**Trabalho de Conclusão de Curso**

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

**ALUNO: André Pesati Revoredo**

**ORIENTADOR: Nome do Orientador**



Sumário

´

[**1. RESUMO 2**](#_Toc83580266)

[**2. INTRODUÇÃO 3**](#_Toc83580267)

[**3. TRABALHOS RELACIONAIS 4**](#_Toc83580268)

[**4. METODOLOGIA 5**](#_Toc83580269)

[**5. RESULTADOS 6**](#_Toc83580270)

[**6. DISCUSSÃO 7**](#_Toc83580271)

[**7. CONCLUSÃO E TRABALHOS FUTUROS 8**](#_Toc83580272)

**ARTIGO CIENTÍFICO – MODELAGEM DE TÓPICOS DA PLATAFORMA RELACLAME AQUI RELACIONADA A EMPRESA NO SETOR DE ENERGIA ELÉTRICA.**

# RESUMO

O *eWoM* consiste no compartilhamento de avaliações sobre produtos e serviços por meio de plataformas on-line. Estas avaliações são baseadas em textos e vêm acompanhadas de uma pontuação e comentários que visam refletir o nível de satisfação do cliente para com a empresa.

O projeto tem como etapa desenvolver um *webscraping* capaz de capturar os comentários, datas, localização, dentre outras informações atribuídas aos comentários realizados pelos usuários para a empresa em específico.

Por conseguinte, será realizado o pré-processamento dos dados para conseguir realizar algumas etapas, tais como tokenização, lematização e radicalização das palavras inseridas nos comentários.

Desse modo, será feito a aplicação do modelo de LDA, o modelo estatístico de aprendizado de máquina, para descobrir os tópicos ocultos em uma coleção de documentos. Em outras palavras, extrair informações relevantes de grandes volumes de texto e organizar em grupos temáticos.

Neste contexto, este projeto tem como finalidade responder as questões: É possível adotar o modelo de LDA para análise de *eWoM* do site Reclame Aqui? Qual as principais palavras chaves utilizadas nos tópicos?

Portanto, os objetivos do projeto é analisar as reclamações submetidas pelos clientes da empresa Enel Distribuição - Ceará no Reclame Aqui, verificar se condiz com o que é dito no texto dos comentários, identificar os tópicos e as palavras mais frequentes e analisar a eficiência do algoritmo LDA para as reclamações diante do contexto.

**Palavras-chave:** *eWoM* (*Electronic Word-of-Mouth*); LDA (*Latent Dirichlet Allocation*); *webscraping*; modelagem de tópicos.

# INTRODUÇÃO

O cenário digital atual está em constante evolução, com um aumento significativo no volume de opiniões dos clientes, expressas através de reclamações e avaliações. Essa abundância de informações oferece às empresas uma perspectiva mais profunda sobre seus clientes e a oportunidade de aprimorar seus produtos.

Neste contexto, o *eWoM* (*Electronic Word-of-Mouth*) desempenha um papel crucial, pois os clientes têm acesso a uma quantidade cada vez maior de informações sobre produtos, serviços e reputação das marcas online. O *eWoM* pode ser geralmente definido como o compartilhamento e troca de informações dos consumidores sobre um produto ou empresa através da Internet, mídias sociais e comunicação móvel (OXFORD, 2021). Na Figura 1, a imagem mostra os tipos de *eWoM* nas plataformas de mídia social.

**Figura 1 – Tipos de *eWoM* na mídia social**

A diagram of a social media network

Description automatically generated

**Fonte - Mahmoud Alghizzawi**

No mundo digital de hoje, as empresas que desejam prosperar precisam obrigatoriamente saber analisar o feedback dos seus clientes. Plataformas como Reclame Aqui, Denuncio e Bondfaro são sistemas de reputação conhecidos no Brasil. Com isso, foi utilizado a plataforma do Reclame Aqui com o foco na coleta e análise de dados das reclamações.

A modelagem de tópicos é uma ferramenta essencial que permite compreender e categorizar automaticamente as reclamações dos consumidores. Trata-se de uma técnica que se baseia em um conjunto de algoritmos para revelar, descobrir e anotar a estrutura temática em uma coleção de documentos. (BLEI, 2012).

