Trabalho de Conclusão de Curso

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

APLICANDO O MODELO LDA PARA A EXTRAÇÃO DE TÓPICOS DE AVALIAÇÕES DE EMPRESAS DE ENERGIA NA PLATAFORMA RECLAME AQUI

ALUNO: André Pesati Revoredo
ORIENTADOR: César Augusto Fonticielha De Rose

PUCRS online uol edtech

Sumário

1. RESUMO

O *eWoM (Electronic word of mouth)* 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. Pode ser definido como qualquer grau ou combinação de comentários, recomendações, ou declarações positivas, negativas ou neutras, sobre empresas, marcas, produtos ou serviços, discutidos ou partilhados entre consumidores em formatos digitais ou eletrônicos.

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.

2. 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.

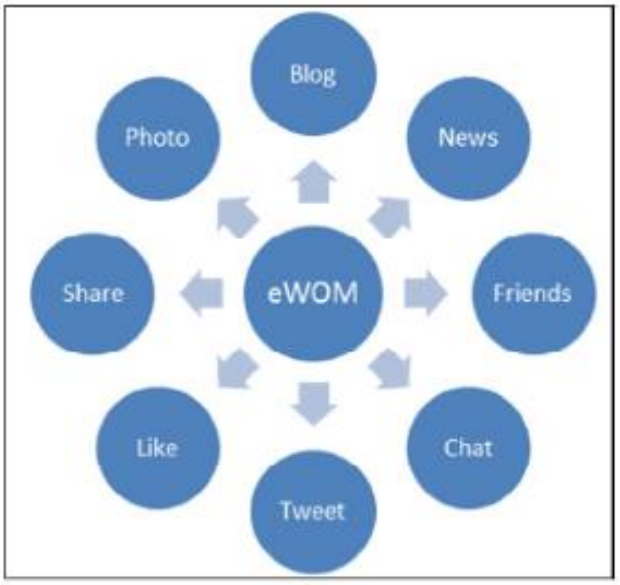

Figura 1 – Tipos de *eWoM* na mídia social
(Mahmoud Alghizzawi)

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).

A Modelagem de Tópicos possibilita realizar a tarefa de resumir e organizar corpus de dados por meio de algoritmos de *Machine Learning* e métodos estatísticos de forma que seja possível descobrir temas e suas relações, como as mudanças dos termos ao longo dos anos (BLEI, 2012; KASZUBOWSKI, 2016).

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 figura 2 explica a modelagem de tópico com LDA no qual cria K tópicos que podem ser vistos como clusters de palavras. Assim, cada documento da coleção é representado como uma distribuição dos tópicos que são eles próprios, distribuições de palavras.

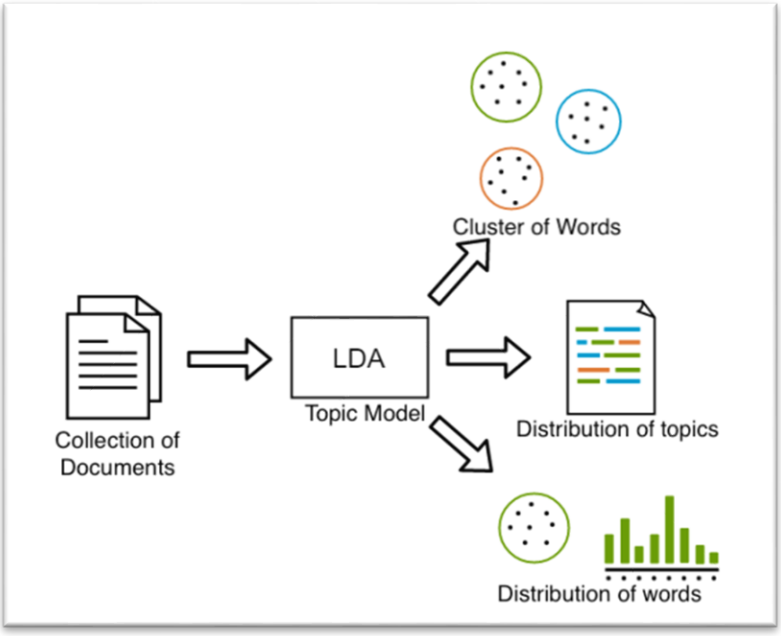

Figura 2 – Modelagem de Tópico com LDA (Yilma)

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*.

