UMinho

Mestrado Engenharia Informática
Mineração de Dados (2022/23)

Topic Modeling on Product Reviews

PG49169 Barbara Freixo
PG51239 Cátia Cardoso
PG49998 Cláudia Ribeiro
Braga, 16 de junho de 2023

Conteúdo

1. Introdução

1.1 Contextualização

Com o surgimento da internet e do comércio eletrónico, os consumidores têm agora acesso a uma vasta quantidade de informações sobre produtos e serviços antes de tomarem uma decisão de compra. Uma das principais fontes de informações são as avaliações de produtos, que são deixadas por utilizadores em plataformas *online*, como *websites* de comércio eletrónico, fóruns e redes sociais. Estas avaliações refletem a opinião dos consumidores sobre os produtos e podem influenciar as decisões de compra de outros consumidores.

Contudo, à medida que a quantidade de avaliações disponíveis aumenta, torna-se cada vez mais desafiante extrair informações úteis e relevantes desses dados. É aqui que a mineração de dados entra em jogo. A mineração de dados é uma área da ciência da computação que se foca em descobrir padrões e informações valiosas em conjuntos de dados grandes e complexos.

Uma técnica poderosa utilizada na mineração de dados é a modelação de tópicos (*topic modeling*). A modelação de tópicos é uma abordagem que tem como objetivo identificar e extrair automaticamente tópicos subjacentes num conjunto de documentos. No contexto das avaliações de produtos, a modelação de tópicos permite descobrir os principais temas discutidos pelos consumidores, revelando *insights* sobre as opiniões, necessidades e preferências dos utilizadores.

Ao compreender os tópicos abordados nas avaliações de produtos, as empresas podem obter *insights* valiosos para aprimorar os seus produtos, serviços e estratégias de marketing. Além disso, os consumidores também beneficiam desta análise, pois podem tomar decisões de compra mais informadas, com base nas opiniões de outros utilizadores.

1.2 Motivações e Objetivos

Com o crescimento do comércio eletrónico, tornou-se fundamental compreender as opiniões dos consumidores sobre os produtos e serviços disponíveis online. As avaliações de produtos, deixadas por utilizadores em plataformas como *websites* de comércio eletrónico, fóruns e redes sociais, desempenham um papel crucial nesse contexto. No entanto, a análise de um grande volume de avaliações torna-se um desafio. É aqui que a mineração de dados e, mais especificamente, a modelação de tópicos, entram em ação.

Neste projeto, temos como objetivos:

- **Identificar tópicos relevantes nas avaliações de produtos:** O nosso principal objetivo é aplicar técnicas de modelação de tópicos para identificar os principais temas abordados nas avaliações de produtos. Pretendemos extrair informações valiosas dos dados, revelando os temas predominantes discutidos pelos consumidores.

- **Classificar opiniões expressas nas avaliações:** O objetivo é utilizar a modelação de tópicos para classificar as opiniões presentes nas avaliações de produtos. Através desta classificação, será possível categorizar as avaliações de acordo com diferentes tópicos, ajudando-nos a compreender as opiniões positivas e negativas dos consumidores.

- **Fornecer *insights* relevantes para as empresas:** Pretendemos fornecer *insights* valiosos para as empresas com base na análise dos tópicos extraídos das avaliações de produtos. Estes *insights* ajudarão as empresas a melhorar os seus produtos, serviços e estratégias de marketing, ajustando-os de acordo com as necessidades e preferências dos consumidores.

- **Verificar a eficácia do *GPT-3.5* na modelação de tópicos:** Pretendemos, também avaliar a eficácia do modelo de linguagem GPT na realização da tarefa de modelação de tópicos em *reviews*

de produtos. Desejamos explorar o potencial do *GPT-3.5* como uma ferramenta automatizada para analisar grandes volumes de dados textuais e fornecer *insights* relevantes.

A motivação surge devido a vários fatores:

- **Crescimento do comércio eletrónico:** O comércio eletrónico tem registado um crescimento significativo nos últimos anos. Neste contexto, a compreensão das avaliações de produtos disponíveis *online* torna-se crucial para as decisões de compra dos consumidores.
- **Importância das opiniões dos consumidores:** As opiniões dos consumidores têm um impacto significativo nas decisões de compra. A análise dessas opiniões e a extração de *insights* relevantes podem ajudar as empresas a compreender melhor o mercado, aumentar a satisfação do cliente e alcançar uma vantagem competitiva.
- **Desafios da análise de grandes conjuntos de dados:** O aumento do volume de dados disponíveis requer técnicas avançadas para extrair informações relevantes. A modelação de tópicos surge como uma abordagem promissora para lidar com grandes conjuntos de dados de avaliações de produtos, superando os desafios da análise manual e proporcionando uma visão abrangente dos dados.
- **Melhoria contínua da experiência do cliente:** A compreensão das necessidades e preferências dos consumidores é fundamental para melhorar a experiência do cliente. A aplicação da modelação de tópicos nas avaliações de produtos permite às empresas identificar áreas de melhoria, adaptar produtos e serviços às demandas dos consumidores e fortalecer o relacionamento com os clientes.

