de informação.
- Centralizar o conhecimento no DW: Definir e implementar um DW escalável com um modelo dimensional otimizado, que sirva como repositório único para a informação integrada e os perfis de clientes.

Objetivos de Análise e Modelação:

- Aplicar análise de sentimentos: Implementar um motor de processamento capaz de aplicar a Análise de Sentimentos ao texto das opiniões e comentários, quantificando e qualificando o sentimento dos clientes sobre produtos, serviços e a marca.

- Estabelecer perfis de alta qualidade: Utilizar a informação para criar e caracterizar perfis de clientes mais detalhados e preditivos, permitindo uma melhor segmentação.
- Suportar modelação preditiva: Implementar processos de segmentação e de modelação preditiva que utilizem a variável de sentimento para melhorar a capacidade de previsão do comportamento do cliente.

Objetivos de Suporte à Decisão:

- Alimentar dashboards e KPIs: Desenvolver uma fonte de visualização, utilizando o Power BI, para apresentar os resultados da análise e os indicadores-chave de desempenho relevantes aos agentes de decisão do retalho.
- Personalizar experiências: Capacitar a empresa para tomar decisões, como a personalização de ofertas, a gestão de *stock* e a melhoria do serviço ao cliente, baseadas no sentimento detectado.

1.4 Viabilidade

A viabilidade do projeto é alta, sendo fundamentada tanto pela sua sustentabilidade técnica como pelos ganhos operacionais esperados.

Viabilidade Técnica (Exequibilidade)

O projeto é tecnicamente viável porque a arquitetura proposta e as ferramentas identificadas são maduras e compatíveis:

- Tecnologias de integração: O sistema utiliza uma arquitetura moderna que integra bases de dados relacionais (Postgres SQL) e NoSQL (MongoDB) para trabalhar com dados estruturados e não estruturados, respectivamente. Isto garante a capacidade de angariar (*Ingestion*) e extrair (*Extraction*) dados heterogéneos.
- Processamento de Big Data/NLP: A análise de sentimentos é suportada por um *pipeline* de processamento bem definido (limpeza, lematização, tradução, etc.) e pela utilização de um modelo robusto.
- Modelagem dimensional: A modelagem do DW, com as tabelas de factos e dimensões, é escalável e eficiente, o que garante a viabilidade de hospedar e consultar grandes volumes de dados.

Ganhos Operacionais e Vantagens

A implementação do sistema resultará em vantagens estratégicas e ganhos operacionais diretos para a empresa de retalho de vinho:

Vantagem Estratégica	Ganho Operacional Específico
Melhoria da qualidade da decisão	A conciliação de dados de diferentes fontes num único repositório assegura a correção e veracidade dos dados, melhorando a qualidade dos dados que suportam o processo de tomada de decisão.
Personalização de ofertas	Os perfis de clientes com qualidade, enriquecidos com o sentimento, permitem personalizar mensagens, produtos e serviços de forma mais eficaz e orientada.
Previsão e redução de rotatividade (*Churn*)	A integração do sentimento negativo como um fator preditivo permite identificar clientes em risco e lançar campanhas de retenção, impulsionando a lealdade a longo prazo.
Descoberta de tendências	A visão holística e centralizada dos dados facilita a descoberta de tendências ocultas no comportamento e nas preferências dos clientes.
Experiências consistentes	O sistema garante experiências consistentes do cliente em todos os canais de comunicação.

Tabela 1: Vantagens e ganhos operacionais

A viabilidade do projeto é assegurada pela robustez da arquitetura proposta e pelos ganhos operacionais, como a melhoria da fidelidade, a personalização e a descoberta de tendências ocultas.

2. Desenvolvimento de um Plano de Execução e Caracterização Global

2.1 Plano para a execução do projeto

O projeto seguiu um ciclo de vida estruturado, dividido nas seguintes macro-fases:

Fase	Objetivo Principal
I. Análise e Desenho	Definir requisitos, modelar a arquitetura e os esquemas de dados
II. Implementação (ETL e Processamento)	Construir as fontes de dados, o *pipeline* de ETL e a análise de sentimentos
III. Análise e Relatório	Aplicar modelos preditivos, desenvolver *dashboards* e validar o sistema
IV. Conclusão e Documentação	Finalizar o relatório, documentar as ferramentas e realizar a análise crítica

Tabela 2: Plano de execução do projeto

2.2 Caracterização Global do Sistema (Arquitetura)

O sistema de dados é um ambiente centralizado, concebido para processar dados de *big data* (estruturados e não estruturados) num fluxo ELT.

A arquitetura geral é composta pelas seguintes componentes e serviços:

Fontes de Dados (Sources)

O sistema consolida dados de natureza heterogénea:

- Dados Não Estruturados/Semi-Estruturados: Feeds RSS, ficheiros CSVs (Comentários) e website feedback (opiniões).
- Dados Estruturados (Transacionais/CRM): Sistemas ERP/CRM da empresa de retalho (vendas - em formato .csv - e utilizadores).

Processo de Angariação, Transformação e Integração

O fluxo de dados segue a metodologia ETL, sendo dividido em etapas principais:

Fase	Repositório / Componente	Processos a Implementar
Angariação (Ingestion)	Document Store (MongoDB)	Receber os dados não estruturados (*Feeds, Feedback*) no seu formato original para processamento
Extração (Extraction)	Base de dados Relacional Postgres SQL	Ingestão e armazenamento temporário dos dados estruturados e de metadados de texto para o processamento
Transformação (Transformation)	Módulo de processamento	Aplicação do *pipeline* de processamento (limpar HTML, lematizar, remover *stopwords, traduzir*) e execução da análise de sentimentos
Integração (Load)	Data Warehouse (DWH - esd_wine_dw)	Carregamento dos dados tratados e enriquecidos (com *score* de sentimento) para o modelo dimensional

Tabela 3: Processo ETL

Plataformas e serviços de análise

As plataformas permitiram aceder aos processos de visualização e análise:

- Data Warehouse: Plataforma central para hospedar o modelo dimensional (tabelas ft_vendas e ft_sentimento) .
- Profiling e Modelação: Serviços para executar a segmentação e a modelação preditiva (baseados nos dados do DW).
- Visualização e *Reporting*: Utilização de ferramentas de Business Intelligence (Power BI) para alimentar um conjunto de *dashboards* e Indicadores de Desempenho (KPI) para os diversos agentes de decisão.

3. Levantamento de requisitos e caracterização de utilizadores

Esta fase visa identificar e caracterizar os diversos perfis de utilização envolvidos no problema e associá-los a cada um dos processos de manipulação de dados.

Perfis de Utilização e Associação aos Processos

Identificamos quatro perfis principais de utilização que interagem com o sistema de dados no retalho de vinho, e associamos o seu interesse aos processos da arquitetura:

Perfil de Utilização	Responsabilidades	Processos Suportados
Gestor de Marketing	Lançamento de campanhas, fidelização, segmentação	Análise e Modelação Preditiva, Visualização (KPIs de fidelidade e *score* de rotatividade)
Analista	Análise de tendências de sabor/região, gestão de *stock*	Análise de sentimentos, Visualização (Sentimento por produto/região)
Gestor de Vendas/CRM	Monitorização do desempenho de vendas e acompanhamento de clientes individuais	Visualização (filtros por perfil, histórico de sentimento, valor do tempo de vida)
Engenheiro de Dados	Manutenção da qualidade e integridade do *pipeline* de dados	Angariação, transformação e integração de dados

Tabela 4: Perfis de utilização

Identificamos também quatro tipos de clusters (grupos [0, 1, 2, 3]), grupo de clientes (utilizadores) que são semelhantes entre si (comportamentos semelhantes) para o nicho de retalho de vinho, com ênfase no seu comportamento de compra e no seu padrão de sentimento:

- Cluster 0 - Clientes novos e pouco frequentes
    - baixa experiência
    - poucos comentários
    - baixo ticket médio
- Cluster 1 - Clientes fidelizados
    - compras regulares
    - ticket médio alto
    - baixa recência
- Cluster 2 - Clientes de oportunidade

    - compram apenas em promoções
    - alta recência
- Cluster 3 - Clientes que interagem bastante
    - muitos comentários
    - envia *feedback*
    - forte *score* de sentimento

Para a associação de clusters a clientes foi utilizado o modelo K-Means (K = número de grupos; Means = média/centro do grupo).

O K-Means segmenta os clientes em grupos com comportamentos semelhantes usando dados das tabelas de vendas e sentimento do DW, permitindo criar perfis úteis para marketing, churn e personalização.

A partir das tabelas factos e dimensões, são gerados features por cliente, por exemplo: Comportamento de compra (ft_vendas) - número total de compras; valor total gasto; ticket médio; número de categorias diferentes compradas; número de produtos distintos; última data de compra (recência); sentimento (ft_sentimento) - média do *score* de sentimento; número total de interações de sentimento; dados do cliente (dim_cliente); idade; região; género;

Estes dados são agregados por utilizador_key (a chave do utilizador/cliente no DW).

O algoritmo tenta dividir os utilizadores em K grupos, onde: dentro de cada grupo, os clientes são muito parecidos; entre grupos, são bem diferentes. Ele faz isso calculando distâncias entre clientes num espaço multidimensional (onde cada feature é um eixo). Assim cada utilizador recebe um grupo.

3.1 Requisitos Específicos do Sistema

Os requisitos funcionais são agrupados por processo de manipulação de dados:

Requisitos de ingestão e extração

1. O sistema deve ser capaz de extrair dados estruturados e não estruturados de fontes heterogêneas;
2. Deve suportar a ingestão de dados não estruturados de fontes externas e internas para repositórios como MongoDB e Postgres;
3. A extração de dados transacionais (venda) deve incluir o comentário associado à transação para posterior processamento.

