ESTUDO COMPRATIVO DE SITEMAS DE RECOMENDAÇÃO PARA CONSUMIDORES DE *E-COMMERCE* NO BRASIL.

Felipe R. Oliveira
Programa de pós-graduação em Engenharia Civil
Universidade Federal do Rio de Janeiro
Rio de Janeiro,Brasil

***Resumo*—É crucial para empresas e fabricantes que desejam se manter competitivos no mercado digital, o estudo de técnicas que otimizem o atendimento de seus consumidores. Neste contexto o presente trabalho tem como objetivo a criação e avaliação de um sistema de recomendação de produtos para clientes de lojas virtuais do banco de dados da *Olist Store*. Através da Análise Exploratória de Dados (AED), é possível observar que o atraso é o maior causador de insatisfação dos clientes. O sistema de recomendação baseado em filtragem colaborativa apresenta desempenho superior ao do sistema baseado em popularidade.**

Palavras Chave—e-commerce, agrupamento, satisfação, recomendação.

I. INTRODUÇÃO

O termo e-commerce, em português "comércio eletrônico", refere-se à modalidade de vendas realizadas de forma virtual. A recente pandemia do COVID-19 acarretou na paralisação de um grande número de atividades comerciais presenciais, provocando um crescimento sem precedentes no número de transações em mercados digitais [1].Dessa forma, é crucial para empresas e fabricantes que desejam se manter competitivos, o estudo de técnicas que otimizem o atendimento de seus consumidores.

Nesse contexto os Sistemas de Recomendação (SR), pertencentes a área de pesquisa de Sistemas de Filtragem de Informação (SFI), propõe o desenvolvimento de sistemas capazes de identificar as preferências de consumidores e a partir delas recuperar produtos que sejam do interesse dos mesmos. Contudo, produzir recomendações adequadas é uma tarefa usualmente intricada devido à grande quantidade de informação e/ou sua alta complexidade. O excesso de registros faltantes e a baixa qualidade dos dados costumam comprometer a eficiência dos SR [2].

Para lidar com algumas das limitações dos SR baseados em conteúdo, os Sistemas de Recomendação Colaborativos (SRC) usam as semelhanças entre usuários e itens simultaneamente para fornecer recomendações. Isso permite recomendações circunstanciais, ou seja, os modelos de filtragem colaborativa podem recomendar um item ao usuário A com base nos interesses de um usuário semelhante B. Além disso, a extração de informação pode ser realizada de forma automática, poupando trabalho de um processo de engenharia de dados manual [3].

Este trabalho dedica-se a análise de um banco de dados da *Olist Store*, disponível no repositório *Keggale*, que possui informações de 100 mil compras realizadas, entre 2016 e 2018, feitas no Brasil, através de plataformas digitais. Seus recursos permitem visualizar um pedido de várias dimensões: do status do pedido, preço, meio de pagamento e frete, atributos do produto e os comentários escritos pelos compradores. O conjunto de dados também possui recursos de geolocalização que relacionam os códigos postais brasileiros às coordenadas (latitude e longitude) dos consumidores e vendedores.

O banco de dados da *Olist Store* permite uma abordagem multidisciplinar e favorece a criação de modelos que, com base nas informações disponibilizadas, possam melhorar o desempenho de lojas virtuais e/ou o nível de satisfação dos consumidores. Contudo, os registros (por motivos de segurança) apresentam pouca granularidade de informação referente aos seus usuários. Dessa forma, a recomendação de produtos torna-se uma tarefa particularmente desafiadora.

A. Apresentação do problema

Este trabalho tem como objetivo a criação e avaliação de diferentes Sistemas de Recomendação (SR) de produtos para clientes de lojas virtuais do banco de dados da *Olist Store*.

B. Apresentação da tecnologia

Para o armazenamento e concatenação de dados é utilizada a linguagem SQL, por meio do ambiente de desenvolvimento *MySQL*. Para análise e visualização de dados, modelagem, otimização e avaliação dos modelos criados, é utilizada a linguagem Python na versão 3.0, por meio do ambiente de desenvolvimento *PyCharm*.

São adotadas as seguintes bibliotecas em Python neste trabalho:

- **Pandas:** biblioteca utilizada na manipulação de dados matriciais na forma de tabelas;

- **GeoPandas:** biblioteca utilizada na manipulação de dados georreferenciados, permitindo também a criação de mapas;
- **Seaborn:** biblioteca utilizada na criação de gráficos, em especial os dedicados a representações estatísticas, como visualização de histogramas, matrizes de correlação, etc.;
- **NLTK:** um conjunto de bibliotecas utilizada na manipulação do corpus, responsável pelo processamento simbólico e estatístico da linguagem natural;
- **Scikit-Learn:** biblioteca utilizada na criação de modelos de aprendizagem de máquina;
- **Suprise:** biblioteca utilizada na construção e análise de sistemas de recomendação que lidam com dados explicitamente classificados.

Para o desenvolvimento de grafos é utilizada a linguagem Java, por meio do ambiente de desenvolvimento *Gephi*.

II. Análise Exploratória de Dados

A. *Caracterização do banco de dados*

O banco de dados é formado por 8 tabelas relacionadas entre si através de chaves (não interpretadas como variáveis). A Figura 1 ilustra o esquema relacional do banco de dados utilizados neste trabalho.

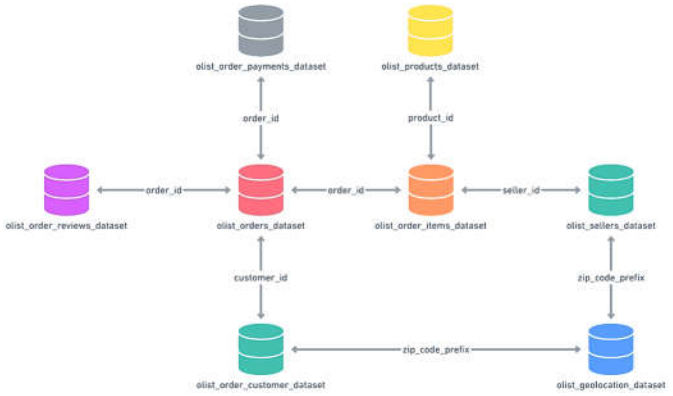

Fig. 1: Esquema relacional do banco de dados. [4]

Como medida de segurança para casos de comentários direcionados aos lojistas, os nomes das lojas virtuais foram substituídos por nomes das grandes casas da série *Game of Thornes*. No total as tabelas somam 36 colunas (desconsiderando informações duplicadas), das quais 5 são chaves encriptadas, 9 são variáveis qualitativas e 22 são variáveis quantitativas. A Tabela 1 apresenta a descrição e tipo das variáveis.

Utilizando as variáveis originais no banco de dados foram agregadas as seguintes novas variáveis:

- **Distância de entrega:** possuindo as coordenadas (latitude e longitude) dos compradores e vendedores é possível determinar a distância entre ambos (variável quantitativa contínua, medida em quilômetros);
- **Tempo de entrega:** diferença entre a data de compra e a data de entrega (variável quantitativa contínua, medida em dias);
- **Tempo de resposta da loja**: diferença entre a data de criação do comentário do comprador e a data de resposta do vendedor (variável quantitativa contínua, medida em dias);
- **Tempo do comentário**: diferença entre a data da compra e a data do comentário do consumidor (variável quantitativa contínua, medida em dias);
- **Atraso na entrega:** diferença entre o tempo de entrega previsto e o tempo de entrega real (variável quantitativa contínua, medida em dias);
- **Período do dia:** possuindo a variável data de compra é possível definir se a mesma foi realizada pela manhã, tarde, noite ou madrugada (variável qualitativa ordinal);
- **Avaliação:** considerando a escala de notas de 1 a 5 utilizado no banco de dados, foi atribuído as notas menores ou iguais a 2 avaliação negativa, iguais a 3 a avaliação regular e maiores que 3 a avaliação positiva (variável qualitativa ordinal).

