MODELOS LINEARES E NÃO LINEARES APLICADOS À ANÁLISE DE SENTIMENTOS DE CONSUMIDORES DE *E-COMMERCE* NO BRASIL.

Felipe R. Oliveira

Resumo:

É crucial para empresas e fabricantes que desejam se manter competitivos no mercado digital, o estudo da satisfação dos seus consumidores. Neste contexto o presente trabalho tem como objetivo estudar o banco de dados da *Olist Store,* contendo 100 mil registros, utilizando processamento de linguagem natural (PNL) para criação de modelos lineares e não lineares de classificação da avaliação de compradores de lojas virtuais através dos seus comentários. Diferentes técnicas de análise indicam que o atraso é o principal fator de insatisfação dos clientes. Dentre os modelos experimentados, os não lineares alcançam os melhores desempenhos. Em todas as análises realizadas observa-se a dificuldade em prever a classe menos frequente do banco de dados.

1. 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 na conjuntura atual, o estudo da satisfação dos seus consumidores.

Nesse contexto a análise de sentimentos apresenta um conjunto de métodos que auxiliam na identificação da satisfação dos consumidores através de opiniões expressas textualmente (comentários, mensagens, etc.). Pretendendo extrair o máximo de informação destes dados não estruturados é normal que a análise de sentimentos necessite da aplicação de técnicas de processamento de linguagem natural (PLN) [2]. Ao longo deste trabalho diferentes ferramentas serão discutidas à medida que são apresentadas.

O PLN tem como objetivo realizar a tradução matemática da linguagem humana. O PLN, tradicionalmente, trabalha em três escalas de granularidade, são elas: (**i**) documento, (**ii**) sentença e (**iii**) entidade. Na primeira escala cada documento (cada comentário, no caso deste trabalho) é considerado uma unidade básica de informação que, após a vetorização, pode ser utilizada em modelos de Aprendizado de Máquinas (frequentemente em problemas de classificação) [3].

1.1. Apresentação do problema

Este trabalho dedica-se a análise de um banco de dados da *Olist Store*, disponível no repositório *Keggale*, que possui registros de 100 mil compras realizadas, entre 2016 a 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.

A fim de delimitar o escopo do relatório e aplicar os conceitos da disciplina Inteligência Computacional, esse trabalho tem como objetivo estudar o banco de dados da *Olist Store* utilizando PNL para criação de modelos lineares e não lineares de classificação da avaliação (*output*) de compradores de lojas virtuais através dos seus respectivos comentários (*input*).

1.2. 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:

i. **Pandas:** biblioteca utilizada na manipulação de dados matriciais na forma de tabelas;
ii. **GeoPandas:** biblioteca utilizada na manipulação de dados georreferenciados, permitindo também a criação de mapas;
iii. **Seaborn:** biblioteca utilizada na criação de gráficos, em especial os dedicados a representações estatísticas, como visualização de histogramas, matrizes de correlação, etc.;
iv. **NLTK:** um conjunto de bibliotecas utilizada na manipulação do corpus, responsável pelo processamento simbólico e estatístico da linguagem natural;
v. **Scikit-Learn:** biblioteca utilizada na criação de modelos de aprendizagem de máquina, também auxiliando na obtenção e otimização de modelos.

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

2. CARATERIZAÇÃO E VISUALIZAÇÃO DE DADOS

O banco de dados é formado por 8 tabelas relacionadas entre si através de chaves (não interpretadas como variáveis) contendo 100 mil registros. A Figura 1 ilustra a organização do banco de dados utilizado neste trabalho.

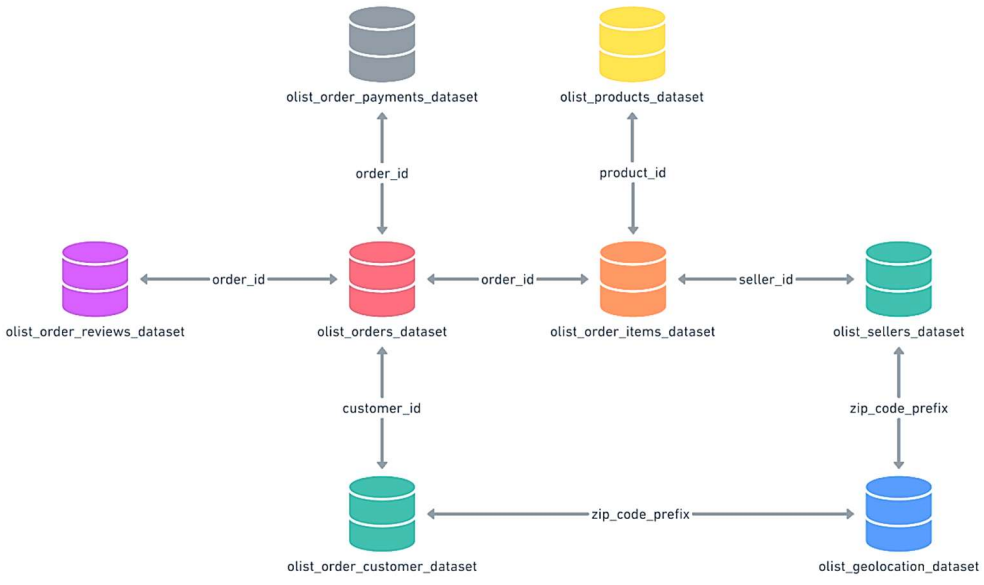

Figura 1: Esquema relacional do banco de dados [4].

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 descrição e tipo das variáveis.

Tabela 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

Utilizando as variáveis originais do banco de dados foram agregadas as seguintes novas variáveis:

i. **Distância de entrega:** possuindo as coordenadas (latitude e longitude) dos compradores e vendedores é possível determinar a distância entre ambos (variável quantitativa continua, medida em quilômetros);
ii. **Tempo de entrega:** diferença entre a data de compra e a data de entrega (variável quantitativa contínua, medida em dias);

iii. **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).

iv. **Tempo do comentário:** diferença entre a data da compra e a data do comentário do comprador (variável quantitativa contínua, medida em dias).

v. **Atraso na entrega:** diferença entre o tempo de entrega previsto e o tempo de entrega real (variável quantitativa contínua, medida em dias).

vi. **Avaliação (*output*):** 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).

2.1. Tratamento de Informações Faltantes

A aplicação das técnicas de PNL pretendidas requerem um corpus, que neste trabalho compreende todos os textos-comentários dos compradores. Dessa forma o banco de dados original, contendo 100 mil registros referentes a compras realizadas, foi subtraído de 521171 entradas que possuem o campo '*review comment message*' faltante, mantendo aproximadamente 49% dos registros originais, referentes a compradores que fizeram comentários.

2.2. Visualização dos Dados Quantitativos

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

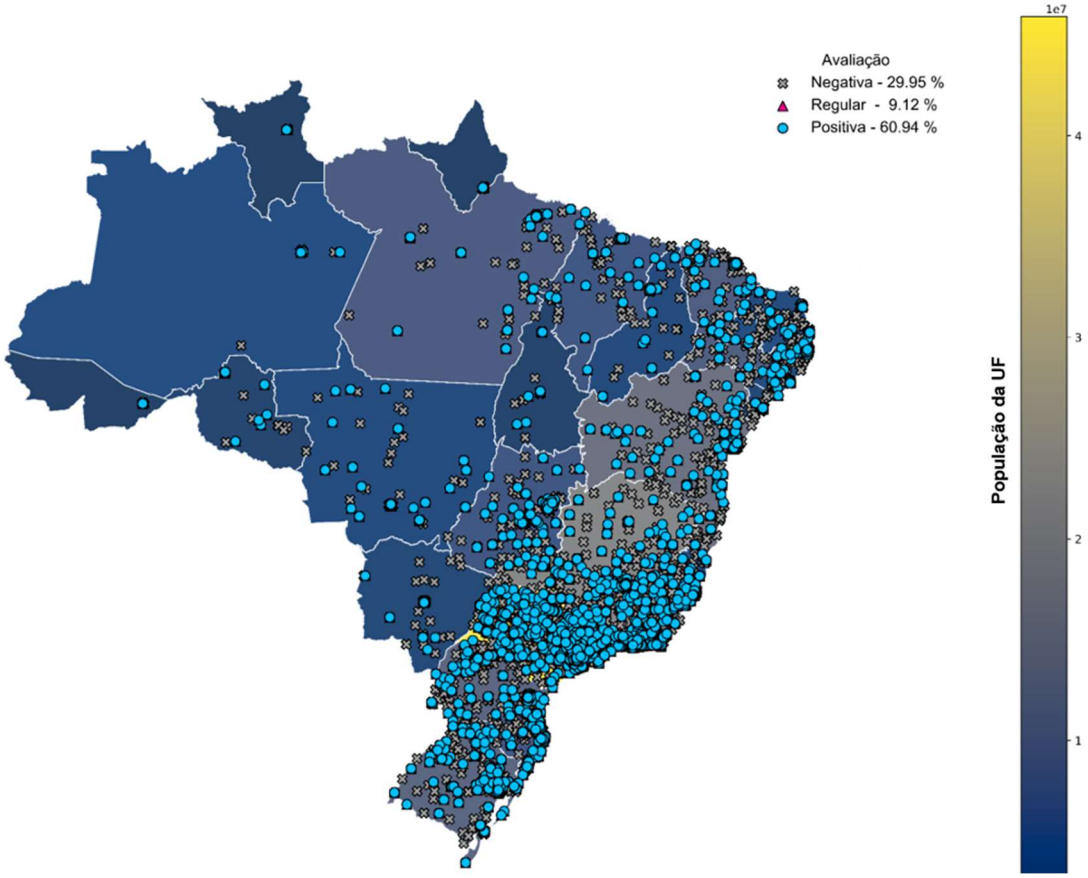

Figura 2: Ditribuição das compras via *e-commerce* no Brasil.

A as distribuições e agrupamentos podem ser melhor visualizadas através dos mapas interativos disponíveis em: https://www.kaggle.com/feliperoliveira/mapas.

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 na Figura 2 que a maioria (60.94%) das avaliações com comentários é positiva, logo em seguida das avaliações negativas (29.95%).Isto indica a polarização das avaliações dos compradores e caracteriza um desbalanceamento de classes no banco de dados (mais desfavorável para classe regular, que possui apenas 9.10% 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 e o desbalanceamento de classes.

