Data Scientist II Technical Challenge

Task 3: Natural Language Processing

Cristina Sánchez Maíz | csmaiz@gmail.com | LinkedIn

Table of Contents

Using the Customer Reviews dataset:

- Preprocess the text data (e.g., tokenization, stopword removal, stemming/lemmatization).
- Perform sentiment analysis on the reviews.
- Visualize the distribution of sentiment scores.

Deliverables:

- Preprocessed text data.
- Sentiment analysis results.
- Visualizations of sentiment distribution.

Preprocessed text data

The steps for preprocessing the text and make it ready to sentiment analysis are shown in Figure 1.

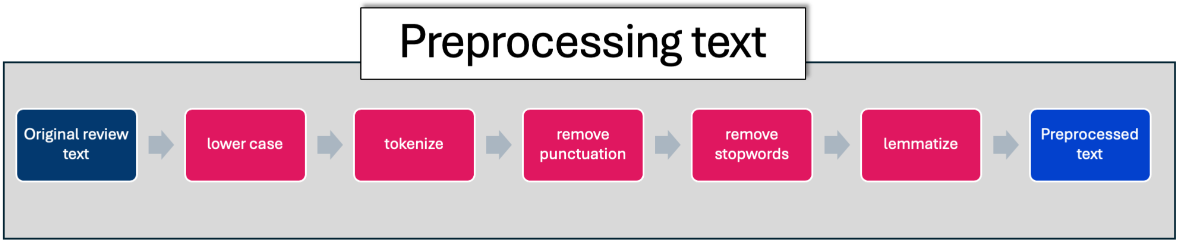


Figure 1: Process to preprocess text

As an example, Table 1 displays 10 rows with the original customer review and the preprocessed text after all the steps of Figure 1:

	review_text	preprocessed_review
1	*Terrible service, will not buy from here again.*	*terrible service buy*
2	*Average quality, you get what you pay for.*	*average quality get pay*
3	*Great product, very satisfied with the quality and performance.*	*great product satisfied quality performance*
4	*Very disappointed with the product, not as described.*	*disappointed product described*
5	*Excellent service, highly recommend!*	*excellent service highly recommend*
6	*The item arrived damaged and customer service was unhelpful.*	*item arrived damaged customer service unhelpful*
7	*Fast delivery and the product works perfectly!*	*fast delivery product work perfectly*
8	*The item arrived damaged and customer service was unhelpful.*	*item arrived damaged customer service unhelpful*
9	*The service was acceptable, but could be improved.*	*service acceptable could improved*
10	Decent product, but there are better options available.	decent product better option available

Table 1: Examples of text transformation

For more information, refer to the complete file in GitHub.

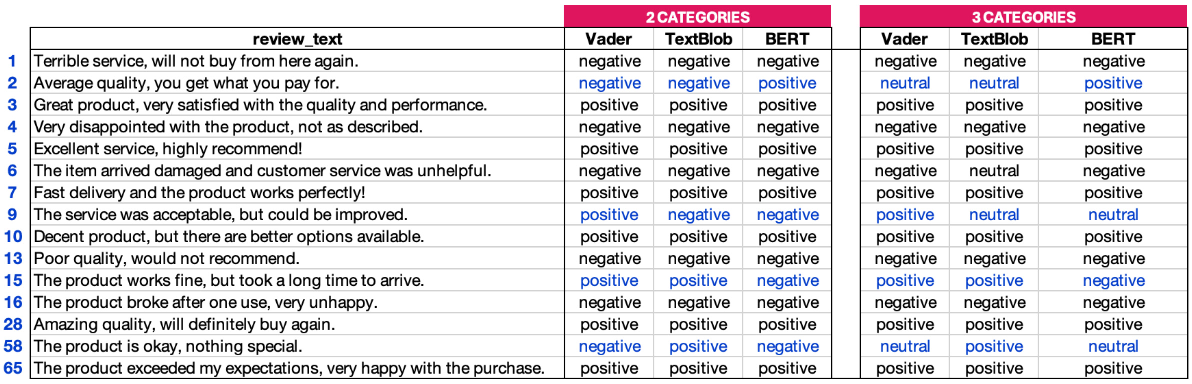
Sentiment analysis results

The goal is to perform sentiment analysis over a customer reviews dataset. To that purpose, I applied lexicon-based methods (Vader and TextBlob) and a Large Language Model (distilled BERT) on the provided data. I started with a binary classification (positive or negative) and then, I added the neutral category.

Lexicon-based methods have a significant dependence on preprocessing while LLMs are less sensitive.

Since the reviews are repeated, I removed duplicates before running the analysis as duplicated reviews can skew the results by giving more weight to the repeated texts.

