**Comparação de abordagens de modelagem de tópicos em reclamações relacionadas ao setor de comércio eletrônico**

Matheus Ribeiro Cerqueira¹\*;Mauricio Evandro Eloy2

1 Universidade de São Paulo. Bacharel em Ciências Biomédicas. Rua Louro Rosa – Itaim Paulista; 08141-670 São Paulo, São Paulo, Brasil

2 Escola Superior de Agricultura “Luiz de Queiroz” (USP-ESALQ). Mestre em Matemática. Av. Pádua Dias, 235,

13.418-900 Piracicaba, São Paulo, Brasil

\*autor correspondente: mattheus.ribeiro000@gmail.com

**Comparação de abordagens de modelagem de tópicos em reclamações relacionadas ao setor de comércio eletrônico**

**Resumo**

O setor eletrônico enfrenta um grande desafio em lidar com um grande volume de reclamações dos consumidores, que podem ter diversas origens, desde problemas como prazo de entrega, até problemas de cancelamento de serviços. Essas reclamações podem impactar negativamente a reputação de uma empresa e afetar sua capacidade de reter e atrair novos clientes. A modelagem de tópicos torna-se fundamental neste contexto porque pode ajudar empresas a compreender e categorizar melhor as reclamações dos consumidores de forma automatizada. Neste contexto, este projeto teve como objetivo comparar e avaliar o desempenho de três algoritmos populares de modelagem de tópicos: LDA, BERTopic e LSI, na extração de tópicos de reclamações de clientes no site Reclame Aqui e a identificação de tópicos relevantes. O projeto utilizou um conjunto de dados de aproximadamente doze mil reclamações de clientes relacionados a produtos e serviços das principais plataformas de comércio eletrônico do Brasil. Os dados foram pré-processados e utilizados no treinamento e otimização de modelos de modelagem de tópicos e o desempenho avaliado com base no valor médio de coerência. Os resultados da avaliação mostraram que os modelos LDA e BERTopic foram capazes de extrair tópicos informativos dos dados. Em termos de pontuação, o BERTopic apresentou o melhor desempenho entre os modelos utilizando o valor médio de coerência como métrica de comparação. As descobertas deste projeto sugerem que as empresas que procuram obter informações valiosas relacionados a reclamações de clientes podem se beneficiar do uso de algoritmos de modelagem de tópicos, como LDA e BERTopic, para entender melhor as preocupações de seus clientes e, portanto, tomar ações baseadas em dados para melhorar a satisfação do cliente.

**Palavras-chave: Transformadores; Alocação de Dirichlet Latente; Indexação Semântica Latente; BERTopic; Indexação Semântica Latente**

**Comparison of topic modeling approaches in complaints related to the e-commerce industry**

**Abstract**

The electronics sector faces a great challenge in dealing with a large volume of consumer complaints, which can have various origins, from issues such as delivery time to problems with service cancellations. These complaints can negatively impact a company's reputation and affect its ability to retain and attract new customers. Topic modeling becomes fundamental in this context because it can help companies better understand and categorize consumer complaints in an automated way. In this context, this project aimed to compare and evaluate the performance of three popular topic modeling algorithms: LDA, BERTopic, and LSI, in extracting topics from customer complaints on the Brazilian website "Reclame Aqui" and identifying relevant topics. The project used a dataset of approximately twelve thousand customer complaints related to products and services from the major e-commerce platforms in Brazil. The data was preprocessed and used in training and optimizing topic modeling models, and the performance was evaluated based on the average coherence value. The evaluation results showed that the LDA and BERTopic models were able to extract informative topics from the data. In terms of score, BERTopic had the best performance among the models using the average coherence value as a comparison metric. The findings of this project suggest that companies seeking to obtain valuable information related to customer complaints can benefit from the use of topic modeling algorithms, such as LDA and BERTopic, to better understand their customers' concerns and take data-driven actions to improve customer satisfaction.

**Keywords** ou **Palabras Clave: Transformers; Latent Dirichlet Allocation; Latent Semantic Indexing; BERTopic; Latent Semantic Indexing**

**Introdução**

Na relação entre empresa e consumidor, as reclamações exercem um papel fundamental na manutenção de um vínculo próspero e duradouro. Elas surgem quando as necessidades e expectativas do consumidor não são plenamente satisfeitas, ou mesmo quando divergem do esperado.

Essas demandas podem variar desde defeitos no produto e atendimento insatisfatório, até problemas relacionados a preços, cobranças e prazos de entrega. O impacto dessas insatisfações pode variar desde o descontentamento do cliente até situações mais graves, como a deterioração da reputação da empresa e prejuízos financeiros significativos (Luo, X. 2009; Claro et al., 2014).

Claro et al. (2014) apresenta uma discussão de que as reclamações dos clientes podem ter um efeito não linear sobre o valor de mercado de uma empresa e que empresas determinadas a evitar reclamações de consumidores a todo custo podem experimentar tanto consequências positivas quanto negativas. Segundo os autores, após um limiar o volume de reclamações pode levar a redução acentuada no valor de mercado de uma empresa.

Ainda segundo Claro et al. (2014) as empresas precisam encontrar um equilíbrio entre abordar as reclamações dos clientes e evitar ser excessivamente atentas para manter seu valor de mercado e reputação, ou seja, otimizar o uso de recursos que dispõe na resolução das reclamações somente quando necessário.

De acordo com Filip A. (2013) as reclamações devem ser vistas como um indicador de avaliação do desempenho organizacional, apontando problemas ou falhas nos processos internos que precisam ser corrigidos rapidamente para evitar a perda de clientes rentáveis.

O autor afirma que aqueles que apresentam reclamações estão dando às empresas a chance de melhorar seus processos operacionais, aprender com situações negativas e, assim, promover de forma positiva a satisfação e confiança dos clientes.

Segundo Filip A. (2013) a reação rápida de uma organização e o tratamento das situações negativas envolvendo reclamações podem garantir a continuidade do relacionamento empresa-consumidor em 95% dos casos.

Bastani et al. (2019) evidenciaram que a utilização de modelagem de tópicos configura um procedimento mais eficaz no que tange à revisão e análise das narrativas veiculadas em reclamações de consumidores.

A modelagem de tópicos do inglês “Topic Modeling” é uma técnica não supervisionada aplicada na análise de coleções de documentos para extrair estruturas ocultas a partir de uma coleção de documentos em uma coleção de tópicos (Barde e Bainwad, 2017).

Ainda, segundo Bastani et al. (2019) a revisão manual de um grande volume de documentos por seres humanos é inviável em razão do tempo exigido e da suscetibilidade destes profissionais a tomadas de decisão desprovidas de imparcialidade.

Outros trabalhos anteriores demonstram a capacidade da modelagem de tópicos na resolução de diferentes tipos de problemas como a automação de revisão literária Asmussen e Møller (2019), desenvolvimento de sistemas de recomendação Pennacchiotti e Gurumurthy (2011) e na produção de relatórios de segurança de aviação (Robinson, 2019).

No contexto de reclamações, a modelagem de tópicos pode ser usada para identificar automaticamente os problemas pelos quais os clientes estão reclamando, permitindo que as empresas possam adotar medidas focalizadas para abordar esses problemas (Bastani et al., 2019).