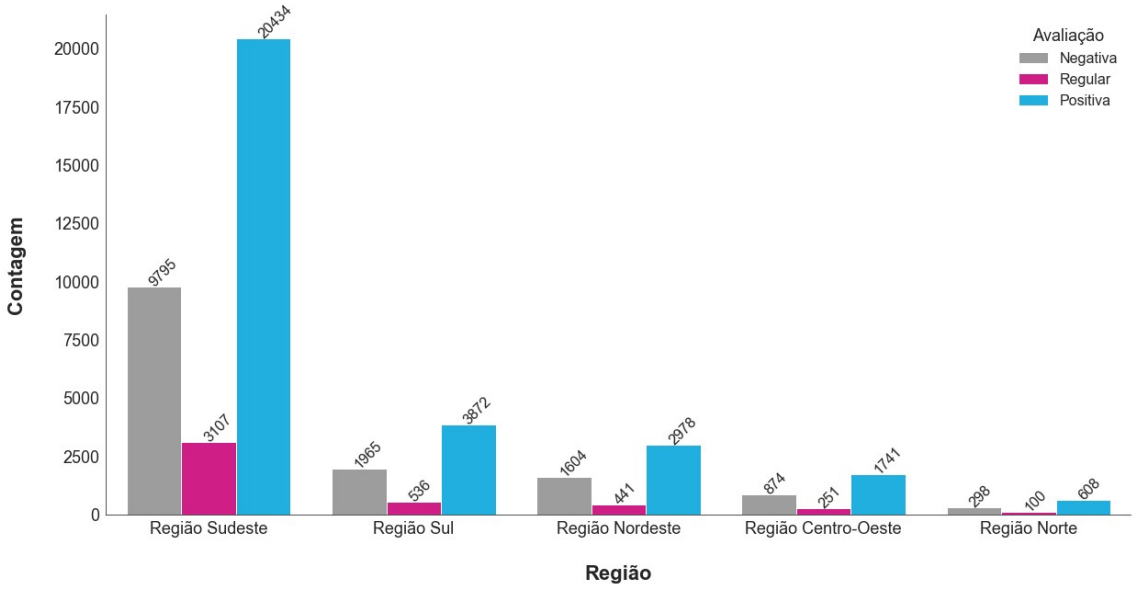

Figura 3: Distribuição das avaliações por região do Brasil.

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 alta concentração de compradores na Região Sudeste, e suas respectivas UF's, era um fato esperado devido à alta densidade populacional dessa região.

Apesar de apresentar mais de 20 variáveis numéricas, poucas delas 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 breve análise exploratória das variáveis quantitativas consideradas mais relevantes da perspectiva da satisfação do consumidor. A Tabela 2 apresenta as estatísticas descritivas dessas variáveis.

Tabela 2: Estatística descritiva das variáveis quantitativas.

Estatística	Nota	Preço (R$)	Frete (R$)	Distância (Km)	Entrega (Dias)	Resposta da Loja (Dias)	Tempo do Comentário (Dias)	Atraso (Dias)
Contagem	48829	48829	48829	48829	47206	48829	47206	47206
Média	3.54	124.86	20.54	623.99	13.38	3.14	-0.05	0.03
Desv. Pad.	1.66	193.84	16.51	609.01	10.71	9.09	2.45	3.69
Mínimo	1.00	0.85	0.00	0.00	0.86	0.09	-7.20	-17.24
1° Quartil	2.00	39.99	13.23	208.91	6.93	0.99	0.11	-2.13
Mediana	4.00	76.00	16.55	446.74	10.68	1.65	0.24	-0.68
3º Quartil	5.00	138.00	21.78	817.55	16.47	3.11	0.38	1.09
Máximo	5.00	6735.00	375.28	3378.71	209.63	446.87	106.38	14.02

Na Tabela 2 é possível observar que apesar da eliminação de informações faltantes referentes aos comentários dos compradores, outras variáveis (não relacionadas diretamente aos modelos pretendidos) apresentam informações faltantes, fato evidenciado nas diferenças de contagens. É possível também observar alguns valores singulares, como entregas com aproximadamente 7 meses de demora, vendedores que demoram mais de 1 ano pra responder os consumidores e comentários feitos após 100 dias da entrega do produto.

A Figura 4 ilustra as correlações entre as variáveis quantitativas do banco de dados. 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.

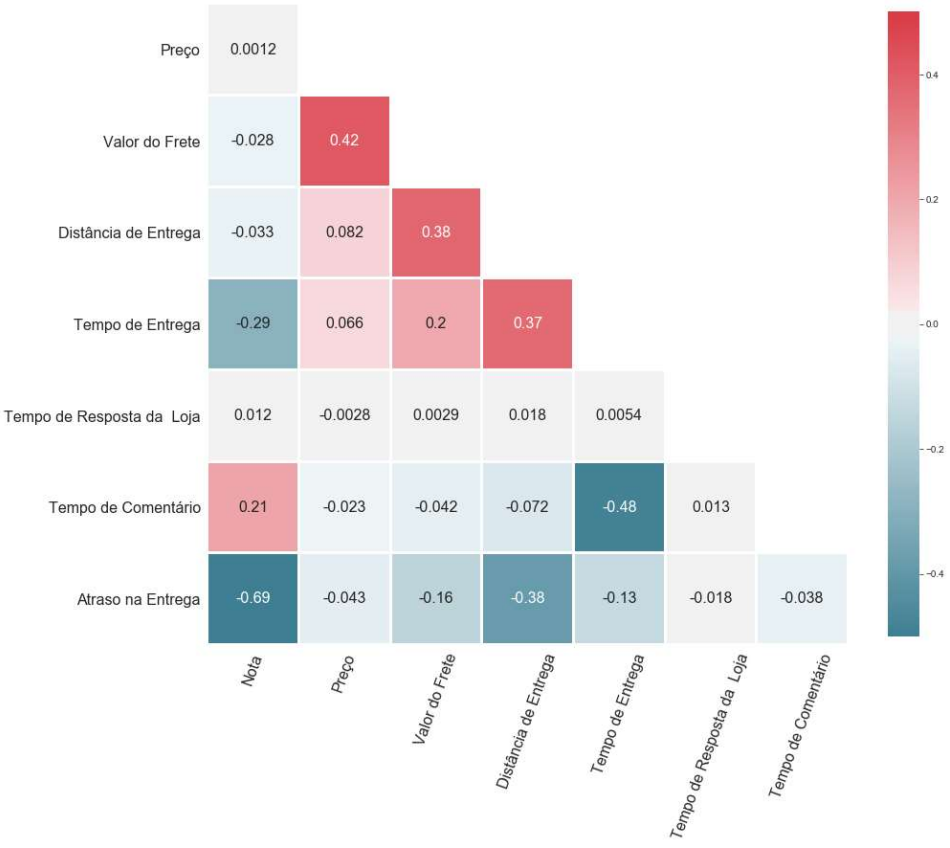

Figura 4: Correlação das variáveis quantitativas.

A Figura 5 apresenta a distribuição do tempo de atraso de entrega de acordo com a avaliação dos compradores. É possível observar que a maioria das avaliações negativas estão relacionadas à pedidos que sofreram atrasos. Os pedidos com avaliações positivas, em sua maioria, foram entregues antes do prazo estimado. Os pedidos com avaliação regular tiveram mediana próxima a zero, ou seja, o número de entregas atrasadas quase se iguala as entregas adiantadas.

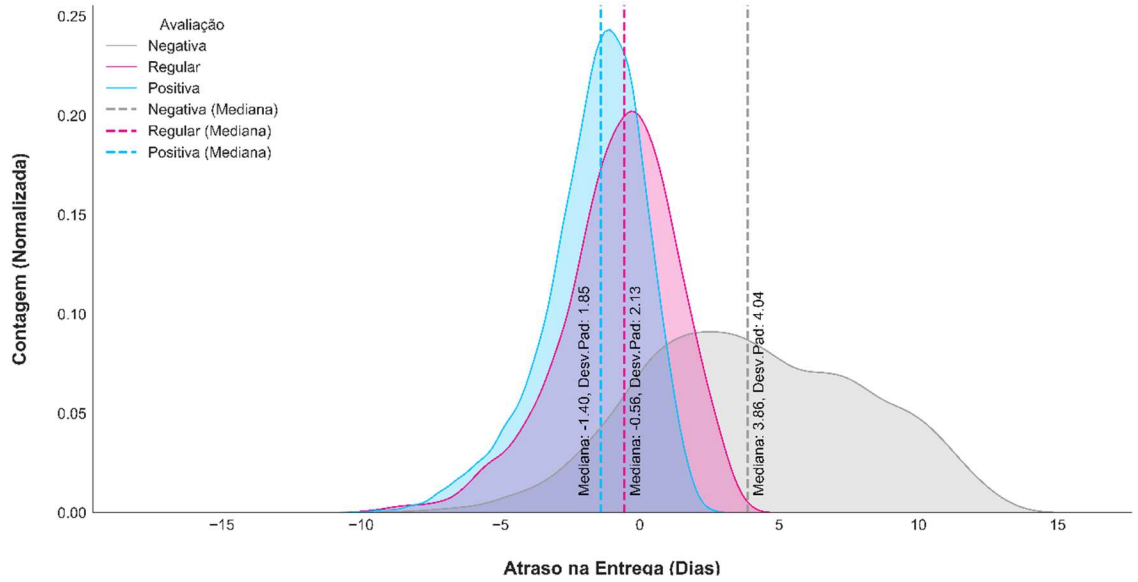

Figura 5:Distribuição do atraso nas entregas de acordo com a avaliação.

A Figura 6 ilustra o diagrama de caixa da variável atraso na entrega em relação a avaliação do comprador.

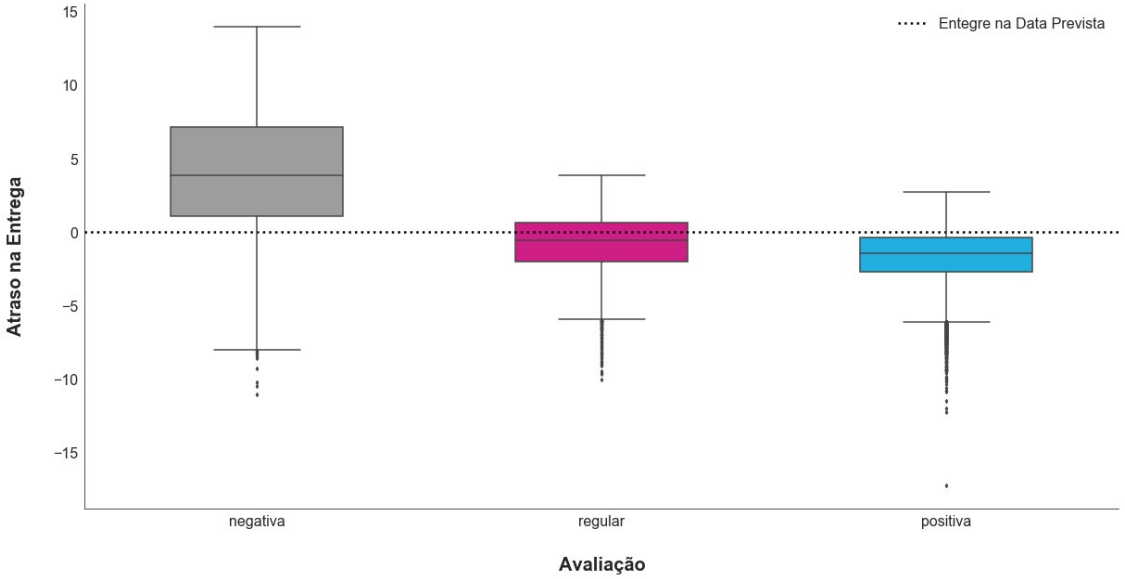

Figura 6: Diagrama de caixa da variável atraso na entrega.