O presente trabalho se propõe a investigar a aplicação da modelagem de tópicos para a análise de dados de *eWoM* coletados na plataforma Reclame Aqui, focando nas reclamações da empresa Enel Distribuição - Ceará. O objetivo principal é extrair perspectivas relevantes sobre as principais causas de insatisfação dos clientes da Enel, identificar os tópicos mais frequentes e analisar a eficiência do modelo de LDA para a análise de reclamações nesse contexto.

As principais contribuições do estudo deste trabalho incluem a identificação de temas relevantes, em que a análise revelou os principais tópicos relacionados às reclamações, fornecendo insights sobre as áreas de

maior insatisfação dos clientes. A compreensão da experiência do cliente que permitiu a empresa Enel entender quais aspectos de seus serviços estavam gerando mais problemas e quais áreas necessitam de maior atenção. A validação da técnica, que confirmou a eficiência do modelo LDA para análise de dados de *eWoM*, incentivando sua aplicação em outros contextos e setores. Em suma, o potencial para o desenvolvimento de ferramentas em que os resultados obtidos podem servir como base para o desenvolvimento de ferramentas que automatizem a análise de *eWoM*, auxiliando empresas a monitorar a satisfação dos clientes em tempo real.

Concluindo, 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.

3. TRABALHOS RELACIONAIS

Neste capítulo serão apresentados os trabalhos encontrados sobre Processamento de Linguagem Natural (PLN) 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*.

No trabalho de Gestão de Relacionamento com clientes, os autores (ALMEIDA; CIRQUEIRA; LOBATO, 2017) apresentam uma análise de sentimentos e localização de reclamações de consumidores utilizando técnicas de PLN e visualização de dados. A análise deste trabalho 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 a arquitetura do sistema. Como conclusão, o trabalho dos alunos da UFOPA é mais abrangente e voltado para o desenvolvimento de uma solução tecnológica para análise de dados de *eWoM*, enquanto o objetivo destra trabalho é mais focado em uma análise específica de reclamações de uma empresa utilizando o LDA. Ambos os trabalhos demonstram a utilidade das técnicas de processamento de linguagem natural e modelagem de tópicos para extrair insights de dados de plataformas online.