Este trabalho tem por objetivo fornecer uma solução de classificação mais precisa para consumidores e empresas. Ao invés de confiar apenas em avaliações numéricas, o sistema considera diversos aspectos da experiência de escrita do usuário, garantindo uma classificação mais abrangente e realista. Desse modo, o método para modelagem de tópico LDA (*Latent Dirichlet Allocation*) foi utilizado, sendo uma técnica para modelar a distribuição de probabilidade de tópicos em um conjunto de documentos. O LDA é um modelo probabilístico generativo para coleções de dados discretos, como corpora de texto, em que cada item de uma coleção é modelado como uma mistura finita sobre um conjunto subjacente de tópicos (BLEI; NG; JORDAN, 2003).

A indisponibilidade de *APIs* para acesso aos dados da maioria das plataformas de *eWoM* demanda a adoção de técnicas alternativas para a coleta de dados, como o *Web Crawling.* O conceito de *crawling*, inicialmente desenvolvido por (GRAY, 1993) no MIT (*Massachusetts Institute of Technology*) utilizando *Perl*, visava mensurar o tamanho da *web*. Neste mesmo contexto, (HUGGINS, 2004) criou o *Selenium* como uma ferramenta de automação de testes para aplicações *web open-source* e que foi incorporando ferramentas como o *Selenium WebDriver*, que possibilitam a prática de *webscraping*.

Por fim, o trabalho desse projeto encontra-se organizado da seguinte maneira: a Seção 2 apresenta a Introdução. Na Seção 3 apresenta alguns trabalhos relacionados ao Processamento de Linguagem Natural. Na Seção 4, apresenta a metodologia usada para resolver o problema incluindo arquitetura desenvolvida, recursos e materiais utilizados. As Seções 5 e 6 descrevem os resultados experimentais e discussões. Por fim, a conclusão e trabalhos futuros são apresentados na Seção 7.

# TRABALHOS RELACIONAIS

Neste capítulo serão apresentados os trabalhos encontrados sobre Processamento de Linguagem Natural e o boca a boca eletrônico (*eWoM*), foco deste projeto. Para isso, apresenta uma análise cronológica de trabalhos relevantes na área de PLN, organizada em seções que representam diferentes abordagens para lidar com os desafios específicos do *eWoM*.

1. **Melhorando a Gestão de Relacionamento Com Clientes Por Meio de Mídias Sociais: um estudo sobre o boca-a-boca eletrônico:**

Este trabalho apresenta uma análise de sentimentos e localização de reclamações de consumidores utilizando técnicas de Processamento de Linguagem Natural (PNL) e visualização de dados (ALMEIDA; CIRQUEIRA; LOBATO, 2017). Através de *webscraping*, dados de duas empresas foram extraídos da plataforma Reclame Aqui e armazenados em formato CSV. Uma interface web interativa, desenvolvida com *Python* e *Flask*, permite visualizar um painel de indicadores com métricas e gráficos gerados a partir da análise dos dados. A figura 2 apresenta o funcionamento básico da plataforma.

A análise identificou os tópicos e termos mais significativos nas reclamações de cada empresa. Além disso, um mapa de calor foi utilizado para visualizar a distribuição geográfica das reclamações, revelando insights sobre a satisfação dos clientes em diferentes regiões. A figura 3 apresenta o diagrama arquitetural do sistema e o seu esquema de funcionamento.

Os resultados demonstram o potencial da plataforma Reclame Aqui como ferramenta para empresas monitorarem a satisfação dos seus clientes, identificarem áreas problemáticas e direcionarem estratégias de melhoria.

Como trabalho futuro, o autor pretende aprimorar a análise com a inclusão de um mapa de calor focado no tópico "garantia". Esta funcionalidade permitirá identificar regiões com maior índice de reclamações relacionadas à garantia de serviços ou produtos específicos, fornecendo informações valiosas para a gestão logística e o relacionamento com o cliente.

**Figura 2: Esquema de funcionamento da plataforma Reclame Aqui**

A diagram of a product

Description automatically generated

**Fonte: Antonio, Fabio e Gustavo**

**Figura 3: Arquitetura do Sistema**

A diagram of a computer process

Description automatically generated

**Fonte: Antonio, Fabio e Gustavo**

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

Este trabalho tem como objetivo realizar a modelagem de tópicos na análise de reclamações de consumidores no setor de e-commerce brasileiro (CERQUEIRA; ELOY, 2023). Através de técnicas de Processamento de Linguagem Natural (PNL) e *webscraping* utilizando o *framework BeautifulSoup*, dados públicos foram coletados da plataforma Reclame Aqui e armazenados em um banco de dados estruturado.