A Figura 5 e a Figura 6 ilustram 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 6 que as entregas com avaliação regular apresentam a menor dispersão do ponto 0 (entrega no prazo).

2.3. 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 7 ilustra a matriz de co-ocorrência de palavras do corpus. É possível observar a presença de grupos de palavras frequentemente associadas.

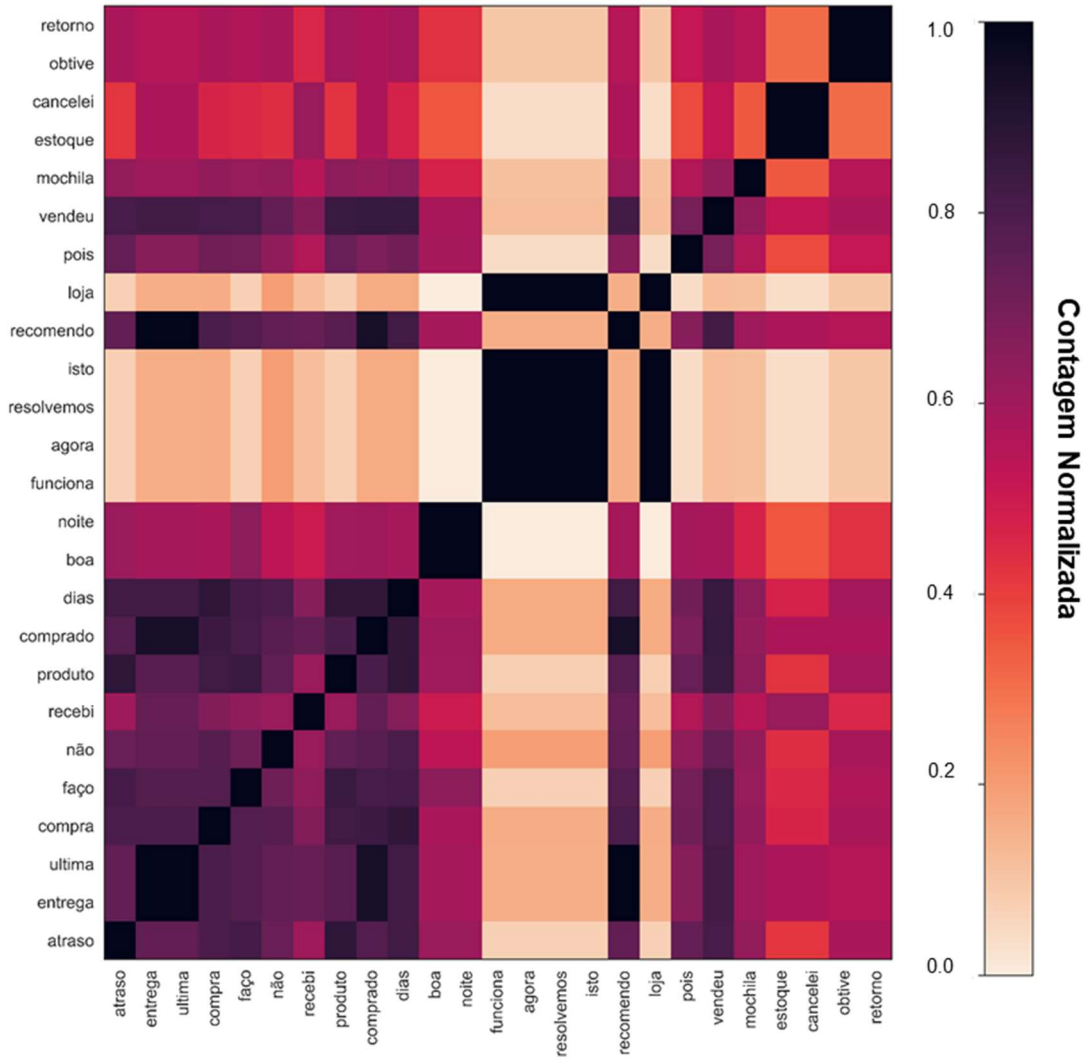

Figura 7: Matriz de co-ocorrência de vinte e cinco palavras do corpus.

A Figura 8 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).

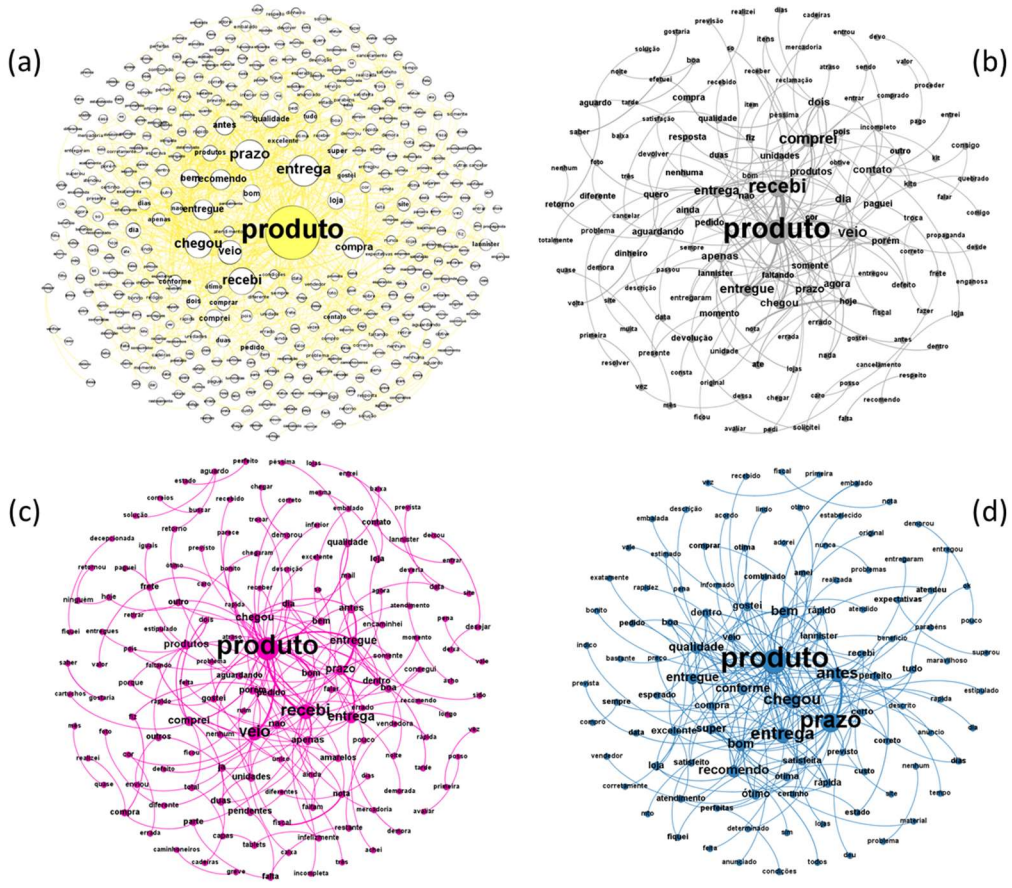

Figura 8: Associações mais frequentes de palavras. (a)Geral. (b)Avaliações negativas. (c)Avaliações regulares. (d)Avaliações positivas.

A Figura 9 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.

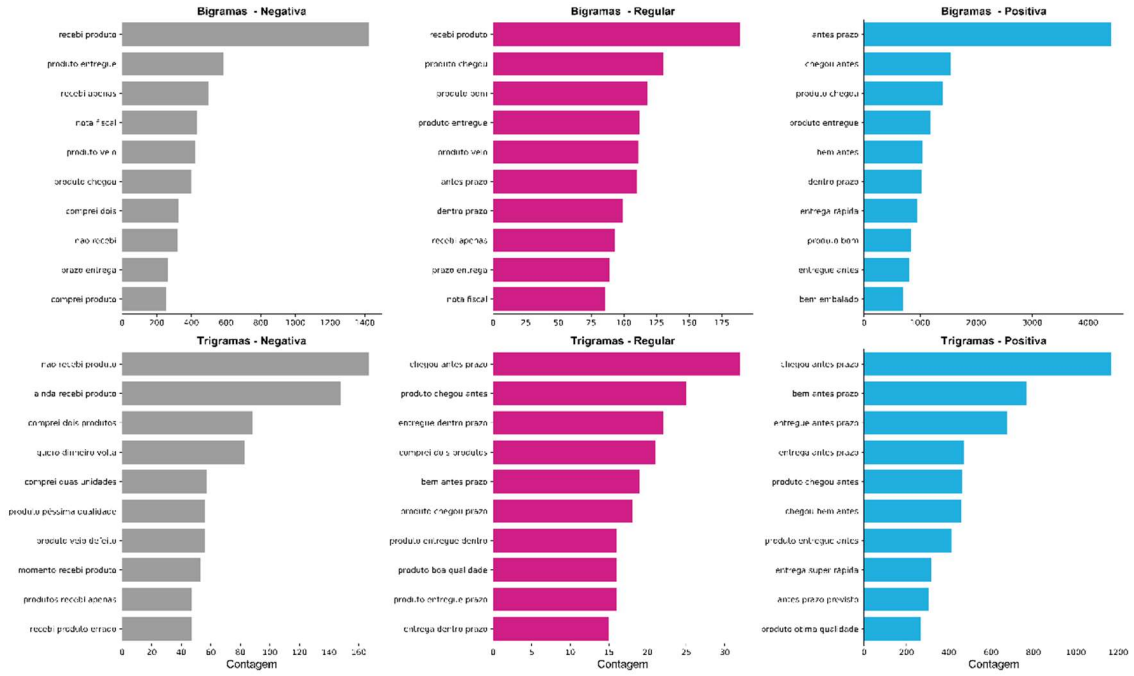

Figura 9: N-gramas mais frequentes de acordo com a avaliação dos compradores.

A Figura 10 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.