3.2 Requisitos de Transformação e Processamento

1. O sistema deve aplicar um *pipeline* de pré-processamento (incluindo limpeza, lematização, tradução e remoção de *stopwords*) ao texto*;*
2. Deve extrair e analisar o sentimento dos clientes a partir das opiniões, comentários ou avaliações, gerando um *score* numérico e uma classificação;
3. O processo de transformação deve mapear as fontes de dados não estruturados ao utilizador_id, para permitir a ligação correta ao perfil de cliente no DW.

3.3 Requisitos de Integração e Carregamento (Load)

1. O sistema deve centralizar os dados e a informação dos perfis dos clientes num DW escalável (esd_wine_dw), garantindo a correção e veracidade dos dados;
2. O processo de carregamento deve mapear os dados processados para o modelo dimensional, povoando as Tabelas de Facto (ft_vendas, ft_sentimento) e Dimensão (dim_utilizador, dim_produto, etc.)

3.4 Requisitos de Análise e Visualização

1. O sistema deve permitir a implementação de processos de segmentação e de modelação preditiva;
2. O sistema deve alimentar um conjunto de dashboards e Indicadores de Desempenho (KPI) para suportar a decisão em áreas como a personalização de ofertas;
3. O *dashboard* deve permitir filtrar a satisfação (sentimento) por atributos do vinho (região), permitindo a descoberta de tendências.

4. Definição e Caracterização dos Sistemas de Dados

Esta etapa define e caracteriza os diferentes tipos de sistemas de dados que modelamos e a implementamos: a base de dados transacional (para ingestão) e o DW (para análise).

4.1 Pipelines

Foram criados oito pipelines que serão orquestrados de modo a correrem em momentos diferentes. Cada pipeline trata de um workflow específico, podendo conter ETL, LET, LE, etc., retratado em seguida através do respectivo diagrama BPMN.

4.1.1 Pipeline Carregamento de dados no chat/website (MongoDB)

Um utilizador acede ao website wine_esd e publica uma mensagem no chat. O serviço associado estabelece ligação a uma document store MongoDB e guarda o comentário numa *collection*.

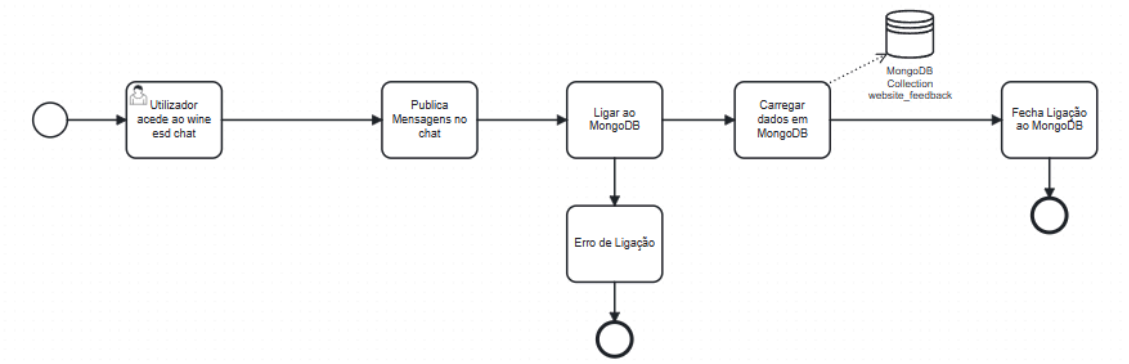

Imagem 1: workflow carregamento de dados do chat

4.1.2 Pipeline Extração de dados de feeds (MongoDB)

O sistema recolhe informação de feeds RSS, definidos num ficheiro JSON. Estabelece ligação ao respectivo feed RSS, recolhe as “notícias” e carrega cada uma delas em *collections* independentes.

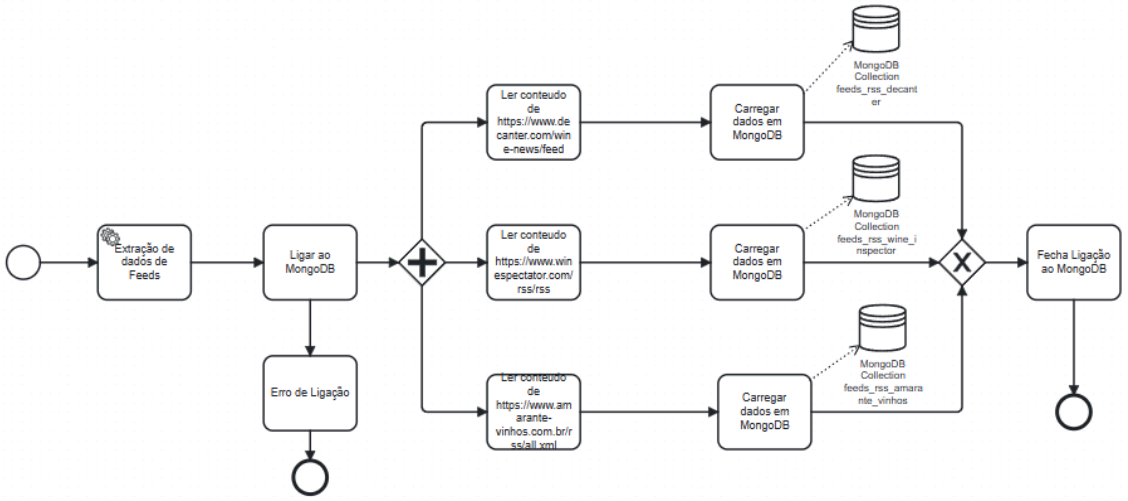

Imagem 2: workflow extração de dados dos feeds

4.1.3 Pipeline Extração de dados de CSV de vendas (MongoDB)

O sistema recolhe dados provenientes de uma loja em formato CSV, uma vez que não existe a possibilidade de comunicação com o ERP. Assim os dados são lidos e carregados do ficheiro CSV numa *collection* em MongoDB.

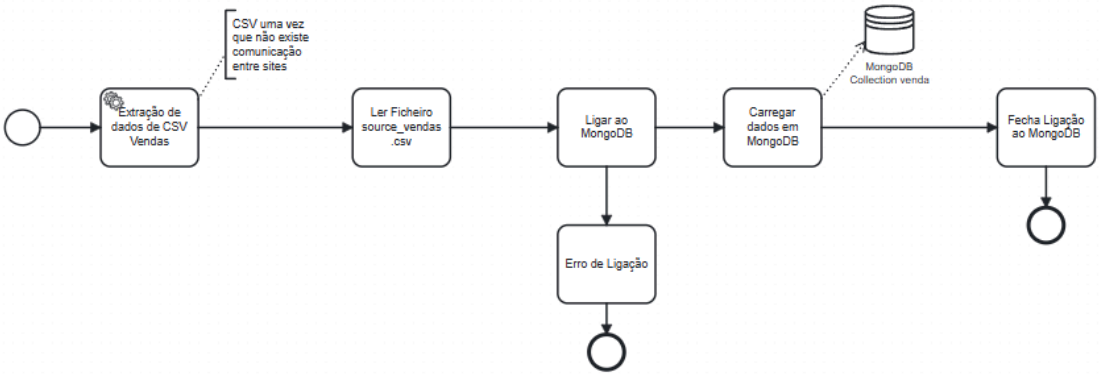

Imagem 3: workflow extração de dados de vendas

4.1.4 Pipeline Extração de dados de MongoDB para Postgres

O sistema extrai os dados existentes na document store MongoDB e carrega-os numa base de dados Postgres, não existindo qualquer transformação de dados (exceto o update nas *collections* informando que os já foram processados).

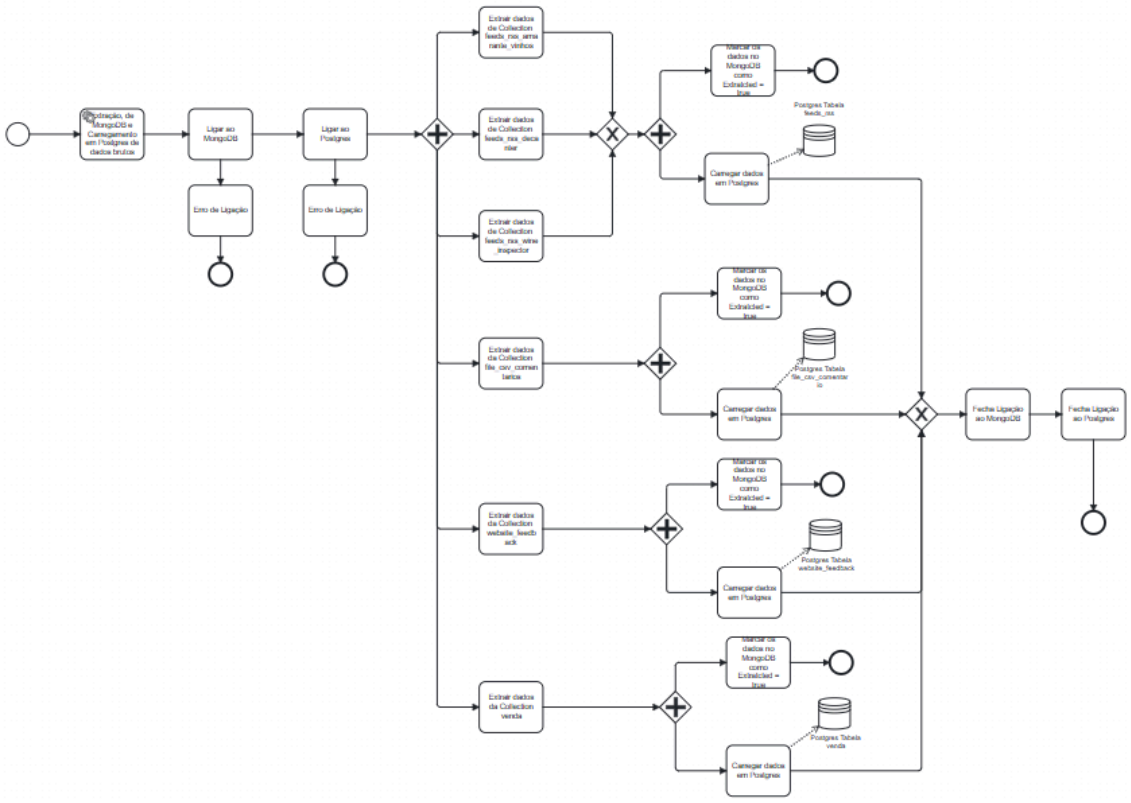

Imagem 4: workflow extração de dados do mongodb para postgres

4.1.5 Pipeline Extração e normalização de Sentimentos e ingestão em Postgres

O sistema percorre cada um dos um dos registos de comentários, e submete-o a um processamento (limpeza, tradução, etc.), por fim carrega os dados na mesma base de dados Postgres.