Para isso, três modelos de modelagem de tópicos foram implementados e comparados: *Latent Semantic Indexing* (LSI), *Latent Dirichlet Allocation* (LDA) e *BERTopic*. A avaliação dos modelos considerou a relevância dos termos extraídos, a visualização da relação entre os tópicos através de mapas de distância e a análise da coerência dos tópicos gerados. A figura 4 mostra a relevância de cada palavra para cada tópico.

Os resultados demonstram que a modelagem de tópicos, especialmente com os modelos *BERTopic* e LDA, é uma ferramenta poderosa para identificar os principais motivos de reclamações no *e-commerce* brasileiro. Embora o modelo LSI tenha apresentado limitações na identificação precisa de tópicos, os modelos *BERTopic* e LDA se mostraram eficazes na extração de *insights* relevantes sobre a satisfação dos consumidores. A figura 5 apresenta a média para a pontuação de coerência para cada modelo obtida a partir dos valores de coerência.

As descobertas deste trabalho oferecem às empresas do setor de *e-commerce* uma oportunidade valiosa para entender as principais causas de insatisfação entre seus clientes. Ao abordar proativamente essas preocupações, as empresas podem melhorar a experiência do cliente, fortalecer sua reputação e proteger seu valor de mercado.

Como pesquisa futura, o autor pretende se aprofundar na análise comparativa, incluindo empresas com diferentes níveis de reputação (alto e baixo ranking) para identificar padrões específicos e melhores práticas.

**Figura 4 - Palavras mais relevantes para cada tópico do modelo**

A group of bar graph

Description automatically generated with medium confidence

**Fonte: Resultados originais da pesquisa**

**Figura 5 - Média dos valores de coerência**

A white rectangular object with black text

Description automatically generated

**Fonte: Resultados originais da pesquisa**

# METODOLOGIA

O desenvolvimento do projeto foi separado em duas etapas. Na primeira parte, foi feita a realização do *webscraping*, a técnica utilizada para extrair dados de páginas da web e transformá-los em um formato estruturado. Em seguida, foi realizado a modelagem de tópico, com o processamento da linguagem natural até o modelo de LDA.

Nesta primeira parte, foi desenvolvido um *script* que fosse capaz de percorrer as páginas dos comentários do site Reclame Aqui através da Empresa Enel. Foi utilizado a linguagem *Python* e as bibliotecas *time*, *random*, *pandas* e *selenium*.

Detalhando melhor sobre a utilização de cada uma delas durante o processo de *webscraping*, a biblioteca “*time*” foi usada para controlar a execução do código, definindo pausas e atrasos (*sleep*) entre as ações do navegador, ou seja, evitando sobrecarga do servidor, espera no carregamento do conteúdo e simulação do comportamento humano. Em conjunto, a biblioteca “*random*” é para variar o tempo de espera, tornando o comportamento mais imprevisível e menos detectável para evitar bloqueios de sites que identificam padrões de solicitações de um mesmo *user-agent.*

A biblioteca “*Selenium*” é para automatizar as interações com o navegador, ou seja, interagir com elementos da página, clicar em botões, rolar a página e esperar o carregamento completo do conteúdo. Dentro do *selenium*, foram importados alguns componentes importantes. O “*webdriver*”, que permite a interação com o navegador *web*. O “*ChromeOptions*”, que personaliza o comportamento do navegador, que identifica o navegador e sistema operacional do usuário. O “*WebDriverWait*” permite esperar até que um elemento específico da página esteja disponível para interação e o “*expected\_conditions*” define o critério para que a espera termine, ou seja, até que um elemento seja clicável ou visível. Dessa maneira, ajuda a evitar erros devido a elementos que ainda não foram carregados. Por fim, o “*ActionChains*” permite simular ações do *mouse*, como movimentos, cliques e rolagem.

As exceções, tais como “*TimeoutException*”, “*NoSuchElementException*”, dentre outras, permitem que o código lide com situações que podem ocorrer durante o *scraping*, como *timeout* de espera, elementos não encontrados ou elementos que se tornaram inválidos. A biblioteca “*Pandas*” foi utilizada para manipulação de “*dataframes*” após a extração dos dados.

Este *webscraping* criado navega por várias dessas páginas fazendo a coleta e salvando as informações necessárias para tomar como base a análise realizada.