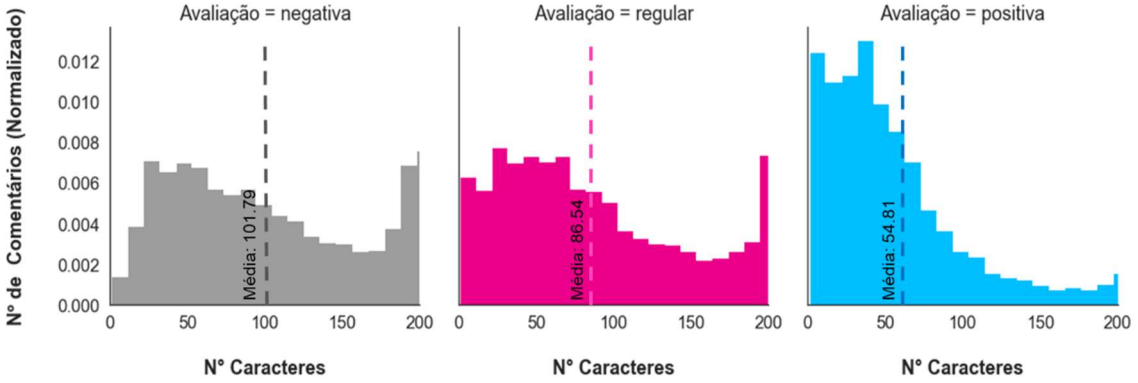

Figura 10: Distribuição de caracteres por comentários de acordo com avaliação.

A Figura 11 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.

Figura 11: Nuvem de palavras das avaliações negativas.

Observa-se na Figura 11 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.

A Figura 12 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”.

Figura 12: Nuvem de palavras das avaliações regulares.

A Figura 13 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”.

Figura 13: Nuvem de palavras das avaliações positivas.

3. METODOLOGIA

3.1. Pré-processamento

Como descrito anteriormente o conjunto de textos (corpus) utilizado neste trabalho compreende mais de 48 mil comentários feitos por compradores de lojas virtuais, totalizando 10236 termos únicos (palavras, pontuação e *emojis*). O corpus reflete a heterogeneidade da língua portuguesa e de seu constante processo de transformação ao longo do tempo. Dessa maneira antes de desenvolver qualquer modelo de classificação, é necessário caracterizar o corpus, padronizar os textos e extrair o máximo de sentido das palavras utilizadas.

A Figura 14 ilustra o número de ocorrências das 20 palavras mais frequentes no corpus.

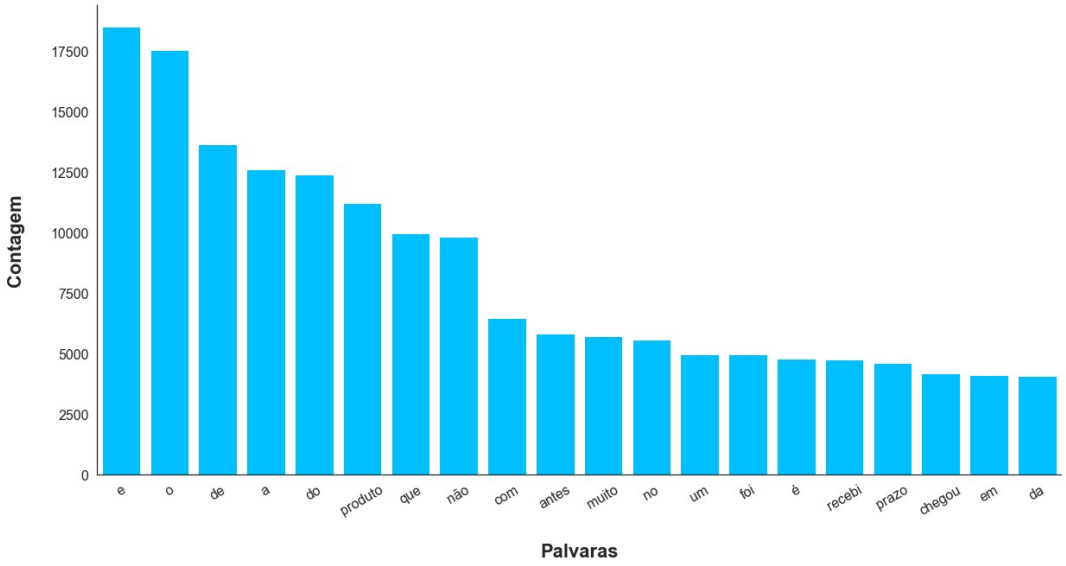

Figura 14: Vinte palavras mais frequentes no corpus.

É possível observar na Figura 14 que entre as palavras mais frequentes encontram-se artigos, preposições e verbos de ligação que, isolados, pouco contribuem no sentido do comentário. Destaca-se também a presença de palavras acentuadas que, se não tratadas, podem ser contadas de forma duplicada devido a ocorrência de palavras acentuadas de forma incorreta (um erro gramatical bem frequente no português). Para corrigir estes e outros problemas foram executados os seguintes tratamentos no corpus:

i. **Remoção de *stop words*:** processo de eliminação de termos que não contribuem na extração de sentido dos comentários (artigos e preposições, por exemplo);
ii. **Remoção de pontuação:** processo de regularização do corpus através da eliminação de toda pontuação;
iii. **Conversão de *emojis* em texto:** processo de regularização do corpus através da transformação de símbolos especiais em textos;
iv. **Remoção de acentuação:** processo de regularização do corpus através da remoção de acentos de todas as palavras;

A Figura 15 ilustra o número de ocorrências das 20 palavras mais frequentes no corpus após a primeira rodada de tratamentos.

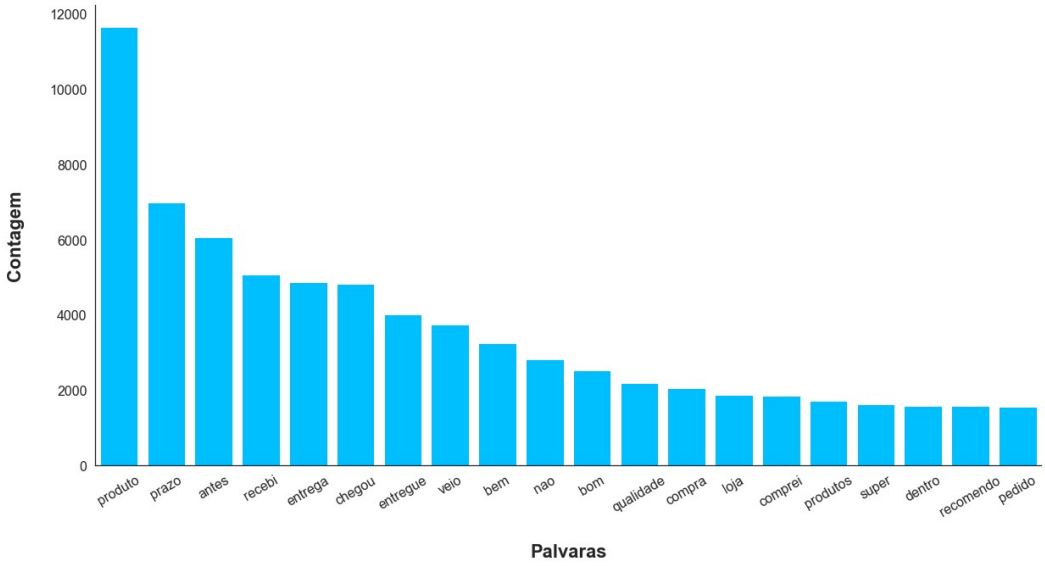

Figura 15 :Vinte palavras mais frequentes no corpus após a regularização.

Quando comparada com a Figura 14, a Figura 15 exibe um notório ganho de informação, e também o aumento na contagem de palavras que anteriormente possuíam acentos , destacando a junção das variantes da palavra “não”. A fim de evitar a inflexão de palavras, um problema evidenciado pela Figura 15 , foi adotada técnica de *Stemming*, que consiste no processo de regularização do corpus através da conservação somente do prefixo das palavras.

A Figura 16 ilustra as 20 palavras mais frequentes após aplicação da técnica de *Stemming*. É possível notar que a redução da verborragia alterou a contagem de palavras, destacando o agrupamento das inflexões da palavra “entrega”. Ao final da etapa de tratamento do conjunto de textos, o número de termos únicos foi reduzido para 8662 (aproximadamente 85% da quantidade original).

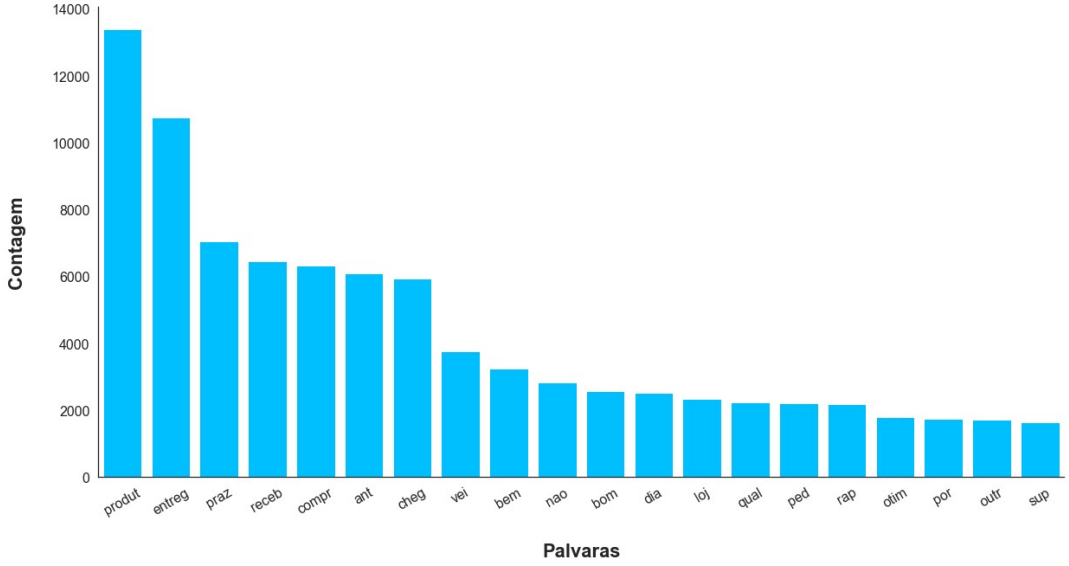

Figura 16: Vinte palavras mais frequentes no corpus após o *Stemming*.

Ainda na etapa de pré-processamento é necessário definir qual técnica de vetorização do conjunto de texto será utilizada. É comum a adoção de métodos de ponderação simples ou *Word Embedding*. Define-se como *Word Embedding* como um

conjunto de técnicas que mapeia a sintática e a semântica do corpus em um espaço vetorial através de métodos estatísticos [3].

Neste trabalho foi adotada a técnica de vetorização de *Term Frequency–Inverse Document Frequency* (TF-IDF). O TF-IDF é uma estatística quantitativa cujo objetivo é refletir a importância de uma palavra (ou uma sequência de palavras) para um documento do corpus [6] (a descrição matemática do processo será apresentada no tópico seguinte).

Combinando a vetorização 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 [7].