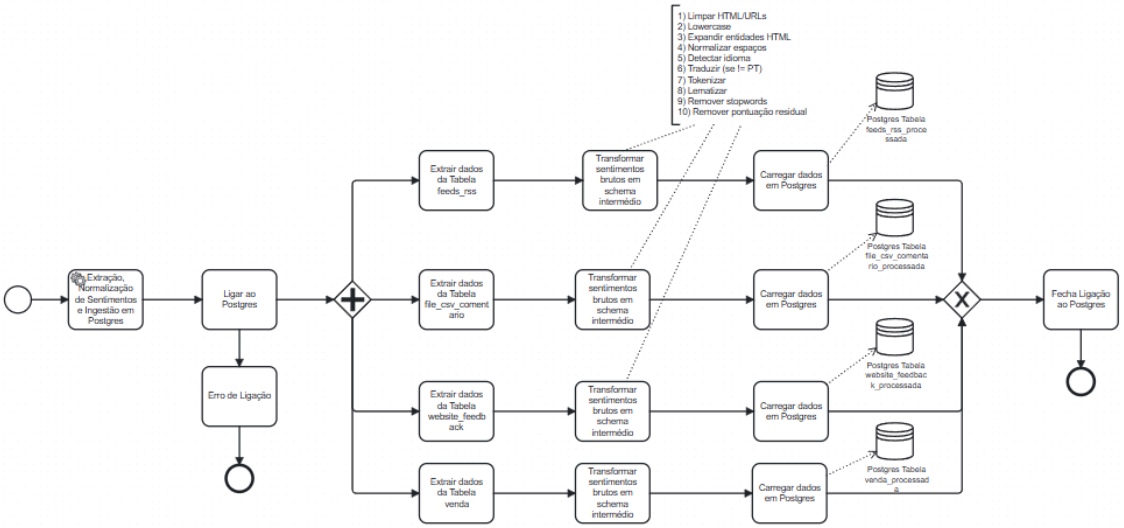

Imagem 5: workflow de normalização de dados de sentimentos

4.1.6 Pipeline Análise Sentimental

O sistema carrega os dados dos comentários previamente processados, e invoca uma técnica de análise de sentimento para obter a classificação (Positivo, Negativo e Neutro) e score (long precision de 0 a 1), utilizado para esse fim o modelo *cardiffnlp/twitter-xlm-roberta-base-sentiment.*

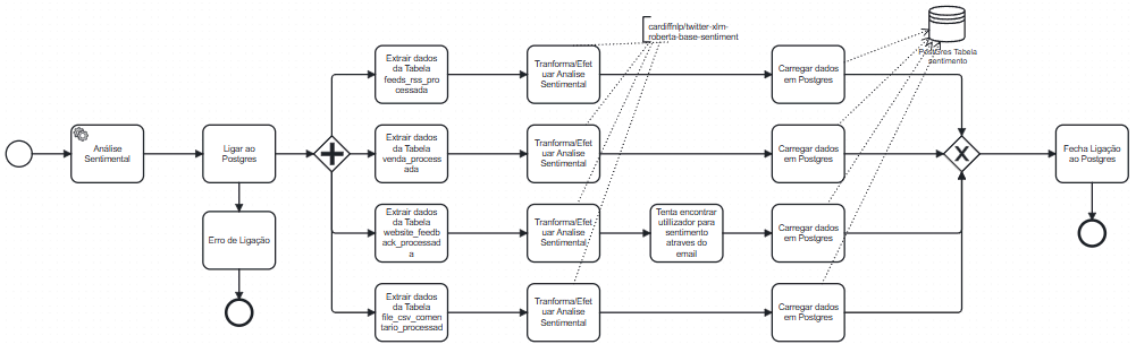

Imagem 6: workflow do processamento da análise de sentimentos

4.1.7 Pipeline de Ingestão em Postgres Data Warehouse

O sistema carrega, e prepara os dados necessários para posteriores análises no sistema de data warehouse para posteriores análises. Os dados são armazenados em factos e dimensões (adiante detalhados).

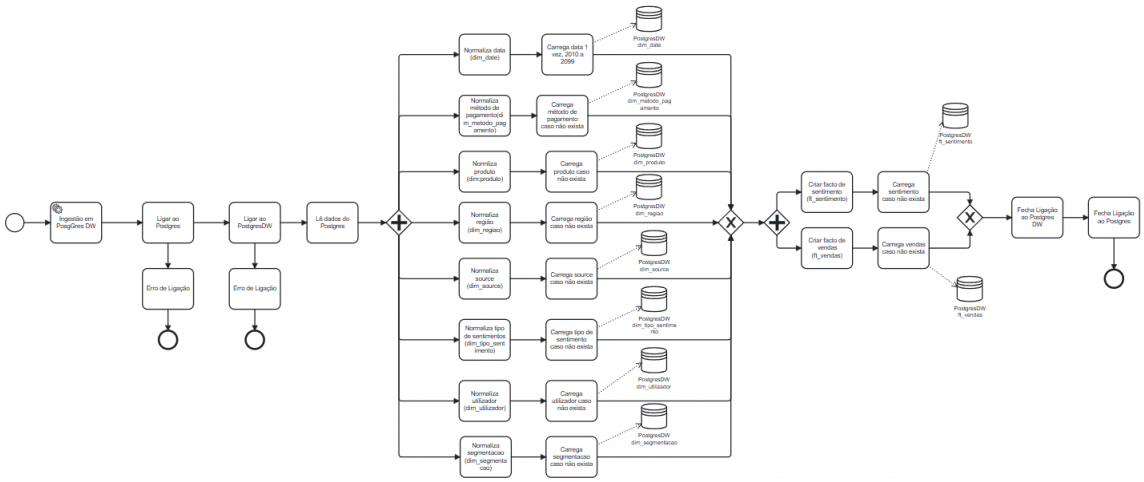

Imagem 7: workflow do processamento de carga do data warehouse

4.1.8 Pipeline Segmentação de Utilizadores

O sistema agrupa os utilizadores por via da utilização do modelo K-Means que, baseado num conjunto de parâmetros e dados, devolve uma segmentação (*clustering*).

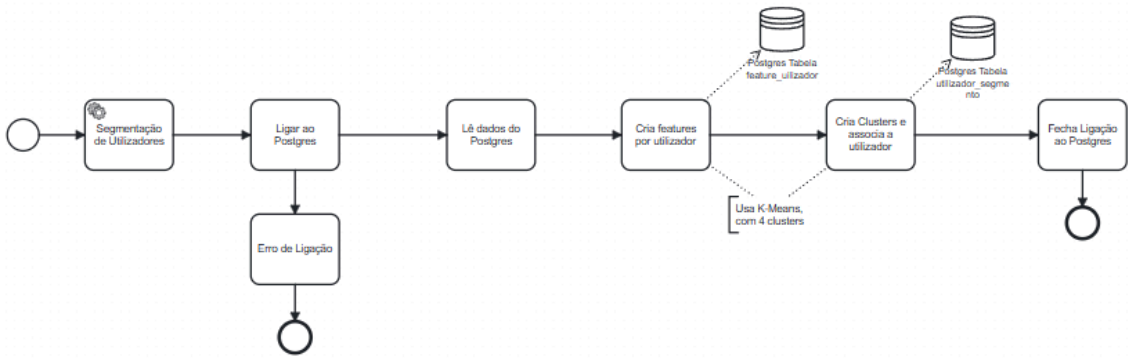

Imagem 8: workflow de segmentação de utilizadores

4.2 Base de Dados não estruturada (*documento store*)

A base de dados esd_wine criada no MongoDB visa armazenar todos os documentos informativos não estruturados, funcionando como uma document store, área inicial para o fluxo de dados. Esta abordagem é essencial para a flexibilidade, uma vez que o MongoDB acomoda facilmente diferentes estruturas de *schema* (schema-less), como feeds RSS e transações com comentários.