Tab. 1: Descrição das variáveis do trabalho.

Variável	Descrição	Tipo
customer_id	Identificador do comprador	Chave
geolocation_zip_code_prefix	Todos os CEP's	Chave
order_id	Identificador da compra	Chave
product_id	Identificador do produto	Chave
seller_id	Identificador do vendedor	Chave
customer_city	Cidade do comprador	Qualitativa Nominal
customer_state	Estado do comprador	Qualitativa Nominal
payment_type	Forma de pagamento	Qualitativa Nominal
review_comment_title	Título do comentário	Qualitativa Nominal
commen	Comentário (***Input***)	Qualitativa Nominal
product_category_name	Categoria do produto	Qualitativa Nominal
seller_zip_code_prefix	CEP do vendedor	Qualitativa Nominal
seller_city	Cidade do vendedor	Qualitativa Nominal
seller_state	Estado do vendedor	Qualitativa Nominal
order_status	Status da entrega (Vou F)	Qualitativa Ordinal
geolocation_lat	Latitude do CEP	Quantitativa Contínua
geolocation_lng	Longitude do CEP	Quantitativa Contínua
price	Preço do produto (R$)	Quantitativa Contínua
freight_value	Preço do frete (R$)	Quantitativa Contínua
payment_value	Valor da parcela	Quantitativa Contínua
review_creation_date	Data de criação do comentário (Dias)	Quantitativa Contínua
review_answer_timestamp	Data de resposta da loja (Dias)	Quantitativa Contínua
order_purchase_timestamp	Data da compra (Dias)	Quantitativa Contínua
order_approved_at	Data aprovação da compra (Dias)	Quantitativa Contínua
order_delivered_carrier_date	Data de envio do produto (Dias)	Quantitativa Contínua
order_delivered_customer_date	Data de chegada do produto (Dias)	Quantitativa Contínua
order_estimated_delivery_date	Data prevista de entrega (Dias)	Quantitativa Contínua
product_photos_qty	Quantidades de fotos do produto	Quantitativa Contínua
product_weight_g	Peso do produto (Kg)	Quantitativa Contínua
product_length_cm	Comprimento do produto (cm)	Quantitativa Contínua
product_height_cm	Altura do produto (cm)	Quantitativa Contínua
product_width_cm	Largura do produto (cm)	Quantitativa Contínua
customer_zip_code_prefix	CEP do comprador	Quantitativa Discreta
shipping_limit_date	Tempo previsto de transporte (Dias)	Quantitativa Discreta
payment_installments	Número de parcelas	Quantitativa Discreta
review_score	Nota do comprador (1 a 5)	Quantitativa Discreta

B. *Visualização dos dados quantitativos*

Utilizando as coordenadas (latitude e longitude) dos compradores-comentaristas é possível visualizar como ocorre a distribuição geográfica das compras virtuais no Brasil, como a Figura 2 ilustra.

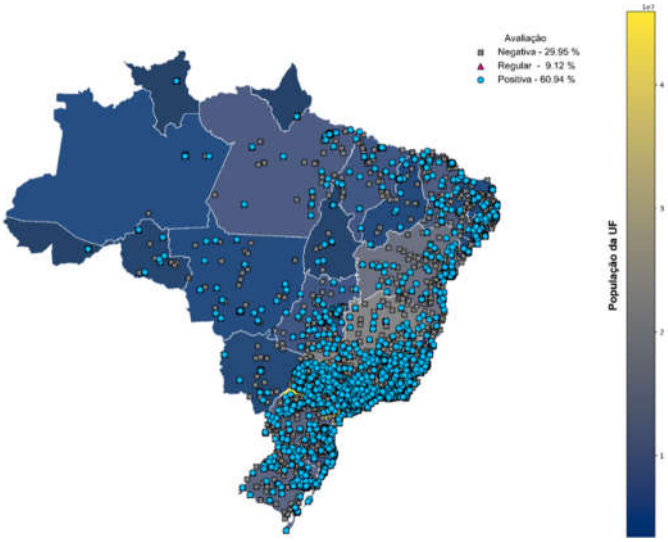

Fig. 2 : Distribuição das compras via e-commerce no Brasil.

As distribuições e agrupamentos podem ser melhor visualizadas através dos mapas interativos disponíveis em (https://www.kaggle.com/feliperoliveira/mapas)[1].

Como esperado, a distribuição dos compradores possui uma relação direta a com a densidade populacional da Unidade Federativa (tipicamente maior na região costeira do país).Também é possível observar que a maioria (60,94%) das avaliações é positiva, logo em seguida das avaliações negativas (29,95%).Isto indica a polarização das avaliações dos consumidores e caracteriza um desbalanceamento de classes no banco de dados (mais acentuado para classe regular , que possui apenas 9,12% das avaliações).

A Figura 3 ilustra a distribuição das avaliações de compradores por região do Brasil. É possível notar de maneira clara a polarização das avaliações dos compradores-comentaristas e o desbalanceamento de classes de avaliações.

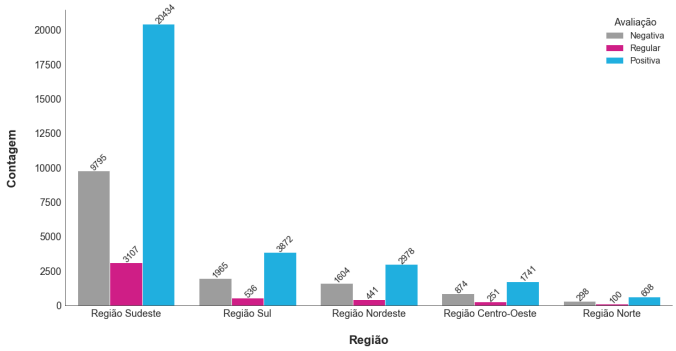

Fig.3: Distribuição das avaliações por região do Brasil.

A Figura 3 evidencia que a Região Sudeste possui o maior número de compradores (68,59% do total do banco de dados), o Estado de São Paulo sozinho é responsável por cerca de 40% de todas as compras no Brasil.

A Figura 4 ilustra a série temporal de compras virtuais. É possível observar que o pico de compras ocorreu no mês de novembro de 2017, provavelmente associado a *black friday*.

Fig.4: Série temporal de compras virtuais.

Os produtos vendidos são divididos em 73 categorias, algumas podem ser consideradas redundantes (Tabela 2). A Figura 5 ilustra a participação percentual das 10 classes mais frequentes. É possível observar que os produtos de cama mesa e banho foram os mais comprados (10,28%), durante o período de aquisição de dados.

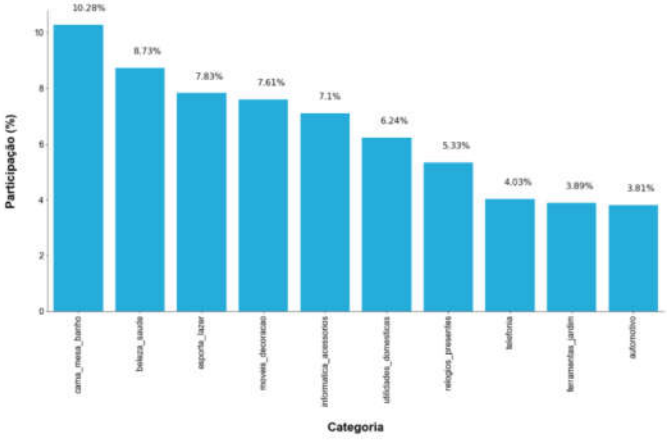

Fig.5:Dez classes de produtos mais comprados.