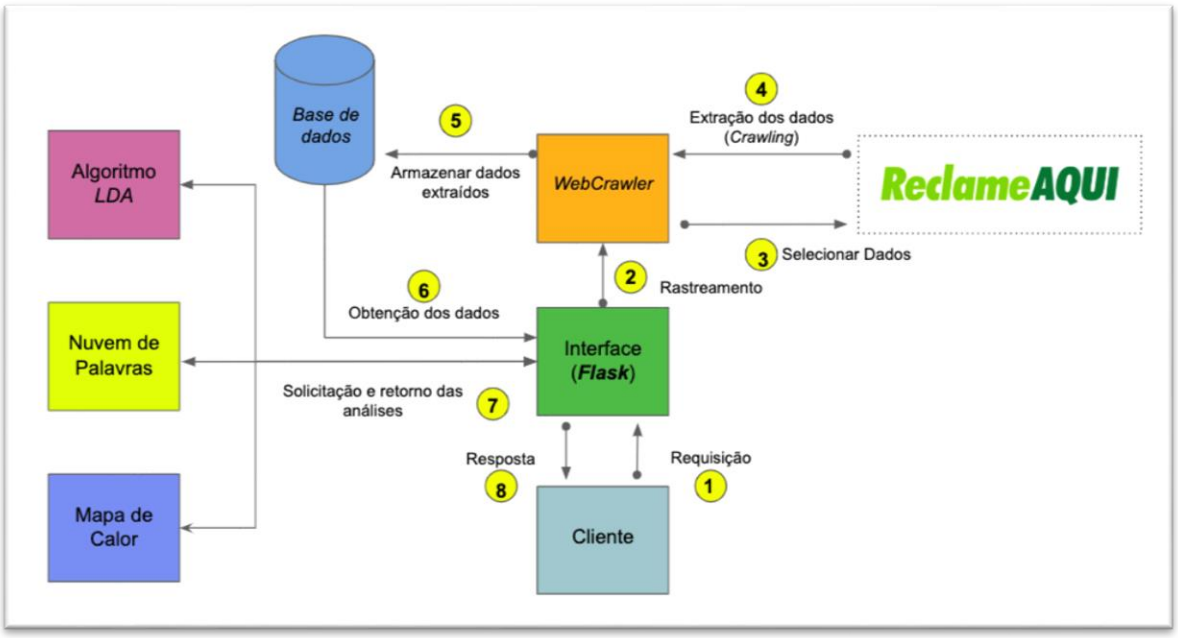

Figura 3: Arquitetura do Sistema
(Antonio, Fabio e Gustavo)

Neste outro trabalho, tem como objetivo realizar a modelagem de tópicos na análise de reclamações de consumidores no setor de e-commerce brasileiro (CERQUEIRA; ELOY, 2023). Para isso, três modelos de modelagem de tópicos foram implementados e comparados: *Latent Semantic Indexing* (LSI), *Latent Dirichlet Allocation* (LDA) e *BERTopic*. A figura 4 apresenta a média para a pontuação de coerência para cada modelo obtida a partir dos valores de coerência. Por fim, o trabalho do Matheus visa comparar o desempenho de três modelos de modelagem de tópicos (LDA*, BERTopic e* LSI) para extrair tópicos relevantes de reclamações no Reclame Aqui, focando em empresas de comércio eletrônico enquanto este trabalho tem como objetivo analisar a aplicabilidade do modelo LDA para análise de *eWoM* em avaliações de empresas de energia no Reclame Aqui.

Modelo	Coerência (Média)
LDA	0.4599
LSI	0.5154
BERTopic	0.6828

Figura 4 - Média dos valores de coerência
(Resultados originais da pesquisa)

Neste próximo trabalho (SOUZA; ROCHA, 2021), tem como objetivo identificar os tópicos de maior relevância nas teses e dissertações produzidas pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, utilizando a modelagem de tópicos com o modelo de LDA (*Latent Dirichlet Allocation*). O trabalho combinou técnicas de processamento de linguagem natural, modelagem de tópicos, análise de frequência e *machine learning* para mapear o conhecimento científico presente nas teses e dissertações do PPGCI da UFMG. Como diferença, o trabalho do Marcos e Renato é focado na análise de documentos acadêmicos e na identificação de temas de pesquisa, enquanto o objetivo deste trabalho se concentra em analisar dados de reclamações de clientes para entender as necessidades e expectativas dos consumidores. Ambos os trabalhos demonstram a utilidade da modelagem de tópicos para extrair insights de grandes volumes de dados.

A seguir, será apresentado a metodologia deste trabalho.

4. METODOLOGIA

Como vimos na seção anterior, este trabalho visa compreender a percepção dos clientes sobre a empresa Enel Distribuição - Ceará através da análise de avaliações na plataforma do Reclame Aqui aplicando o modelo de LDA para a extração de tópicos de comentários.

Para atingir o objetivo do trabalho, foram realizadas algumas etapas:

- Coleta de Dados: foi feita a realização do *webscraping*, uma técnica utilizada para extrair dados de páginas da web e transformá-los em um formato estruturado
- Pré-processamento: O texto bruto dos comentários é limpo e preparado para a análise. Isso inclui remover caracteres especiais, números, stop words (palavras comuns que não carregam significado), realizar a lematização (reduzir as palavras à sua forma base) e a tokenização (separar as palavras em tokens).
- Modelagem de Tópicos com LDA: O modelo LDA é aplicado ao conjunto de comentários pré-processados. Ele identifica os temas latentes (tópicos) presentes nos dados textuais.

Tendo em vista as etapas mencionadas, será detalhado cada uma delas.

4.1 Coleta de Dados

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 (Autor do projeto)

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

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 5 mostra o esquema realizado das etapas de *scraping* dos dados.