A seguir, é detalhada a estrutura e o propósito de cada coleção essencial para a angariação (Ingestion) e para a integração no *pipeline* ELT:

A. Coleção `feeds_rss` (Dados Externos/Tendências)

As *collections* "*feeds_rss_amarante_vinhos*", "*feeds_rss_decanter*" e "*feeds_rss_wine_inspector*" apresentam a mesma estrutura/organização de informação. Esta coleção armazena dados de fontes externas (como blogs de vinho ou notícias do setor), que podem ser usados como *input* para a análise de sentimentos.

campo	descrição	exemplo
_id	id gerado pelo Mongo DB	691faf7a489c0fab6c16ff7f
fonte_nome	nome da fonte	Amarante vinhos
fonte_url	url do feed	https://www.amarante-vinhos.com.br/rss/all.xml
idioma	idioma do feed	PT
titulo	título da notícia	Gin on Mondays - Outubro 2025
link	link da notícia	https://www.amarante-vinhos.com.br/news/gin-on-mondays-outubro-2025/
sumario	sumário da notícia	Organizada por mim e realizada no Espaço Vip do Empório Frei Caneca, …
data_publicacao	data de publicação da notícia (array[9])	2025 10 12 23 0 0 6 285 0
extracted	extraído do mongo por serviço de carregamento na base de dados postgres	true/false
datahora	data e hora da inserção no Mongo DB	2025-11-21 00:16:58
extracted_on	data de extracção	2025-11-21T00:17:10.204+00:00

Tabela 5:. "feeds_rss_amarante_vinhos", "feeds_rss_decanter" e "feeds_rss_wine_inspector"

B. Coleção file_csv_comentarios (Comentários de Clientes)

A *collection* “file_csv_comentarios” apresenta a seguinte estrutura/organização. Esta coleção é crucial para o perfil de cliente, pois armazena o *feedback* de clientes que está associado a produtos e lojas.

campo	descrição	exemplo
_id	id gerado pelo Mongo DB	691faf79489c0fab6c16fd57
id	id do comentário	1
utilizador_id	id do utilizador	1
produto_id	id do produto	35
loja_id	id da loja	1
datahora	data e hora da inserção no Mongo DB	2024-09-24 15:04:24
texto	comentário deixado	Textura sedosa e final longo.
extracted_on	data de extracção	2025-11-21T00:17:10.204+00:00
extracted	extraído do mongo por serviço de carregamento na base de dados postgres	true/false

Tabela 6: “file_csv_comentarios”

C. Coleção vendas (Dados Transacionais com Comentário)

A *collection* “vendas” apresenta a seguinte estrutura/organização. Embora contenha muitas informações estruturadas (IDs, valores), esta coleção é tratada como "documento" no MongoDB e é importante por incluir o comentário associado à transação, ligando o ato de compra ao sentimento.

campo	descrição	exemplo
_id	id gerado pelo Mongo DB	691faf800730363d3597bb04
venda_id	id da venda	1
utilizador_id	id do utilizador	2
datahora	data e hora da inserção no Mongo DB	2023-12-28 10:13:36
loja_id	id da loja	2
produto_id	id do produto	5
quantidade	quantidade adquirida pelo cliente	1
valor_unitario	valor unitário em EUR, sem IVA	14.11
metodo_pagamento_id	id do método de pagamento utilizado na transacção	1

comentario	comentário deixado	Textura sedosa e final longo.
extracted	extraído do mongo por serviço de carregamento na base de dados postgres	true/false
extracted_on	data de extracção	2025-11-21T00:17:10.456+00:00

Tabela 7: Collection “vendas”

Esta etapa confirmou o papel essencial do MongoDB como area inicial para o projeto. Ao armazenar coleções como feeds_rss, file_csv_comentarios e vendas no seu formato de documento, a base de dados não estruturada garante a flexibilidade e a capacidade de ingestão de dados heterogéneos e brutos, antes de serem submetidos ao processamento e carregamento final no DW.

4.3 Base de Dados Operacional

Esta secção detalha a estrutura da sua base de dados operacional (OLTP), que é o ambiente central para a integração, o *staging* e o pré-processamento de dados antes do DW.

A base de dados esd_wine (PostgreSQL) serve como o sistema operacional principal do projeto, acolhendo tanto os dados estruturados transacionais quanto a área de staging para o texto não estruturado antes da análise de sentimentos.

- Finalidade: Armazenar dados brutos de venda e informação de utilizador, gerir a integridade das entidades de negócio e integrar a informação temporária de texto (staging) para o fluxo de dados não estruturados.
- Modelo: Modelo Relacional (OLTP-like), garantindo a integridade e consistência dos dados (Chaves primárias e estrangeiras).

A. Tabelas de Staging e Processamento de Sentimento (Text Staging)

A base de dados Postgres SQL apresenta a seguinte estrutura:

Tabela: **feeds_rss**
Integra notícias externas ou artigos de vinho, sendo uma fonte de texto para a análise de sentimentos.

campo	tipo dados	descrição	chave
id	integer	id da tabela feeds_rss	PK

fonte_nome	text	fonte da notícia	
fonte_url	text	url da notícia	
idioma	text	idioma da notícia	
titulo	text	título da notícia	
link	text	link da notícia	
sumario	text	sumário da notícia	
datahora	timestamp	timestamp corresponde data registo/notícia	
extracted_on	timestamp	timestamp correspondente data extracção	
processed	boolean	se já foi processado. default = false	

Tabela 8: feeds_rss

Tabela: **feeds_rss_processada**

Armazena os feeds/rss processados (limpos, traduzidos, etc.).

campo	tipo dados	descrição	chave
id	integer	id da tabela feeds_rss_processada	PK
r_id	integer	id do feeds_rss.id relacionado	FK
titulo	text	titulo (limpo)	
sumario	text	sumário da notícia (limpo)	
processed_on	timestamp	timestamp data processamento	
applied_tas	boolean	se já foi aplicada TAS. default = false	

Tabela 9: feeds_rss_processada

Tabela: **file_csv_comentarios**

Armazena comentários recolhidos de ficheiros CSV, tipicamente *reviews* de produto.

campo	tipo dados	descrição	chave
id	integer	id da tabela file_csv_comentarios	PK
utilizador_id	integer	id do utilizador que comentou	FK
produto_id	integer	id produto caso comentário inclua produto	FK
loja_id	integer	id loja caso comentário inclua loja	FK
datahora	timestamp	timestamp da data/hora do comentário	
texto	text	comentário do utilizador	

extracted_on	timestamp	timestamp correspondente data extracção	
processed	boolean	se já foi processado. default = false	

Tabela 10: file_csv_comentarios

Tabela: **file_csv_comentarios_processada**

Armazena os comentários processados (limpos, traduzidos, etc.).

campo	tipo dados	descrição	chave
id	integer	id da tabela file_csv_comentarios_processada	PK
r_id	integer	id do feeds_rss.id relacionado	FK
texto	text	comentário (limpo)	
processed_on	timestamp	timestamp data processamento	
applied_tas	boolean	se já foi aplicada TAS. default = false	

Tabela 11: file_csv_comentarios_processada

Tabela: **website_feedback**

Armazena os comentários recebidos por via do interface web.

campo	tipo dados	descrição	chave
id	integer	id	PK
object_id	text	id gerado pelo Mongo DB	
email	text	email de quem comentou	
comentarios	text	comentário do utilizador	
classificacao	integer	classificacao de 1 a 10 no website	
datahora	timestamp	timestamp da data/hora do comentário	
source	text	source do comentários. “website feedback”	
extracted	boolean	se já foi extraído do Mongo. default=false	

Tabela 12: website_feedback

Tabela: **website_feedback_processada**

Armazena os comentários processados (limpos, traduzidos, etc.).

campo	tipo dados	descrição	chave
id	integer		PK
r_id	integer	id do website_feedback.id relacionado	FK

comentarios	text	comentário (limpo)	
processed_on	timestamp	timestamp data processamento	
applied_tas	boolean	se já foi aplicada TAS. default = false	

Tabela 13: website_feedback_processada

Tabela: **loja**

Armazena as lojas (on-lines e físicas).

campo	tipo dados	descrição	chave
loja_id	integer	id da loja	PK
loja	text	nome da loja	

Tabela 14: loja

Tabela: **metodo_pagamento**

Armazena as lojas (on-lines e físicas).

campo	tipo dados	descrição	chave
metodo_pagamento_id	integer	id do método pagamento	PK
metodo_pagamento	text	descrição do método pagamento	

Tabela 15: metodo_pagamento

Tabela: **utilizador**

Armazena os utilizadores / clientes.

campo	tipo dados	descrição	chave
utilizador_id	integer	id do utilizador	PK
nome	text	nome	
email	text	último email conhecido do utilizador	
data_nascimento	timestamp	timestamp correspondente à data nascimento	
genero	text	género do utilizador	
regiao	text	região onde reside	

Tabela 16: utilizador

Tabela: **produto**

Armazena os produtos comercializados nas lojas.

campo	tipo dados	descrição	chave

produto_id	integer	id do produto	PK
produto	text	descrição do produto/vinho	
regiao	text	regiao a que pertence o vinho	
safra	text	safra	

Tabela 17: produto

Tabela: **utilizador_email**

Armazena o histórico de emails dos utilizadores / clientes.

campo	tipo dados	descrição	chave
utilizador_id	integer	id do utilizador	PK
email	text	último email conhecido do utilizador	
created_on	timestamp	timestamp correspondente à data criação	

Tabela 18: utilizador_email

Tabela: **sentimento**

Armazena os sentimentos; ou seja, os comentários processados onde aplicamos a técnica de análise de sentimentos.

campo	tipo dados	descrição	chave
sentimento_id	integer	id do sentimento	PK
utilizador_id	integer	id do utilizador.utilizador_id relacionado	FK
text	text	texto submetido à TAS	
datahora	timestamp	timestamp em que foi aplicado TAS	
modelo	text	registo do modelo utilizado para TAS	
sentimento	text	classe devolvida pelo TAS (NEG, NEU, POS)	
score	double	*score* devolvido pelo TAS	
created_on	timestamp	timestamp correspondente à data criação	