A distribuição das avaliações seguiu a frequência de ocorrência dos produtos, ou seja, em números absolutos, a maior quantidade de avaliações negativas, regulares e positivas foram dadas aos produtos mais comprados. Proporcionalmente, os itens mais bem avaliados foram os livros (cerca de 70% das avaliações dessa classe de produto é positiva). Por outro lado, os produtos de festa natalina são os mais mal avaliados (cerca de 48% das avaliações dessa classe é negativa.)

No total 20446 registros (aproximadamente 20,5% do total do banco de dados) são de compradores recorrentes, aqueles que realizaram uma ou mais compras no site. Dentro do grupo de compradores recorrentes somente 502 (aproximadamente 0,5% do total do banco de dados) realizaram três ou mais compras. A Figura 7 ilustra todas as associações de compras realizadas. É possível observar que as categorias de itens mais comprados

[1] As legendas podem ser alteradas de acordo com o navegador

também são as que possuem maior recorrência de associação entre si.

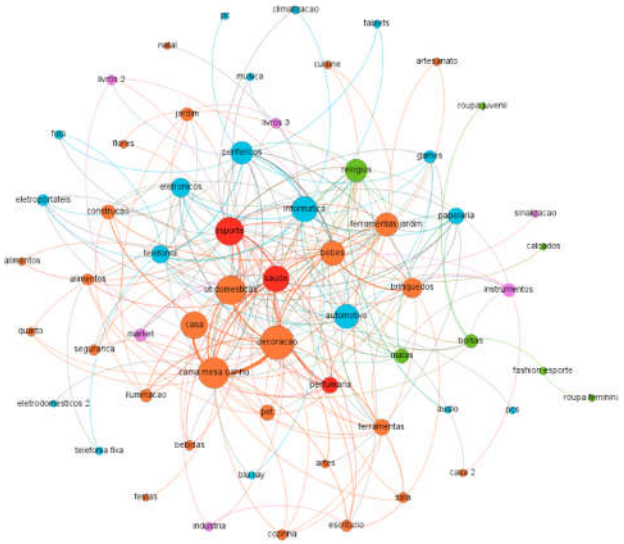

Fig.6: Associação de compras realizadas.

Cerca de 80% das transações reincidente são de itens iguais a da compra anterior mais recente, ou seja, é mais provável que um cliente que compre um certo produto *A* volte a comprar este mesmo produto em uma ocasião futura. Baseado nas distâncias dos nós (métrica que avalia o menor caminho entre dois nós) do grafo ilustrado na Figura 6, observa-se que, em média, os consumidores recorrentes compram até 2 produtos diferentes

A Figura 7 ilustra a distribuição percentual dos produtos por faixa de preço. É possível observar que a maioria das transações (64,92%) foram de valores abaixo de R$100. Os computadores, em média, são os itens mais caros (R$1097,34). No outro extremo, os produtos de casa e conforto apresentam o menor preço médio (R$25,34). O produto mais caro comprado durante o período de aquisição de dados pertence à classe de utilidades domésticas (R$6735).

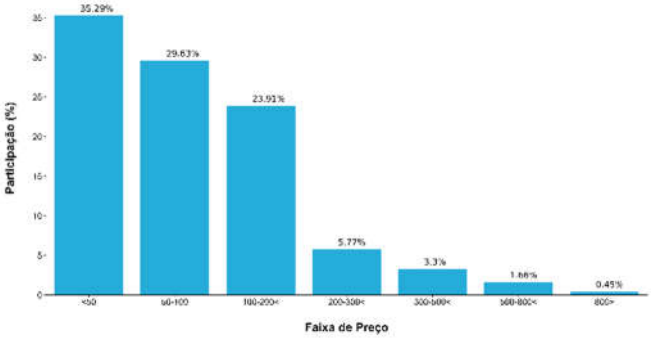

Fig.7: Distribuição das faixas de preço dos itens.

Apesar de apresentar mais de 20 variáveis quantitativas, nem todas podem realmente contribuir com o objetivo desse trabalho (por exemplo, peso e dimensões são irrelevantes). Dessa forma, pretendendo extrair mais informações sobre o banco de dados foi realizada uma análise de correlação das variáveis quantitativas consideradas mais relevantes da perspectiva de recomendação do produto, como a Figura 8 ilustra.

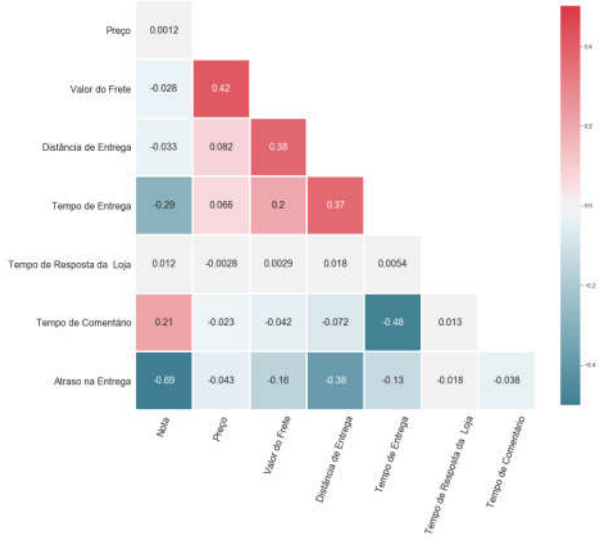

Fig.8 Correlação das variáveis quantitativas.

É possível observar que o tempo de entrega e o tempo de atraso se destacam pela relação inversa com a nota dada pelo comprador, e consequentemente na avaliação, indicando estas como possíveis causas principais da insatisfação.

A Figura 9 apresenta a distribuição do tempo de atraso de entrega de acordo com a avaliação dos compradores. Os pedidos com avaliações positivas, em sua maioria, foram entregues antes do prazo estimado.

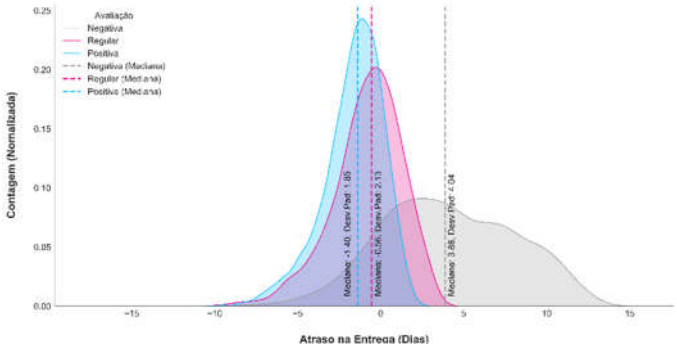

Fig.9: Distribuição do atraso de acordo com a avaliação.

A Figura 9 também caracteriza um problema da perspectiva logística, pois tanto o atraso quanto a antecedência excessiva são indícios de erros de planejamento das entregas (evidente que o atraso é mais desagradável ao comprador).Também é possível observar na Figura 9 que as entregas com avaliação regular apresentam a menor dispersão do ponto 0 (entrega no prazo).

C. Visualização do corpus

O item anterior deste trabalho antecipou algumas tendências esperadas após o processamento de linguagem natural (PNL). Em especial o atraso como um dos principais causadores de avaliações negativas. Contudo, a análise dos dados quantitativos ainda deixou uma série de questionamentos que podem ser elucidados pela análise do conjunto de comentários (corpus).

A Figura 10 ilustra a matriz de co-ocorrência de palavras do corpus. É possível observar a presença de grupos de palavras frequentemente associadas.