**Tabela 1 – Descrição dos dados coletados no website Reclame Aqui.**

|  |  |
| --- | --- |
| Campos | Descrição |
| location | Cidade e Estado do usuário |
| date | data e hora da reclamação |
| id | id de referência do usuário |
| category | categoria do problema |
| product | tipo do produto |
| problem | tipo do problema |
| complaint | texto da reclamação do usuário |
| response | tipo da resposta |

**Fonte – Autor do projeto**

O conjunto de dados possui temas como: luz, transformador, gerador, fios, pagamentos, atendimento, dentre outros. A Tabela 1 descreve os dados que foram utilizados na extração da plataforma e como eles foram descritos. Nessa etapa de *webscraping* dos dados, foram extraídos em torno de 3000 linhas de comentários referentes a empresa e em seguida armazenados em um arquivo *Excel Spreadsheet* (XLSX). A Figura 6 mostra o esquema realizado das etapas de *scraping* dos dados.

**Figura 6 – Scraping dos dados**

A diagram of a webpage

Description automatically generated

**Fonte – Autor do projeto**

A segunda etapa a é do pré-processamento dos dados, para preparar o texto bruto (os comentários dos usuários), que foram extraídos pelo *webscraping*, para ser usado em tarefas de processamento de linguagem natural. Foi utilizado a linguagem *Python* e algumas bibliotecas para realizar os seguintes procedimentos: manipulação de *dataframes*, PLN, normalização dos textos, lematização, modelo LDA e gráficos. Será detalhado cada parte do pré-processamento até a finalização do modelo.

Esta etapa de pré-processamento é importante realizar todos os passos para que elimine os ruídos e as inconsistências, tornando os dados mais limpos e precisos. Para isso, foi utilizado a biblioteca do kit de ferramentas de linguagem natural (NLTK - *Library Natural Language Toolkit*) para prepará-los para análise. Foi importado o componente “*stopwords*” das palavras portuguesas para que não sejam utilizadas no processamento.

O primeiro passo foi converter todas as palavras da coluna “*complaint*” do *dataframe* para minúsculo. Foi utilizado uma biblioteca de expressão regular para remover os caracteres especiais e números. Em seguida, fui utilizado a biblioteca “*enelvo*” para normalizar os textos para corrigir abreviações, gírias, erros ortográficos, dentre outras funcionalidades. Para remover caracteres *Non-ASCII* (acentuações) das palavras, foi utilizado a biblioteca “*unicodedata*”.

Com as *stopwords* importadas, foi realizado a remoção dessas palavras no tratamento dos dados. Em seguida, foi feito a lematização. Para isso, foi utilizado a biblioteca “*spacy*” que oferece modelos pré-treinados para várias línguas, inclusive em português. Dessa maneira, foi feita a transformação dos dados em palavras individuais, conhecidas como *tokens*. A Figura 7 mostra o *dataframe* das primeiras linhas comparando o texto original das reclamações com o pré-processamento dos dados.

**Figura 7 – Reclamação do texto original e o pré-processamento**

A close-up of a newspaper

Description automatically generated

**Fonte – Autor do projeto**

O próximo passo feito no pré-processamento foi a criação do corpus e o dicionário. Neste contexto, o dicionário é como um mapa que associa cada palavra única no seu conjunto de dados (corpus) a um identificador numérico, facilitando o processamento do LDA, pois transforma as palavras em números, que são mais eficientes para o modelo. O corpus é a representação numérica do seu conjunto de dados, ou seja, é uma lista de documentos, com cada documento representado como uma lista de pares (ID da palavra, contagem) e para realizar a análise de tópicos, foi utilizado a biblioteca “*gensim*.

Na sequência, foi feita a modelagem de bigramas e trigramas das palavras. A biblioteca *gensim* apresenta o componente *Phrases* que realiza essa implementação.

Em seguida, foi realizado a criação do modelo de LDA. Alguns parâmetros são importantes para definir o melhor modelo a ser criado. O número de tópicos representa uma combinação de palavras-chave e cada uma possui uma contribuição com um nível de importância diferente (peso). Na Tabela 2 apresenta os valores utilizados pelo modelo LDA.

**Tabela 2 – Valores do LDA**

|  |  |
| --- | --- |
| Campos | Valores |
| num\_topics | 3 – 27 |
| random\_state | 60 |
| update\_every | 1 |
| chunksize | 100 |
| passes | 2 |
| alpha | auto |
| per\_word\_topics | True |

**Fonte – Autor do Projeto**

É preciso incluir o corpus e o dicionário que foram criados anteriormente. O *random\_state* é um parâmetro que define a semente aleatória para o modelo LDA. Ao definir uma semente, garante que os resultados do modelo sejam repetíveis. O *chunksize* define o tamanho do bloco de documentos que será usado para atualizar o modelo. O *passes* define quantas vezes o modelo será treinado sobre o corpus inteiro. O *alpha* controla a concentração de tópicos definindo o parâmetro de Dirichlet para a distribuição de tópicos de cada documento. O *per\_word\_topics* indica se deseja obter a distribuição de tópicos para cada palavra no corpus. Caso *true*, o modelo também fornecerá a probabilidade de cada palavra pertencer a cada tópico.