O LSI, desenvolvido por Deerwester et al. (1990), é um modelo não probabilístico que usa uma matriz termo-documento para representar o conteúdo dos documentos.

Já o LDA, desenvolvida por Blei et al. (2003), é um modelo probabilístico que assume que cada documento é uma mistura de tópicos, e cada tópico é uma mistura de palavras. A característica probabilística do LDA faz deste algoritmo uma técnica mais vantajosa do que o LSI principalmente na separação de palavras em tópicos distintos utilizando métodos estatísticos.

Por fim, o BERTopic ou Representações de Codificador Bidirecional de Transformadores para Tópicos é uma técnica que aplica transformadores pré-treinados para gera representações vetoriais densas do inglês “document embeddings” para cada documento (Grootendorst, 2022).

Os transformadores são um tipo de arquitetura de aprendizagem profunda usada para tarefas de processamento de linguagem natural como classificação, tradução e modelagem de tópicos.

Este tipo de arquitetura foi desenvolvida por Vaswani et al. (2017) com o objetivo de substituir o uso de redes neurais recorrentes [RNRs] por mecanismos de auto-atenção devidos às limitações da redes recorrentes como o gradiente de fuga e capacidade de capturar informações de sequências longas.

Os métodos de modelagem de tópicos, ao contrário das abordagens de clusterização como o K-Means e o HDBSCAN, que agrupam documentos com base em métricas de distância, têm como propósito identificar os tópicos temáticos latentes interpretáveis presentes no texto. Esses tópicos capturam as relações das palavras em um contexto específico por meio de uma modelagem probabilística das distribuições de palavras (Lossio-Ventura et al., 2021).

Neste contexto, este projeto tem como finalidade responder às seguintes questões: Quais os principais motivos/temas das reclamações? e Quais métodos são capazes de extrair estes temas em forma de tópicos claros, distintos e interpretáveis?

Portanto, este trabalho tem como objetivo comparar os métodos LDA, LSI e BERTopic aplicados a reclamações do site Reclame Aqui relacionadas ao setor de comércio eletrônico obtidas por meio de técnicas de “web scraping” e a análise do dados gerados, assim como, a comparação dos modelos.

**Material e Métodos**

Utilizando como critérios a presença da empresa no site Reclame Aqui, atuação no comércio eletrônico e volume de publicações mencionando a empresa este projeto realizou a coleta de aproximadamente 12,000 reclamações por meio de técnicas de “web scraping”com as bibliotecas Selenium e BeautifulSoup ambas disponíveis na linguagem Python.

Os dados coletados foram pré-processados com o uso da biblioteca do kit de ferramentas de linguagem natural [NLTK, em inglês] (http://www.nltk.org/) para prepará-los para análise. O primeiro passo foi converter as palavras no minúsculo e transformar os dados em palavras individuais, conhecidas como tokens.

Sequencialmente, espaços em branco, endereços de internet, pontuações, números, datas, palavras frequentes (Figura 1) e tokens com menos de quatro caracteres foram removidos, a fim de otimizar a qualidade dos dados.

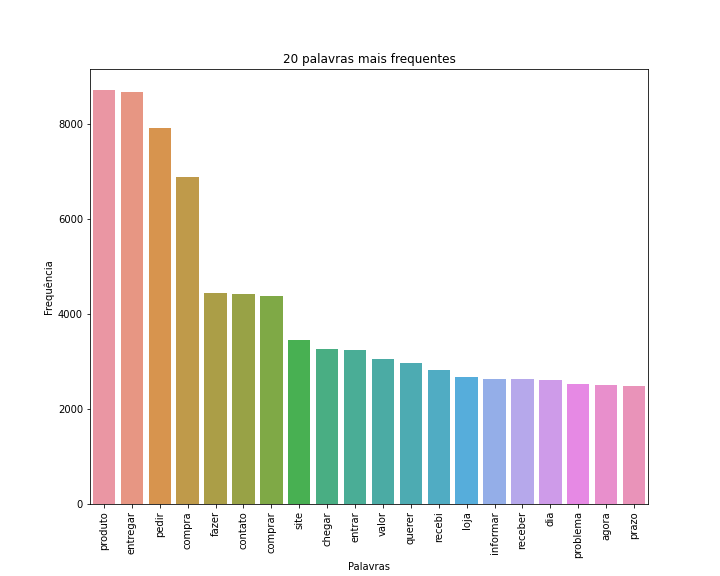


Figura 1. As vinte palavras mais frequentes

Fonte: Dados originais da pesquisa

Adicionalmente, "stop words", isto é, palavras que apresentam pouco valor semântico e que foram obtidas a partir de uma lista pré-definida em português da biblioteca NLTK, bem como outras "stop words" identificadas manualmente foram removidas (Tabela 1).

Tabela 1. Palavras removidas

| amazon |
| --- |
| americanas |
| casas bahia |
| magazine luiza |
| shein |
| kabum |
| samsung |
| mercado livre |
| banco brasil |
| apple |
| magazine |
| luiza |
| luizar |
| casas |
| bahia |
| casa |
| mercado |
| livre |

Fonte: Dados originais da pesquisa

Em seguida, os tokens foram normalizados para sua forma raiz, ou seja, foram reduzidos à sua forma básica com o propósito de padronizar os dados através da eliminação das variações das palavras com uso da biblioteca Spacy (https://spacy.io/). Por fim, os tokens foram filtrados por marcação de parte do discurso [POS, em inglês] para as classes gramaticais: substantivos, verbos, adjetivos e advérbios.

Além dos tokens em formato singular ou unigrama, foram considerados também pares de palavras (bigramas) e trios de palavras (trigramas) com frequência de ocorrência igual ou superior a vinte vezes, utilizando uma medida estatística chamada Informação Mútua Pontual (IMP) para quantificar a associação entre dois termos em um corpus linguístico.

Para fins de avaliação, a coerência form. (1) foi utilizada como métrica de performance dos modelos. A coerência é uma das métricas mais comuns na avaliação e comparação de modelos para técnicas de modelagem de tópicos (Terragni et al., 2021).

​

(1)

onde, k é um índice que se refere a um dos tópicos do modelo, k {1, 2, …, K}; n é o índice de uma palavra em um tópico, n {1, 2, …, K}; n,k vetor para representar a palavra no índice n no tópico k, onde | n,k | = N e o conjunto das N palavras mais prováveis no tópico k.

De acordo com Röder et al. (2015), existe uma correlação entre a pontuação de coerência e o julgamento humano de tópicos. Portanto, quanto maior a pontuação de coerência, melhor será a distinção dos tópicos e mais clara será a sua interpretação.

Para cada conjunto (unigramas, unigramas + bigramas e unigramas + bigramas + trigramas) de tokens foi obtido o intervalo de número ótimo de tópicos dos modelos LSI e LDA através da análise de coerência por número de tópicos utilizando um modelo não otimizado, ou seja, com hiperparâmetros padrões do modelo.

O intervalo do modelo BERTopic foi definido manualmente, porém dentro do intervalo obtido a partir dos outros modelos como no mínimo 8 tópicos e no máximo 14 tópicos.