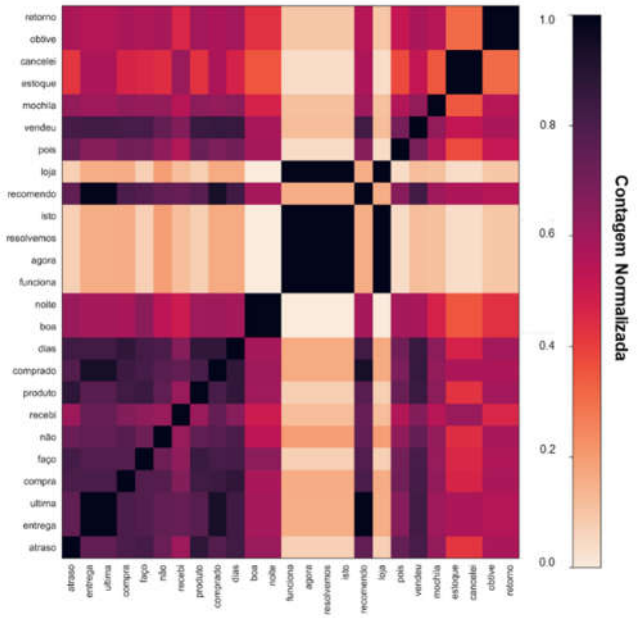

Fig.10: Matriz de co-ocorrência de vinte e cinco palavras do corpus.

A Figura 11 ilustra as associações mais frequentes de palavras. De maneira geral a associação mais frequente observada é entre as palavras "produto" + "entrega" + "prazo". Destaca-se a associação "produto" + "antes" + "prazo" na classe positiva, que indica que uma das principais causas das avaliações positivas é entrega antes do prazo. Por outro lado, também se destaca a associação recorrente entre as palavras "não" + "recebi" + "produto" + "prazo" na classe negativa, que indica que uma das principais causas das avaliações negativas é entrega fora do prazo (ou mesmo a não realização da entrega).

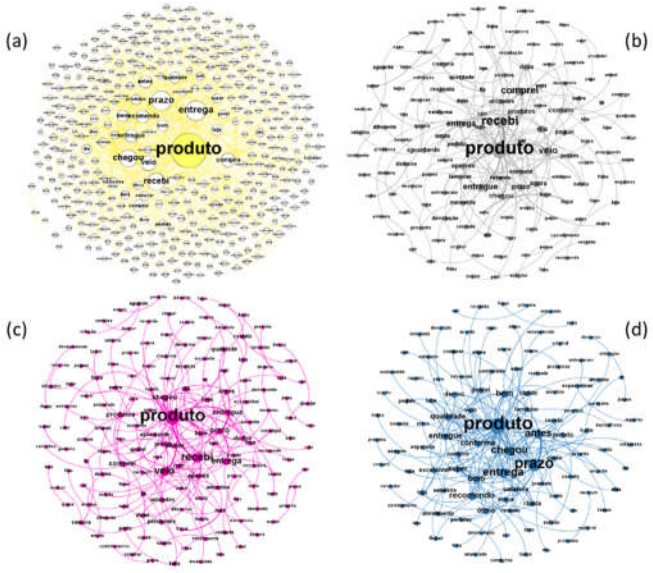

Fig.11: Associações mais frequentes de palavras. (a)Geral. (b)Avaliações negativas. (c)Avaliações regulares. (d)Avaliações positivas.

A Figura 12 ilustra os n-gramas (sequência de palavras) mais frequentes de acordo com a classe de avaliação. Observa-se, novamente, a influência do prazo de entrega na satisfação dos compradores. Vale destacar que os n-gramas mais frequentes da classe regular também são recorrentes nas demais classes.

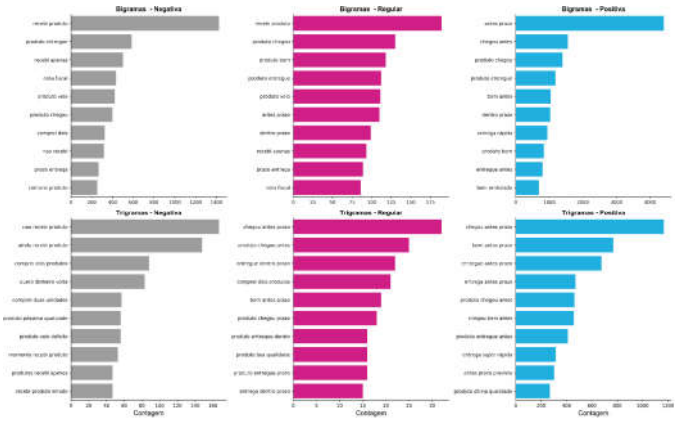

Fig.12: N-gramas mais frequentes de acordo com a avaliação dos compradores.

A Figura 13 ilustra a distribuição de caracteres por comentário (incluindo pontuação e *emojis*) de acordo com a avaliação do comprador. É possível notar que, em média, quanto melhor a avaliação, mais conciso é o comentário. Isso indica que os compradores menos satisfeitos expressam de forma mais detalhada suas motivações, facilitando a identificação de diferentes causas para as avaliações negativas.

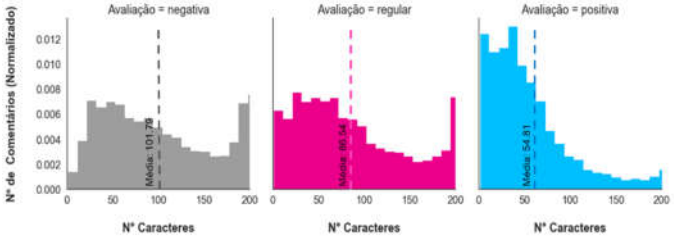

Fig.13: Distribuição de caracteres por comentários de acordo com avaliação.

A Figura14 ilustra a nuvem de palavras relativa as avaliações negativas. A nuvem permite observar a frequência dos termos mais relevantes dessa classe através da diferença de tamanhos das palavras. Observa-se na Figura 14 o destaque de palavras que ,quando associadas, atribuem sentido (entendendo por sentido a capacidade de atribuir significado a frase [5]) de "atraso" e "não recebimento do produto", como esperado após a análise dos dados quantitativos. Também se destacam palavras que remetem a insatisfação com lojistas (lembrando que "*lannister*" é um pseudônimo de uma loja virtual), atendimento e produtos/marcas.

Fig.14: Nuvem de palavras das avaliações negativas.

A Figura 15 ilustra a nuvem de palavras relativas as avaliações regulares. Destacam-se na nuvem termos referentes ao recebimento do produto, a produtos específicos e conjunções adverbiais concessivas. Quando associadas, tais expressões apresentam sentido de "aprovação com ressalvas".

Fig.15: Nuvem de palavras das avaliações regulares.

A Figura 16 ilustra a nuvem de palavras relativa as avaliações positivas. Destacam-se na nuvem termos referentes ao prazo de entrega e elogios. Quando associadas, tais expressões apresentam sentido de "satisfação com o prazo de entrega".

Fig.16: Nuvem de palavras das avaliações positivas.

Combinando a vetorização de *Term Frequency–Inverse Document Frequency* (TF-IDF) e a técnica de redução de dimensionalidade *T-distributed Stochastic Neighbor Embedding* (t-SNE) é possível visualizar a distribuição dos comentários em uma projeção tridimensional, como a Figura 17 ilustra. O t-SNE reduz um vetor de alta dimensionalidade a um ponto, bidimensional ou tridimensional, de tal forma que vetores semelhantes são representados como pontos próximos e vetores diferentes são representados com pontos distantes [6].