2. Metodologia de Desenvolvimento e Fontes de Dados

2.1 Fontes de Dados

2.1.1 Fontes de Dados utilizadas

Neste projeto, fizemos uso de dados derivados de duas empresas da indústria de e-commerce: Amazon e Walmart. Os referidos dados são obtidos a partir das *reviews* fornecidas pelos consumidores após a aquisição dos produtos, o que confere um valor importante à nossa análise.

- ***Amazon:*** A Amazon, sendo uma das maiores plataformas de venda online a nível global, dispõe de um vasto conjunto de *reviews* de produtos. Tais *reviews*, que refletem diretamente as opiniões dos consumidores sobre os produtos, proporcionam-nos uma visão clara e precisa acerca das preferências e insatisfações dos consumidores.
- ***Walmart:*** A Walmart, outra plataforma de venda online, dispõe também de um vasto conjunto de *reviews* de diversos produtos. As *reviews* dos produtos provenientes da Walmart servem para enriquecer as da Amazon, introduzindo assim uma maior diversidade ao nosso conjunto de dados.

Basicamente, estas duas fontes de dados oferecem-nos uma perspectiva completa do ponto de vista dos consumidores. Recolhemos *reviews* dos produtos, classificações e outros pormenores pertinentes e aplicamos esses dados nas nossas análises e deduções. Ao combinar os dados da Amazon e da Walmart, asseguramos que os nossos resultados sejam mais completos e precisos.

2.1.2 Método utilizado para recolha de Dado

Para obter os dados necessários, dado que não existe uma API que forneça esse tipo de informações, optamos por utilizar o método de *web scraping*. Utilizamos a renomada *framework Scrapy* para realizar a recolha de dados de plataformas como a Amazon e o Walmart.

O *web scraping* permitiu extrair informações relevantes dos websites alvo, como detalhes de produtos e comentários de clientes.

2.1.3 Processamento e Analise de Dados

Para o processamento e análise de dados, utilizamos o modelo *GPT-3.5*, que é um modelo avançado de geração de texto desenvolvido pela OpenAI. O *GPT-3.5* é treinado em grandes quantidades de dados de texto da internet e possui a capacidade de gerar respostas coerentes e contextuais com base nas entradas fornecidas. Esse modelo é baseado na arquitetura Transformer, que permite capturar relacionamentos de longo alcance entre palavras.

Utilizamos o *GPT-3.5* para realizar a técnica de *topic modeling* nos comentários dos produtos, utilizando a abordagem de *prompting*. Esta técnica envolve o envio dos comentários como entrada para o modelo, que então gera respostas com base nos tópicos e conteúdos presentes nos dados analisados.

Essa aplicação do *GPT-3.5* permitiu-nos extrair informações relevantes e identificar os principais tópicos abordados nos comentários dos produtos, auxiliando na compreensão e análise dos dados recolhidos.

2.2 Desenvolvimento da Aplicação

O primeiro passo no desenvolvimento da aplicação consistiu em verificar quais ferramentas seriam necessárias. Após a decisão das ferramentas a serem utilizadas, prosseguimos com a implementação.

Inicialmente, criamos um projeto utilizando a *framework Scrapy* para realizar o *web scraping* das fontes de dados. Devido às ferramentas anti-bots utilizadas pelas fontes, o *web scraping* tornou-se mais complexo, exigindo o uso do ScrapeOps para evitar bloqueios de proxy impostos por essas fontes.

Após a conclusão do processo de *web scraping*, avançamos para o desenvolvimento de uma aplicação utilizando o *framework Flask*. Essa aplicação foi responsável por criar nossa API e nosso *website*. Além de se comunicar com a nossa aplicação de *web scraping*, a aplicação também se conectou à API da OpenAI para aplicar o modelo *GPT-3.5* e realizar a análise de tópicos nos nossos dados.

Para estabelecer a comunicação com a API da OpenAI, utilizamos a *framework LangChain*. Essa escolha foi motivada pelo grande volume de comentários a serem enviados, pois a framework permite o envio desses dados em "chunks".

Após a conclusão do desenvolvimento da aplicação e a implementação das ferramentas requeridas, prosseguimos para a integração do modelo *GPT-3.5* da OpenAI utilizando a técnica de *prompting*. Além disso, foi necessário adicionar a estrutura desejada no *prompt* para que nossa aplicação pudesse interpretar adequadamente as respostas geradas pelo modelo.