Tabela 19: sentimento

Tabela: **venda**

Armazena os comentários processados (limpos, traduzidos, etc.).

campo	tipo dados	descrição	chave
venda_id	integer	id da venda	PK

utilizador_id	integer	id do utilizador.utilizador_id relacionado	FK
datahora	timestamp	timestamp corresponde à datahora de criação da venda	
loja_id	integer	id do loja.loja_id relacionado	FK
produto_id	integer	id do produto.produto_id relacionado	FK
quantidade	integer	quantidade adquirida	
valor_unitario	double	valor unitário (sem IVA)	
metodo_pagamento_id	integer	id do metodo_pagamento.metodo_pagamento_id relacionado	FK
comentario	text	comentário criado aquando venda	
extracted	boolean	extraído de Mongo DB	

Tabela 20: venda

Tabela: **venda_processada**

Armazena as vendas com o campo comentário processado (limpo/traduzido).

campo	tipo dados	descrição	chave
venda_processada_id	integer	id da venda_processada	PK
venda_id	integer	id do venda.venda_id relacionado	FK
comentario	text	comentário (processado)	
processed_on	timestamp	timestamp da data/hora do processamento	
applied_tas	boolean	se já foi aplicado TAS. default=false	

Tabela 21: venda_processada

Tabela: **utilizador_segmento**

Armazena o histórico de segmentação dos utilizadores.

campo	tipo dados	descrição	chave
utilizador_key	integer	id do utilizador.utilizador_id relacionado	PK
cluster	text	nome atribuído à segmentação	
timestamp	timestamp	timestamp correspondente à data criação	

Tabela 22: utilizador_segmento

Tabela: **feature_utilizador**

Armazena as vendas com o campo comentário processado (limpo/traduzido).

campo	tipo dados	descrição	chave

utilizador_id	integer	id do utilizador.utilizador_id relacionado	PK
idade	integer	idade do utilizador	
n_compras	bigint	número total de compras	
quantidade_total	bigint	quantidade total de produtos comprados	
valor_total	double	valor total da mercadoria, sem IVA	
ticket_medio	double	valor total dividido pelo núm de compras	
recencia_dias	integer	número de dias desde a última compra	
n_produtos	bigint	número distintos de produtos comprados	
n_lojas	bigint	número distintos de lojas onde comprou	
n_regioes	bigint	número distintos de regiões onde comprou	
score_sentimento_medio	double	*score* médio do utilizador	
n_registos_sentimento	bigint	números de registos de sentimentos	
n_comentarios	bigint	número de comentários feitos	

Tabela 23: feature_utilizador

4.4 Análise do Diagrama de Entidade-Relacionamento

O diagrama apresentado é o núcleo do seu sistema OLTP (Online Transaction Processing) e da área de staging.

4.4.1 Entidades de negócio e dimensões de referência

O modelo define as entidades primárias do negócio de retalho e as suas relações:

- Entidade Central do Cliente (utilizador): Esta tabela armazena a caracterização do cliente (nome, email, género, região) . É ligada à tabela auxiliar utilizador_email, que regista o histórico de emails, o que é crucial para resolver a identidade do cliente ao receber *feedback* de diferentes fontes
- Entidades de Suporte: As tabelas loja, produto, e metodo_pagamento atuam como dimensões de referência, fornecendo o contexto para as transações de venda. A tabela produto inclui atributos descritivos importantes como região e safra.

4.4.2 O Fluxo Crítico de Dados e Staging

O diagrama demonstra que o sistema foi concebido para gerir dados estruturados e o fluxo de processamento de texto:

- Tabela de Fatos Transacionais (venda): É a tabela principal das transações. Está diretamente ligada às entidades (utilizador, loja, produto, metodo_pagamento) e inclui o campo comentário, que é a fonte do texto para o perfil de clientes. A tabela auxiliar venda_processada é usada para o controle de quais comentários já foram processados pela análise de sentimento.
- Gestão de Comentários (file_csv_comentarios): Esta tabela armazena os comentários recolhidos de ficheiros CSV. Está ligada à venda e a outras entidades, indicando que os comentários podem estar relacionados a produtos específicos ou lojas.. A tabela file_csv_comentarios_processada atua como *staging area* para o texto limpo, pronto para a análise de sentimentos.
- Tabela central de sentimento (sentimento): Esta entidade armazena o resultado da análise de sentimento. O campo utilizador_id com chave estrangeira é fundamental, pois garante que o *score* de sentimento e a classe (sentimento, *score*) são corretamente atribuídos a um perfil de cliente conhecido.

4.4.3 Conclusão do DER

O Diagrama de Entidade-Relacionamento é sólido e coerente com os objetivos do projeto. Ele não só cumpre a função de uma base de dados operacional (OLTP), como também integra de forma eficiente a área de staging e os mecanismos de rastreamento (tabelas _processada) que são necessários para a transformação (T) no *pipeline* ELT. A estruturação das relações, especialmente em torno do utilizador_id e da tabela sentimento, assegura que o DW será povoado com dados completos para a análise preditiva.

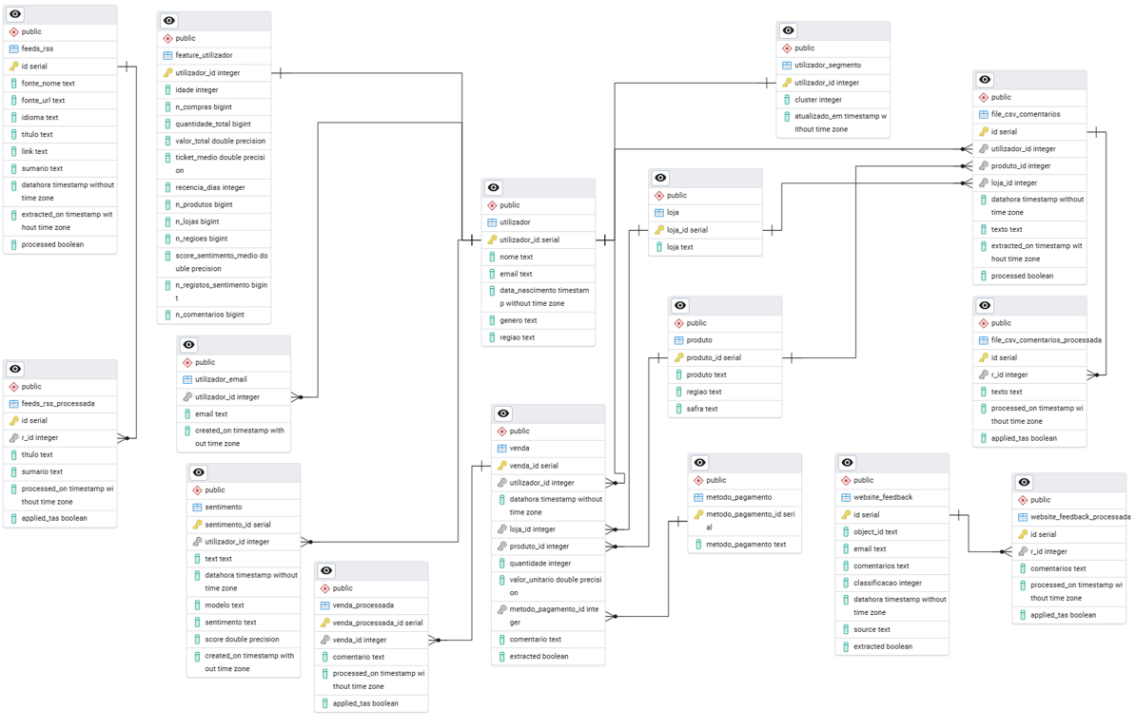

Imagem 9: Diagrama de Entidade-Relacionamento

4.5 Tabelas de Dados Estruturados

As tabelas centrais representam as entidades de negócio:

- utilizador e utilizador_email: Contêm os perfis de clientes e os emails associados/histórico, para ligar o feedback não estruturado ao cliente;
- venda: Regista os logs de transação, incluindo a datahora e o comentário (texto não estruturado associado à compra);
- produto, loja, metodo_pagamento: Tabelas de suporte com dados de referência.

4.6 Tabelas de Staging e Processamento (Dados Não Estruturados)

Estas tabelas geriram o fluxo de texto para a análise de sentimento, atuando como um *staging area*:

- feeds_rss, website_feedback, file_csv_comentarios: Acolhem os dados brutos angariados.

- Tabelas _processada (ex: venda_processada, feeds_rss_processada): Armazenam os registos que já foram submetidos ao *pipeline* de pré-processamento;
- sentimento: Armazena o resultado da análise de sentimento (o *score* e classificação), com ligações ao utilizador_id e o modelo utilizado("*cardiffnlp/twitter-xlm-roberta-base-sentiment*")

4.7 Data Warehouse

O DW (esd_wine_dw) é o sistema analítico principal, modelado dimensionalmente para suportar os requisitos de modelação preditiva e visualização.

4.7.1 Tabela de factos: ft_vendas (Facto de vendas)

O primeiro diagrama mostra a Tabela de Factos central para a análise transacional .

- Finalidade: Medir o desempenho das vendas e ligar as transações aos atributos do cliente e do produto.
- Métricas (Measures):
    - valor_venda e valor_iva: Os valores financeiros das transações.
    - valor_desconto: Medida para análise de promoções.
- Dimensões Ligadas:
    - dim_date: Permite analisar as vendas por dia, mês, semestre, ano e identificar a sazonalidade (flag_feriado, flag_fim_de_semana).
    - dim_utilizador: Permite segmentar as vendas pelo perfil do cliente (nome, regiao, sexo).
    - dim_produto: Permite analisar as vendas por atributos do vinho (nome, categoria, safra, regiao).
    - dim_metodo_pagamento: Permite analisar as vendas pelo tipo de pagamento.
    - dim_regiao: Uma dimensão auxiliar ligada à região da loja (ou dimensão de cliente), útil para análise geográfica.

