**A blue square with white lines

AI-generated content may be incorrect.**

Projeto Prático - Especificação da 1ª Etapa: Solução Simbólica

Disciplina: SCC0633-SCC5908 - PLN Linguagem Natural

**Os Segredos da Análise de Sentimentos:**

**Um Estudo Aplicado as Lojas Americanas**

Érica Ribeiro[[1]](#footnote-1), Júnior Fernandes Marques[[2]](#footnote-2), Luís Vogel[[3]](#footnote-3), Marlon José Martins[[4]](#footnote-4), Raphael Franco Chaves[[5]](#footnote-5), Thiago Ambiel[[6]](#footnote-6)

ICMC-USP

# **1 Introdução**

A análise de sentimentos tem se tornado uma ferramenta essencial para compreender as opiniões e percepções dos consumidores em relação a produtos e serviços, especialmente no contexto do e-commerce. Neste trabalho, será realizada uma investigação aprofundada utilizando o Córpus "B2W-Reviews01.csv", disponibilizado pelas Lojas Americanas. Através da análise das avaliações contidas neste Córpus, buscamos não apenas classificar os sentimentos expressos pelos consumidores, mas também explorar as nuances dessas opiniões, a fim de identificar padrões e tendências que podem informar estratégias de marketing e desenvolvimento de produtos. Este estudo pretende contribuir para uma compreensão mais profunda das dinâmicas de consumo na era digital, revelando insights que podem beneficiar tanto as empresas quanto os consumidores.

# **2 Informações sobre o Córpus Escolhido**

O Córpus escolhido para o desenvolvimento do presente trabalho corresponde a base de dados e informações “B2W-Reviews01.csv” disponibilizada pelas Lojas Americanas. Tal base de dados e informações encontra-se disponível no Github oficial das Americanas-tech[[7]](#footnote-7), e corresponde a um Córpus aberto de avaliações de produtos ofertados pelas Lojas Americanas na sua plataforma de e-commerce. Este Córpus contempla mais de 130 mil avaliações de clientes, as quais foram coletadas do site Americanas.com entre os meses de janeiro e maio de 2018. Adicionalmente, este Córpus oferece informações importantes para a análise exploratória de dados contemplando o perfil do avaliador, como gênero, idade e localização geográfica. O Córpus também apresenta duas diferentes taxas de avaliação providas por clientes, dentre as quais se destacam:

1. Uma escala de avaliação entre 1 e 5 pontos, em que: 1 = Ruim, 2 = Regular, 3 = Bom, 4= Ótimo e 5 = Excelente; e
2. Uma pergunta com as alternativas "Sim“ ou “Não" que representa a disposição do cliente em recomendar o produto a outra pessoa.

# **3 Arquitetura da Solução Proposta**

O diagrama ilustrado abaixo, contempla a arquitetura da solução proposta pelo grupo para a Fase 1. Adicionalmente, apresentamos um resumo da arquitetura planejada referente a Fase 2.

A screenshot of a computer

AI-generated content may be incorrect.

Fonte - Elaborado pelos autores.

# **4 Coleta de Dados**

A coleta de dados corresponde a um processo de extrema importância para o desenvolvimento de qualquer modelo simbólico eficaz. Neste capítulo, detalharemos o processo através do qual os dados foram obtidos, processados e preparados para análise. Com base no que foi discutido no capítulo anterior, utilizamos o repositório oficial da Americanas-tech no Github como a principal fonte de dados. Para garantir a precisão e a eficiência na extração dos dados, foram desenvolvidos códigos específicos na linguagem Python, proporcionando uma coleta automatizada e sistemática. Esses códigos, juntamente com a documentação detalhada, estão disponíveis no material anexo, permitindo a reprodução e a verificação dos métodos utilizados.

# **5 Pré-processamento dos Dados**

O pré-processamento dos dados pode ser considerada uma das fases mais importantes no desenvolvimento de modelos preditivos, a qual envolve várias atividades essenciais para garantir a qualidade e a eficácia do modelo. Os capítulos a seguir, detalham as principais etapas efetuadas.