O próximo passo é calcular a perplexidade e coerência do modelo. Neste contexto, a perplexidade é uma medida de quão bem o modelo LDA se ajusta aos dados. Basicamente, ela mede a capacidade do modelo de prever novos documentos. Uma perplexidade baixa indica que o modelo se ajusta bem aos dados e é capaz de prever novos documentos com maior precisão. Uma perplexidade alta sugere que o modelo não está se ajustando bem aos dados. A coerência é uma medida da interpretabilidade e inteligibilidade dos tópicos gerados pelo modelo LDA. Ela avalia se as palavras dentro de um tópico são semanticamente relacionadas e se o tópico como um todo faz sentido. Uma pontuação de coerência alta indica que os tópicos são bem definidos e compreensíveis. Uma pontuação baixa pode indicar que os tópicos são confusos ou que as palavras dentro de um tópico não estão relacionadas semanticamente.

A escolha do número de tópicos é crucial para a qualidade do modelo. O modelo pode ser muito genérico e não capturar nuances importantes nos seus dados caso o número seja baixo. O modelo pode ser muito específico e criar tópicos irrelevantes ou sobrepostos com um valor alto. Foi criado uma função que tem função como objetivo automatizar o processo de testar diferentes números de tópicos para encontrar o melhor número de tópicos para o modelo LDA e que calcula a coerência para cada modelo.

**Figura 8 – Gráficos de melhor valor de coerência**

A line graph with numbers and a white background

Description automatically generated

**Fonte – Autor do Projeto**

Pelo gráfico da Figura 8, apresenta os valores de coerência. Podemos entender que quanto maior o número de tópicos, maior é o valor de coerência.

Os algoritmos e a métricas utilizadas estão disponíveis em: https://github.com/Pesati/tcc\_pucrs.

# RESULTADOS

A modelagem de tópicos também permite determinar de qual tópico o documento trata. A Figura 9 apresenta as palavras mais frequentes criadas com o uso da biblioteca Spacy de todos os modelos apresentados.

**Figura 9 – Wordcloud do ngrama**

A close-up of words

Description automatically generated

**Fonte – Autor do Projeto**

Analisando os resultados em relação aos dados coletados e análises feitas, é possível responder as perguntas de pesquisa levantadas anteriormente. Com isso, é possível adotar o modelo de LDA para análise de *eWoM* do site Reclame Aqui? O LDA é uma ferramenta poderosa para análise de texto, oferecendo insights sobre temas latentes e relações entre palavras e documentos. Ele atribui probabilidades de cada documento pertencer a cada tópico e, ao mesmo tempo, define a probabilidade de cada palavra pertencer a cada tópico. Desse modo, o modelo encontrou a combinação de tópicos e palavras que melhor se ajustaram aos dados do *webscraping* e do pré-processamento.

Para o próximo resultado, foi definir quais as principais palavras chaves utilizadas nos tópicos? A palavra é representada por uma área proporcional à sua probabilidade no tópico. Desse modo, através das *wordclouds*, ajuda a compreensão e exploração dos resultados do modelo, fornecendo uma representação intuitiva e visualmente atraente dos temas encontrados no corpus de texto.

**Figura 10 – Tópicos e termos mais relevantes**

A group of words in different languages

Description automatically generated

**Fonte – Autor do Projeto**

A Figura 10 consegue separar por tópico os termos mais frequentes para assim poder identificar sobre o que está relacionado a cada problema apresentado pelos usuários.

Dessa forma, os dados da Plataforma Reclame aqui juntamente com todo o pré-processamento e criação do modelo LDA foram significativos para entender o comportamento das reclamações feitas pelos consumidores.

# DISCUSSÃO

O projeto foi realizado baseado somente no modelo de tópicos *Latent Dirichlet Allocation*. O principal ponto de limitação é não ter como base uma comparação com outros modelos para verificar qual seria o mais eficiente para este caso. No entanto, mesmo com esse ponto, o modelo foi muito significativo para a análise e com os resultados apresentados.