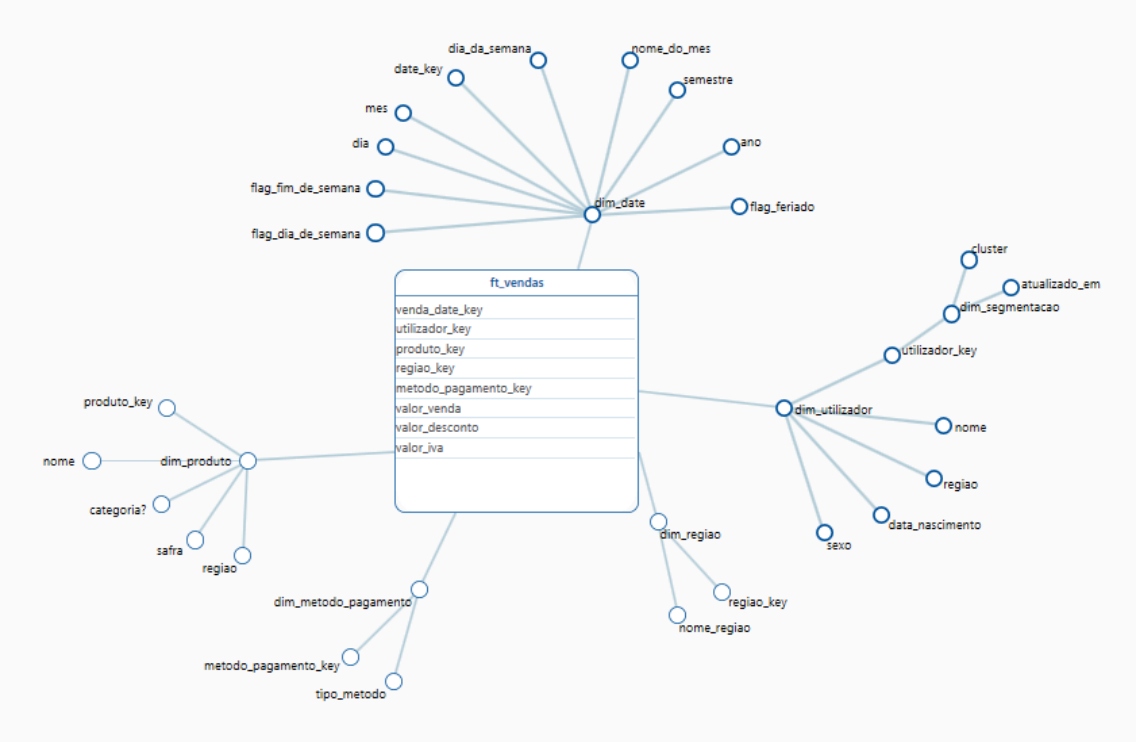

Imagem 10: ft_vendas

4.7.2 Tabela de Factos: ft_sentimento (Facto de Sentimento)

O segundo diagrama mostra a Tabela de Factos para a análise de *feedback*, que é o repositório do conhecimento emocional do cliente.

- Finalidade: Medir o resultado da análise de sentimentos, ligando o *score* diretamente ao cliente e ao contexto do *feedback*.
- Métricas (Measures):
    - *score*: O valor numérico (e.g., entre -1 e 1) é devolvido pela análise de sentimento, representando a intensidade do sentimento. Esta é a métrica chave para o perfil de clientes.
- Dimensões Ligadas:
    - dim_date: Contextualiza o sentimento no tempo (quando o *feedback* foi deixado).

- dim_utilizador: Crucial para o perfil de clientes, permitindo que o *score* de sentimento seja ligado ao perfil do cliente.
- dim_tipo_sentimento: Dimensão que classifica o *score* (e.g., Negativo, Neutro, Positivo, conforme o campo classificacao).
- dim_source: Identifica a origem do *feedback* (Website, RSS, CSV), permitindo analisar a qualidade do sentimento por canal.

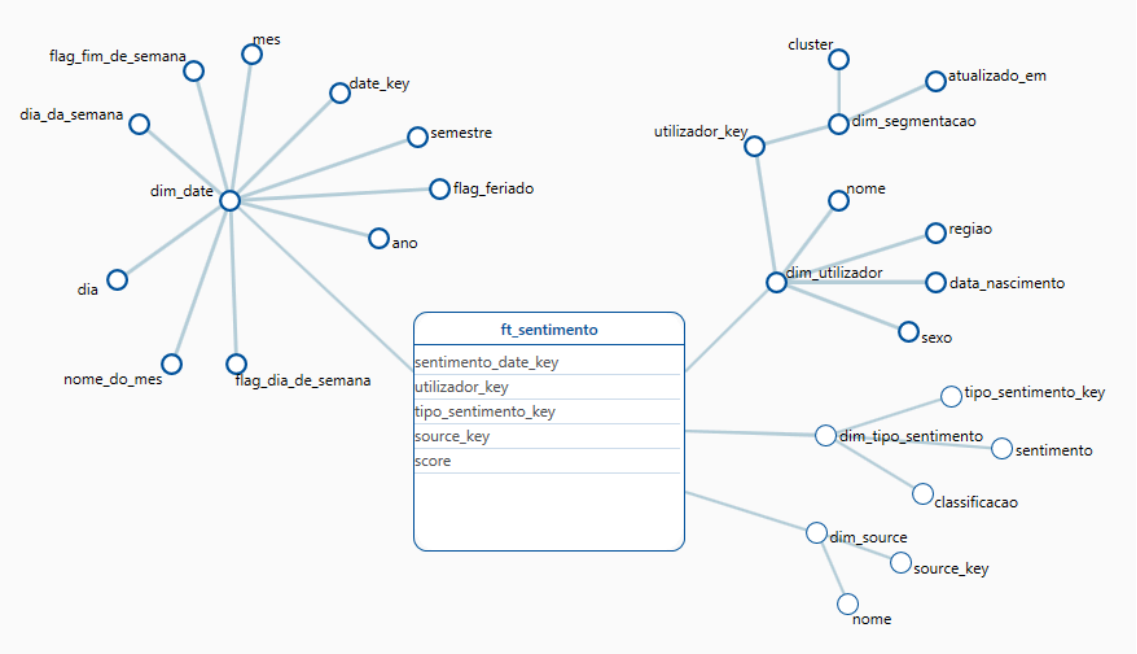

Imagem 11: ft_sentimento

4.8 Análise do Diagrama Dimensional (Modelo Estrela Dupla)

O diagrama mostra duas tabelas de facto centrais (public_ft_vendas e public_ft_sentimento) rodeadas por múltiplas tabelas de dimensões.

As dimensões são entidades constituídas por atributos descritivos, essenciais para filtrar e segmentar a análise:

- dim_utilizador: Contém a caracterização do cliente
- dim_produto: Contém atributos do vinho
- dim_date: Suporta a análise temporal e sazonal

- dim_tipo_sentimento: Classifica o resultado da análise de sentimento
- dim_source: Identifica a origem do *feedback*
- dim_metodo_pagamento: Identifica o tipo de pagamento
- dim_regiao: Identifica a região produtora do vinho

4.9 Dimensões Compartilhadas (Conformed Dimensions)

O diagrama confirma a utilização de Dimensões Conformadas (ou compartilhadas), que são essenciais para que as análises de vendas e sentimento possam ser ligadas:

- public_dim_utilizador: Contém a caracterização do cliente, como nome, regiao e sexo. Está ligada tanto à ft_vendas (para saber quem comprou) quanto à ft_sentimento (para saber quem sentiu) .
- public_dim_date: Suporta a análise temporal e sazonal, com atributos como ano, mês, e *flags* de tempo como flag_feriado. Está ligada a ambas as tabelas de fato.

Esta arquitetura de Estrela Dupla é crucial para a análise preditiva e o *reporting*, pois permite ligar de forma eficiente as tabelas de facto de vendas e sentimento, quantificando o impacto da emoção do cliente nas métricas financeiras. A estrutura otimiza as consultas (*queries*), tornando o DW ideal para o suporte à decisão e para a implementação de modelos preditivos.

Imagem 12: Diagrama de relacionamento das dimensões e factos

Esta etapa resultou na definição de duas arquiteturas de base de dados para o sistema de retalho de vinho:

1. A base de dados (esd_wine) no PostgreSQL, que serve como área de *staging* e hospeda as tabelas transacionais e de texto processado (sentimento) .
2. O DW (esd_wine_dw) que adota o modelo dimensional, sendo o repositório central para análise .

O sucesso da modelagem reside na definição de duas tabelas de facto centrais, que, ligadas pelas sete dimensões, permitem a análise cruzada do comportamento de compra e do sentimento do cliente.

5. Definição e Caracterização do Sistema de Integração

5.1 Mapeamento e Povoamento das Tabelas de Dimensão

A composição do DW depende de como os dados da base operacional (esd_wine) e das fontes não estruturadas são mapeados para as sete dimensões.

5.1.1 Dimensões de baixa volatilidade (dados de suporte)

Estas dimensões são relativamente estáticas e podem ser carregadas a partir de dados de referência ou inserções diretas.