### **5.1 Limpeza dos Dados**

No processo de limpeza dos dados, visamos garantir que as informações utilizadas no desenvolvimento do nosso modelo simbólico estivessem em um formato adequado para análise, sem ruídos que pudessem comprometer os resultados finais da nossa análise. Dentre os processos de limpeza de dados efetuados, destacam-se:

1. Valores Nulos: Primeiramente, removemos todas as observações que continham valores nulos em qualquer uma das colunas. A função dropna() foi utilizada para este propósito, garantindo que apenas registros completos fossem considerados na análise.
2. Dados Duplicados: Em seguida, valores duplicados foram removidos com o objetivo de evitar a redundância de informações. Isso foi feito utilizando a função drop\_duplicates(), mantendo apenas a primeira ocorrência de cada registro duplicado.
3. Padronização: Para padronizar os textos das avaliações dos clientes, todos os caracteres foram convertidos para minúsculas. Isso ajuda a evitar a distinção entre palavras que deveriam ser consideradas iguais, independentemente de estarem em maiúsculas ou minúsculas.
4. Pontuações: Os caracteres de pontuação dos textos das avaliações foram substituídos por espaços. Isso foi feito para garantir que apenas palavras e espaços fossem mantidos, facilitando a análise subsequente.
5. Números: Por fim, todos os números presentes nos textos das avaliações foram removidos. Números geralmente não contribuem para a análise de sentimentos e podem introduzir ruídos nos dados.

### **5.2 Nuvem de Palavras**

A Nuvem de Palavras corresponde a uma representação visual que destaca as palavras mais frequentes em um conjunto de textos, onde o tamanho de cada palavra é proporcional à sua frequência. A imagem a seguir, contribuiu para que o nosso grupo pudesse identificar rapidamente as palavras mais comuns nos comentários dos clientes, proporcionando insights visuais sobre os temas e sentimentos predominantes:

A close up of words

AI-generated content may be incorrect.

Fonte - Elaborado pelos autores.

Para o desenvolvimento da nuvem de palavras ilustrada acima utilizamos a biblioteca WordCloud do Python.

### **5.3 Stop Words**

Em projetos de análise de sentimentos, a etapa de remoção de stop words é uma etapa fundamental do pré-processamento de texto. Via de regra, as stop words correspondem a palavras que, em geral, não agregam significado relevante ao contexto do texto quando analisado em um modelo de análise de sentimentos. No Córpus utilizado para o desenvolvimento deste trabalho foram identificadas 177 diferentes stop words. Ao efetuarmos uma análise detalhada de cada stop word, entendemos que a única stop word que deveria ser mantida correspondia a palavra “Não”, dado que tal stop word foi comumente utilizada pelos consumidores para expressar um sentimento negativo referente a um determinado produto. No processo de remoção de stop words foram consideradas as funcionalidades providas pela biblioteca NLTK.

### **5.4 Tokenização**

A Tokenização corresponde ao processo de dividir um texto em unidades menores, chamadas de "tokens". Esses tokens podem ser palavras, frases ou até mesmo caracteres, dependendo do nível de tokenização que se deseja aplicar. No nosso trabalho utilizamos a função word\_tokenize da biblioteca NLTK para dividir os comentários providos pelos clientes em palavras individuais. Tal processo permitiu ao nosso grupo manipular e analisar os comentários dos clientes de maneira mais granular, o que é fundamental para extrair informações significativas e relevantes.

# **6 Análise Exploratória dos Dados**

A análise exploratória dos dados (EDA) é uma etapa fundamental em um projeto de análise de sentimentos, pois permite uma compreensão profunda e inicial do conjunto de dados disponível. Primeiramente, 2 diferentes visualizações de dados foram desenvolvidas, as quais nos auxiliaram a entender a distribuição dos ratings atribuídos pelos clientes, bem como as 10 palavras mais frequentes utilizadas.

|  |  |
| --- | --- |
| A graph with different colored bars  AI-generated content may be incorrect. | A graph with a line going up  AI-generated content may be incorrect. |

Fonte - Elaborado pelos autores.

Adicionalmente, foi desenvolvido um histograma contendo a distribuição do comprimento dos comentários após o pré-processamento. Este histograma nos permitiu identificar como os tamanhos dos comentários variam, fornecendo uma visão clara da densidade e dispersão dos dados. Ao visualizarmos a distribuição dos tamanhos dos comentários, pudemos identificar padrões, como a prevalência de comentários curtos ou longos. Isso pode indicar a natureza das interações dos clientes, como comentários mais curtos talvez indicando feedback rápido e direto, enquanto os mais longos podem conter opiniões mais detalhadas. A linha de média adicionada ao gráfico fornece uma métrica estatística fundamental, permitindo comparar a média do tamanho dos comentários com a distribuição geral. Isso ajuda a contextualizar os dados e pode ser útil para identificar desvios significativos.