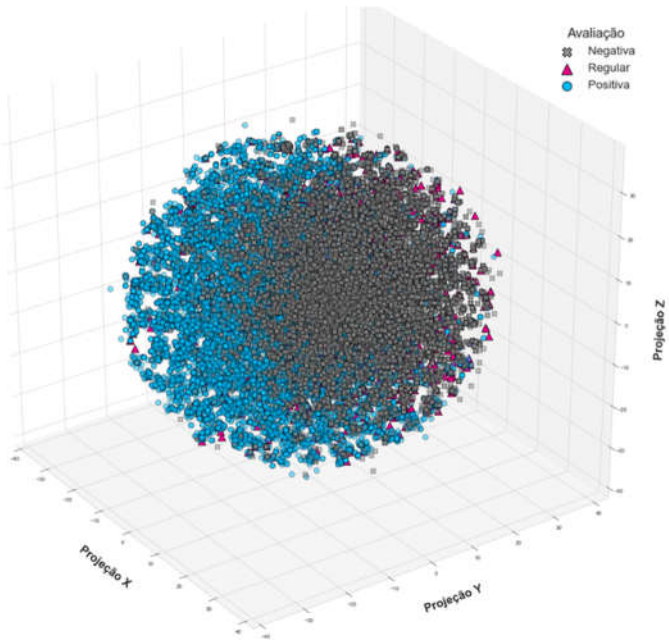

Fig.17: Projeção 3D da distribuição dos comentários.

Na Figura 17, é possível observar diferentes graus de emaranhamento entre as classes. As classes negativa e positiva são, visualmente, mais separáveis entre si. A classe regular é, visualmente, a mais complexa de ser isolada das demais classes.

Dessa forma, em um possível problema de classificação baseado nos comentários dos clientes, haverá grande dificuldade de separar a classe regular das demais classes, o que elevaria os erros de predição.

III. Agrupamento de Classes

Como descrito anteriormente, existe uma aparente redundância de classes de produtos. A pesar de provavelmente haver uma razão comercial para separar tais classes, para os objetivos deste trabalho não há motivo para segregação de alguns itens ("casa e conforto" e "casa e conforto 2", por exemplo).Dessa forma, optou-se pelo agrupamento de classes a fim de (**i**) reduzir a complexidade do problema de recomendação , (**ii**) reduzir a esparsidade da matriz usuário-item e (**iii**) aproximar as classes de itens baseado na similaridade das compras realizadas.

Dessa forma, utilizando o algoritmo *K-Means* e tomando como base a matriz de co-ocorrência de compras (matriz usuário-item), foram avaliados diferentes agrupamentos

possíveis para os produtos. A Figura 18, ilustra a aplicação do *Elbow Method* para determinação do número ideal de grupos.

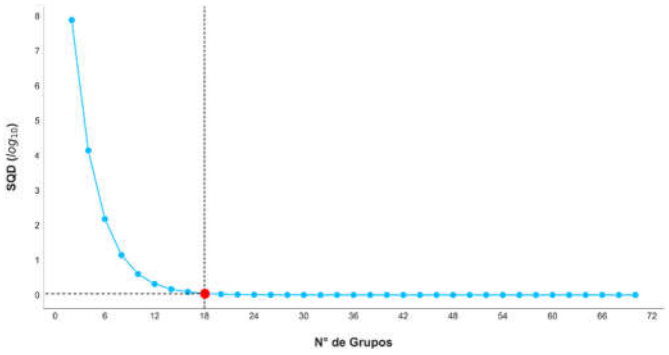

Fig.18: Elbow Method.

A Figura 19 ilustra a análise de silhueta para diferentes números de agrupamento do *K-Means*. A projeção bidimensional dos grupos foi possível através do t-SNE.

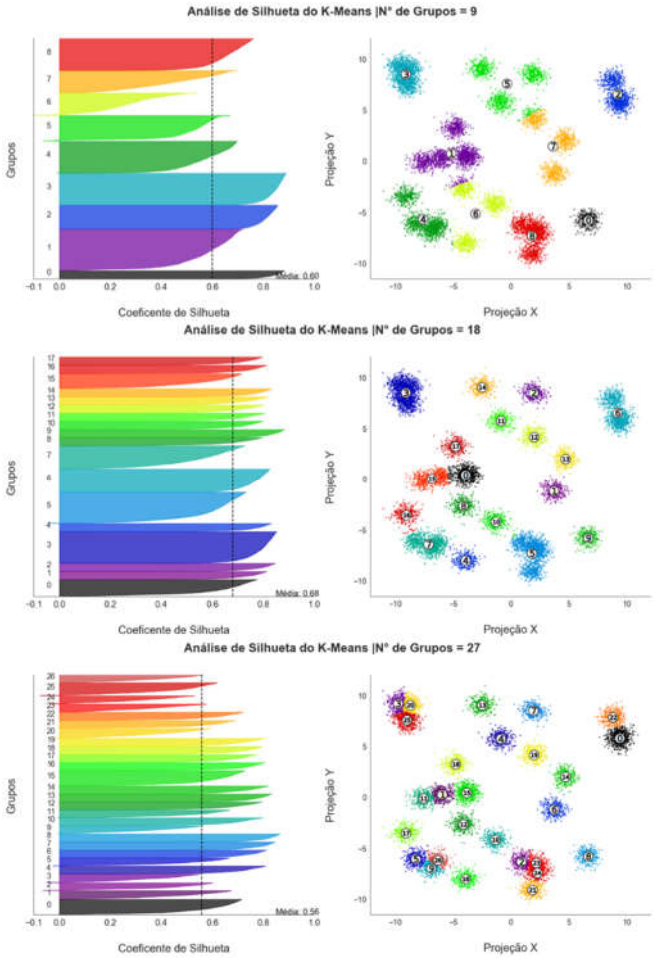

Fig.19: Análise de silhueta.

Tanto a análise visual (*Elbow Method*), quanto a média do índice de silhueta indica que os produtos podem ser agrupados em 18 classes. A Tabela 2 apresenta os produtos e as respectivas classes atribuídas pelo *K-Means*.

Tab.2: Agrupamento de produtos.

Produto	Classe
informatica_acessorios	0
automotivo	1
agro_industria_e_comercio	2
industria_comercio_e_negocios	2
market_place	2
cine_foto	3
pc_gamer	3
pcs	3
seguros_e_servicos	3
tablets_impressao_imagem	3
climatizacao	4
cama_mesa_banho	5
casa_conforto	5
casa_conforto_2	5
casa_construcao	5
la_cuisine	5
telefonia_fixa	5
utilidades_domesticas	5
artigos_de_festas	5
artigos_de_natal	5
esporte_lazer	6
fashion_bolsas_e_acessorios	6
fashion_calcados	6
fashion_esporte	6
fashion_roupa_feminina	6
fashion_roupa_infanto_juvenil	6
fashion_roupa_masculina	6
fashion_underwear_e_moda_praia	6
malas_acessorios	6
beleza_saude	7
fraldas_higiene	7
perfumaria	7
bebes	8
brinquedos	8
consoles_games	8
pet_shop	8
telefonia	8
construcao_ferramentas_construcao	9
construcao_ferramentas_ferramentas	9
construcao_ferramentas_iluminacao	9
construcao_ferramentas_jardim	9
construcao_ferramentas_seguranca	9
ferramentas_jardim	9
flores	9
portateis_casa_forno_e_cafe	9
relogios_presentes	9
moveis_colchao_e_estofado	10
moveis_cozinha_area_de_servico	10
moveis_decoracao	10
moveis_escritorio	10
moveis_quarto	10
moveis_sala	10
alimentos	11
alimentos_bebidas	11
bebidas	11
portateis_cozinha	12
livros_importados	13
livros_interesse_geral	13
livros_tecnicos	13
papelaria	13
artes	14
artes_e_artesanato	14
audio	15
cds_dvds_musicais	15
cool_stuff	15
dvds_blu_ray	15
eletrodomesticos_2	15
instrumentos_musicais	15
musica	15
sinalizacao_e_seguranca	16
eletrodomesticos	17
eletronicos	17
eletroportateis	17

De maneira geral, observa-se que o agrupamento realizado de forma automatizada possui coerência (principalmente pela redução da redundância de produtos). Contudo, alguns grupos incomuns também foram gerados, destacando a classe 8 que engloba itens de pet shop e telefonia em uma mesma classe.

