Data Scientist II Technical Challenge

Task 4: Real-world Scenario

Cristina Sánchez Maíz | csmaiz@gmail.com | LinkedIn

Table of Contents

Tasks:
Consider the following business problem:
Your company wants to improve customer satisfaction by understanding the main topics and sentiments expressed in customer reviews. Your task is to:

- Use topic modeling to identify the main topics in the customer reviews.
- Summarize the findings and suggest actionable insights for business improvements.

Deliverables:

- Explanation of the topic modeling process.
- Summary of the main topics identified.
- Actionable insights based on the analysis.

Explanation of the topic modeling process

Topic modeling is applied to find the main topics within the customer reviews data. Figure 1 shows the steps I followed to answer the requests in the Challenge brief. By combining topic modeling and sentiment analysis (Task 3), we can gain valuable insights into customer feedback and drive business improvements.

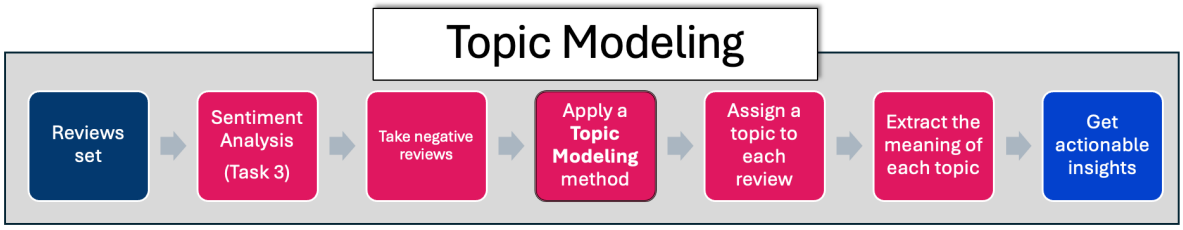


Figure 1: Topic Modeling process

As we want to find issues and suggest actionable insights for business improvements, I took the reviews classified as **negative** in Task 3 as text corpus from which to extract the main themes.

I applied Latent Dirichlet Allocation (LDA), one of the most popular topic modeling methods, and a Large Language Model (LLM) followed by clustering. I explain both processes and then I compare the results.

Latent Dirichlet Allocation (LDA)

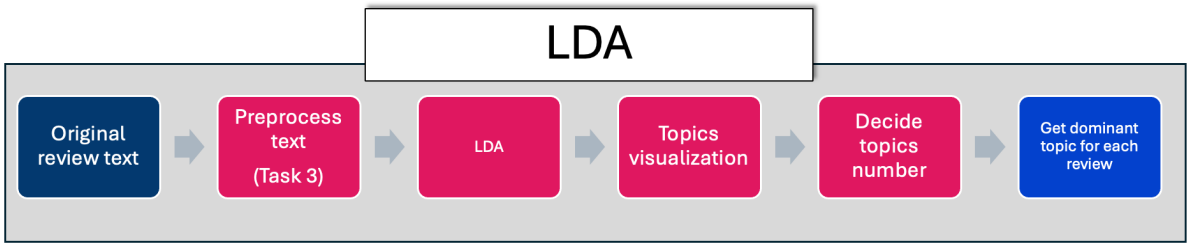


Figure 2: LDA process

Starting from the raw reviews:

1. **Preprocess the text reviews**: Please, refer to Task 3 document for more details.

2. **Appy LDA**: LDA is an unsupervised machine learning technique used to discover abstract topics within a collection of documents. It operates by modeling each document as a mixture of multiple topics and each topic as a distribution of words.

3. **Topics visualization**: After fitting the LDA model, I used pyLDAvis to visualize the topics. Steps 2 and 3 were an iterative process until I managed to tune the hyperparameters.

4. **Topics number**: as it is shown in Figure 3, the four topics represented by the circles do not overlap and make sense to keep them to explain the main themes within the reviews. In the jupyter notebook of Task 4, it is available an interactive visualization.

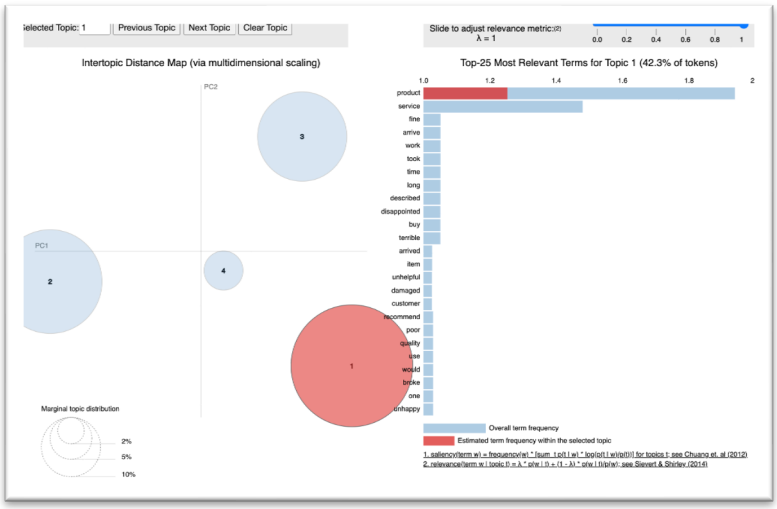


Figure 3: pyLDA visualization

5. Finally, the dominant topic is assigned to each review. Note that the model assigns a probability distribution over these topics for each document. So, we take the one with the highest probability.