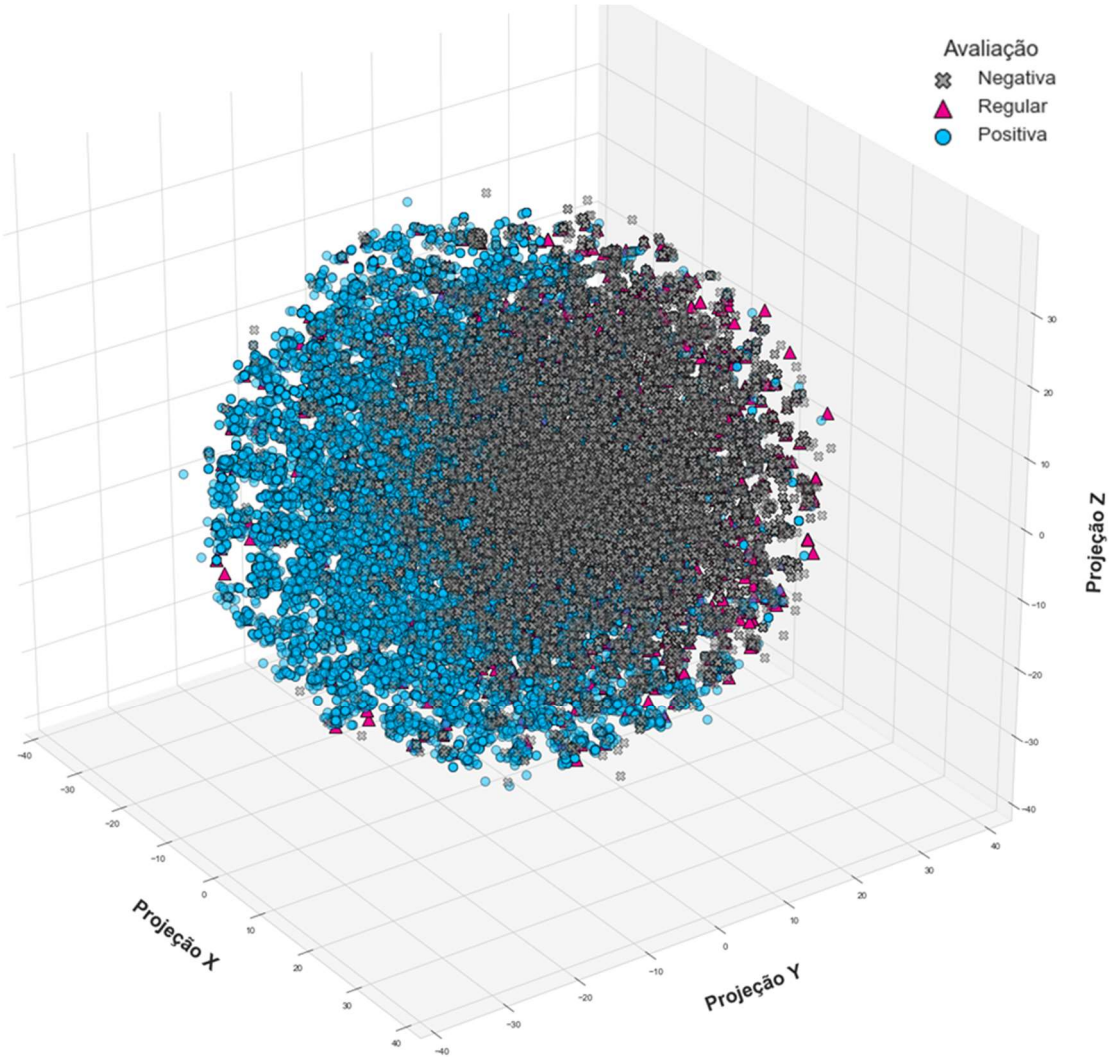

Figura 17: Projeção 3D da distribuição dos comentários.

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.

3.2. Descrição Matemática da Metodologia

Esse tópico tem como objetivo unicamente a descrição matemática da metodologia adotada. O critério utilizado na escolha dos modelos será apresentado no tópico seguinte. Independente do modelo discutido, quando apresentado um problema de otimização (maximização ou minimização), deve se assumir que o mesmo será resolvido através do método dos gradientes.

3.2.1. Term Frequency–Inverse Document Frequency

O *Term Frequency–Inverse Document Frequency* (TF-IDF) é dividido em duas etapas: (**i**) o cálculo da frequência do termo (tf) (**ii**) o cálculo da frequência inversa dos documentos (idf). A computação da frequência do termo $tf(t,d)$, usualmente, consiste na contagem da ocorrência de um determinado termo em um documento, ou seja, o número de vezes que esse termo t ocorre no documento d, de acordo com a seguinte equação:

$$tf(t,d) = f_{t,d} \quad (\mathbf{1}).$$

A frequência inversa dos documentos é uma medida de quanta informação um termo fornece, ou seja, se é comum ou raro em todos os documentos. O $idf\ smooth$, adotado neste trabalho, é calculado da seguinte forma:

$$idf\ smooth(n_t, D) = \left(\frac{N}{1+n_t}\right) + 1 \quad (\mathbf{2}),$$

onde é N o número total de documentos e n_t o número de documentos com o termo. É pertinente destacar que o TF-IDF é frequentemente criticado pela desconsideração da semântica das palavras, em contraponto aos *Word Embedding*, e pela produção de matrizes excessivamente esparsas.

Neste trabalho o TF-IDF foi aplicado em conjunto a técnica de tokenização (particionamento do texto) em bigramas. Ou seja, ao invés de procurar, quanto importante é uma palavra para o corpus, procura-se a importância de uma sentença de duas palavras. Esse método foi adotado esperando um ganho semântico na interpretação dos resultados.

3.2.2. Regressão Logística

A Regressão Logística, semelhante a regressão linear, pode ser aplicada a problemas de classificação assumindo que a variável objetivo y , é um valor discreto. O modelo de Regressão Logística utiliza uma função sigmoide de achatamento que descreve uma previsão do modelo como a probabilidade σ de uma dada uma entrada x pertencer a uma das classes de y. A função logística é dada por:

$$\sigma(x) = \frac{1}{1+e^{-\lambda \ldots x}} \quad (\mathbf{3})$$

onde λ é uma constante de ponderação. O princípio geral do classificador logístico é minimização da função erro E ao longo de um número máximo de iterações ou convergência da função. O erro é dado por:

$$E(w,b) = \frac{1}{N}\sum_{i=1}^{N} L\big(y_{(i)}, \sigma(x_i)\big) + \alpha . R(w)\big) \quad (4),$$

onde N é o número total de registros, L é a função de perda do classificador (entropia cruzada neste trabalho), onde $\sigma(x)$ é a função de probabilidade dada por $\sigma(x) = w^T . x + b = w$ (onde w é o vetor de parâmetros e b o coeficiente linear), α é uma constante não negativa e R é o termo de regularização definida pela função de penalidade .

Neste trabalho foi adotada a função de regularização "L2", descrita por:

$$R(w) = \frac{1}{2}\sum_{i=1}^{N} w_{(i)}^2 \quad (5).$$

3.2.3. Máquinas de Vetor de Suporte de Núcleo Linear (SVM-Linear)

As Máquinas de Vetores de Suporte (SVM) pertencem a uma classe de modelos de aprendizagem supervisionada de base estatística, que se diferenciam entre si pela função de núcleo (*kernel*) adotada. A função de *kernel,* $k(x_i, x_j)$, transforma o espaço n dimensional das variáveis em um espaço d dimensional (sendo $d \geq n$), que possibilita a separação das classes do problema através de um ou mais hiperplanos.

Os SVM, de forma genérica, podem ser entendidos como problemas de otimização cuja função objetivo é a maximização da margem de separação entre as classes. É natural associar uma boa separação ao hiperplano que possui a maior distância dos dados de treinamento mais próximos, pois menor é o erro do classificador. A forma primal do objetivo dos SVM é dada por:

$$\begin{array}{rl} minimizar: & \frac{1}{2}.w^T.w + C.\sum_{i=1}^{m}\zeta_{(i)} \\ w,b,\zeta & \\ sujeito\ a: & y_{(i)}.\left(w^T.w_{(i)} + b\right) \geq 1 - \zeta_{(i)} \ \ e\ \ \zeta_{(i)} \geq 0\ para\ i = 1,2,\dots,n \end{array} \quad (6),$$

onde b é o termo de polarização, w é o vetor de peso das variáveis de entrada, ζ é a variável de folga e C é um hiperparâmetro que define o limite superior da função. A forma dual dos problemas de SVM é dada por:

$$\begin{array}{rl} minimizar: & \alpha^T.y_{(i)}.y_{(j)}.k(x_{(i)}.x_{(j)}).\alpha + e^T.\alpha \\ \alpha & \\ sujeito\ a: & y^T.\alpha = 0\ e\ 0 < \alpha_{(i)} < C_{(i)} \end{array} \quad (7),$$

onde α é o é o vetor dos multiplicadores de Lagrange e e é um vetor de uns. O SVM de núcleo linear é caraterizado pela utilização da função de *kernel* $x_{(i)}.x_{(j)}^T$.Cuja solução é dada por:

$$f(x) = \sum_i \alpha_{(i)} . y_{(i)} . k\big(x_{(i)}.x_{(j)}\big) + b \quad (8).$$

3.2.4. Árvores de decisão

Os modelos de Árvores de Decisão representam as variáveis do problema de classificação como nós, cada aresta (ramificação) representa uma decisão (regra) e cada folha representa um possível resultado. Esses modelos permitem uma fácil

compressão visual em problemas de classificação envolvendo variáveis qualitativas e/ou de múltiplas classes. As árvores que crescem muito profundamente tendem a aprender padrões altamente irregulares sobreajustando (*overfitting*) seus conjuntos de treinamento, consequentemente perdendo a capacidade de generalização.

Um modelo de árvore de decisão parte de uma variável de origem, denominada raiz, que é particionada em subconjuntos sucessores. A divisão é baseada em um conjunto de regras que varia de acordo com as características dos problemas de classificação. Esse processo é repetido em cada subconjunto derivado de uma maneira sucessiva, chamada particionamento recursivo. A recursão é concluída quando o subconjunto de um nó apresenta os mesmos valores da variável objetivo ou quando a divisão não agrega mais valor às previsões.

Algoritmos utilizados na construção de Árvores de Decisão geralmente funcionam de cima para baixo, escolhendo uma variável em cada etapa que melhor divide o conjunto de dados. As três principais métricas para escolha dos nós de uma árvore de decisão são: (**i**) funções de impureza, (**ii**) ganho de informação e (**iii**) variância. Neste trabalho foi adotada a técnica de minimização de impureza Ginni para J classes, descrita por:

$$I_G(p) = 1 - \sum_{i=1}^{J} p_{(i)}^2 \qquad (\mathbf{9}),$$

onde $p_{(i)}$ é a probabilidade de um item com o rótulo i ser classificado de forma errada.