IV. Sistemas de Recomendação

Os SR costumam enfrentar dificuldades em gerar recomendações de qualidade para novos usuários, devido à falta de informações prévias para modelar suas preferências (criar um perfil). Esse problema é conhecido como inicialização a frio (*cold start*) [7]. Dessa forma, dada a baixa proporção de clientes que realizaram mais de uma compra no banco de dados adotado optou-se pela estratégia de recomendação baseada em intens.

Neste trabalho são avaliados diferentes modelos de filtragem para Sistemas de Recomendação, são eles: (**i**) modelo de popularidade, utilizado como referência, e a (**ii**) filtragem colaborativa. Os tópicos seguintes tratarão destes modelos.

A. Modelos de popularidade

Os modelos baseados em popularidade estão entre os mais utilizados, devido sua fácil implementação e lógica. Esse tipo de modelo simplesmente recomenda a um usuário os itens mais populares observados, sem nenhum tipo de customização. Esse modelo parte do princípio que a popularidade reflete o "conhecimento coletivo", e geralmente costuma fornecer boas recomendações para a maioria dos consumidores [8].

B. Filtragem colaborativa

A filtragem colaborativa (CF) tem duas estratégias principais de implementação: (**i**) memória, que determina a semelhança entre itens baseado nas aquisições prévias dos usuários e (**ii**) modelos, que utilizam diferentes algoritmos de aprendizado de máquina para recomendar itens aos usuários.

Os modelos de fatores latentes (LMF) estão entre os mais aplicados em sistemas de filtragem colaborativa. Os LMF comprimem a matriz usuário-item em uma representação de baixa dimensionalidade em termos de fatores latentes. Uma vantagem desse paradigma é que, em vez de trabalhar com uma matriz esparsa de alta dimensão, lida-se com uma matriz menor em um espaço dimensional inferior [9].

Neste trabalho foi adotado o modelo de fator latente denominado Decomposição de Valores Singulares (SVD). O SVD é um método que decompõe uma matriz A $(n \times p)$ em três outras matrizes, como as seguintes equações descrevem:

$$A_{n\times p} = U_{n\times n} \cdot S_{n\times p} \cdot V^T_{p\times p} \quad (1)$$

$$U^T \cdot U = I_{n\times n} \quad (2)$$

$$V^T \cdot V = I_{p\times p} \quad (3)$$

Onde I representa a matriz identidade, as colunas U são os vetores singulares (vetores de coeficiente), S é a matriz diagonal de valores singulares (amplitudes de modo) e as linhas de V^T são os vetores singulares (similaridade entre intens. e valores latentes). Dessa forma, o cálculo do SVD resume-se a encontrar os autovalores e autovetores AA^T e A^TA.

No contexto dos Sistemas de Recomendação, o SVD é aplicado a matriz usuário-item de forma que cada linha n representa um usuário e cada coluna p representa um item. Os elementos dessa matriz são os números de itens comprados ponderados pelas avaliações que são dadas aos itens pelos usuários. A figura 20 exemplifica a aplicação do SVD na previsão de notas dadas a produtos.

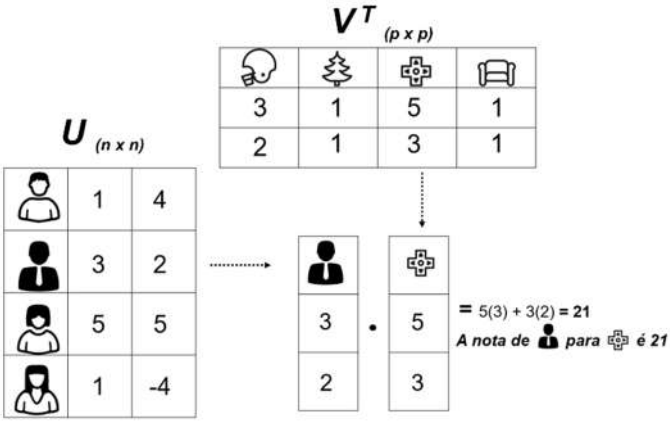

Fig.20 : Exemplo de aplicação do SVD.

C. Métricas de avaliação

A acurácia é uma métrica de avaliação de modelos de recomendação que informa a fração das previsões que o SR acertou. A acurácia A é dada por:

$$A = \frac{TP+TN}{TT} \quad (4)$$

Onde TP corresponde ao número de previsões corretas de uma determinada classe (vamos chamá-la de $C1$), TN corresponde as predições corretas das classes diferentes de $C1$.

Vale destacar que a acurácia é uma métrica pouco interessante quando o conjunto de dados de treino é muito desbalanceado. Por outro lado, a revocação *(recall*) é uma métrica mais adequada para esse tipo de situação, pois indica a proporção de classes que foi recomendada corretamente. A revocação R é dada por:

$$R = \frac{TP}{TP+FN} \quad (5)$$

Onde FN é o número de recomendações de produtos diferentes dos realmente adquiridos. A precisão, que indica a proporção de previsões que está realmente correta, é dada por:

$$P = \frac{TP}{TP+FP} \quad (6)$$

Onde FP é o número de recomendações incorretas feitas pra uma classe específica, mas que na verdade deviam ter sido feitas a outras classes. A métrica $F1$, que indica o nível de significância da acurácia, é dada por:

$$F1 = 2.\left(\frac{P.R}{P+R}\right) \quad (7)$$

Até então as técnicas de avaliação apresentadas foram extraídas de problemas de classificação. Contudo, quando tratamos de Sistemas de Recomendação, ao invés de avaliar a qualidade de uma previsão, avalia-se as k melhores recomendações para um determinado cliente. Dessa forma, as métricas de avaliação dos SR são popularmente conhecidas como *recall*@k, precisão@k e F1@k.

V. Procedimento Experimental

A Figura 21 ilustra a partição do banco de dados que foi adotada neste trabalho.

Fig.21: Particionamento do banco de dados.

A separação dos registros de compras isoladas (**i**), treino-teste (**ii**) e a posterior segregação treino (**iii**) e teste (**iv**) foi adotada partindo do princípio que as experiências passadas podem "ensinar" um SR a gerar recomendações de qualidade para usuários recorrentes. Dessa forma, o problema da inicialização a frio pode ser contornado.

A. SR baseado em popularidade

Para o desenvolvimento do modelo de popularidade foi adotado a técnica de *dummy classifier* presente na biblioteca Scikit-Learn, em *Python*.O *dummy classifier* recomenda os k itens mais frequentes no banco de dados de treino.

B. SR baseado em filtragem colaborativa

Para o desenvolvimento do sistema de classificação baseado em filtragem colaborativa (FC) foi adotada a biblioteca *Suprise*, em *Python*, utilizando como variável de entrada a matriz usuário-item. O *Suprise* utiliza os vetores singulares produzidos pelo SVD, como parâmetros de um regressor que estima as notas que um cliente dará a um determinado produto, como a Figura 20 ilustra [10]. Dessa forma, o objetivo do algoritmo adotado é minimizar (através do método dos gradientes) a raiz quadrada do erro médio quadrático (RMSE) das previsões. O RMSE é dado por:

$$RMSE = \sqrt{\frac{\sum_i^N (\hat{y}_i - y_i)^2}{N}} \qquad (8)$$

Onde $\hat{y}$ é a nota estimada, y é o valor real da nota e N é o número de estimativas. Dessa maneira, o sistema de recomendação colaborativo implementado recomenda os produtos com as k melhores previsões de avaliação. Vale ressaltar que o SVD é apenas um dos diversos algoritmos de LMF disponíveis na biblioteca *Suprise*.