Dimensão: dim_metodo_pagamento

Armazena os dados de método de pagamento.

campo	tipo dados	Mapeamento	chave
metodo_pagamento_key	integer	surrogate key	PK
tipo_metodo	text	Coluna metodo_pagamento da tabela metodo_pagamento	

Tabela 24: dim_metodo_pagamento

Dimensão: dim_tipo_sentimento

Armazena os tipos de sentimentos resultado da análise de sentimentos.

campo	tipo dados	Mapeamento	chave
tipo_sentimneot_key	integer	surrogate key	PK
sentimento	text	Coluna sentimento da tabela sentimento	
classificacao	text	Valor hard-coded “”	

Tabela 25: dim_tipo_sentimento

Dimensão: dim_source

Armazena o nome do sistema da origem onde houve o registo do dado.

campo	tipo dados	Mapeamento	chave
source_key	integer	surrogate key	PK

nome	text	Coluna source da tabela website_feedback	

Tabela 26: dim_source

Dimensão: dim_date

Armazena o nome do sistema da origem onde houve o registo do dado.

campo	tipo dados	Mapeamento	chave
date_key	integer	Tabela carregada a partir de loop resultado do range de intervalos de datas entre 01-01-2010 a 31-12-2099. Data em formato YYYYMMDD, onde YYYY representa o ano, MM representa o mês e DD representa o dia.	PK
dia	integer	Número do dia.	
dia_da_semana	integer	Número do dia da semana.	
mes	integer	Número do mês.	
nome_do_mes	text	Nome do mês.	
semestre	integer	Número do semestre.	
ano	integer	Número do ano.	
flag_feriado	boolean	Indicador se é feriado ou não.	
flag_fim_de_semana	boolean	Indicador se é fim de semana ou não.	
flag_dia_de_semana	boolean	Indicador se é dia de semana ou não.	

Tabela 27: dim_date

Dimensão: dim_regiao

Armazena o nome do sistema da origem onde houve o registo do dado.

campo	tipo dados	Mapeamento	chave
regiao_key	integer	surrogate key	PK
nome_regiao	text	Coluna regiao da tabela produto.	

Tabela 28: dim_regiao

5.2.2 Dimensões de média/alta volatilidade (Dimensões principais)

Estas dimensões requerem um processo de ETL mais dinâmico, pois dependem dos dados que chegam constantemente.

Dimensão: dim_utilizador

Armazena o nome do sistema da origem onde houve o registo do dado.

campo	tipo dados	Mapeamento	chave
utilizador_key	integer	surrogate key	PK
nome	text	Coluna nome da tabela utilizador	
regiao	text	Coluna regiao da tabela utilizador	
data_nascimento	date	Coluna data_nascimento da tabela utilizador	
sexo	text	Coluna genero da tabela utilizador	

Tabela 29:dim_utilizador

Dimensão: dim_produto

Armazena o nome do sistema da origem onde houve o registo do dado.

campo	tipo dados	Mapeamento	chave
produto_key	integer	surrogate key	PK
nome	text	Coluna produto da tabela produto	
categoria	text	Valor fixo NA	
safra	text	Coluna safra da tabela produto	
regiao	text	Coluna regiao da tabela produto	

Tabela 30:dim_produto

Dimensão: dim_segmento

Armazena o nome do sistema da origem onde houve o registo do dado.

campo	tipo dados	Mapeamento	chave
utilizador_key	integer	Chave do utilizador	PK
cluster	integer	Coluna produto da tabela produto	
atualizado_em	date	Data de classificação do utilizador no seguemento	

Tabela 31: dim_segmento

Esta etapa estabeleceu o *pipeline* ELT do projeto, essencial para concretizar o perfil do cliente no retalho do vinho.

5.2 Mapeamento e Povoamento das Tabelas de Fatos

As tabelas fatos precisam de processamento específico, uma vez que não se deve permitir atualizações nos dados, o mapeamento há também a necessidade de se recuperar as chaves das dimensões.

Dimensão: ft_sentimento

Armazena o resultado da análise de sentimentos do utilizadores a partir de comentário fornecido na aplicação, de modo a identificar aspetos positivos ou o que se deve melhorar.
sentimento_date_key,utilizador_key,tipo_sentimento_key,source

campo	tipo dados	Mapeamento	chave
sentimento_key	integer	Coluna sentimento_id da tabela Sentimento	PK
sentimento_date_key	integer	Data em formato YYYYMMDD da coluna datahora da tabela Sentimento	PK
utilizador_key	integer	Recupera o utilizador_key da dim_utilizador por meio de lookup a usar as colunas nome e data_nascimento da tabela Sentimento	PK
tipo_sentimento_key	integer	Recupera o tipo_sentimento_key da dim_tipo_sentimento por meio de lookup a usar as coluna sentimento da tabela Sentimento	PK
source_key	integer	Recupera o source_key da dim_source para a origem website feedback	PK
score	float	Coluna score da tabela Sentimento	

Tabela 32: ft_sentimento

Dimensão: ft_vendas

Armazena os factos de vendas registados, permite análise temporal da evolução das vendas e agrupamento pelas dimensões.

campo	tipo dados	Mapeamento	chave
venda_date_key	integer	Data em formato YYYYMMDD da coluna datahora da tabela Venda	PK
utilizador_key	integer	Recupera o utilizador_key da dim_utilizador por meio de lookup a usar as colunas nome e data_nascimento da tabela Sentimento	PK
produto_key	integer	Recupera o producto_key da dim_produto por meio de lookup a usar a coluna produto da tabela Produto	PK
regiao_key	integer	Recupera o regiao_key da dim_regiao por meio de lookup a usar a coluna regiao da tabela Produto	FK
metodo_pagamento_key	integer	Recupera o metodo_pagamento_key da dim_metodo_pagamento por meio de lookup a usar a coluna metodo_pagamento da tabela Venda	PK
valor_venda	numeric	Coluna quantidade * Coluna valor_unitario da tabela Venda	
valor_desconto	numeric	Valor fixo 0	

valor_iva	numeric	0,23 * Coluna quantidade * Coluna valor_unitario da tabela Venda	

Tabela 33: ft_vendas

A metodologia definida garante a integração de dados heterogêneos, através de dois fluxos principais: o fluxo transacional e o fluxo de sentimento. Este último é o mais crítico, pois utiliza o módulo de processamento LN para aplicar a análise de sentimentos ao texto bruto, gerando o score que enriquece a tabela de facto de sentimento (ft_sentimento) .

O sucesso desta etapa assegura que o DW (esd_wine_dw) está povoado com informação limpa, integrada e acionável, cumprindo o objetivo de centralizar os dados dos perfis de clientes num repositório escalável.

6. Construção, Validação e Documentação do Sistema de Análise de Dados

Esta etapa envolve as atividades de implementação listadas nesta secção.

6.1. Aplicação de Modelação Preditiva e Segmentação

Utilizando os dados integrados e enriquecidos no DW (combinando ft_vendas e ft_sentimento), o foco é construir os perfis de qualidade e os modelos de suporte à decisão:

- Segmentação de Clientes: Implementar algoritmos de *clustering* para segmentar os clientes com base na combinação de fatores transacionais (valor de compra, frequência) e emocionais (*score* de sentimento).
    - *Exemplo:* Identificar o segmento de "Clientes de Alto Valor com Sentimento Negativo" (alvo para a retenção, conforme a Persona Sofia).
- Modelagem Preditiva (*Churn Prediction*): Construir modelos de classificação para prever o risco de rotatividade do cliente (probabilidade de *Churn*).
    - *Validação:* O modelo deve demonstrar que a inclusão da variável sentimento melhora significativamente a precisão da previsão do que apenas o histórico transacional.

6.2. Desenvolvimento e Alimentação do Reporting (BI)

A informação fornecida pelos modelos e pela análise cruzada deve ser disponibilizada aos agentes de decisão (gestor de marketing, analista).

- Construção de Dashboards: Desenvolver *dashboards* interativos no Power BI que consomem dados do DW (esd_wine_dw).
    - *KPIs-Chave:* Taxa de sentimento positivo/negativo por categoria de vinho, Customer Lifetime Value vs. sentimento, e matriz de risco de *churn*.
- Validação da experiência do utilizador: Garantir que as *Personas* conseguem obter as respostas aos seus requisitos.
    - *Exemplo:* O analista deve conseguir filtrar facilmente o sentimento para um vinho específico e a sua região.

Nesta fase, os dados integrados no DW foram utilizados para a criação de perfis de clientes de alta qualidade, através da aplicação de técnicas de segmentação e modelagem preditiva.

O sistema foi validado ao alimentar um conjunto de Dashboards e KPIs no Power BI , garantindo que os resultados da análise de sentimento são traduzidos em informação para os agentes de decisão.

7. Construção, Validação e Documentação

7.1 Implementação e validação da etapa de Visualização (Power BI)

O desenvolvimento do *Business Intelligence (BI)* é a fase final do *pipeline ELT*, onde os dados analíticos do DW são analisados pela ferramenta de visualização *Power BI*. Esta etapa pretende traduzir o perfil do consumidor enriquecido com sentimento em indicadores-chave de desempenho (KPIs) e insights para os agentes de decisão.

A validação do sistema ocorre ao demonstrar que os dashboards refletem a correta integração dos dados de vendas (ft_vendas) e sentimento (ft_sentimento) e permitem a análise cruzada das dimensões (dim_utilizador, dim_produto, dim_date) e apresentar os utilizadores segmentados (dim_segmentacao).