3.2.5. Floresta Aleatória

As Florestas Aleatórias são modelos de classificação que operam através do agrupamento de um grande número de Árvores de Decisão na etapa de treinamento, e devolvendo a moda estatística dos resultados obtidos pelas árvores individuais. As Florestas Aleatórias corrigem a tendência de sobreajuste que as Árvores de Decisão apresentam. Isso ocorre às custas de um pequeno aumento no viés e perda da capacidade de interpretação visual, mas geralmente aumentando muito o desempenho no modelo final.

Uma floresta pode ser entendida coma a união de esforços de Árvores de Decisão que isoladas seriam mal sucedidas, mas juntas produzem bons resultados. Embora não sejam muito semelhantes, as florestas fornecem os efeitos de uma Validação Cruzada *k-fold*. O algoritmo de treinamento para Florestas Aleatórias aplica a técnica de agregação de *bootstrap*, também conhicda como *bagging*. Dado um conjunto de treinamento X com respostas y , o *bagging* é aplicado B vezes selecionando uma amostra aleatória que substitui o conjunto de treinamento e ajusta as árvores a esta amostra. Após o treinamento, as previsões para amostras desconhecidas x_u são realizadas através do cálculo da moda estatística das previsões de todas as árvores individuais que formam a floresta. Assim como tópico anterior deste trabalho, optou-se pela técnica de minimização da impureza Ginni (Equação 9) paras Árvores de Decisão que formam a Floresta Aleatória.

3.2.6. Métricas de Avalição

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

$$A = \frac{TP+TN}{TT} \quad (\mathbf{10}),$$

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 é uma métrica mais adequada para esse tipo de situação, pois indica a proporção de classes que foi identificada corretamente. A revocação R é dada por:

$$R = \frac{TP}{TP+FN} \quad (\mathbf{11}).$$

onde FN é o número de elementos de uma classe que foram classificados como pertencentes a outras. A precisão, que indica a proporção de previsões está realmente correta, é dada por :

$$P = \frac{TP}{TP+FP} \quad (\mathbf{11}).$$

onde FP é o número de elementos classificados incorretamente como pertencentes a uma classe específica, mas na verdade pertencem a outras. A métrica $F1$, que indica o nível de significância da acurácia, é dada por:

$$F1 = 2.\left(\frac{P.R}{P+R}\right) \quad (\mathbf{12})$$

3.3. Descrição do Procedimento Experimental

O procedimento experimental adotado neste trabalho, em uma perspectiva geral, segue as seguintes etapas:

i. **Particionamento de dados:** divisão dos dados que serão utilizados para treino, teste e avaliação dos modelos pretendidos;
ii. **Escolha dos modelos:** baseado no desempenho dos modelos na sua configuração original (não otimizada), define-se quais melhores se adaptam ao conjunto de treino e teste;
iii. **Otimização dos modelos escolhidos:** uma vez definidos quais modelos serão estudados, é ideal obter os seus respectivos melhores desempenhos;
iv. **Avaliação dos modelos:** por fim é necessário definir as qualidades e limitações dos modelos escolhidos, baseado no resultado de diferentes métricas de avaliação no conjunto de dados de validação.

3.3.1. Particionamento dos dados

O banco de dados foi dividido em três partes, como a Figura 18 ilustra. Esta configuração foi adotada com o propósito de simular a situação real de desenvolvimento

de um modelo, através dos dados de treino e teste, e aplicação em dados nunca antes observados. Em todos os grupos ocorre o desbalanceamento de classes (mais acentuado nos dados de validação).

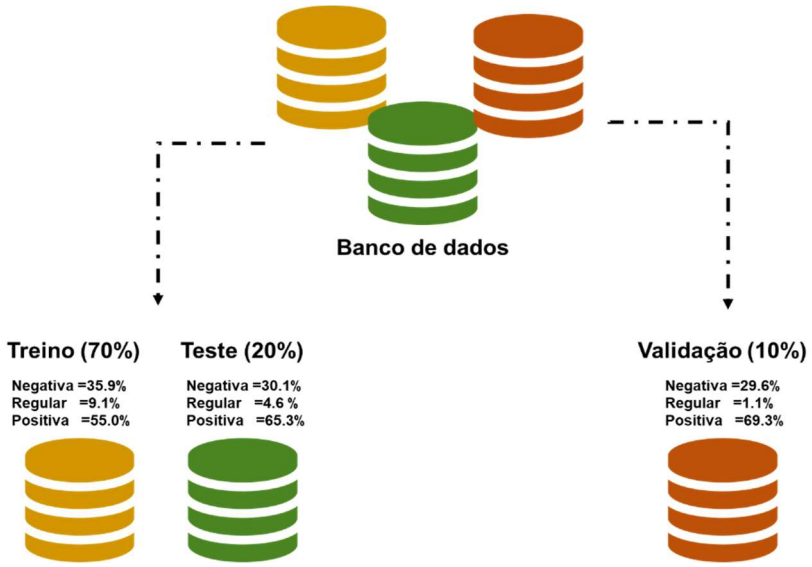

Figura 18: Particionamento do banco de dados.

3.3.2. Validação cruzada e a escolha dos modelos de classificação

Existe um grande número de modelos que podem ser utilizados em problemas de classificação, como o apresentado neste trabalho. Dessa forma, como critério de escolha dos modelos de estudo (2 lineares e 2 não lineares), é utilizada a acurácia média e o desvio padrão de uma série de classificadores após o processo de validação cruzada em 10 *folds* do conjunto de dados treino ∪ teste. A Tabela 3 apresenta os resultados obtidos pelos diferentes modelos avaliados.

Tabela 3: Acurácia média pós validação cruzada.

Modelo	Tipo	Acurácia Média	Desv. Pad.
Floresta Aleatória	Não Linear	86.48%	± 0.68%
Regressão Logística	Linear	84.43%	± 0.56%
SVM [a] - Linear	Linear	84.12%	± 0.57%
Árvore de decisão	Não Linear	81.74%	± 0.83%
SVM [a] – Polinomial (grau=3)	Não Linear	80.80%	± 0.21%
SVM [a] - RBF	Não Linear	80.76%	± 0.50%
Bayesiano Simples (*Naive*)	Pode Variar	80.35%	± 0.97%
K-Vizinhos Próximos	Não Linear	73.53%	± 0.62%

a-Máquinas de Vetor de Suporte

Baseado nos resultados apresentados na Tabela 3, foram adotados os modelos de Máquinas de Vetor de Suporte de núcleo linear (SVM-Linear) e a Regressão Logística, como os classificadores lineares deste estudo .E como classificadores não lineares foram adotados os modelos de Árvores de Decisão e Floresta Aleatória.

3.3.3. Otimização de Hiperparâmetros

Uma vez determinados quais modelos serão utilizados, é ideal obter o melhor desempenho possível de cada um. Dessa forma, é adotada técnica de otimização de hiperparâmetros de busca em grid (busca exaustiva), tomando como função objetivo a maximização da acurácia, em todos os modelos selecionados.

A otimização de hiperparâmetros foi realizada tomando como base o conjunto de dados de treino. Os modelos e seus respectivos hiperparâmetros otimizados foram:

i. **Regressão Logística:** foram testados valores de C, parâmetro relacionado a λ na Equação 3, entre 10^{-3} e 1 (1, por padrão no *Scikit-Learn*). Também foram experimentados valores entre 10^2 e 10^4 , para o número máximo de interações do algoritmo (500, por padrão no *Scikit-Learn*);
ii. **SVM-Linear:** foram experimentados valores entre 10^{-3} a 10^3 para o parâmetro C da Equação 6 (1, por padrão no *Scikit-Learn*);
iii. **Árvore de Decisão:** foram experimentados valores de profundidades de árvore entre 2 e 16 (3, por padrão no *Scikit-Learn*);
iv. **Floresta Aleatória:** foram experimentadas florestas com número de árvores entre 10^2 e 10^4 (10^2 , por padrão no *Scikit-Learn*), e profundidades de árvores individuais entre 2 e 8.

4. RESULTADOS E DISCUSSÃO

4.1. Impactos da Otimização

A Tabela 4 apresenta os resultados da acurácia antes e depois da otimização dos hiperparâmetros dos modelos. Observa-se que os classificadores obtiveram um desempenho maior na etapa de treino do que na etapa de teste. Também é possível observar, nos modelos otimizados, o aumento da acurácia no conjunto de treino em detrimento de uma redução da acurácia de teste, caracterizando uma perda de capacidade de generalização.

Tabela 4: Desempenho após otimização de hiperparâmetros.

Modelo	Original		Otimizado	
	Acurácia Treino	Acurácia Teste	Acurácia Treino	Acurácia Teste
Floresta Aleatória	98.17%	86.99%	99.56%(+)	86.42%(-)
Regressão Logística	86.43%	84.44%	87.25%(+)	83.80%(-)
SVM - Linear	87.15%	84.04%	87.98%(+)	83.13%(-)
Árvore de decisão	98.18%	82.73%	99.72%(+)	82.01%(-)

Baseado nos resultados é possível notar o aumento do sobreajuste (*overfitting*) pós otimização. Dessa forma, optou-se por seguir os estudos utilizando a configuração de parâmetros padrão dos modelos, a fim de manter o maior grau de generalização obtido.

4.2. Resultados dos Modelos Lineares

A Regressão Logística e o SVM-Linear obtiveram acurácias no conjunto de dados de validação de 82.87% e 80.67%, respectivamente (ambas abaixo da acurácia média de

validação cruzada). A Figura 19 ilustra a matriz de confusão dos classificadores quando aplicados no conjunto de dados de validação.