The results of the sentiment analysis are summarized in Table 2:



		2 CATEGORIES			**3 CATEGORIES**		
	review_text	**Vader**	**TextBlob**	**BERT**	**Vader**	**TextBlob**	**BERT**
1	Terrible service, will not buy from here again.	negative	negative	negative	negative	negative	negative
2	Average quality, you get what you pay for.	negative	negative	positive	neutral	neutral	positive
3	Great product, very satisfied with the quality and performance.	positive	positive	positive	positive	positive	positive
4	Very disappointed with the product, not as described.	negative	negative	negative	negative	negative	negative
5	Excellent service, highly recommend!	positive	positive	positive	positive	positive	positive
6	The item arrived damaged and customer service was unhelpful.	negative	negative	negative	negative	neutral	negative
7	Fast delivery and the product works perfectly!	positive	positive	positive	positive	positive	positive
9	The service was acceptable, but could be improved.	positive	negative	negative	positive	neutral	neutral
10	Decent product, but there are better options available.	positive	positive	positive	positive	positive	positive
13	Poor quality, would not recommend.	negative	negative	negative	negative	negative	negative
15	The product works fine, but took a long time to arrive.	positive	positive	negative	positive	positive	negative
16	The product broke after one use, very unhappy.	negative	negative	negative	negative	negative	negative
28	Amazing quality, will definitely buy again.	positive	positive	positive	positive	positive	positive
58	The product is okay, nothing special.	negative	positive	negative	neutral	positive	neutral
65	The product exceeded my expectations, very happy with the purchase.	positive	positive	positive	positive	positive	positive

Table 2: Sentiment analysis results

The three libraries manage to classify the reviews that have a strong positive/negative sentiment. The differences arise (highlighted in blue in Table 2) when the review

- has a more impartial tone:
    - #2. *Average quality, you get what you pay for.*
    - #58. *The product is okay, nothing special.*
- is composed of two parts, one positive and the other negative:

  - #9. *The service was acceptable, but could be improved*.
  - #15. *The product works fine, but took a long time to arrive*.

With the binary classification, BERT model correctly categorizes the reviews, even the four aforementioned. The two lexicon-based methods struggle with at least one of them.

Regarding the 3-classes categorization:

- Vader correctly classifies reviews #2 and #58 as neutral, while keeping #6 and #9 as positive.
- TextBlob assigned score 0 to reviews two reviews #6 and #9. Adding the new threshold for multiclass classification, review #6 is assigned a neutral sentiment what is clearly incorrect.
- The BERT model I applied was specifically trained for binary sentiment analysis, so it would be better to use another model trained for multiclass classification. In any case, the performed classification is correct.

Future work

In order to improve the sentiment analysis, suggestions as the following ones could be considered:

- It should be convenient to experiment with different thresholds since the classification is strongly dependent on thresholds.
- Use a LLM specifically trained on multiclass sentiment analysis.
- Get labeled data and run a ML model (Naive Bayes, SVMs, RNNs, LSTM).

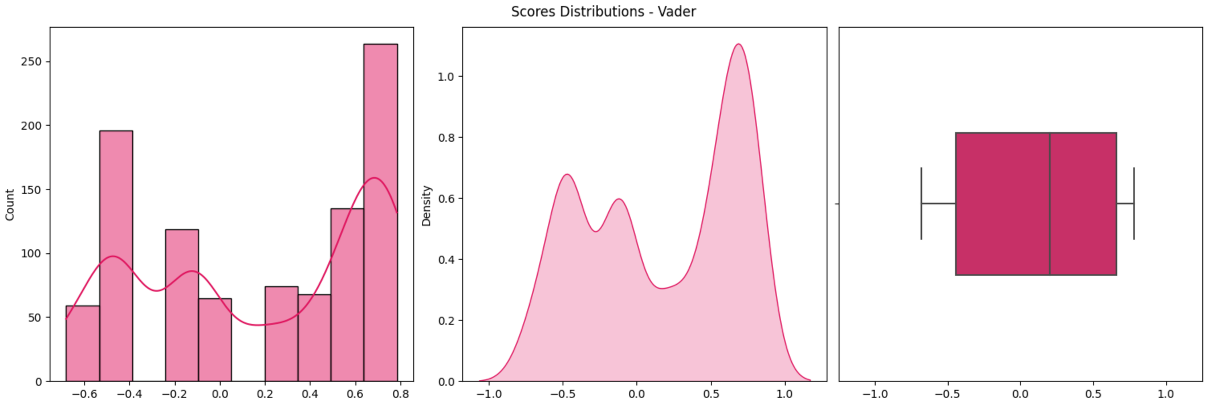
Visualizations of sentiment distribution

Analysis of the Scores Distribution Plot

I have plotted the histogram, the density estimation and the boxplot for each scores distribution:

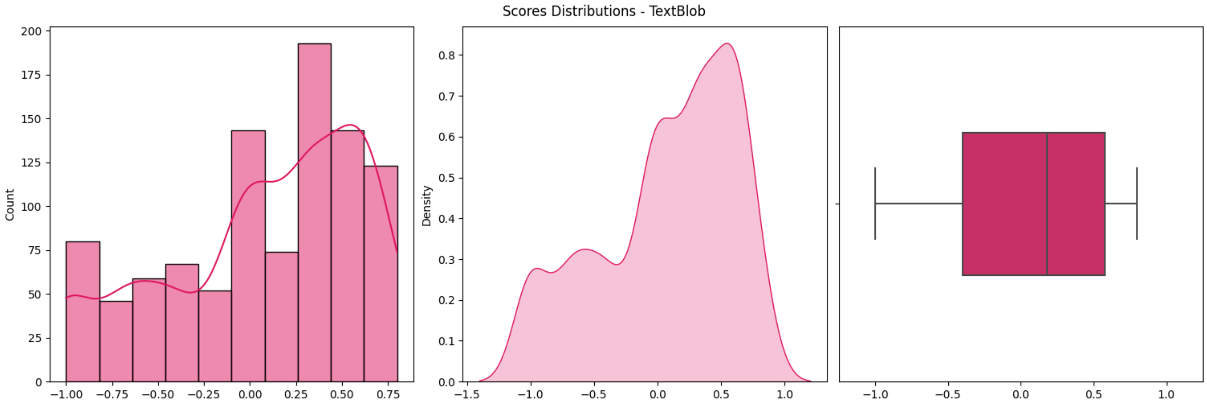
- The histogram (left graph) represents the frequency of occurrences for each score value.
- The plot in the middle shows the probability density function (PDF) estimated using Kernel Density Estimation (KDE). It provides a smoothed representation of the data distribution.
- On the right side, a boxplot offers a visual summary of the data distribution, including median and quartiles.

Vader



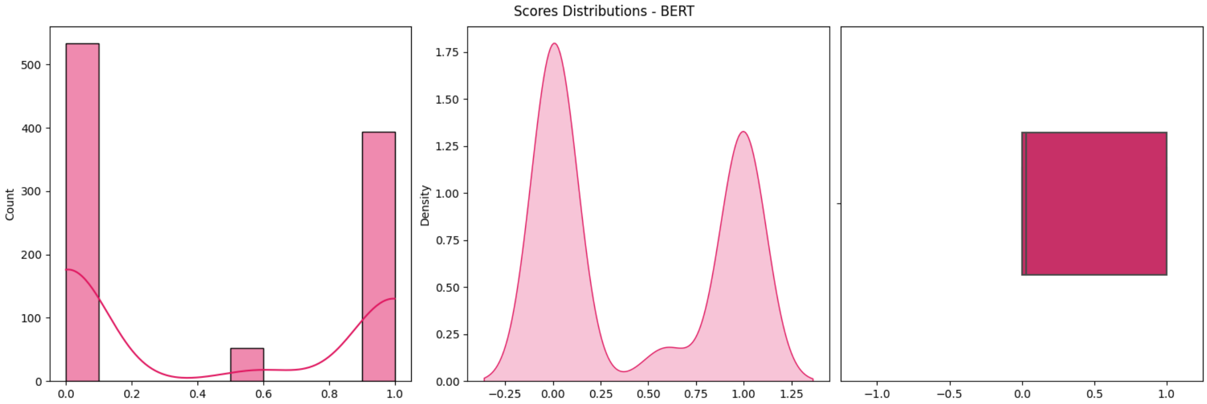
- The KDE plot has two distinct modes, allowing to separate positive reviews (right mode) from negative reviews (left mode).
- Score ranges from -1 to 1.

TextBlob



- The separation between modes is less clear with TextBlob.
- Score ranges from -1 to 1.

BERT



- The KDE plot has two separated modes, allowing to classify positive reviews (right mode) from negative reviews (left mode).
- Score ranges from 0 to 1.

WordCloud

I have created a visualization of the most common words for positive and negative reviews considering VADER method and BERT model.

VADER

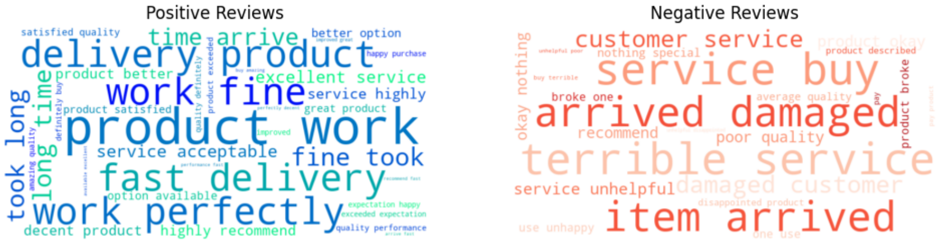


Figure 2: WordCloud from VADER sentiment analysis

BERT

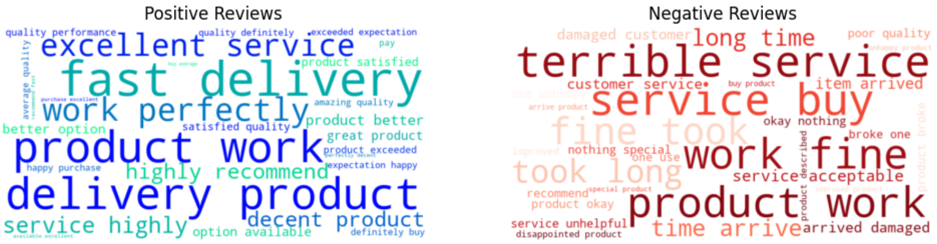


Figure 3: WordCloud from BERT sentiment analysis