A graph with a black line

AI-generated content may be incorrect.

Fonte - Elaborado pelos autores.

# **7 Processo de Lematização**

O processo de lematização tem como principal objetivo reduzir as palavras em um texto à sua forma base ou raiz, conhecida como *lemma*, enquanto preserva o contexto e a integridade semântica das palavras. A lematização busca retornar palavras que realmente existem na língua, levando em consideração a classe gramatical de cada termo. Entendemos que este aspecto é de extrema importância para o nosso projeto de análise de sentimentos, onde a precisão na compreensão dos significados das palavras é crucial para a interpretação correta dos sentimentos expressos nos textos. Para o desenvolvimento desta etapa utilizamos a biblioteca spaCy do Python,

# **8 Definição da Variável Dependente**

No contexto da análise de sentimentos, a variável dependente binária adotada neste modelo é representada pelo atributo "recommend\_to\_a\_friend", que classifica as respostas em duas categorias: "Yes" e "No". Para a construção do modelo, essas categorias foram mapeadas para valores numéricos binários, onde "Yes" foi codificado como 0 e "No" como 1. Essa abordagem permite que o modelo identifique e analise a polaridade das opiniões dos usuários, facilitando a interpretação dos resultados. A nova variável 'target' foi então incorporada ao conjunto de dados, possibilitando a realização de análises mais aprofundadas sobre a disposição dos consumidores em recomendar um produto ou serviço a um amigo. A seguir, apresentamos a distribuição desta variável:

A graph with a bar and a number

AI-generated content may be incorrect.

Fonte - Elaborado pelos autores.

# **9 Amostra de Desenvolvimento**

A amostra de desenvolvimento corresponde a uma etapa crucial no processo de construção de modelos de análise de sentimentos, pois define a base sobre a qual o modelo será treinado e testado. Ao término de todas as etapas mencionadas nos capítulos anteriores, dispomos agora de uma base de dados e informações processada e pronta para ser consumida, a qual representa um passo significativo na construção do nosso modelo de análise de sentimentos. Esta base de dados e informações foi utilizada para a construção do modelo simbólico, bem como para a construção do modelo de aprendizado de máquina supervisionado referente a fase II.

# **10 Modelo Simbólico**

# **11 Performance do Modelo Simbólico**

# **12 Exemplo de Execução**

# **13 Planejamento da Fase II**

Conforme previamente ilustrado no capítulo 3, o nosso grupo já dispõe de uma arquitetura planejada para a Fase II. Ressaltamos que alguns testes já foram realizados com diferentes algoritmos de aprendizado de máquina supervisionado, com o objetivo principal de aumentar a acurácia do nosso modelo de análise de sentimentos. Nossa expectativa é que o nosso modelo seja capaz de classificar se o conteúdo de uma frase fornecida por um(a) determinado(a) usuário(a) apresenta um sentimento positivo ou negativo, além de fornecer a probabilidade associada a tal classificação. As imagens abaixo, demonstram o nosso planejamento para a publicação em produção utilizando o pacote Streamlit do Python, na qual a primeira imagem corresponde a uma frase que obteve uma classificação final referente a um sentimento positivo, e a segunda imagem a um sentimento negativo:

A screenshot of a computer

AI-generated content may be incorrect.

Fonte - Elaborado pelos autores.

A screenshot of a computer

AI-generated content may be incorrect.

Fonte - Elaborado pelos autores.

1. ericaribeiro@usp.br [↑](#footnote-ref-1)
2. junior.marques@usp.br [↑](#footnote-ref-2)
3. luisvlopes@usp.br [↑](#footnote-ref-3)
4. mjmartins@alumni.usp.br [↑](#footnote-ref-4)
5. raphaelchaves@usp.br [↑](#footnote-ref-5)
6. thiago.ambiel1@gmail.com [↑](#footnote-ref-6)
7. Dados disponíveis em: <https://github.com/americanas-tech/b2w-reviews01/blob/main/B2W-Reviews01.csv> [↑](#footnote-ref-7)