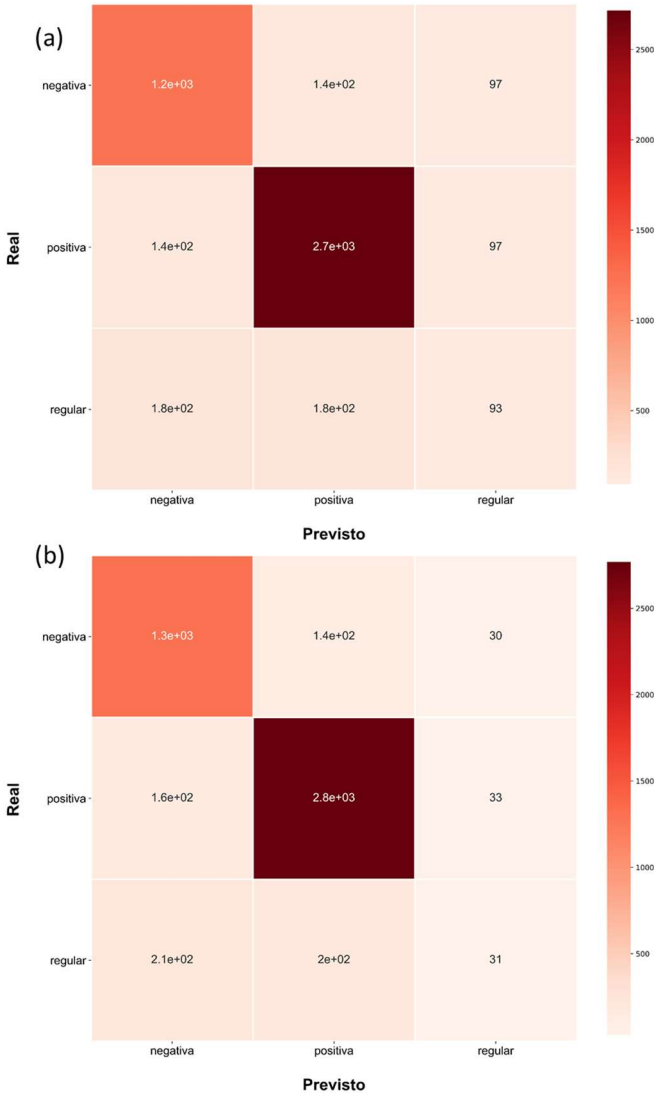

Figura 19: Matrizes de confusão. (a)Regressão Logística. (b)SVM-Linear

Como esperado em problemas de classificação desbalanceados, foram observados mais erros de predição na classe menos frequente. Vale ressaltar que independente da proporção de registros, existe uma dificuldade intrínseca em separar a avaliação regular das demais, como a caracterização do corpus indica (sendo uma tarefa complexa até para um ser humano). Isso também vale para os demais classificadores não lineares.

A Tabela 5 apresenta as métricas dos classificadores lineares quando aplicados aos dados de validação. Observa-se que a Regressão Logística obteve o melhor desempenho em todas as métricas testadas. Destaca-se que as classes de avaliação positiva e negativa, mesmo desbalanceadas entre si, alcançaram taxas de Revocação (*Recall*) altas.

Tabela 5:Métricas de avalição dos modelos lineares.

Classe	Regressão Logística			SVM-Linear		
	Revocação	Precisão	F1	Revocação	Precisão	F1
negativa	83.56%	79.22%	81.33%	78.90%	77.37%	78.13%
regular	20.95%	32.40%	25.44%	18.69%	27.76%	22.34%
positiva	91.82%	89.49%	90.64%	90.74%	87.32%	89.00%

A Figura 20 ilustra as curvas de Característica de Operação do Receptor (ROC) dos classificadores lineares testados. Observa-se que, como esperado, a tarefa de classificação da avaliação regular é que obtém o pior desempenho, sendo a mais próxima de uma escolha aleatória.

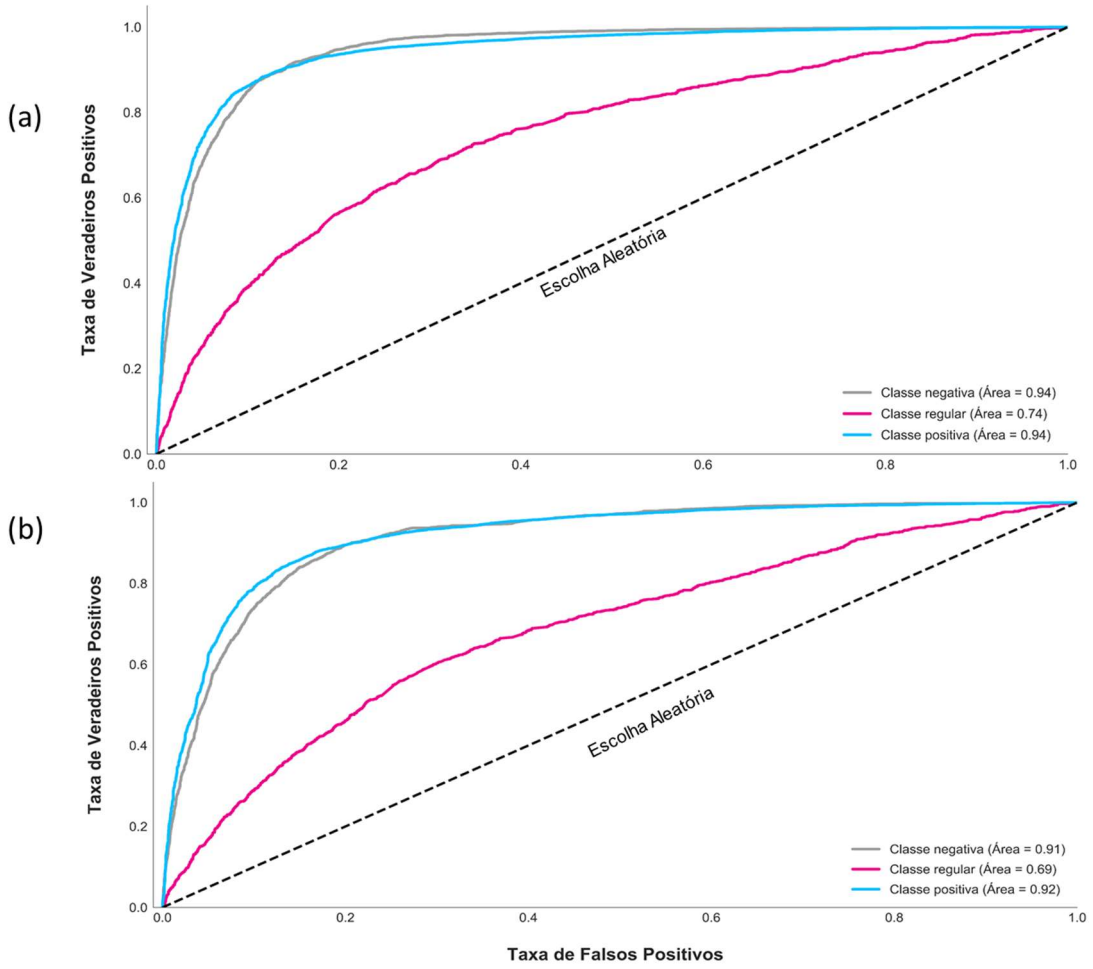

Figura 20: Curvas ROC dos classificadores lineares. (a)Regressão Logística. (b)SVM-Linear.

Observa-se na Figura 20 que as áreas sob as curvas ROC da Regressão Logísticas denotam um melhor desempenho, quando comparado ao SVM-Linear. Dessa forma baseado nas métricas apresentadas é possível afirmar que a Regressão Logística é o classificador linear que mais se ajusta ao problema proposto.

A Figura 21 ilustra os bigramas com a maior relevância para determinação de pertencimento (em verde) ou não pertencimento (em vermelho) de um comentário a uma das classes de avaliação. As palavras “excelente”, “perfeito” e “rápido” são extremamente influentes na determinação do não pertencimento a classe negativa, em ambos os classificadores.

(a)

Negativa		Regular		Positiva	
Peso	Bigrama	Peso	Bigrama	Peso	Bigrama
+2.489	nao recom	+2.313	perfeit	+2.582	dia outr
+2.352	pres dat	+2.275	parab	+1.603	produt inferi
+2.260	pess	+2.092	surpreend	+1.547	pad est
+2.130	soluca	+2.002	cheg problem	+1.543	talv
+1.987	desd dia	+1.916	ant	+1.441	vei incomplet
+1.947	resolv questa	+1.841	falt ped	+1.430	pont
+1.873	insatisfeit	+1.837	barr	+1.426	dev entreg
+1.795	impossi	+1.816	maravilh	+1.403	contrari
+1.782	compr nad	+1.720	nenhum problem	+1.402	cobr fret
+1.768	falt respeit	+1.712	cheg conform	+1.381	unidad nao
+1.732	envelop	+1.681	rap	+1.375	fot parec
+1.727	de unidad	+1.662	contat lannist	+1.361	no dia
+1.709	nao gost			+1.324	dia difer
+1.692	der			+1.306	no moment
+1.684	efetu troc	-1.663	inferi	+1.302	perc
+1.673	apen unidad	-1.737	falt respeit	+1.293	ult dia
		-1.755	falsific		
		-1.804	providenc		
-1.706	nenhum problem	-1.816	insatisfeit	-1.294	procon
-1.728	agradec	-1.828	pres dat	-1.314	der
-1.732	efici	-1.836	cancel produt	-1.331	surpreend
-1.854	produt excel	-1.916	dat receb	-1.365	dinh
-1.938	dia outr	-1.964	boa tard	-1.409	prejuiz
-1.953	satisfeit entreg	-1.967	desd dia	-1.417	unidad falt
-1.968	parab	-2.006	decepca	-1.420	falt ped
-2.112	perfeit	-2.327	nao recom	-1.431	cinz
-2.313	rap	-2.374	pess	-1.788	tod produt

(b)

Negativa		Regular		Positiva	
Peso	Bigrama	Peso	Bigrama	Peso	Bigrama
+3.495	pess	+3.629	perfeit	+2.494	dia outr
+3.186	nao recom	+3.408	parab	+2.063	por
+2.421	nao gost	+3.352	rap	+1.861	pont
+2.251	insatisfeit	+3.341	excel	+1.749	produt inferi
+2.200	dinh	+3.299	ador	+1.624	talv
+2.099	nao	+3.295	ant	+1.392	fal
+2.095	soluca	+2.799	satisfeit	+1.346	contrari
+2.019	horri	+2.797	lind	+1.329	hp
+1.947	der	+2.620	otim	+1.291	entreg parc
		+2.570	ame	+1.267	cobr fret
		+2.397	maravilh	+1.260	vei incomplet
-1.953	tud	+2.269	bem embal	+1.260	kg
-2.048	parab	+2.233	otim produt	+1.235	dev entreg
-2.051	ame	+2.099	recom	+1.228	part
-2.074	satisfeit	+2.084	surpreend	+1.227	colcha
-2.095	bom			+1.218	vei cor
-2.128	obrig			+1.211	pouc
-2.168	boa	-2.132	nao gost	+1.210	encaminh
-2.215	cert	-2.167	decepcion		
-2.266	ador	-2.218	gost sab		
-2.376	recom	-2.222	receb apen	-1.225	dinh
-2.446	lind	-2.270	ruim	-1.268	maravilh
-2.583	excel	-2.367	inferi	-1.278	otim produt
-2.726	ant	-2.450	nao recom	-1.304	sup
-2.851	perfeit	-2.503	unidad	-1.360	parab
-3.078	otim	-2.717	nao	-1.392	satisfaca
-3.621	rap	-3.055	pess	-1.710	sempr

Figura 21: Bigramas mais influentes nos modelos. (a)Regressão Logística. (b) SVM-Linear.