Também foram definidos a frequência mínima “min\_count” (Tabela 2 e Tabela 3) de bigramas e trigramas, em um intervalo entre 2 a 20, dentro do corpus de documentos.Portanto, bigramas e trigramas com frequência abaixo do valor de min\_count foram ignorados, ou seja, não foram considerados como possíveis bigramas e/ou trigramas.

Além disso, palavras com frequência maior definido pelo hiperparâmetro filter\_extremes (no\_above) (Tabela 2 e Tabela 3) que um limiar variando entre 0.70 a 0.95 presentes no corpus de documentos foram filtradas para ajudar a reduzir o ruído nos dados.

Este hiperparâmetro tem como função controlar a exclusão ou preservação de palavras. Para valores muitos baixos muitas palavras relevantes são excluídas, enquanto um valor muito alto muitas palavras irrelevantes são preservadas. Portanto, foi definido um intervalo de ajuste deste parâmetro, a fim de se observar como isso afetaria a qualidade dos tópicos gerados e o valor de coerência.

No modelo LDA os hiperparâmetros alpha e eta também foram considerados. O alpha tem como função controlar a distribuição de tópicos em documentos, enquanto eta é responsável pelo controle da distribuição de palavras sobre tópicos .

A Tabelas 2 e 3 contém um resumo dos hiperparâmetros usados na otimização dos modelos LSI e LSA.

Tabela 2. Intervalo de hiperparâmetros do modelo Indexação Semântica Latente [LSI, em inglês]

| Indexação Semântica Latente [LSI, em inglês] | |
| --- | --- |
| ngram\_range | (1,1), (1,2), (1,3) |
| num\_topics | 2 - 4 |
| min\_count (bigramas/trigramas) | 2 - 20 |
| filter\_extremes (no\_above) | 0.70 - 0.95 |

Fonte: Dados originais da pesquisa

Tabela 3. Intervalo de hiperparâmetros do modelo Alocação de Dirichlet Latente [LDA, em inglês]

| Alocação de Dirichlet Latente [LDA, em inglês] | |
| --- | --- |
| ngram\_range | (1,1), (1,2), (1,3) |
| num\_topics | 4 - 12 |
| alpha | 0.01 - 1.0 |
| eta | 0.01 - 1.0 |
| min\_count (bigramas/trigramas) | 2 - 20 |
| filter\_extremes (no\_above) | 0.70 - 0.95 |

Fonte: Dados originais da pesquisa

No treinamento do BERTopic foram usados o seguintes modelos: Representações Codificadoras Bidirecionais de Transformadores [BERT, em inglês] para incorporação de palavras/documentos, Análise de Componentes Principais [PCA, em inglês] e Aproximação de Coletor Uniforme e Projeção [UMAP, em inglês] para redução de dimensionalidade, K-means e Agrupamento espacial de aplicativos baseado em densidade hierárquica com ruído [HDBSCAN, em inglês] para agrupamento.

A Tabela 4 mostra de forma detalhada os hiperparâmetros dos algoritmos e seus correspondentes valores ou intervalos utilizados.

Tabela 4. Intervalo de hiperparâmetros do modelo Representações de Codificador Bidireccional de Transformadores para Tópicos [BERTopic]

| Representações de Codificador Bidirecional de Transformadores para Tópicos [BERTopic] | |
| --- | --- |
| BERTopic | nr\_topics = [8, 14], n\_gram\_range= (1,1), (1,2), (1,3) |
| HDBSCAN | min\_cluster\_size = [10, 12], cluster\_selection\_method = "leaf" ou "oem", cluster\_selection\_epsilon = [0.0, 1.0], alpha = [0.1, 1.0], min\_samples = [5, 10] |
| K-MEANS | n\_clusters = [12, 20], max\_iter = 200, n\_init = 10, random\_state = 42 |
| UMAP | n\_neighbors = [10, 13], n\_components = [2, 3], metric = "cosine", random\_state = 42 |
| PCA | n\_components = [5, 6], random\_state=42 |

Fonte: Dados originais da pesquisa

Nota: Os colchetes [] denotam um intervalo

**Resultados**

A Tabela 5 resume as principais vantagens e desvantagens de cada um dos modelos.

Tabela 5. Principais vantagens e desvantagens de cada modelo

| Características | BERTopic | LDA | LSI |
| --- | --- | --- | --- |
| Pode lidar com grande variedade de idiomas | Sim | - | - |
| Permite ajuste fino dos hiperparâmetros | Sim | Sim | Sim |
| Requer mais recursos computacionais | Sim | - | - |
| Requer um pré-processamento cuidadoso dos dados | Sim\* | Sim | Sim |
| Fácil interpretação do modelo | - | Sim | Sim |
| Pode identificar distribuição de tópicos por documento | Sim | Sim | - |
| Requer o número de tópicos | Sim\* | Sim | Sim |
| Fonte: Resultados originais da pesquisa | | | |
| Nota: \*Opcional | | | |

A Figura 3 apresenta os 20 mais frequentes bigramas e trigramas criados com o uso da biblioteca Spacy. A Figura 3 revela a frequência com que as palavras aparecem no conjunto de documentos. É observado que as palavras produto, compra e loja aparecem com frequência formando pares e trios de palavras.

****

Figura 3. Nuvem de palavras com os 20 mais frequentes pares e trios de palavras. Bigramas (Figura à esquerda) e trigramas (Figura à direita)

Fonte: Resultados originais da pesquisa

Na Figura 4 observa-se que o valor ótimo de coerência do modelo LSI para unigramas está entre 10 a 12 tópicos, porém também é observado um outro pico entre 2 a 4 tópicos indicando que este modelo são foi capaz de definir de forma clara o número de tópicos onde o valor de coerência é o melhor possível.

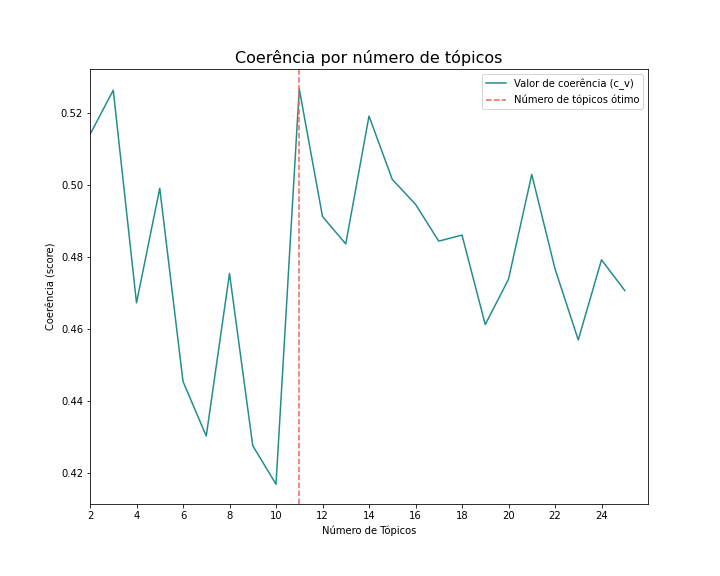
****

Figura 4. Coerência versus número de tópicos para unigramas gerada a partir do modelo Indexação Semântica Latente [LSI, em inglês]

Fonte: Resultados originais da pesquisa

Nas Figuras 5 (unigramas + bigramas) e 6 (unigrama + bigrama + trigramas) observa-se que o valor de coerência apresenta um declínio considerável, ou seja, o modelo teve dificuldade em extrair tópicos quando incluído as formas bigramas e trigramas das palavras. Em outras palavras, os tópicos gerados pelo modelo apresentam uma sobreposição entre eles.