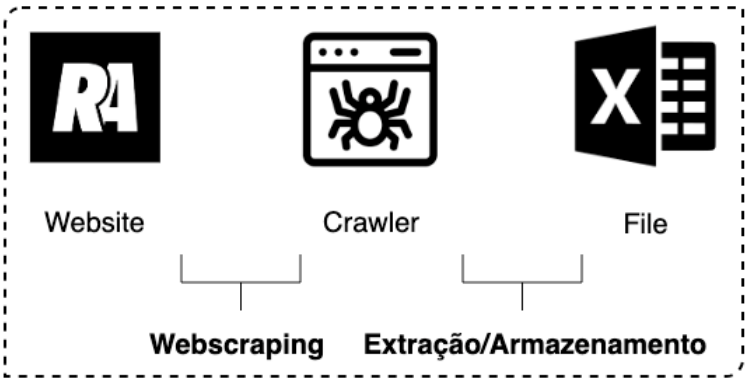

Figura 5 – Scraping dos dados
(Autor do projeto)

4.2 Pré-processamento

A segunda etapa é a 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 e lematização.

Nesta etapa 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 6 mostra o *dataframe* das primeiras linhas comparando o texto original das reclamações com o pré-processamento dos dados.

complaint	preprocessing complaints
Na madrugada de sexta para sábado uma árvore caiu em cima da fiação elétrica, causando fogo e explosão na rede. Deixando moradores sem energia elétrica, uma equipe somente apareceu no domingo, tirou fotos e foram embora alegando não ter equipamento. Na segunda outra equipe veio e somente cortou o tronco da árvore, e foram embora também alegando não ter equipamentos. Desde então não mais retornaram. Estamos sem energia elétrica desde sábado, hoje é quarta e não veio mais nenhuma equipe restabelecer a energia.	madrugar sexta sabar arvore cair cima fiacao eletrico causar fogo explosao rede deixar morador energia eletrico equipe somente aparecer domingo tirar foto embora alegar nao ter equipamento segundo outro equipe vir somente cortar tronco arvore embora tambem alegar nao ter equipamento desde entao nao retornar energia eletrico desde sabar hoje quarto nao vir nenhum equipe restabelecer energia
Já faz mais mais de 72 horas praticamente sem energia pq nada funciona pois e energia fraca desde segunda feira 7 horas da manhã sem energia e hoje já e quarta feira e nada de resolverem já liguei tantas vezes que nem conto mais um total descaso não vou mais me estressar com isso a partir de amanhã irei atrás dos meus direitos vc paga um absurdo por um serviço que não presta.	ja fazer hora praticamente energia porque nada funcionar pois energia fraco desde segundo feira hora mar energia hoje ja quarto feira nada resolver ja liguei tanto vez conto total descaso nao ir estressar partir amanha irar atra direito voce pagar absurdo servico nao presta
Ja procurei no app enel pra tirar essa conta de fatura por imail e nao acho por favor quero receber em maos o minha conta todos os meses na minha recidrncia msm e um debito qur esta no serasa no meu nome que nem devoo nao vou paga algo qur nao devo assim me complica neh botando meu nome no serasa por uma conta qur nao devoo e por favor que ajeitem essas duas questoens logo ..	ja procur appplication anel pra tirar conta fatura underline nao achar favor querer receber mao contar todo mês residenciar debito serasa nome devoo nao ir pagar algo nao devo assim complico ne botar nome serasa contar nao devoo favor ajeitar dois questoe logo
Entrei em contato com a Enel,pela primeira vez fiz a solicitação de primeira ligação,eles deram o prazo de 5 dias úteis,mais eles não apareceram, entrei em contato novamente,mim deram mais 5 dias úteis,sendo que já tinha mim dado 5 dias , deram mais 5 novamente, resumindo estou a quase 15 dias no escuro,com minha esposa,meu filho de 8 anos que é neuro divergente, estamos passando uma barra no escuro sem energia elétrica,ainda sem nenhum resposta, queria pelo amor de deus que alguém ou algum órgão pudesse mim ajudar.	entrar contato anel primeira vez fiz solicitacao primeiro ligacao dar prazo dia utel nao aparecer entrar contato novamente eu dar dia utel ser ja eu dar dia dar novamente resumir quase dia escuro esposo filho ano neuro divergente passar barra escuro energia eletricar ainda nenhum resposta querer amor deus alguir algum orgao poder eu ajudar
Precisei fazer uma consulta no serasa e vi que estava com uma dívida de pouco mais de 18 reais que não lembro de que foi (acredito que foi cobrança [Editado pelo Reclame Aqui]). Paguei essa dívida pelo site do governo federal no programa desenrola. Esperei os cinco dias úteis e simplesmente a Enel não retirou meu nome do Serasa. Isso é um absurdo! Uma empresa que faz uma cobrança de nem sei o que e ainda não retira o nome da inadimplência mesmo depois de pago. Empresa irresponsável e desorganizada.	precisar fazer consulta serasa ver divir pouco real nao lembro acreditar cobranca editar reclame aqui paguei divir site governo federal programa desenrolar esperei cinco dia utei simplesmente anel nao retirar nome serasa absurdo empresa fazer cobranca saber ainda nao retirar nome inadimplencia pagar empresa irresponsavel desorganizar