Esta etapa de personalização do *prompting* foi essencial para garantir que as respostas do modelo estivessem alinhadas com a estrutura desejada pela nossa aplicação. Esta imagem a baixo apresenta o *prompt* final utilizado:

```
Analyze only the following collection of reviews and employ topic modeling techniques to categorize the feedback into specific features of the product.
Divide each feature in positive characteristics and in negative characteristics.
Response format provided in a json format like this: {Features:[{
                -name: x
                -Positive Reviews:(full reviews only the ones about this feature)
                -Negative Reviews:(full reviews only the ones about this feature)
                }]}

Do not repeat the same review twice.
If there are no positive or negative characteristics, write "Not applicable".
Give at least 6 Features.
The product is: """ + product + """
Provide it in JSON format.
```

Figura 2.1: *Prompt* utilizado no *GPT-3.5*.

Através dessa personalização do *prompt*, nossa aplicação foi capaz de interpretar as respostas geradas pelo modelo *GPT-3.5* de acordo com a estrutura estabelecida. Isso permitiu obter resultados mais precisos e adequados aos objetivos e requisitos da aplicação.

Essa integração do modelo *GPT-3.5* utilizando a técnica de prompting, juntamente com a definição da estrutura desejada, foi fundamental para a eficiência e o sucesso da nossa aplicação na geração de respostas relevantes e coerentes com base nas entradas fornecidas.

3. Análise de Resultados

O sistema desenvolvido demonstrou um bom desempenho na obtenção e tratamento de dados provenientes das nossas duas fontes de dados, garantindo assim um bom método de recolha e análise de *reviews* de produtos.

Com a abordagem seguida conseguimos transformar opiniões dispersas e fragmentadas em *insights* estruturados e úteis, categorizados segundo as características específicas do produto e discriminados em termos de percepções positivas e negativas.

O sistema foi concebido para tentar maximizar a eficiência na recolha de dados utilizando técnicas de multithreading. Isto permitiu o lançamento de processos em simultâneo, garantindo a obtenção de dados de forma paralela e significativamente mais rápida, reduzindo o tempo total de execução. Mesmo assim o tempo de execução obtido não foi o ideal, uma vez que ainda apresentou alguma lentidão na apresentação dos resultados.

Simultaneamente, recorremos ao *webscraping* para extrair as *reviews* dos produtos das páginas web da Amazon e da Walmart, o que permitiu automatizar a recolha de grandes volumes de dados que seriam impraticáveis de recolher manualmente.

Para a construção da nossa aplicação web, escolhemos o Flask, uma *framework* do Python. Apesar da sua simplicidade e minimalismo, o Flask ofereceu uma flexibilidade considerável, permitindo assim um desenvolvimento rápido e eficaz da nossa aplicações web. Complementarmente, a utilização de templates HTML facilitou a criação de uma interface de utilizador, permitindo assim uma melhor visualização e compreensão dos resultados obtidos.

No desenvolvimento do nosso projeto, o do modelo *GPT-3.5* destacou-se apesar de não possuir uma configuração específica para *topic modeling*, a sua capacidade de gerar respostas coerentes e contextualizadas a partir de *prompts* bem elaborados permitiu-nos automatizar a análise das *reviews* dos produtos de maneira eficaz.

Assim, a eficácia do *GPT-3.5*depende em grande parte da qualidade dos *prompts* fornecidos. Os *prompts* atuam como diretrizes para a API, instruindo-a sobre o tipo de resposta esperada. Como tal, a habilidade em formular *prompts* eficazes é crucial para maximizar o desempenho da API. Aperfeiçoar a arte de criar *prompts* claros e eficazes é essencial para obter as respostas desejadas do modelo de linguagem.

A API do *GPT-3.5* demonstrou ser uma ferramenta potente e multifuncional para processar e compreender a linguagem natural. A sua capacidade em fornecer respostas coerentes e contextualizadas foi um dos principais pontos fortes identificados ao longo do projecto.

A utilização da API simplificou de forma notável a análise de texto, uma vez que o modelo de linguagem já se encontrava pré-treinado e pronto a ser aplicado. Isto poupou um tempo muito significativo que teria sido despendido na formação de um modelo de linguagem a partir da base.

Por outro lado, apesar dos benefícios, foram identificados alguns desafios ao utilizar a API do *GPT-3.5*.

Em primeiro lugar, a qualidade da análise depende da qualidade do *prompt*. Portanto, é crucial assegurar que os dados fornecidos à API sejam relevantes e de alta qualidade.

Em segundo lugar, a API pode gerar respostas que são corretas no contexto fornecido, mas que não são necessariamente factualmente corretas. Isto é algo a ter em conta quando se utilizam os resultados gerados pela API.

Por fim, a utilização da API implica custos, o que pode ser um obstáculo para projetos com orçamentos limitados. Como tal, é importante levar em conta estes custos ao planear a utilização da API do *GPT-3.5*GPT 3.5.

Em resumo, a nossa experiência com a API do*GPT-3.5* foi em grande parte positiva, embora esta apresente alguns desafios. Estes são, contudo, superáveis e, de forma geral, acreditamos que os benefícios da API superam os seus pontos negativos.