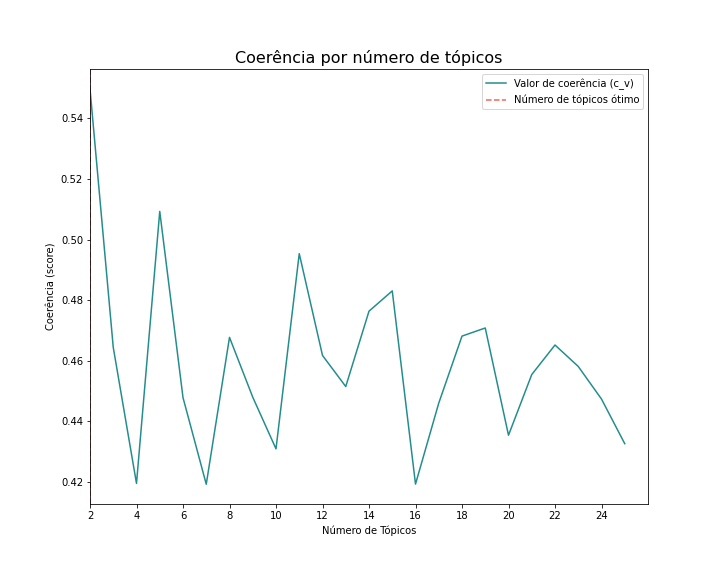


Figura 5. Coerência versus número de tópicos para unigramas + bigramas gerada a partir do modelo Indexação Semântica Latente [LSI, em inglês]

Fonte: Resultados originais da pesquisa

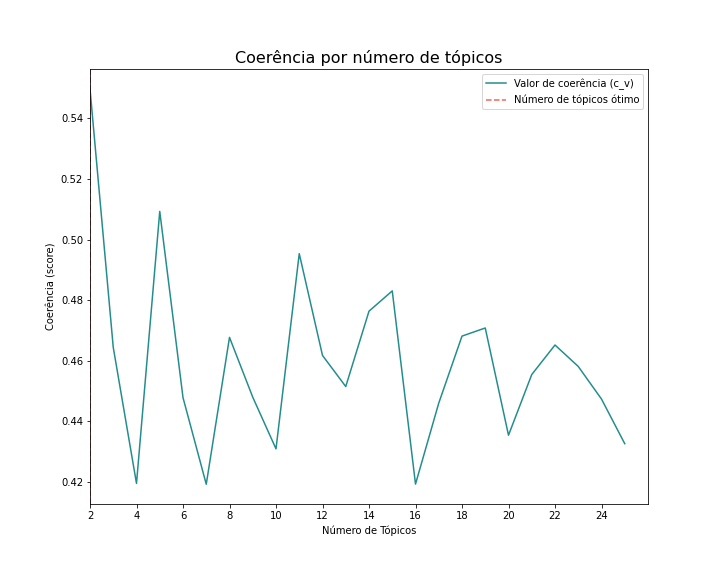


Figura 6. Coerência versus número de tópicos para unigramas + bigramas + trigramas gerada a partir do modelo Indexação Semântica Latente [LSI, em inglês]

Fonte: Resultados originais da pesquisa

O modelo LDA, ao contrário do LSI, demonstrou capacidade de gerar tópicos para todos conjuntos de dados (Figuras 7, Figura 8 e Figura 9). Considerando somente a forma unigrama (Figura 7) observa-se um pico bem definido entre o intervalo entre 4 a 6.

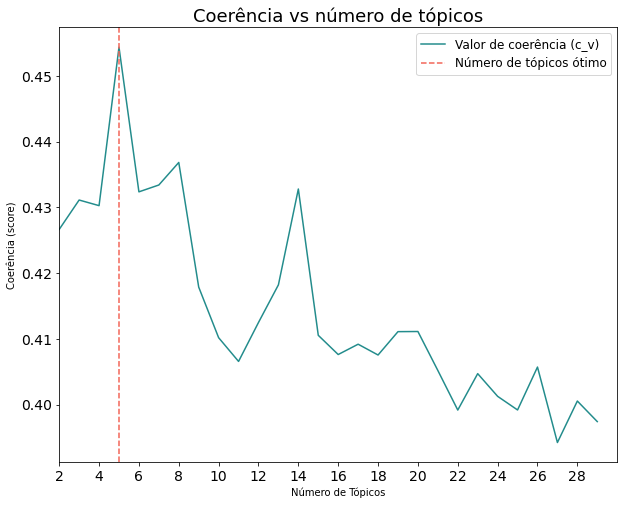


Figura 7. Coerência versus número de tópicos para unigramas gerada a partir do modelo Alocação de Dirichlet Latente [LDA, em inglês]

Fonte: Resultados originais da pesquisa

Ao contrário do modelo LSI, o LDA também foi capaz de extrair tópicos na presença de bigramas (Figura 8) e trigramas (Figura 9). Além disso, observa-se nas Figuras 8 e 9 que o valor de coerência ótimo se desloca para a direita, ou seja, o modelo foi capaz de gerar mais tópicos a partir da inclusão de tokens de pares e trios de palavras.

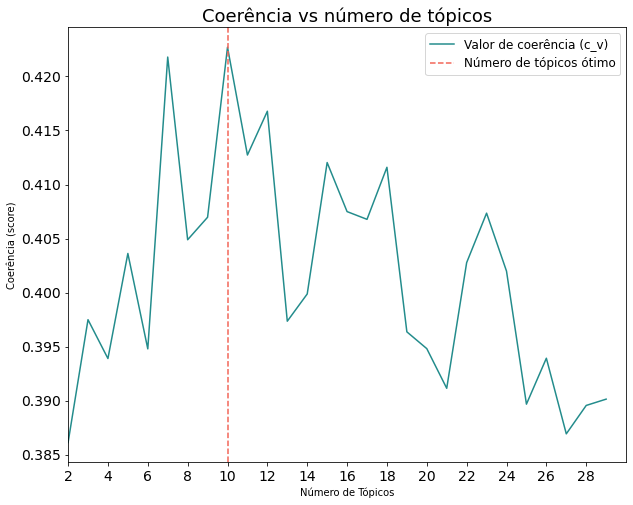


Figura 8. Coerência versus número de tópicos para unigramas + bigramas gerada a partir do modelo Alocação de Dirichlet Latente [LDA, em inglês]

Fonte: Resultados originais da pesquisa

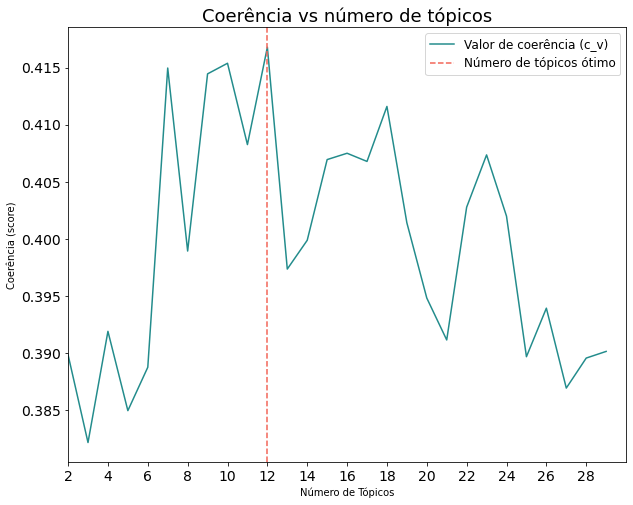


Figura 9. Coerência versus número de tópicos para unigramas + bigramas + trigramas gerada a partir do modelo Alocação de Dirichlet Latente [LDA, em inglês]