Figura 6 – Reclamação do texto original e o pré-processamento (Autor do projeto)

4.3 Modelagem de Tópicos com LDA

O próximo passo 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 (Autor do Projeto)

Campos	Valores
num_topics	3 – 27
random_state	60
update_every	1
chunksize	100
passes	2
alpha	auto
per_word_topics	True

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.

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

5. RESULTADOS

A modelagem de tópicos também permite determinar de qual tópico o documento trata. A Figura 7 apresenta as palavras mais frequentes criadas com o uso da biblioteca *Spacy* de todos os modelos apresentados. A utilização das *wordclouds* organiza as palavras de acordo com sua frequência em cada tópico, mostrando de forma gráfica quais termos são mais importantes para definir cada tema. As palavras mais frequentes aparecem maiores e mais destacadas e facilita a interpretação do modelo LDA, tornando o processo mais intuitivo e fácil de entender. A visualização gráfica ajuda a perceber padrões e relações entre as palavras que podem ser difíceis de detectar apenas com a análise de dados numéricos.

Figura 7 – Wordcloud do ngrama
(Autor do Projeto)

Pelo gráfico da Figura 8, apresenta os valores de coerência. Podemos entender que ocorre alto valor coerência em torno de 15 tópicos, porém o maior pico ocorre por volta de 40 tópicos.

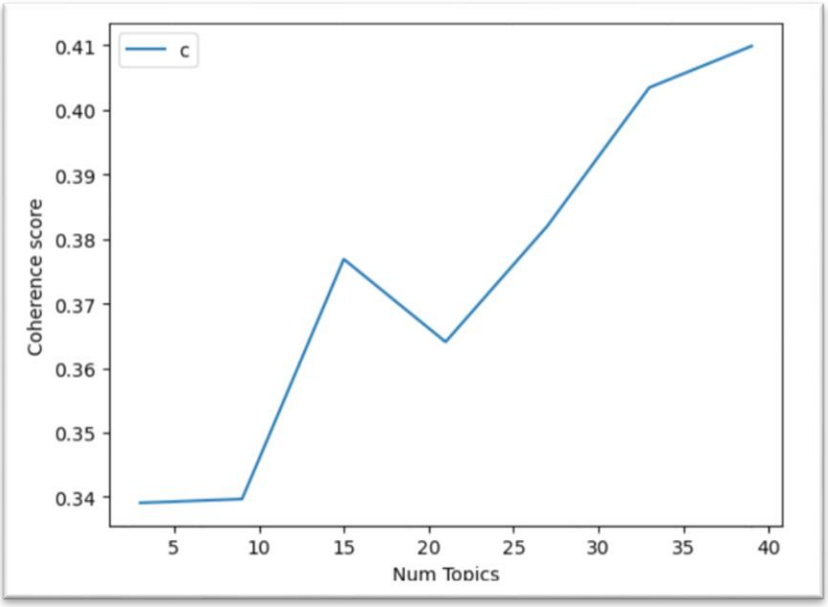

Figura 8 – Gráficos de melhor valor de coerência por número de tópicos (Autor do Projeto)