O modelo LDA se torna extremamente relevante para empresas que buscam entender as reclamações dos usuários, principalmente por causa da capacidade de identificar temas latentes e insights ocultos dentro dos dados textuais. O LDA pode analisar milhares de reclamações e identificar os principais temas que os clientes estão mencionando.

Dessa maneira, o LDA permite que a empresa direcione uma solução mais técnica para um grupo e um atendimento mais empático para o outro, pois é possível auxiliar na segmentação de clientes com base nos temas das reclamações, permitindo que a empresa ofereça soluções e comunicações personalizadas.

# CONCLUSÃO E TRABALHOS FUTUROS

A presente pesquisa demonstrou a eficácia da modelagem de tópicos como técnica de extração de insights estruturados a partir de extensos conjuntos de dados de reclamações de clientes. Os resultados obtidos evidenciaram a capacidade da técnica em identificar temas recorrentes, revelando as principais causas de insatisfação e as áreas de maior necessidade de atenção por parte das empresas.

A aplicação da modelagem de tópicos permite uma análise aprofundada dos dados de reclamações, possibilitando a identificação de padrões e tendências, além de fornecer uma compreensão abrangente das expectativas e necessidades dos clientes. Essa compreensão facilita a tomada de decisões estratégicas, direcionando ações eficazes para o aprimoramento de produtos e serviços, e culminando em uma melhoria significativa na experiência e satisfação do cliente.

Como perspectiva de trabalhos futuros, poderá ser direcionada para a comparação da performance da modelagem de tópicos com outras técnicas de análise de tópicos, visando ampliar o escopo da análise e validar a robustez dos resultados obtidos. A aplicação da metodologia a outros contextos, englobando diferentes empresas e setores, permitirá a generalização dos resultados e o fortalecimento do potencial da técnica para a análise de reclamações de clientes em diversos cenários.

# REFERÊNCIAS

**REFERÊNCIAS**

WEBSCRAPER. **Brief History of Web Scraping**. 2021. Disponível em: <https://webscraper.io/blog/brief-history-of-web-scraping>.

ALGHIZZAWI, M. A survey of the role of social media platforms in viral marketing: The influence of eWOM – IJITLS, p. 54-60, 2019.

OXFORDBIBLIOGRAPHIES. **Electronic Word-of-Mouth (eWOM)**. 2021. Disponível em: < https://www.oxfordbibliographies.com/display/document/obo-9780199756841/obo-9780199756841-0267.xml>.

GROOTENDORST, M. (2022). **BERTopic: Neural topic modeling with a class-based TF-IDF procedure**. arXiv. Disponível em: <https://doi.org/10.48550/arXiv.2203.05794>

Almeida, G. R. T. d., Lobato, F., and Cirqueira, D. (2017). Improving Social CRM th-

rough eletronic word-of-mouth: a case study of ReclameAqui. In XIVWorkshop de

Trabalhos de Iniciac¸ ˜

ao Cient´

ıﬁc.

Almeida, G. R. T. d., Lobato, F., and Cirqueira, D. (2017). Improving Social CRM th-

rough eletronic word-of-mouth: a case study of ReclameAqui. In XIVWorkshop de

Trabalhos de Iniciac¸ ˜

ao Cient´

ıﬁc.

BLEI, D. M. (2012). **Probabilistic topic models**. Communications of the ACM, 55(4): pp.77\_84

BLEI, David M.; NG, Andrew Y.; JORDAN, M. I. (January 2003). Lafferty, John (ed.). "**Latent Dirichlet Allocation**". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

**MIT**. GRAY, M. K. 1993. Disponível em: https://www.mit.edu/~mkgray/index.3.html

SCRAPINGDOG. **What is Web Scraping? Meaning, Uses, and Legality.** 2023. Disponível em: <https://www.scrapingdog.com/blog/what-is-web-scraping/>

SELENIUM. **Selenium History**. 2004. Disponível em: <https://www.selenium.dev/history>.

ALMEIDA, G. R. T. d., LOBATO, F., and CIRQUEIRA, D. (2017). **Improving Social CRM th-rough eletronic word-of-mouth: a case study of ReclameAqui.** In XIVWorkshop de Trabalhos de Iniciação Científica.

CERQUEIRA, M. R., and ELOY, M. E. **Comparação de abordagens de modelagem de tópicos em reclamações relacionadas ao setor de comércio eletrônico.** 2023. MBA – USP/Esalq