Fonte: Resultados originais da pesquisa

A Tabela 7 apresenta os valores de coerência do modelo de Alocação de Dirichlet Latente (LDA) otimizado com a biblioteca Optuna, para diferentes valores de número de tópicos, n-grama, alpha, eta, min\_count e filter\_extremes(no\_above) e a Tabela 8 a descrição dos hiperparâmetros. A partir dos valores apresentados na Tabela 7, é possível observar que diferentes configurações de hiperparâmetros geram diferentes valores de coerência.

Tabela 7. Valores de coerência do modelo Alocação de Dirichlet Latente [LDA, em inglês] optimizado com a biblioteca Optuna

| Index | Número de tópicos | N-grama | Alpha | Eta | min\_count | filter\_extremes (no\_above) | Coerência |
| --- | --- | --- | --- | --- | --- | --- | --- |
| model\_1 | 11 | (1, 1) | 0.0124 | 0.2986 | 5 | 0.7009 | 0.4720 |
| model\_2 | 10 | (1, 1) | 0.4766 | 0.1122 | 16 | 0.7599 | 0.4667 |
| model\_3 | 6 | (1, 3) | 0.3172 | 0.1835 | 11 | 0.9328 | 0.4474 |
| model\_4 | 9 | (1, 1) | 0.8118 | 0.2650 | 2 | 0.7041 | 0.4654 |
| model\_5 | 9 | (1, 3) | 0.2370 | 0.1722 | 3 | 0.7994 | 0.4804 |
| model\_6 | 4 | (1, 1) | 0.5806 | 0.3127 | 12 | 0.8633 | 0.4652 |
| model\_7 | 7 | (1, 1) | 0.1369 | 0.3983 | 5 | 0.8545 | 0.4520 |
| model\_8 | 11 | (1, 3) | 0.5064 | 0.6790 | 11 | 0.9362 | 0.4325 |
| model\_9 | 8 | (1, 3) | 0.7630 | 0.3622 | 3 | 0.9117 | 0.4667 |
| model\_10 | 5 | (1, 3) | 0.0450 | 0.1033 | 2 | 0.9451 | 0.4525 |
| model\_11 | 12 | (1, 2) | 0.9588 | 0.0229 | 20 | 0.7972 | 0.4482 |
| model\_12 | 10 | (1, 2) | 0.0279 | 0.5863 | 6 | 0.7133 | 0.4502 |
| model\_13 | 12 | (1, 3) | 0.2204 | 0.9893 | 7 | 0.7574 | 0.4571 |
| model\_14 | 9 | (1, 1) | 0.2722 | 0.2340 | 8 | 0.8129 | 0.4712 |
| model\_15 | 10 | (1, 2) | 0.0181 | 0.0139 | 9 | 0.7492 | 0.4594 |
| model\_16 | 8 | (1, 1) | 0.1585 | 0.4532 | 4 | 0.7895 | 0.4652 |
| model\_17 | 11 | (1, 3) | 0.3426 | 0.1768 | 5 | 0.7201 | 0.4467 |
| model\_18 | 7 | (1, 1) | 0.1541 | 0.2848 | 9 | 0.7366 | 0.4555 |
| model\_19 | 9 | (1, 3) | 0.3791 | 0.4953 | 14 | 0.7812 | 0.4834 |
| model\_20 | 9 | (1, 3) | 0.3980 | 0.5277 | 15 | 0.7811 | 0.4674 |

Fonte: Resultados originais da pesquisa

Tabela 8. Descrição da tabela de coerência do modelo Alocação de Dirichlet Latente [LDA, em inglês]

| Coluna | Descrição |
| --- | --- |
| Index | Identificador único |
| Número de tópicos | Número de Tópicos |
| N-grama | Tipo de sequência de palavras consideradas para a geração dos tópicos. Unigramas (palavras isoladas), bigramas (pares de palavras) ou trigramas (trios de palavras). |
| alpha | Prior Dirichlet para a distribuição documento-tópico. Determina a distribuição de probabilidade sobre os tópicos de um documento |
| eta | Prior Dirichlet para a distribuição tópico-palavra. Determina a distribuição de probabilidade sobre as palavras de um tópico |
| min\_count | Frequência mínima de bigramas/trigramas para serem incluídos no modelo |
| filter\_extremes (x% dos documentos) | Filtro de palavras que ocorrem em mais de x% dos documentos |
| Coerência | Medida de consistência e interpretabilidade dos tópicos gerados |

Fonte: Resultados originais da pesquisa

Em média o modelo LDA teve uma pontuação de 0.4599 para um intervalo de número de tópicos entre 4 a 12. A Figura 10 mostra a distribuição dos tópicos do modelo identificado como model\_19 na Tabela 7. A Figura 10 mostra que este modelo foi capaz de gerar tópicos sem sobreposição.



Figura 10. Mapa de distância inter tópico do modelo Alocação de Dirichlet Latente [LDA, em inglês] (model\_19)

Fonte: Resultados originais da pesquisa

Na Tabela 9 é observado os termos mais relevantes de cada tópico. As palavras aparecem de forma ordenada considerando como critério a relevância dentro do tópico. Observa-se que o LDA gerou tópicos relacionados a problemas de prazo, devolução, valor, data, etc.

Tabela 9. Tópicos e termos do modelo Alocação de Dirichlet Latente [LDA, em inglês]

| Tópico | Palavras |
| --- | --- |
| 1 | ['prazo', 'resposta', 'reclamação', 'problema', 'dia', 'mensagem', 'unidade', 'caixa', 'coisa', 'momento'] |
| 2 | ['devolução', 'dia', 'troca', 'pacote', 'retorno', 'mensagem', 'defeito', 'código', 'empresa', 'anúncio'] |
| 3 | ['valor', 'problema', 'frete', 'mês', 'mercadoria', 'cliente', 'reembolso', 'número', 'dia', 'tempo'] |
| 4 | ['data', 'atraso', 'mensagem', 'prazo', 'previsão', 'atendimento', 'telefone', 'problema', 'retorno', 'número'] |
| 5 | ['empresa', 'informação', 'novembro', 'nota', 'telefone', 'suporte', 'acesso', 'conta', 'problema', 'número'] |
| 6 | ['pagamento', 'valor', 'promoção', 'correio', 'desconto', 'confirmação', 'endereço', 'erro', 'chat', 'tempo'] |
| 7 | ['loja', 'atendimento', 'compr', 'dia', 'informação', 'encomenda', 'empresa', 'telefone', 'ponto', 'poder'] |
| 8 | ['dinheiro', 'dia', 'cancelamento', 'momento', 'cidade', 'valor', 'poder', 'vendedor', 'estorno', 'prazo'] |
| 9 | ['cartão', 'cancelamento', 'valor', 'fatura', 'cashback', 'crédito', 'serviço', 'atendimento', 'oferta', 'forma'] |

Fonte: Resultados originais da pesquisa (model\_19)

A Figura 11 mostra a relevância de cada termo em um tópico de forma gráfica, assim como a Tabela 9, porém apresenta os valores de relevância de cada termo no eixo X. Para cada tópico o valor máximo de relevância corresponde ao valor próximo do calculado para a palavra mais importante dentro do tópico. Segundo os resultados observados na Figura 11 as palavras mais relevantes são: prazo, devolução, valor, data, empresa e pagamento.