Apresentam-se de seguida os principais componentes do *dashboard* construído:

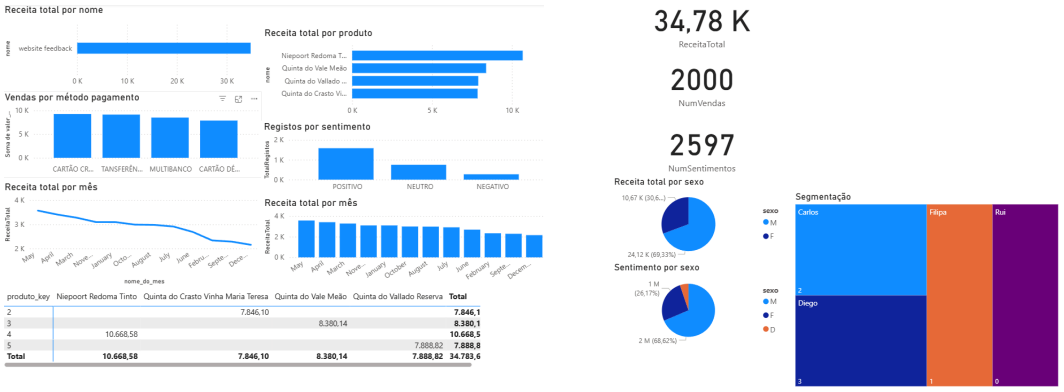

Imagem 13: Visualização (Power BI)

7.1.1 KPIs (34,78 K | 2000 | 2597)

A análise dos KPI fornecem uma visão sumarizada das métricas e validam o sucesso da integração de dados no DW.

A Receita Total de 34,78 K€ valida o carregamento da tabela de facto de Vendas (ft_vendas) e representa o valor financeiro do retalho. Os 2000 registos em NumVendas confirmam o volume transacional processado. Mais importante, os 2597 registos em NumSentimentos validam o

pipeline de análise de sentimentos (ft_sentimento), provando o sucesso na transformação de dados não estruturados em métricas acionáveis.

A coexistência destes três indicadores no dashboard confirma que o sistema fornece uma visão holística e integrada do desempenho financeiro e da satisfação emocional do cliente, sendo a base para todas as análises subsequentes.

7.1.2 Agregação e Validação do Data Warehouse

As métricas de topo confirmam o volume e a integridade dos dados carregados a partir do *pipeline* ELT:

- Receita Total (34,78 K): O total das métricas de vendas valida o carregamento da tabela de fatos ft_vendas e a correta agregação do campo valor_venda. Este é o *benchmark* financeiro do sistema.
- NumVendas (2000): A contagem das transações confirma o volume de dados transacionais processados.
- NumSentimentos (2597): A contagem total dos registos de *feedback* processados pela análise de sentimentos valida o carregamento da tabela ft_sentimento. O elevado número demonstra o sucesso na angariação e processamento de dados não estruturados.

7.1.3 Análise de Vendas, Produto e Fonte

Estes gráficos validam a ligação dos factos às dimensões de produto e fonte de dados.

- Receita total por produto:
    - Validação do Sistema: Confirma o mapeamento da ft_vendas à dimensão dim_produto.
    - Análise dos Dados: Demonstra que o produto "Niepoort Redoma T." gera a maior receita. Este *insight* é crucial, pois indica qual vinho deve ser prioritário em termos de *stock* e promoções.
- Receita Total por *nome* (Origem do Feedback):
    - Validação do Sistema: Valida o mapeamento da ft_vendas à dimensão dim_source.
    - Análise dos Dados: Mostra a distribuição da receita associada à fonte de *feedback* (e.g., *website feedback*). Isto permite medir o impacto financeiro dos canais onde o *feedback* é mais ativo.

7.1.4 Análise Temporal e Transacional

Estes gráficos exploram o contexto da compra, validando dimensões de suporte.

- Vendas por método de pagamento:
    - Validação do Sistema: Confirma a integridade da dimensão dim_metodo_pagamento.
    - Análise dos Dados: Permite a análise da distribuição das transações por tipo de pagamento (e.g., CARTÃO CRÉDITO, MULTIBANCO). Esta informação é importante para otimizar os processos financeiros.
- Receita Total por *nome_do_mes*:
    - Validação do Sistema: Valida a dimensão dim_date e a capacidade de análise de séries temporais.
    - Análise dos Dados: Os gráficos de linhas e colunas mostram que a receita é mais forte nos meses iniciais do ano (Abril, Maio), caindo gradualmente. Esta identificação clara da sazonalidade permite antecipar campanhas promocionais nos meses de menor desempenho (e.g., setembro, dezembro).

7.1.5 Validação do Customer Profiling: Sentimento por Sexo

Este é o resultado analítico mais importante do projeto, pois demonstra o sucesso do enriquecimento de dados.

- Gráfico de Sentimento por sexo:
    - Validação do sistema: Demonstra a ligação cruzada entre o sentimento (ft_sentimento) e o perfil demográfico (dim_utilizador).
    - Análise dos dados: Revela a distribuição da satisfação (*scores*) por género. Uma percentagem significativa (26.17%) pertence ao sexo 'M', que também está associado a uma grande fatia da receita. Isto permite segmentar e priorizar clientes de alto valor/alto risco com base na sua pontuação emocional.

8. Análise Crítica do Trabalho Realizado e Trabalho Futuro

Esta etapa consiste na análise final do sistema de dados implementado, identificando os seus pontos fortes e fracos e propondo melhorias.

8.1 Pontos fortes

O projeto alcançou os seus objetivos centrais e demonstrou a capacidade de integração de dados complexos:

- Integração de dados heterogéneos: Foi estabelecido um *pipeline* (ELT) que integrou com sucesso dados estruturados (vendas/CRM) e dados não estruturados (sentimento), resolvendo o problema central do *customer profiling*.
- Enriquecimento preditivo: A implementação da análise de sentimentos e do seu *pipeline* de pré-processamento provou ser eficaz ao gerar o *score* de sentimento, uma variável de valor para enriquecer os perfis de clientes e suportar a modelagem preditiva (ex: *Churn Prediction*).
- Modelo dimensional otimizado: O DW foi modelado de forma dimensional, com as tabelas de factos (ft_vendas, ft_sentimento) e as sete dimensões, permitindo uma análise cruzada entre o que o cliente compra e o que ele sente.
- Suporte à decisão acionável: A implementação da visualização (Power BI) traduziu dados complexos em KPIs e segmentos de fácil compreensão.

8.2 Pontos fracos

O projeto apresenta algumas oportunidades de melhoria, listadas abaixo:

- Baixa capacidade de recuperação a partir de falhas, o sistema carrega todos os dados sempre que precisa efetuar novo processamento e é capaz de processar incremental, mas não há capacidade de recuperar de onde parou, ou seja, caso bata ao meio é necessário percorrer toda cadeia de processamento.
- O DW foi conceptualizado numa base de dados relacional, o que pode ser um fator de lentidão em caso de relatórios mais complexos.

- A cadeia de processamento está por camadas, uma camada só inicia o processamento após o fim da sua antecessora, fazendo com que tabelas que já estejam prontas a serem processadas aguardem até que a camada anterior esteja completamente finalizada.

9. Conclusão Final do Trabalho

O sistema de modelação e análise de perfis de clientes baseado em sentimentos no retalho do vinho mostrou que os objetivos do projeto foram plenamente alcançados, resultando num sistema de suporte à decisão totalmente funcional e com valor estratégico.

O projeto implementou um DW e utilizou-o como um repositório para integrar dados heterogêneos (vendas transacionais e *feedback* não estruturado). A aplicação da análise de sentimentos serviu como catalisador, enriquecendo o perfil do cliente com a dimensão emocional do cliente.

Os resultados gerados - perfis de cliente de qualidade, modelos de segmentação e *dashboards* no Power BI - capacitam a empresa de retalho para tomar decisões personalizadas, aumentando a lealdade e a eficácia operacional.

Referências

Dimensional Modeling Techniques – Kimball Group. (s.d.). *Kimball Group*. Retirado em [10,Novembro 2025], de kimballgroup.com/wp-content/uploads/2013/08/2013.09-Kimball-Dimensional-Modeling-Techniques11.pdf

High-Level Overview of a Data Migration Approach – Data Dominance. (s.d.). *OMG*. Retirado em [15, Novembro 2025], de omg.org/spec/BPMN/2.0.2/PDF

Kimball Group. (2013). *Kimball Dimensional Modeling Techniques* [White Paper]. Retirado em [2,Novembro 2025], de kimballgroup.com/wp-content/uploads/2013/08/2013.09-Kimball-Dimensional-Modeling-Techniques11.pdf

Kimball, R., & Merz, R. (2000). *Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. John Wiley.

Kimball, R., Reeves, L., Ross, M., & Thornthwait, W. (1998). *The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing, and Deploying Data Warehouses*. John Wiley & Sons.

Oliveira, B., Oliveira, Ó., & Belo, O. (2021). *Using BPMN for ETL Conceptual Modelling: A Case Study*. International Conference on Data Technologies and Applications. DOI:10.5220/0010575702670274

Reis, J., & Housley, M. (2022). *Fundamentals of data engineering: Plan and build robust data systems*. O'Reilly Media.

Lista de Siglas e Acrónimos

BD	Base de Dados
BI	*Business Intelligence*
DW	Data Warehouse
OLAP	*On-Line Analytical Processing*
OLTP	*On-Line Transaction Processing*
ETL	*Extract, Transform and Load*
ELT	*Extract, Load and Transform*
LTE	Load, Transform and *Extract*
CRM	*Customer Relationship Management*
NLP	*Natural Language Processing*
KPI	*Key Performance Indicator*
BPMN	*Business Process Model and Notation*
ERP	*Enterprise Resource Planning*

Anexos

Poderão ser consultados em: https://github.com/MIA-CDFR/ESD.git