A perplexidade foi calculada utilizando o modelo LDA previamente criado, aplicando o corpus à função logarítmica da perplexidade. O valor da perplexidade é apresentado na Figura 9. A Figura 10, por sua vez, ilustra o gráfico de tópicos, mostrando os tópicos e seus respectivos valores.

```
# Calculando a perplexidade e o score de coerência para avaliar a qualidade de um modelo LDA

#print('\nPerplexity: ', lda_model.log_perplexity(corpus))

print('\nPerplexity: ', np.exp(lda_model.log_perplexity(corpus)))

coherence_model_lda = CoherenceModel(model=lda_model, texts = bigrams, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)

Perplexity:  0.0006666800720308504
```

Figura 9 – Valor da Perplexidade (Autor do Projeto)

```
# lista dos valores de coerência, para melhor identificar o ponto de inflexão do gráfico

for m, cv in zip(x, coherence_values):
     print("A quantidade de tópicos =", m, " tem um valor de coerência de ", round(cv, 4))
```

```
A quantidade de tópicos = 3  tem um valor de coerência de  0.3391
A quantidade de tópicos = 9  tem um valor de coerência de  0.3397
A quantidade de tópicos = 15  tem um valor de coerência de  0.3769
A quantidade de tópicos = 21  tem um valor de coerência de  0.3641
A quantidade de tópicos = 27  tem um valor de coerência de  0.382
A quantidade de tópicos = 33  tem um valor de coerência de  0.4034
A quantidade de tópicos = 39  tem um valor de coerência de  0.4099
```

Figura 10 – Tópicos e valores de coerência (Autor do Projeto)

Quando a coerência é alta, as palavras dentro de um tópico são semanticamente relacionadas. A visualização do *wordcloud* para cada tópico mostra que as palavras mais importantes fazem sentido juntas, formando um conceito claro e coeso. Tabela 3 mostra a descrição dos hiper parâmetros do modelo LDA.

Tabela 3 – Descrição da Tabela de coerência do Modelo LDA (Autor do Projeto)

Campos	Valores
num_topics	Número de tópicos
random_state	Define a semente aleatória utilizada para inicializar o modelo LDA.
update_every	Determina a periodicidade com que o modelo é atualizado com base no processamento de um número específico de documentos
chunksize	Define o tamanho do bloco de documentos que será processado em cada iteração de treinamento do modelo LDA
passes	Define quantas vezes o modelo LDA irá iterar sobre todo o corpus de dados durante o treinamento
alpha	Controla a concentração de tópicos em cada documento. É um parâmetro de Dirichlet que define a distribuição de tópicos de cada documento

per_word_topics	Especifica se o modelo LDA deve retornar a distribuição de tópicos para cada palavra no corpus.

Na Tabela 4 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 energia, empresa, problema, pagar, valor, dentre outros.

Tabela 4 – Descrição da Tabela de coerência do Modelo LDA (Autor do Projeto)

Tópico	Palavras
0	Energia, empresa, fazer, hoje, vir
1	Energia, problema, valor, empresa, resolver
2	Energia, pagar, empresa, problema, hoje
3	Energia, pagar, valor, empresa, resolver

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 11 – Tópicos e termos mais relevantes
(Autor do Projeto)

A Figura 11 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.

6. DISCUSSÃO

O projeto foi realizado baseado somente no modelo de tópicos *Latent Dirichlet Allocation*. A aplicação do modelo LDA na análise das reclamações contra a Enel Distribuição - Ceará permitiu identificar os principais tópicos e termos relacionados à insatisfação dos clientes, fornecendo informações cruciais para a empresa. Essa análise revelou as áreas que mais necessitam de atenção, como falhas no fornecimento de energia, problemas com o atendimento e cobranças indevidas. Com base nesses insights, a Enel pode direcionar suas ações para resolver os problemas mais frequentes, melhorando a qualidade do serviço e, consequentemente, sua reputação online. Ao atender as necessidades dos clientes de forma eficiente, a empresa pode fortalecer o relacionamento com eles e construir uma imagem mais positiva no ambiente digital.