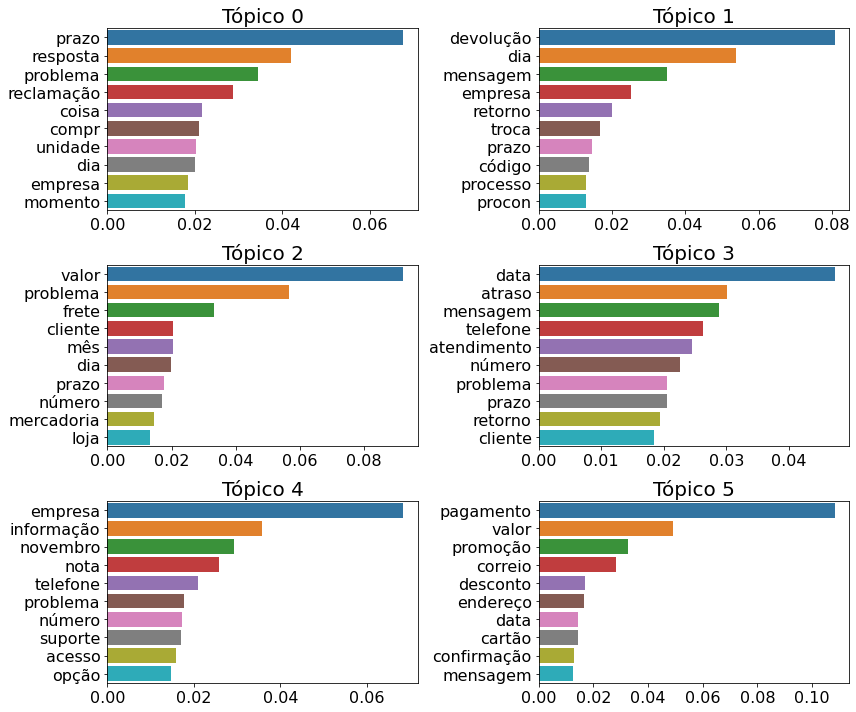


Figura 11. Palavras mais relevantes para cada tópico (model\_19)

Fonte: Resultados originais da pesquisa

A Tabela 10 apresenta os valores de coerência do modelo de Indexação Semântica Latente (LSI) otimizado com a biblioteca Optuna, para diferentes valores de número de tópicos, n-grama, min\_count e filter\_extremes(no\_above) e a Tabela 11 a descrição dos hiperparâmetros. Assim como no modelo LDA as diferentes configurações de hiperparâmetros geram diferentes valores de coerência.

Tabela 10. Valores de coerência do modelo Indexação Semântica Latente [LSI, em inglês] optimizado com a biblioteca Optuna

(contínua)

| Index | Número de tópicos | N-grama | min\_count | Filter\_extremes (no\_above) | Coerência |
| --- | --- | --- | --- | --- | --- |
| model\_1 | 12 | (1, 3) | 3 | 0.8906 | 0.4939 |
| model\_2 | 12 | (1, 2) | 3 | 0.8665 | 0.4922 |
| model\_3 | 9 | (1, 2) | 15 | 0.7835 | 0.5300 |
| model\_4 | 3 | (1, 3) | 11 | 0.9179 | 0.4470 |
| model\_5 | 9 | (1, 3) | 9 | 0.7331 | 0.4856 |
| model\_6 | 13 | (1, 2) | 8 | 0.9267 | 0.5033 |
| model\_7 | 2 | (1, 1) | 11 | 0.7749 | 0.5367 |
| Tabela 10. Valores de coerência do modelo Indexação Semântica Latente [LSI, em inglês] optimizado com a biblioteca Optuna  (conclusão) | | | | | |
| Index | Número de tópicos | N-grama | min\_count | Filter\_extremes (no\_above) | Coerência |
| model\_8 | 5 | (1, 2) | 14 | 0.8534 | 0.4324 |
| model\_9 | 11 | (1, 2) | 2 | 0.8354 | 0.4825 |
| model\_10 | 5 | (1, 3) | 4 | 0.7010 | 0.4621 |
| model\_11 | 2 | (1, 1) | 19 | 0.7814 | 0.5367 |
| model\_12 | 2 | (1, 1) | 19 | 0.7912 | 0.5367 |
| model\_13 | 4 | (1, 1) | 20 | 0.7921 | 0.5718 |
| model\_14 | 5 | (1, 1) | 14 | 0.8234 | 0.5634 |
| model\_15 | 5 | (1, 1) | 16 | 0.8184 | 0.5634 |
| model\_16 | 7 | (1, 1) | 17 | 0.8197 | 0.5406 |
| model\_17 | 7 | (1, 1) | 20 | 0.7482 | 0.5406 |
| model\_18 | 4 | (1, 1) | 14 | 0.8109 | 0.5718 |
| model\_19 | 4 | (1, 1) | 17 | 0.8016 | 0.5718 |
| model\_20 | 6 | (1, 1) | 13 | 0.8562 | 0.4447 |

Fonte: Resultados originais da pesquisa

Tabela 11. Descrição da tabela de coerência do modelo Indexação Semântica Latente [LSI, em inglês]

| Coluna | Descrição |
| --- | --- |
| Index | Identificador único |
| Número de tópicos | Número de Tópicos |
| N-grama | Tipo de sequência de palavras consideradas para a geração dos tópicos. Pode ser unigramas (palavras isoladas), bigramas (pares de palavras) ou trigramas (trios de palavras). |
| min\_count | Frequência mínima de bigramas/trigramas para serem incluídos no modelo |
| filter\_extremes (x% dos documentos) | Filtro de palavras que ocorrem em mais de x% dos documentos |
| Coerência | Medida de consistência e interpretabilidade dos tópicos gerados |

Fonte: Resultados originais da pesquisa

Enquanto o LSI teve um desempenho marginalmente superior ao LDA em termos de coerência (0.5154 vs 0.4599) este modelo teve problemas na distribuição dos termos. A Tabela 12 mostra de forma clara a sobreposição de termos, obtido pelo modelo LSI, como prazo, valor, dia e loja em diferentes tópicos. Este resultado evidencia ainda mais as limitações já observadas nas Figuras 5 e 6 deste modelo em extrair tópicos distintos.