3.1 Exemplos de Uso

Vamos agora mostrar dois exemplos de uso da nossa aplicação web. Quando iniciamos a nossa aplicação obtemos a seguinte página:

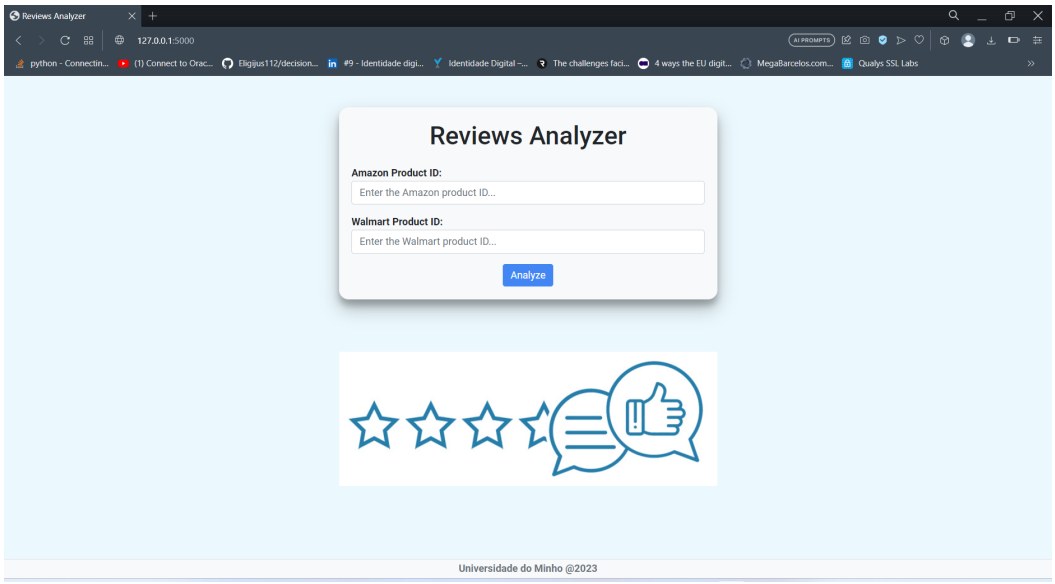

Figura 3.1: Página principal da aplicação web

Imaginemos agora que queremos analisar as *reviews* do jogo *Super Mario Bros. U Deluxe* para a *Nintendo Switch*. Para tal, primeiro vamos à Amazon e ao Walmart procurar pelos *ids* correspondentes a este produto. Depois, inserimos os ids correspondentes na pagina inicial da nossa aplicação web como se pode ver em seguida:

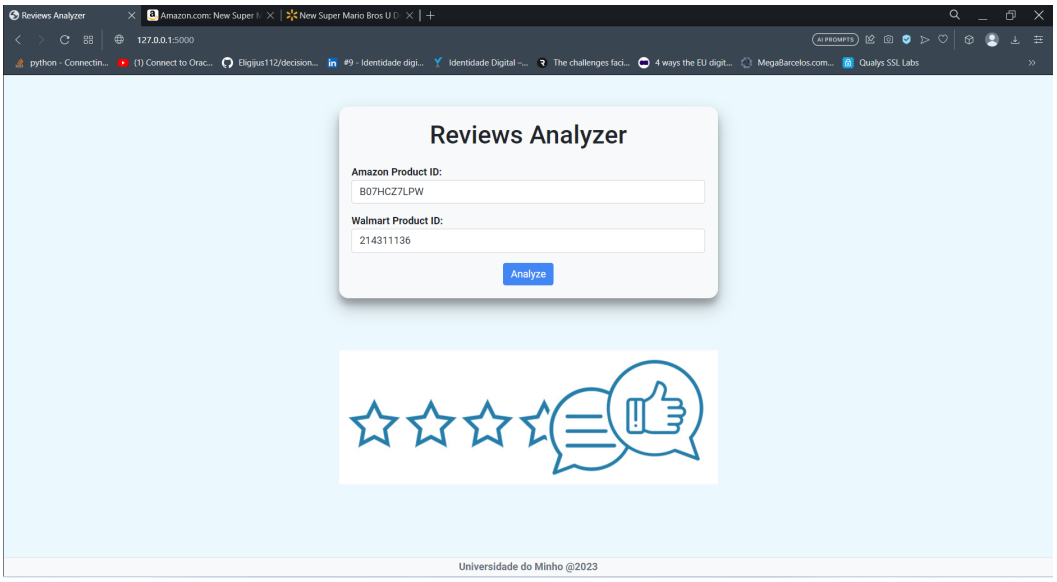

Figura 3.2: Inserir *Ids*

Após a inserção dos *Ids*, basta carregar em *"Analyze"* e aguardar que os resultados sejam mostrados. Quando os resultados forem mostrados aparecerá a seguinte página:

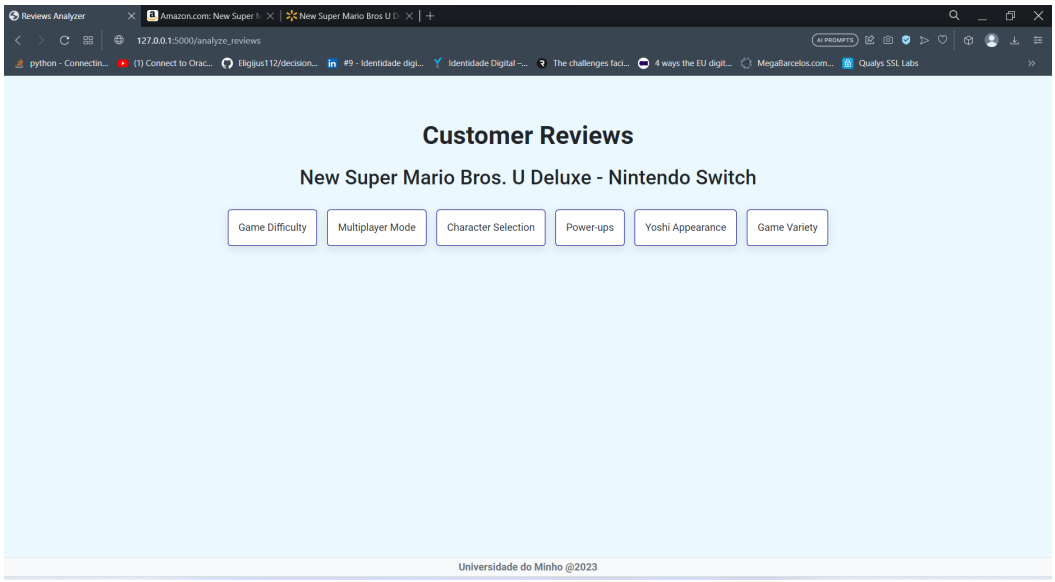

Figura 3.3: Resultados obtidos

Agora para analizar as *reviews* de cada *feature* basta carregar no botão correspondente a essa *feature*. Para as *features Multiplayer Mode, Power-ups e Yoshi Appearence* obtemos o seguinte:

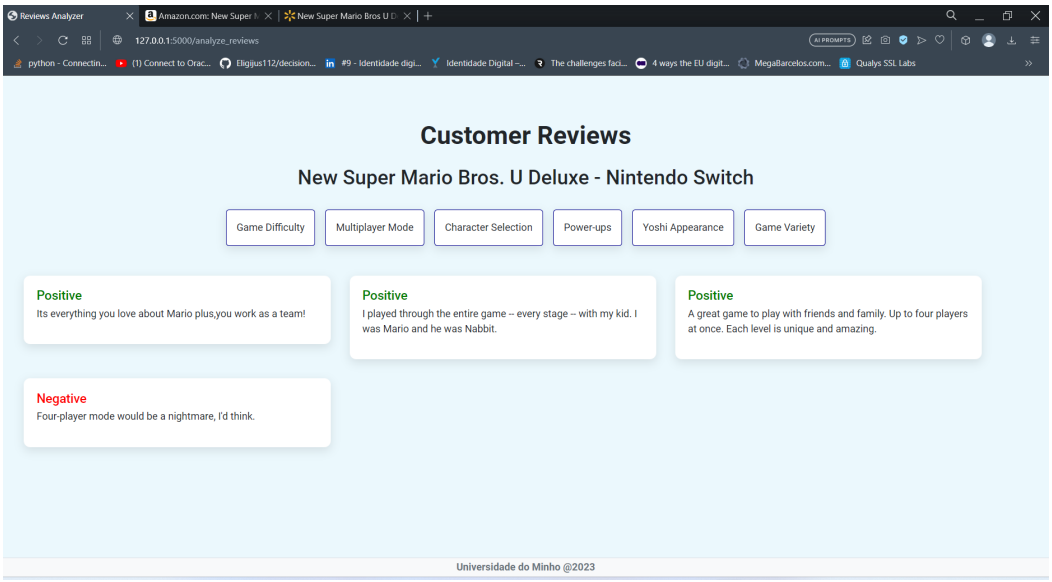

Figura 3.4: *Multiplayer Mode*

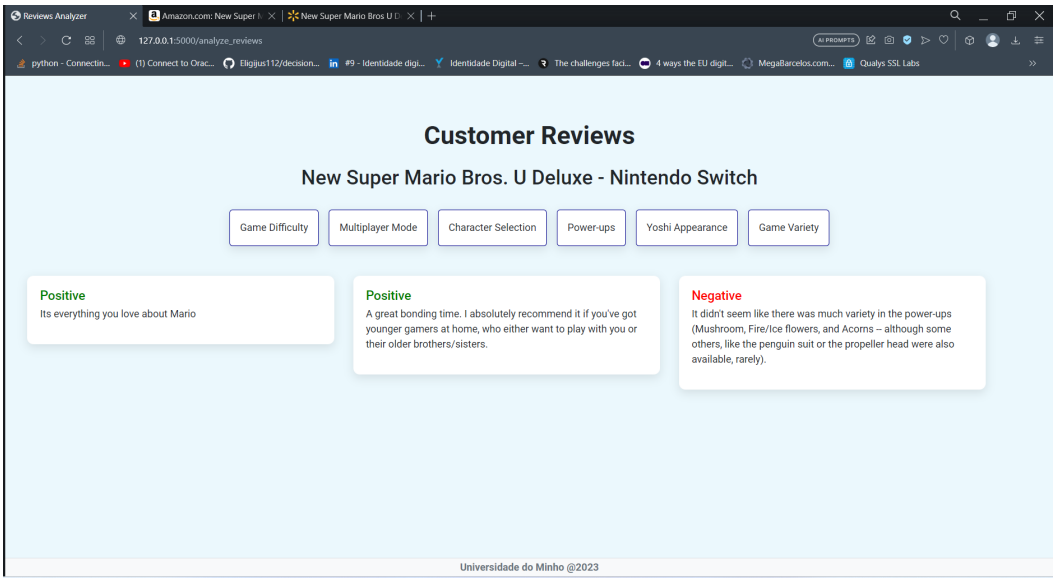

Figura 3.5: *Power-ups*

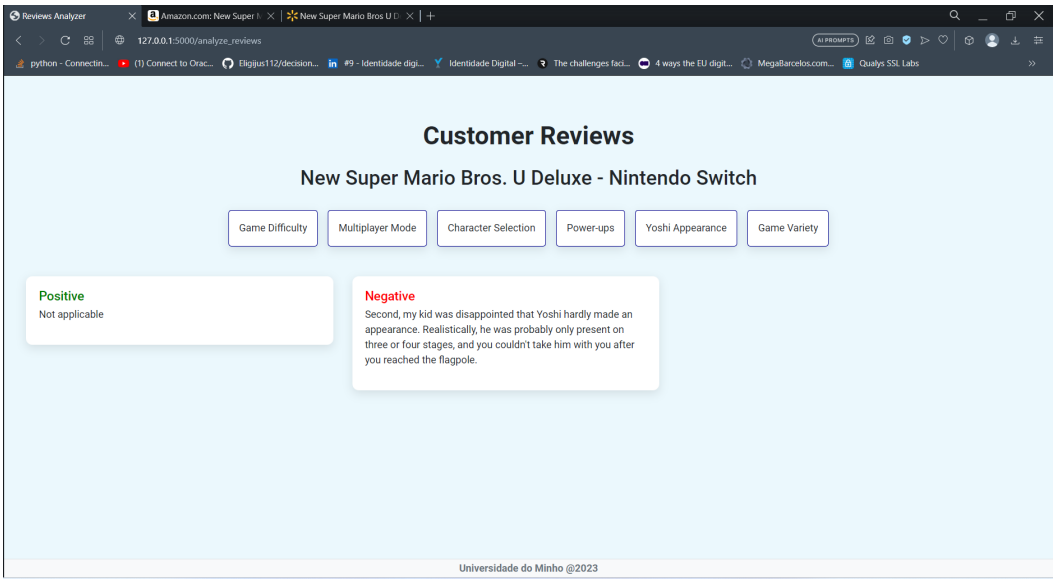

Figura 3.6: *Yoshi Appearence*

Vamos agora considerar outro exemplo. Imaginemos agora que queremos analisar as *reviews* do *smartphone Google Pixel 7*. Novamente, primeiro vamos à Amazon e ao Walmart procurar pelos *ids* correspondentes a este produto. Depois, inserimos os ids correspondentes na pagina inicial da nossa aplicação web como se pode ver em seguida:

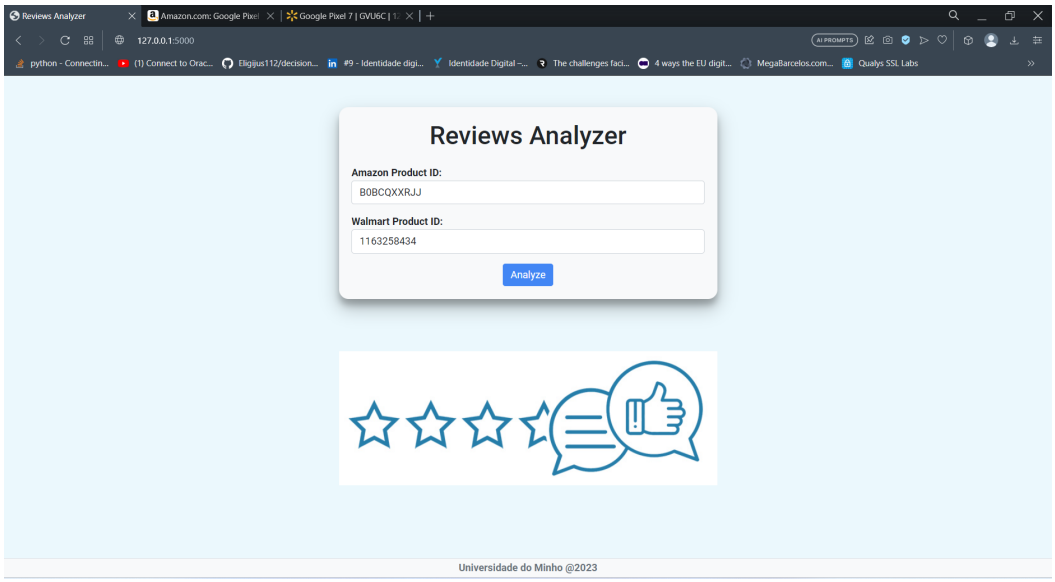

Figura 3.7: Inserir *Ids*

Novamente, após a inserção dos *Ids*, basta carregar em *"Analyze"* e aguardar que os resultados sejam mostrados. Quando os resultados forem mostrados aparecerá a seguinte página:

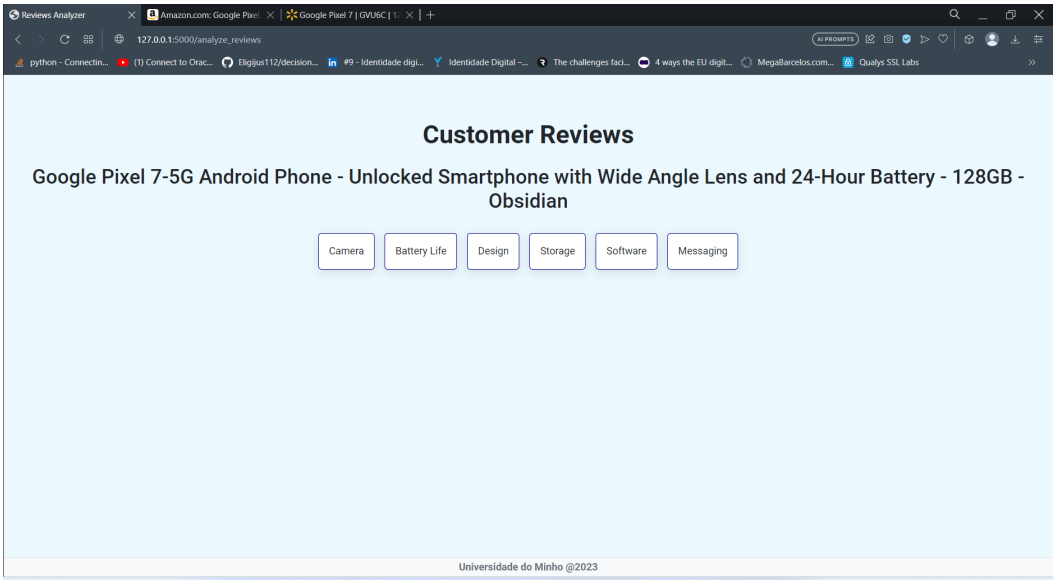

Figura 3.8: Resultados obtidos

Novamente, para analizar as *reviews* de cada *feature* basta carregar no botão correspondente a essa *feature*. Para as *features Camera, Messaging e Battery Life* obtemos o seguinte:

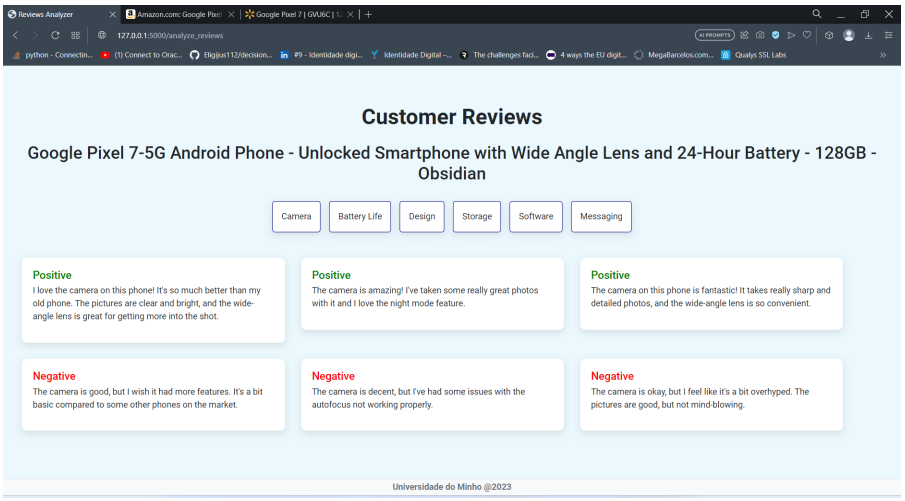

Figura 3.9: *Camera*

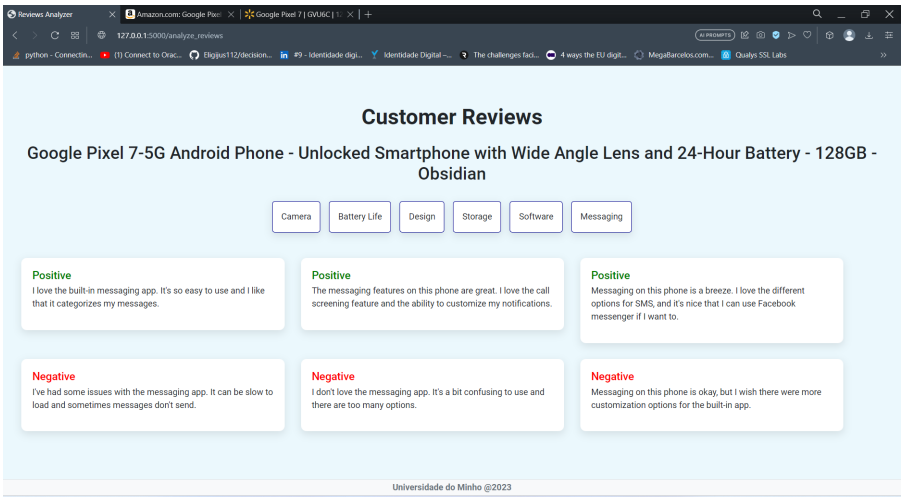

Figura 3.10: *Messaging*

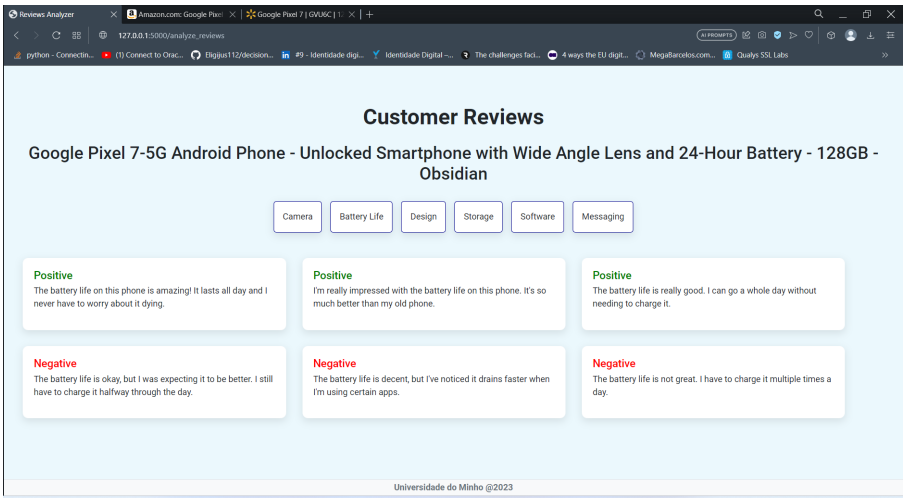

Figura 3.11: *Battery Life*

4. Conclusão

Este projeto destacou a importância do *topic modeling* na análise de avaliações de produtos e demonstrou a viabilidade de aplicar essa técnica por meio de webscraping e da API da *OpenAI*. A combinação dessas abordagens permitiu a extração eficiente de tópicos relevantes em grandes volumes de avaliações de produtos.

Ao longo do desenvolvimento do projeto, verificamos que o *GPT-3.5*, o modelo de linguagem utilizado da *OpenAI*, apresentou um desempenho impressionante na modelação de dados. O *GPT* mostrou-se capaz de compreender a linguagem natural e gerar respostas coerentes e relevantes, contribuindo de forma significativa para a obtenção de *insights* precisos a partir das avaliações de produtos. A sua A eficácia foi uma descoberta surpreendente, fornecendo uma solução poderosa e eficiente para a análise de grandes conjuntos de dados textuais. A capacidade de generalizar e aplicar o conhecimento adquirido em diferentes contextos e fontes de dados permitiu-nos extrair tópicos relevantes com precisão e coerência. No entanto, é importante salientar que o uso do *GPT* como ferramenta de modelação de dados tem limitações, apesar de não requerer dados de treino específicos, pode gerar respostas imprecisas.

No futuro, pretendemos aplicar melhorias significativas ao *website* desenvolvido para análise de avaliações de produtos. Essas melhorias visam aprimorar a precisão e a utilidade das informações fornecidas aos utilizadores, permitindo uma análise mais abrangente e detalhada das opiniões dos consumidores. Nesta seção, apresentaremos algumas das melhorias que planeamos implementar.

- **Aprimorar o Prompt**
  - Refinar o texto do prompt, tornando-o mais claro e específico em relação aos objetivos do estudo.
  - Experimentar diferentes abordagens de prompt, explorando diferentes aspectos das avaliações para uma análise mais abrangente.
  - Refinar o texto do prompt, tornando-o mais claro e específico em relação aos objetivos do estudo.
- **Adicionar Mais Fontes de Dados**
  - Incluir mais sites de avaliações para ampliar a diversidade das opiniões recolhidas.
  - Integrar análise de avaliações e menções em redes sociais para obter *insights* adicionais.