4.3. Resultados dos Modelos Não Lineares

A Árvore de Decisão e a Floresta Aleatória obtiveram acurácias no conjunto de dados de validação de 83.26% e 87.05%, respectivamente (ambas acima da acurácia média de validação cruzada). A Figura 22ilustra a matriz de confusão dos classificadores quando aplicados no conjunto de dados de validação.

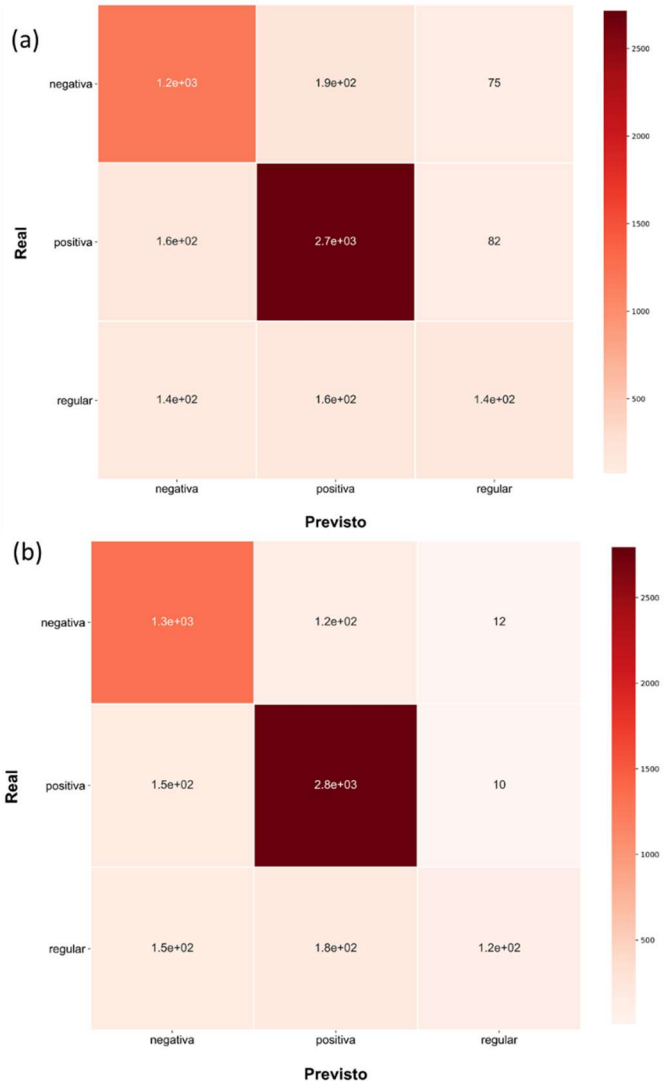

Figura 22 : Matrizes de confusão. (a)Árvore de Decisão. (b)Floresta Aleatória

A Tabela 6 apresenta as métricas dos classificadores não lineares quando aplicados aos dados de validação. Observa-se que a Floresta Aleatória obteve o melhor desempenho em todas as métricas testadas. Destaca-se, novamente, que as classes de avaliação positiva e negativa, mesmo desbalanceadas entre si, alcançaram taxas de revocação (*Recall*) altas.

Tabela 6: Métricas de avalição dos modelos não lineares.

Classe	Árvore de Decisão			Floresta Aleatória		
	Revocação	Precisão	F1	Revocação	Precisão	F1
negativa	81.64%	80.05%	80.84%	90.89%	81.21%	85.78%
regular	32.43%	47.84%	38.66%	25.90%	83.94%	39.59%
positiva	91.79%	88.38%	90.05%	94.46%	90.39%	92.38%

A Figura 20 ilustra as curvas ROC dos classificadores não lineares testados. Observa-se que as áreas sob as curvas ROC da Floresta Aleatória denotam um melhor desempenho, quando comparado as Árvores de Decisão. Dessa forma, baseado nas métricas apresentadas é possível afirmar que a Floresta Aleatória é o classificador não linear que mais se ajusta ao problema proposto.

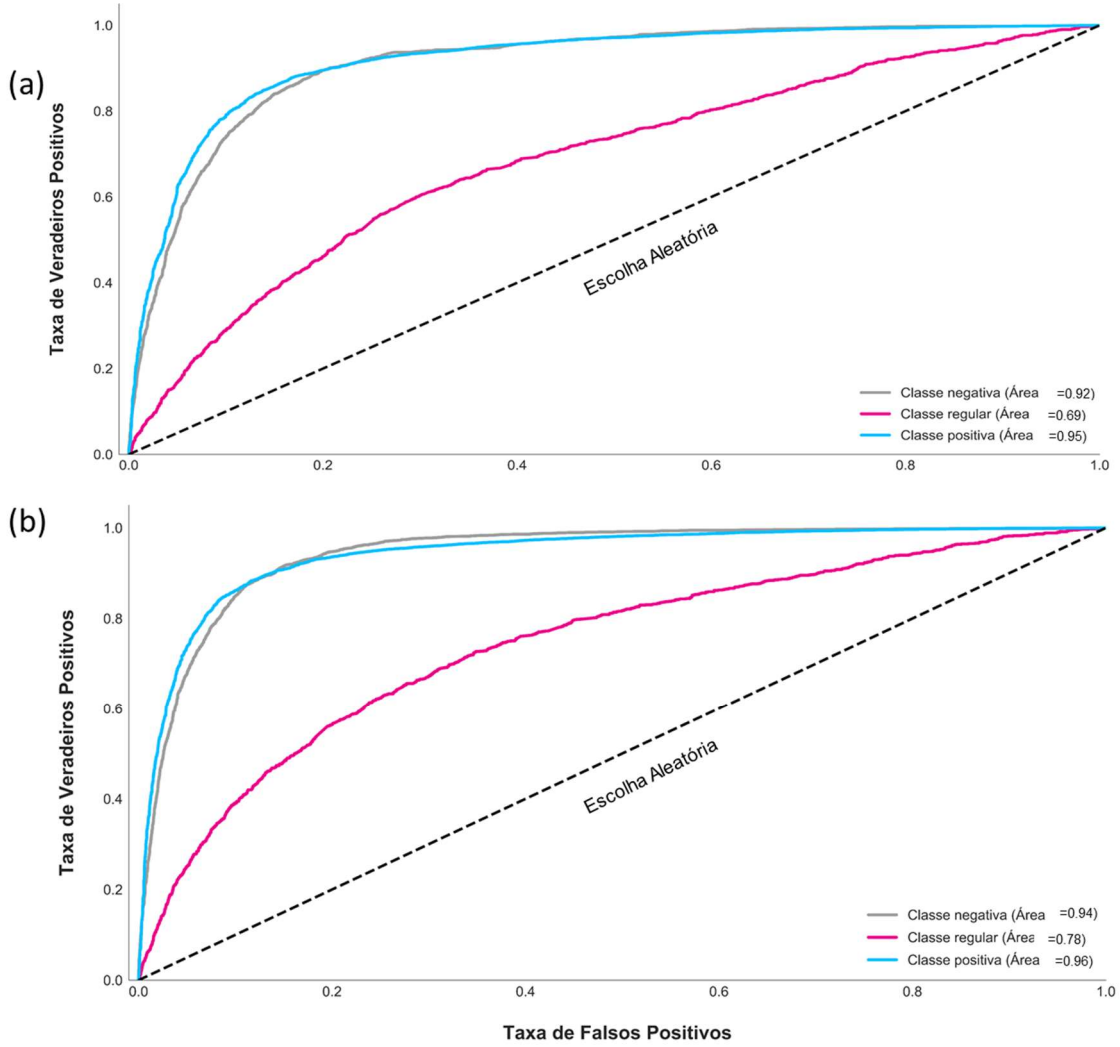

Figura 23: Curvas ROC dos classificadores lineares. (a)Regressão Logística. (b)SVM-Linear.

4.4. Comparação de Resultados

Em geral os classificadores não lineares obtiveram resultados superiores quando comparados aos lineares. Baseado nos resultados anteriores é possível afirmar que, dentre todos os modelos experimentados, a Floresta Aleatória é o que melhor se ajusta ao problema de classificação proposto. A Tabela 7 apresenta os resultados ordenados das métricas avaliadas.

Tabela 7: Resultados ordenados.

Modelo	Revocação (Média)	Precisão (Média)	F1 (Média)	Acurácia (Validação)
Floresta Aleatória	70.42%	85.18%	72.58%	87.13%
Árvore de Decisão	68.62%	72.09%	69.85%	83.26%
Regressão Logística	65.44%	67.04%	65.80%	82.87%
SVM-Linear	62.78%	64.15%	63.16%	80.67%

Sozinho, o desbalanceamento dos dados não justifica a alta taxa de erros dos classificadores em relação a avaliação regular, pois a classe negativa mesmo possuindo menos registros que a classe positiva, obteve altas taxas de revocação. Dessa forma, é importante considerar a dificuldade intrínseca em identificar um comentário regular. Portanto, mesmo aplicando técnicas de balanceamento de registros como a geração de dados sintéticos (SMOTE), poucos ganhos seriam observados.

5. CONCLUSÃO

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.

De forma geral os classificadores não lineares obtiveram desempenho superior aos lineares, em especial o modelo de Floresta Aleatória. Contudo, todos os classificadores apresentaram problemas semelhantes na identificação da classe regular. Tal fato se deu, principalmente, pela dificuldade de identificar termos que separem estas avaliações das demais, uma tarefa relativamente complexa até para um ser humano.

É importante destacar que esse trabalho não possui uma classe de interesse, ou seja, a extração de informação através da análise de sentimentos 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 d reclamações).

A temática escolhida também possibilita os seguintes desenvolvimentos futuros:

i. Avaliação de diferentes técnicas de *Word Embedding* e seus impactos em diferentes modelos de classificação;

ii. Análise de agrupamento dentro das diferentes classes de avaliações, a fim de descobrir suas motivações (em especial das avaliações negativas);

iii. Análise de geo-agrupamento, a fim de entender a influência do fator geográfico no processo de avaliação do comprador;

iv. Estudo da associação de palavras (ou expressões) através de grafos.

6. 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] S. Das, R. K. Behera, M. Kumar, e S. K. Rath, "Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction", *Procedia Comput. Sci.*, vol. 132, nº Iccids, p. 956–964, 2018.

[3] M. C. Hermínio, "Estudo Comparativo dos Métodos de Word Embedding na Análise de Sentimentos", Universidade Federal de Pernambuco, 2018.

[4] Olist Store, "Brazilian E-Commerce Public Dataset by Olist", 2018. [Online]. Available at: https://www.kaggle.com/olistbr/brazilian-ecommerce. [Acessado: 03-ago-2019].

[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] V. Kotu e B. Deshpande, "Chapter 9 – Text Mining", in *Predictive Analytics and Data Mining*, 1º ed, Elsevier Inc., 2015, p. 275–303.

[7] 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.