Tabela 12. Tópicos e termos do modelo Indexação Semântica Latente [LSI, em inglês]

| Tópico | Palavras |
| --- | --- |
| 1 | ['loja', 'valor', 'dia', 'empresa', 'problema', 'prazo', 'cartão', 'dinheiro', 'atendimento', 'mensagem'] |
| 2 | ['valor', 'loja', 'cartão', 'dia', 'prazo', 'empresa', 'pagamento', 'fatura', 'magalu', 'devolução'] |
| 3 | ['loja', 'dia', 'valor', 'prazo', 'empresa', 'cartão', 'problema', 'dinheiro', 'mensagem', 'devolução'] |
| 4 | ['empresa', 'dia', 'valor', 'problema', 'prazo', 'atendimento', 'cliente', 'cartão', 'número', 'poder'] |
| 5 | ['cartão', 'valor', 'empresa', 'pagamento', 'fatura', 'dinheiro', 'dia', 'devolução', 'prazo', 'número'] |

Fonte: Resultados originais da pesquisa

Na Tabela 13 é observado os valores de coerência do modelo Representações de Codificador Bidireccional de Transformadores para Tópicos ou BERTopic otimizado com a biblioteca Optuna, para diferentes valores de número de tópicos, n-grama, algoritmo de clusterização e algoritmo de redução de dimensionalidade e a Tabela 14 a descrição dos hiperparâmetros. Este modelo também gerou diferentes valores de coerência para diferentes configurações de hiperparâmetros. Observe que os valores de coerência foram superiores com a inclusão de trios de palavras (trigramas) no treinamento do modelo.

Tabela 13. Valor de coerência do modelo Representações de Codificador Bidireccional de Transformadores para Tópicos [BERTopic]

(continua)

| Nome do modelo | Número de tópicos | N-grama | Algoritmo de clusterização | Algoritmo de redução de dimensionalidade | Coerência |
| --- | --- | --- | --- | --- | --- |
| model\_trial\_1 | 9 | (1, 3) | HDBSCAN | UMAP | 0.6407 |
| model\_trial\_2 | 10 | (1, 1) | HDBSCAN | UMAP | 0.3694 |
| model\_trial\_3 | 9 | (1, 1) | K-means | UMAP | 0.3754 |
| model\_trial\_4 | 10 | (1, 2) | HDBSCAN | UMAP | 0.5096 |
| model\_trial\_5 | 8 | (1, 2) | HDBSCAN | UMAP | 0.4578 |
| model\_trial\_6 | 10 | (1, 2) | HDBSCAN | UMAP | 0.5598 |
| model\_trial\_7 | 12 | (1, 3) | K-means | UMAP | 0.8287 |
| model\_trial\_8 | 8 | (1, 2) | K-means | UMAP | 0.7641 |
| model\_trial\_9 | 14 | (1, 2) | K-means | UMAP | 0.7412 |
| model\_trial\_10 | 13 | (1, 1) | K-means | PCA | 0.3749 |
| model\_trial\_11 | 11 | (1, 3) | K-means | PCA | 0.8466 |
| model\_trial\_12 | 11 | (1, 3) | K-means | PCA | 0.8324 |
| model\_trial\_13 | 10 | (1, 3) | K-means | PCA | 0.8630 |
| model\_trial\_14 | 10 | (1, 3) | K-means | PCA | 0.8646 |
| model\_trial\_15 | 10 | (1, 3) | K-means | PCA | 0.8635 |
| model\_trial\_16 | 12 | (1, 3) | K-means | PCA | 0.8514 |
| model\_trial\_17 | 10 | (1, 1) | K-means | PCA | 0.3692 |
| Tabela 13. Valor de coerência do modelo Representações de Codificador Bidireccional de Transformadores para Tópicos [BERTopic]  (conclusão) | | | | | |
| Nome do modelo | Número de tópicos | N-grama | Algoritmo de clusterização | Algoritmo de redução de dimensionalidade | Coerência |
| model\_trial\_18 | 10 | (1, 3) | K-means | PCA | 0.8392 |
| model\_trial\_19 | 9 | (1, 3) | K-means | PCA | 0.8581 |
| model\_trial\_20 | 10 | (1, 3) | K-means | PCA | 0.8457 |

Fonte: Resultados originais da pesquisa

Tabela 14. Descrição da tabela de coerência do modelo Representações de Codificador Bidirecional de Transformadores para Tópicos [BERTopic]

| Coluna | Descrição |
| --- | --- |
| Nome do Modelo | Nome da simulação do modelo |
| Número de tópicos | Números de tópicos gerados pelo BERTopic |
| N-grama | Tipo de sequência de palavras consideradas para a geração dos tópicos. Pode ser unigramas (palavras isoladas), bigramas (pares de palavras) ou trigramas (trios de palavras). |
| Número de n-gramas | Número de palavras a ser considerado (uma, duas ou três) como token |
| Modelo de Clusterização | Algoritmo utilizado para agrupar os dados |
| Modelo de Redução de Dimensionalidade | Algoritmo utilizado para reduzir a dimensionalidade dos dados |
| Coerência | Medida de consistência e a interpretabilidade dos tópicos gerados |

Fonte: Resultados originais da pesquisa

Em média o BERTopic obteve a maior pontuação dentre os modelos e foi capaz assim como o LDA de gerar tópicos interpretáveis. A Figura 12 mostra a distribuição dos tópicos obtida a partir do modelo model\_trial\_2 resultante do processo de otimização de hiperparâmetros. Assim como o LDA, o BERTopic foi capaz de extrair tópicos distintos e interpretáveis.

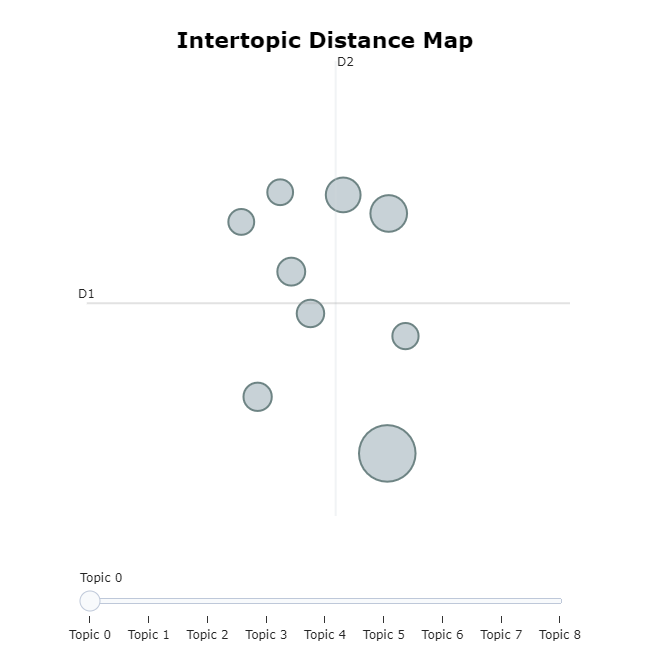
****

Figura 12. Mapa de distância inter tópico do modelo Representações de Codificador Bidirecional de Transformadores para Tópicos [BERTopic] (model\_trial\_2)

Fonte: Resultados originais da pesquisa

A Figura 13 mostra a relevância de cada palavra para cada tópico. Observe que o BERTopic foi capaz de extrair termos relevantes, alguns encontrados também no resultado do LDA, que fornecem informações importantes sobre os motivos das reclamações como cartão, endereço, telefone, prazo e cancelamento.