VI. Resultados e Discussão

As Figuras 22 a 24 ilustram, respectivamente, a revocação média, a precisão média e a média de F1 alcançada pelo modelo baseado em popularidade. Observa-se que os resultados obtidos (a revocação em especial) indicam o baixo desempenho deste modelo.

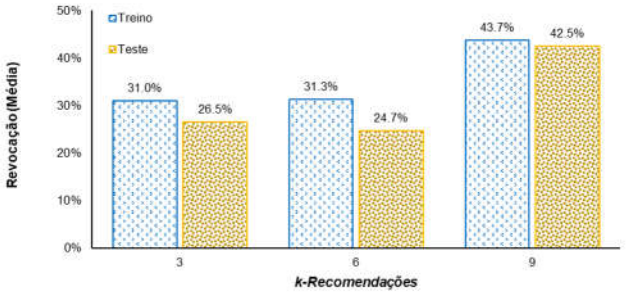

Fig.22: Revocação média - Popularidade.

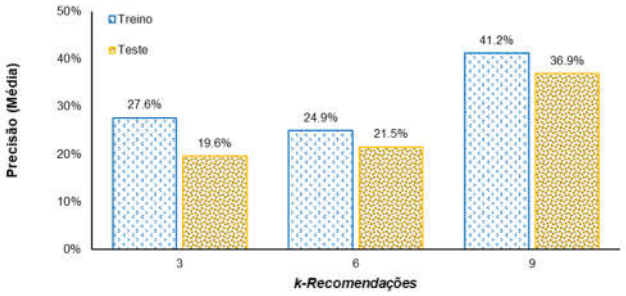

Fig.23: Precisão média - Popularidade.

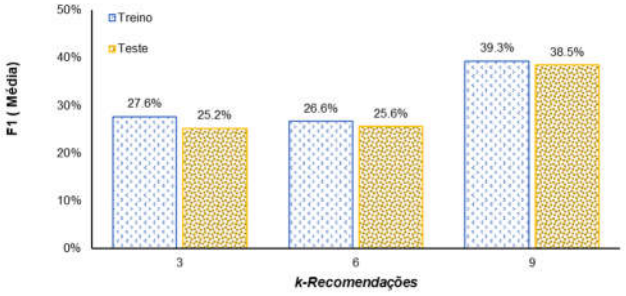

Fig.24: Média de F1- Popularidade

As Figuras 25 a 27 ilustram, respectivamente, a revocação média, a precisão média e a média de F1 alcançada pelo modelo baseado em filtragem colaborativa. Observa-se que os resultados obtidos indicam o desempenho satisfatório desse modelo, em especial quando são realizadas 9 recomendações de produtos.

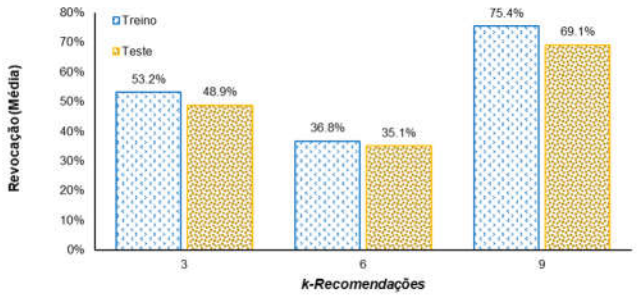

Fig.25: Revocação média - Filtragem Colaborativa.

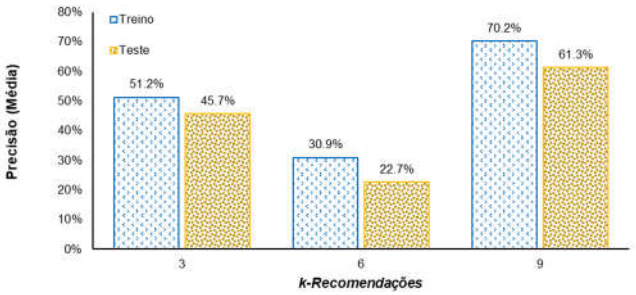

Fig.26: Precisão mdia - Filtragem Colaborativa.

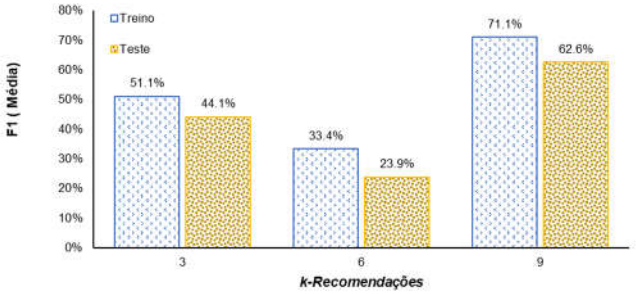

Fig.27: Média de F1 -Filtragem Colaborativa.

A raiz quadrada do erro médio quadrático (RMSE) alcançado pelo sistema de recomendação colaborativo na previsão de notas foi de 0,79, um valor ligeiramente elevado considerando uma escala de notas que varia entre 1 e 5.

Em ambos os SR observa-se que (**i**) o desempenho nos dados de treino é superior ao alcançado nos dados de teste, caracterizando o sobreajuste (*overfitting*) , e (**ii**) à medida que o número de recomendações aumenta, maior é a taxa de revocação média. Existe um sentido probabilístico nesta última observação, pois quanto maior é o número de classes sugeridas maior é a chance de acertar qual produto realmente foi adquirido.

As Tabelas 3 e 4 sumarizam os resultados alcançados pelo modelo de popularidade e pelo modelo de filtragem colaborativa, respectivamente. Quando comparados, independentemente do número de recomendações, a filtragem colaborativa obteve resultados muito superiores em todas as métricas aferidas.

Tab.3:Resultados do modelo de popularidade.

Partição	N° de Recomendações	Revocação (Média)	Precisão (Média)	F1 (Média)
Treino	3	31.0% ± 2.9%	27.6% ± 2.3%	27.6% ± 0.2%
	6	31.3% ± 2.1%	24.9% ± 5.9%	26.6% ± 1.4%
	9	43.7% ± 0.1%	41.2% ± 0.2%	39.3% ± 3.8%
Teste	3	26.5% ± 3.0%	19.6% ± 2.1%	25.2% ± 0.1%
	6	24.7% ± 0.0%	21.5% ± 7.2%	25.6% ± 1.1%
	9	42.5% ± 3.4%	36.9% ± 2.3%	38.5% ± 1.0%

Tab.4: Resultados do modelo de baseado em FC.

Partição	N° de Recomendações	Revocação (Média)	Precisão (Média)	F1 (Média)
Treino	3	53.2% ± 1.3%	51.2% ± 0.1%	51.1% ± 4.4%
	6	36.8% ± 1.6%	30.9% ± 1.3%	33.4% ± 1.9%
	9	75.4% ± 1.1%	70.2% ± 6.5%	71.1% ± 3.3%
Teste	3	48.9% ± 1.1%	45.7% ± 0.2%	44.1% ± 3.7%
	6	35.1% ± 1.4%	22.7% ± 0.1%	23.9% ± 1.6%
	9	69.1% ± 1.2%	61.3% ± 3.0%	62.6% ± 4.1%

As Figuras 28 e 29 ilustram a taxa de revocação para as diferentes classes de produtos, alcançadas pelos SR baseados em popularidade e FC, respectivamente, quando aplicados no conjunto de dados de teste. Em ambos os SR foi observado que as classes que englobam os produtos mais comprados apresentam as maiores taxas de *recall* e precisão (os produtos de cama mês e banho fazem parte de C5, por exemplo). Semelhante a um problema de classificação desbalanceado, onde o classificador apresenta as maiores taxas de acerto nas classes mais frequentes.