Using LDA, the reviews are classified as follows:

Topic 1:
- Very disappointed with the product, not as described.
- Poor quality, would not recommend.

Topic 2:
- The product works fine, but took a long time to arrive.
- The product broke after one use, very unhappy.

Topic 3:
- Terrible service, will not buy from here again.

Topic 4:
- The item arrived damaged and customer service was unhelpful.

Large Language Model + Clustering

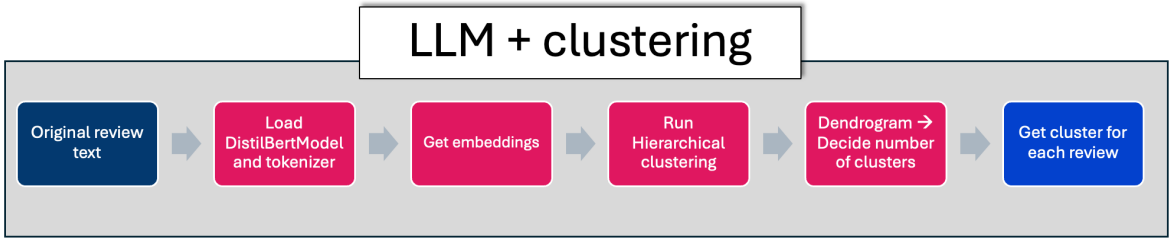


Figure 4: LLM + clustering process

I used a pre-trained LLM like DistilBERT to generate dense vector representations (embeddings) for each review. Then, I run an aglomerative clustering to group reviews based on their similarity in the embedding space. The dendrogram shown in Figure 5 allows us to determine the number of clusters.

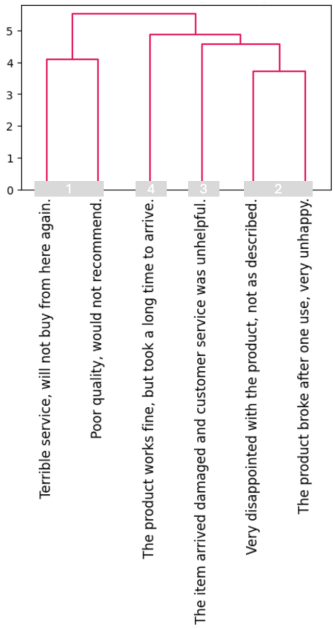


Figure 5: Hierarchical Clustering Dendrogram

Using LLM+Clustering, the reviews are classified as follows:

Topic 1:
- Terrible service, will not buy from here again.
- Poor quality, would not recommend.
Topic 2:
- Very disappointed with the product, not as described.
- The product broke after one use, very unhappy.
Topic 3:
- The item arrived damaged and customer service was unhelpful.
Topic 4:
- The product works fine, but took a long time to arrive.

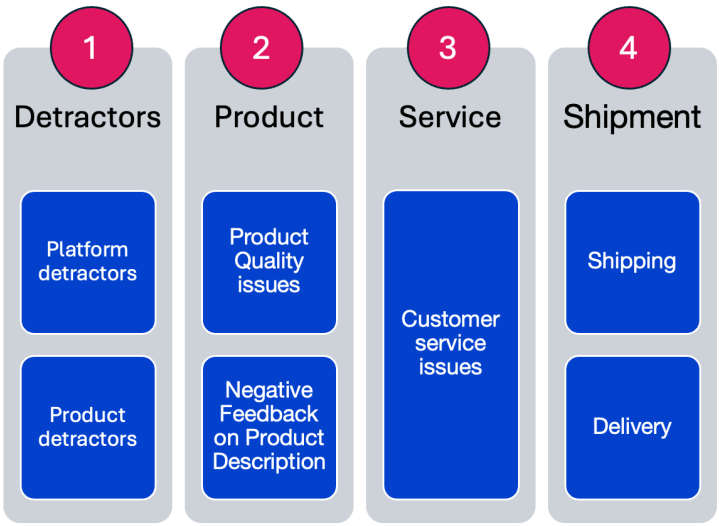
Summary of the main topics identified

In Table 1 *topic_lda* and *topic_llm* columns are the topic IDs assigned by LDA and LLM, respectively.

review_text	topic_lda	topic_llm
Terrible service, will not buy from here again.	2	1
Very disappointed with the product, not as described.	1	2
The item arrived damaged and customer service was unhelpful.	4	3
Poor quality, would not recommend.	2	1
The product works fine, but took a long time to arrive.	1	4
The product broke after one use, very unhappy.	4	2

Table 1: Topic modelling results

The assignment made by LDA is a bit more confusing than with LLM. With this second method, it is easier to find the meaning of the topics. I chose the topic assignment done by LLM.



Actionable insights based on the analysis

Based on the identified topics and sentiments, here are some actionable insights for the business:

Improve Customer Service:

- Conduct training sessions for customer service representatives to improve their communication skills and responsiveness.
- Implement a better tracking system for customer inquiries to ensure timely responses.

Enhance Product Descriptions:

- Review and update product descriptions to ensure they accurately reflect the product features and quality.
- Include more detailed images and customer reviews to set realistic expectations.

Address Product Quality Issues:

- Investigate and address any recurring quality issues in the products.
- Consider quality checks and improvements in the production process.

Streamline Delivery and Shipping:

- Partner with reliable shipping companies to ensure timely and accurate deliveries.
- Implement a tracking system for customers to monitor their orders in real-time.

Additionally, check and compare topics over time for detecting trends and emerging issues.