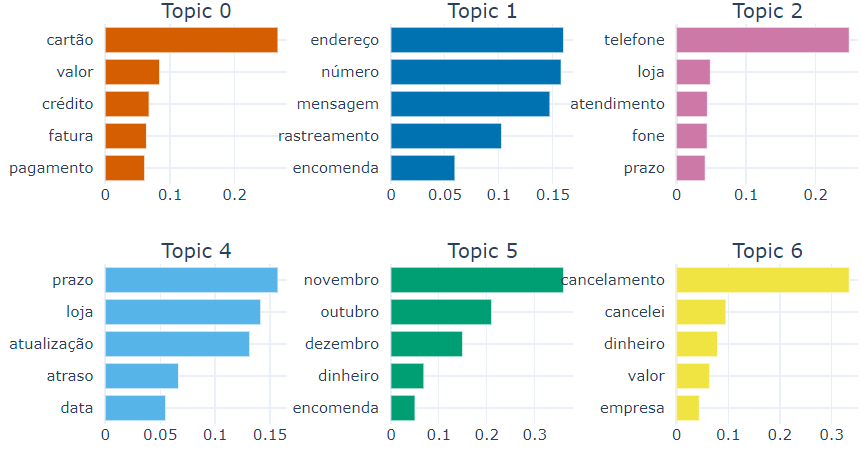


Figura 13. Palavras mais relevantes para cada tópico do modelo (model\_trial\_2)

Fonte: Resultados originais da pesquisa

A Tabela 15 apresenta a média para a pontuação de coerência para cada modelo obtida a partir dos valores de coerência nas Tabelas 7, 10 e 13.

Tabela 15. Média dos valores de coerência

| Modelo | Coerência (Média) |
| --- | --- |
| LDA | 0.4599 |
| LSI | 0.5154 |
| BERTopic | 0.6828 |

Fonte: Resultados originais da pesquisa

Em resumo, os modelos LDA e BERTopic obtiveram o melhor resultado em termos de interpretabilidade e distribuição dos termos em tópicos (Tabela 15). Os modelos LDA e BERTopic foram capazes de identificar temas como prazo de entrega, cancelamento e problemas associados a cartões de crédito. O BERTopic assim como demonstrado em trabalhos anteriores Grootendorst (2022) e Ogunleye et al. (2023) obteve a maior pontuação entre os modelos com uma média de 0.6828.

**Conclusão(ões) ou Considerações Finais**

Em conclusão, o presente projeto revelou que a modelagem de tópicos é uma técnica altamente eficaz para a extração de informações estruturadas a partir de grandes conjuntos de dados de reclamações de clientes. Notadamente, os modelos BERTopic e LDA foram capazes de identificar os principais motivos subjacentes às queixas relacionadas ao comércio eletrônico no Brasil, destacando, com relevância, os temas referentes a problemas com cartões de crédito, pagamentos, atrasos em entregas, prazos de entrega e devoluções, bem como problemas decorrentes do atendimento direto com as empresas.

Por outro lado, o modelo LSI demonstrou limitações na separação das reclamações em tópicos com temas precisos. A análise dos termos dos tópicos gerados indicou a sobreposição de muitos termos, tais como "loja", "valor" e "prazo", em diferentes tópicos, o que sugere a necessidade de maiores investigações.

Os resultados obtidos indicam que, ao identificar as preocupações e os problemas mais comuns entre os clientes, as empresas podem compreender melhor as razões subjacentes à insatisfação destes e tomar medidas efetivas para melhorar seus produtos ou serviços, o que, por sua vez, pode levar a um melhor relacionamento e satisfação do cliente.

Além disso, os achados deste estudo destacam que a modelagem de tópicos pode auxiliar as empresas a lidar de maneira mais efetiva com as reclamações dos clientes, equilibrando o impacto em seu valor de mercado e reputação, e promovendo a tomada de decisões baseadas em dados para reduzir o desgaste do cliente e, por conseguinte, melhorar a experiência do cliente.

Em resumo, espera-se que as contribuições deste estudo possam servir de base para o desenvolvimento de futuras pesquisas na área de modelagem de tópicos aplicadas ao “feedback” dos consumidores e para o desenvolvimento de soluções práticas que possam ser utilizadas por empresas interessadas em melhorar a experiência de seus consumidores e que possa contribuir para futuros trabalhos envolvendo modelagem de tópicos em diferentes áreas de interesses. Os modelos LDA e BERTopic demonstraram-se modelos com ótima performance na extração de termos temáticos de reclamações, abrindo um caminho de oportunidades para novas aplicações. O BERTopic oferece uma ampla gama de usos como análise de evolução de tópicos por tempo e modelagem incremental de tópicos, os quais não foram explorados neste projeto e que poderiam trazer resultados ainda melhores na compreensão dos problemas presentes em reclamações.

**Agradecimentos**

Expresso minha profunda gratidão a minha família e amigos pelo apoio, aos professores do MBA em Data Science e Analytics da USP/ESALQ que dedicaram seu tempo e conhecimento e meu orientador pelo incentivo de construir um excelente projeto.

**Referências**

Asmussen, C. B.; Møller, C. 2019. Smart literature review: a practical topic modeling approach to exploratory literature review. Journal of big data 6.

Barde, B. V.; Bainwad, A. M. 2017. An overview of topic modeling methods and tools. 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). Anais... In: 2017 International Conference On Intelligent Computing And Control Systems (Iciccs). IEEE.

Bastani K.; Namavari H.; Shaffer J. 2019. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, Expert Systems with Applications. 127: 256-271

Blei, D. M.; Ng, A. Y E Jordan M. I. 2003. Latent dirichlet allocation. The journal of machine learning research 3: 993–1022.

Claro, D. P.; Fragoso, A. F. G. R.; Laban Neto, S. A.; & Claro, P. B. de O. 2014. Consumer Complaints and Company Market Value. BAR - Brazilian Administration Review, 11(BAR, Braz. Adm. Rev.) 11(3): 248–263.

Deerwester et al. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41: 391-407.

Filip A. 2013. Complaint Management: A Customer Satisfaction Learning Process, Procedia - Social and Behavioral Sciences, 93: 271-275.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. 2021. Evaluation of clustering and topic modeling methods over health-related tweets and emails. Artificial Intelligence in Medicine, 117, 117:102096.

Luo, X. 2009. Quantifying the Long-Term Impact of Negative Word of Mouth on Cash Flows and Stock Prices. Marketing Science, 28(1): 148–165.

Ogunleye, B.; Maswera, T.; Hirsch, L.; Gaudoin, J.; Brunsdon, T. 2023. Comparison of Topic Modelling Approaches in the Banking Context. Approaches in the Banking Context. Appl. Sci 13: 797.

Pennacchiotti, M. E Gurumurthy, S. 2011. Investigating topic models for social media user recommendation. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India 101–102.

Robinson, S. D. 2019. Temporal topic modeling applied to aviation safety reports: A subject matter expert review. Safety Science 116: 275–286.

Röder, M.; Both, A.; Hinneburg, A. 2015. Exploring the space of topic coherence measures. Em: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China 399-408.

Terragni S. et al. 2021. OCTIS: Comparing and Optimizing Topic Models is Simple!.Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations 263–270.

Vaswani, A. 2017 et al. Attention Is All You Need. Advances in Neural Information Processing Systems 30: 5998-6008.