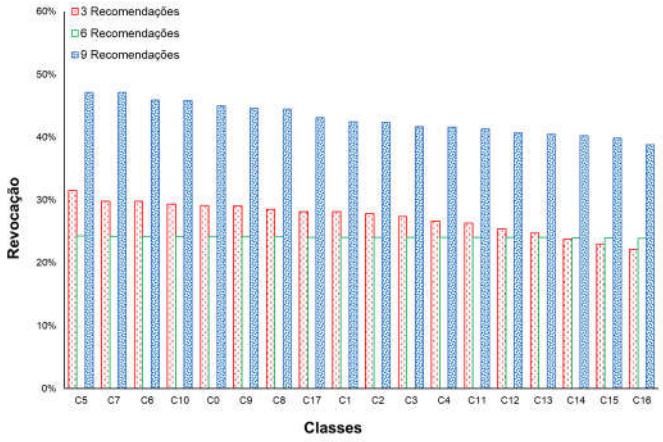

Fig.28:Revocação por classe -Popularidade (dados de teste).

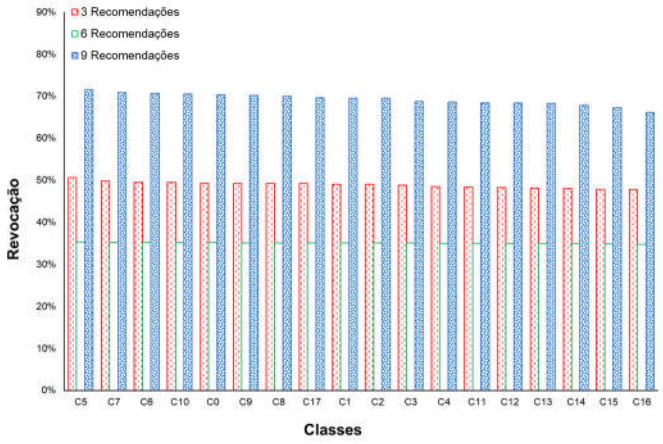

Fig.29: Revocação por classe -FC (dados de teste).

VII. Conclusões e Críticas

A análise exploratória dos dados (AED) quantitativos e do corpus, evidenciou a relação inversa entre o atraso e a satisfação dos clientes. A AED também permitiu reconhecer as dificuldades intrínsecas ao problema de identificação da

satisfação dos consumidores através dos seus comentários. É particularmente interessante observar como duas metodologias tão distintas podem levar a mesma conclusão.

É importante destacar que esse trabalho não possui um único objeto de interesse, ou seja, a extração de informação em si justifica as aplicações das técnicas adotadas. Contudo, em um cenário empresarial hipotético onde o *churn* (taxa de abandono dos clientes) é relevante, a identificação de avaliações negativas e suas motivações, podem ser úteis na tomada de decisão comerciais. Dessa forma, baseado nos resultados apresentados, os vendedores devem focar esforços na melhoria de seus sistemas de entrega (em especial a loja "*lannister*", que concentra o maior número de reclamações).

Inicialmente era pretendido a comparação entre três Sistemas de Recomendação (SR) diferentes , contudo o sistema de filtragem baseado em conteúdo parece ser pouco viável para o banco de dados da *Olist Store* pois (**i**) existem poucas informações para o desenvolvimento do perfil de seus usuários; (**ii**) Os comentários ,apesar de serem úteis para medir a satisfação dos clientes, não apresentam as descrições dos produtos , tornando difícil traçar a similaridade entre os mesmos. Dessa forma, o banco de dados utilizado pode ser caracterizado como propenso ao *cold start,* pouco favorável ao desenvolvimento de SR, independentemente do tipo.

Dadas as características pouco favoráveis do banco de dados utilizado, foram escolhidos os SR baseados em popularidade e filtragem colaborativa (FC), por demandarem poucas informações para seu funcionamento. Dentre as abordagens experimentadas a FC foi a que apresentou o melhor desempenho. Contudo, a estratégia de partição de dados, que apesar de produzir resultados satisfatórios, descartou a maioria dos registros, prejudicando a capacidade de generalização dos SR.

Vale ressaltar que, como descrito na AED, 80% das compras recorrentes são iguais as últimas aquisições mais recentes. Ou seja, um sistema que simplesmente recomenda-se o mesmo produto da compra passada teria um desempenho muito próximo (se não superior) ao melhor modelo de recomendação desenvolvido. Dessa forma, entende-se que mais que a complexidade dos algoritmos de programação utilizados e das métricas aferidas, um SR satisfatório pode ser construído aplicando um sistema logico consistente.

O banco de dados escolhido também possibilita os seguintes desenvolvimentos futuros:

- Análise de agrupamento dentro das diferentes classes de avaliações, a fim de descobrir suas motivações (em especial das avaliações negativas);
- Análise de geo-agrupamento, a fim de entender a influência do fator geográfico no perfil dos compradores [11];
- Estudo de sistemas de agrupamento e recomendação através de grafos;
- A aplicação de sistemas de recomendação híbridos.

VIII. REFERÊNCIAS

[1] J. Lin, L. Li, X. (Robert) Luo, e J. Benitez, "How do agribusinesses thrive through complexity? The pivotal role of e-commerce capability and business agility", *Decis. Support Syst.*, nº June, p. 113342, 2020.

[2] I. R. G. Medeiros, "Estudo sobre Sistemas de Recomendação Colaborativos", UFPE, 2013.

[3] C. Feng, J. Liang, P. Song, e Z. Wang, "A fusion collaborative filtering method for sparse data in recommender systems", *Inf. Sci. (Ny).*, vol. 521, p. 365–379, 2020.

[4] Olist Store, "Brazilian E-Commerce Public Dataset by Olist", 2018. [Online]. Available at: https://www.kaggle.com/olistbr/brazilian-ecommerce. [Acessado: 03-ago-2020].

[5] H. A. Da Fontoura e L. S. Siegel, "Reading, syntactic, and working memory skills of bilingual Portuguese-English Canadian children", *Read. Writ.*, 1995.

[6] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, e D. K. Hartline, "t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis", *Mar. Genomics*, vol. 51, nº September, p. 100723, 2020.

[7] N. Silva, D. Carvalho, A. C. M. Pereira, F. Mourão, e L. Rocha, "The Pure Cold-Start Problem: A deep study about how to conquer first-time users in recommendations domains", *Inf. Syst.*, vol. 80, p. 1–12, 2019.

[8] R. M. Bertani, R. A. C. Bianchi, e A. H. R. Costa, "Combining novelty and popularity on personalised recommendations via user profile learning", *Expert Syst. Appl.*, vol. 146, p. 113149, 2020.

[9] G. Ye e X. Zhao, "Improved SVD algorithm based on Slope One", *Proc. 30th Chinese Control Decis. Conf. CCDC 2018*, nº 1, p. 1002–1006, 2018.

[10] N. Hug, "Surprise' documentation". [Online]. Available at: https://surprise.readthedocs.io/en/stable/index.html. [Acessado: 14-set-2020].

[11] Y. Ma, J. Mao, Z. Ba, e G. Li, "Location recommendation by combining geographical, categorical, and social preferences with location popularity", *Inf. Process. Manag.*, vol. 57, nº 4, p. 102251, 2020.