Alguns pontos de limitações podem ser observados no modelo LDA. Apesar de fornecer informações cruciais, a análise do modelo exige a interpretação humana dos resultados. Foi preciso analisar os tópicos e termos relevantes para entender o contexto e gerar conclusões significativas. O LDA pode ter dificuldade em lidar com nuances da linguagem, como sarcasmo ou ironia, que podem influenciar a percepção do cliente, mas não são capturados pela análise. O LDA considera apenas o texto dos comentários, sem levar em consideração outros dados relevantes, como a data da reclamação, a localização do cliente e a resposta da empresa. Isso limita a profundidade da análise e a capacidade de gerar conclusões mais completas.

Para o projeto, um ponto de limitação importante foi em 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 possui aplicações diversas além da análise de reclamações de empresas de energia. Ele pode ser utilizado para compreender a opinião pública, ou seja, analisando comentários em redes sociais sobre

produtos, serviços, políticos ou eventos. Identificar tendências emergentes em setores ou mercados específicos. Analisar conteúdo de notícias identificando os temas mais relevantes para facilitar o acompanhamento de informações. Organizar literatura científica identificando os temas de pesquisa mais relevantes em uma área específica, auxiliando na organização da literatura e na identificação de gaps de pesquisa.

O modelo 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. 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.

7. CONCLUSÃO E TRABALHOS FUTUROS

No presente trabalho foi analisado os comentários dos clientes da Enel Distribuição - Ceará no Reclame Aqui para entender a percepção deles sobre a empresa. Para isso, foi utilizado uma técnica de modelagem de tópicos LDA (*Latent Dirichlet Allocation*), que identifica os temas principais presentes em um conjunto de textos. O processo se iniciou com a coleta de todos os comentários sobre a empresa na plataforma do Reclame Aqui, usando uma técnica de *webscraping* que extrai dados de sites automaticamente. Em seguida, esses dados passaram por um processo de limpeza para remover caracteres especiais, números, palavras muito comuns e outras informações irrelevantes.

Após a limpeza, os comentários foram analisados pelo modelo LDA, que identificou os temas mais frequentes nas reclamações. Para visualizar esses temas, foi utilizado as *wordclouds*, que são representações visuais das palavras mais frequentes em cada tópico. Com isso, foi possível identificar quais eram as principais áreas de insatisfação dos clientes da Enel, como falhas no fornecimento de energia, problemas com o atendimento e cobranças indevidas. Essas informações são valiosas para a empresa, pois permitem que ela compreenda melhor as necessidades e expectativas dos clientes e direcione suas ações para melhorar a qualidade do serviço e aumentar a satisfação dos clientes.

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, pode ser feito a incorporação de dados adicionais, tais como a data da reclamação, a localização do cliente e a resposta da empresa. Isso permitiria uma análise mais profunda e

contextualizada. Comparar com outros modelos, como o BERTopic e o LSI, poderia validar os resultados e oferecer dados adicionais.

Analisar a evolução temporal dos tópicos, ou seja, realizar em diferentes períodos permitiria identificar a evolução das áreas de insatisfação e a efetividade das ações tomadas pela Enel. Combinar com outras técnicas como a análise de sentimentos poderia fornecer uma visão mais completa sobre a percepção dos clientes.

A aplicação da metodologia a outros contextos pode contribuir para fortalecer as conclusões e aumentar a aplicabilidade do estudo em diferentes cenários.

8. REFERÊNCIAS

REFERÊNCIAS

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

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.

B. A. Yilma, Y. Naudet, and H. Panetto, **Personalised Visual Art Recommendation by Learning Latent Semantic Representations** in Conference: 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP2020) On the Move to Meaningful Internet Systems". Conference Paper, 2020.

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.

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

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>

KASZUBOWSKI, Erikson. **Modelo de tópicos para associações livres**. 2016. 213f. Tese (Doutorado em Psicologia) -Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2016. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/172577/343427.pdf?sequence=1. Acesso em: 1 mar. 2019

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

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

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>.

SOUZA, M., and SOUZA, R. R., **Mapeamento de conhecimento científico: modelagem de tópicos das teses e dissertações do Programa de Pós-Graduação em Ciência da Informação da UFMG.** 2021. UFMG

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