Data Scientist II Technical Challenge

Task 3: Natural Language Processing

Cristina Sánchez Maíz | csmaiz@gmail.com | LinkedIn

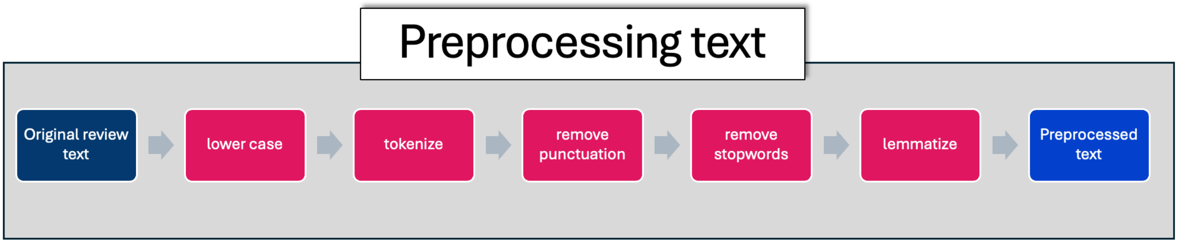
Using the Customer Reviews dataset:

- 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



	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

See complete file in GitHub.

Sentiment analysis results

I use lexicon-based methods (Vader and TextBlob) and a Large Language Model (distilled BERT) on the provided data.

Lexicon-based methods require preprocessing the review text while LLMs are less sensitive.

Since the reviews are repeated, I removed duplicates before running the clustering since duplicate reviews can skew the clustering results by giving more weight to the repeated texts.

The results are summarized in the following table:

		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

The three libraries manage to classify the reviews that have a strong positive/negative sentiment. The differences arise 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 two categories, BERT model correctly classifies the reviews, even the four aforementioned. The two lexicon-based methods struggle with at least one of them.

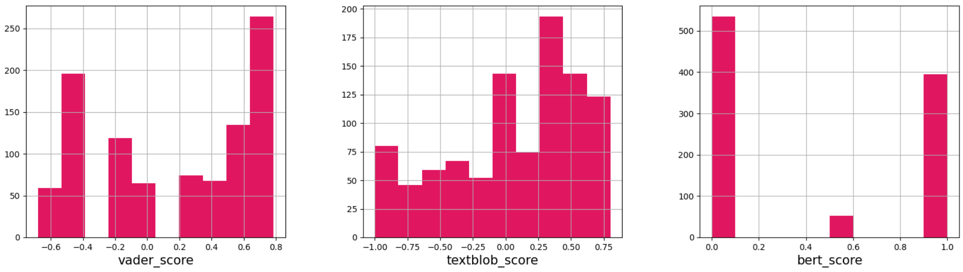
I run a more deeper classification in three categories, adding a new one named "neutral":

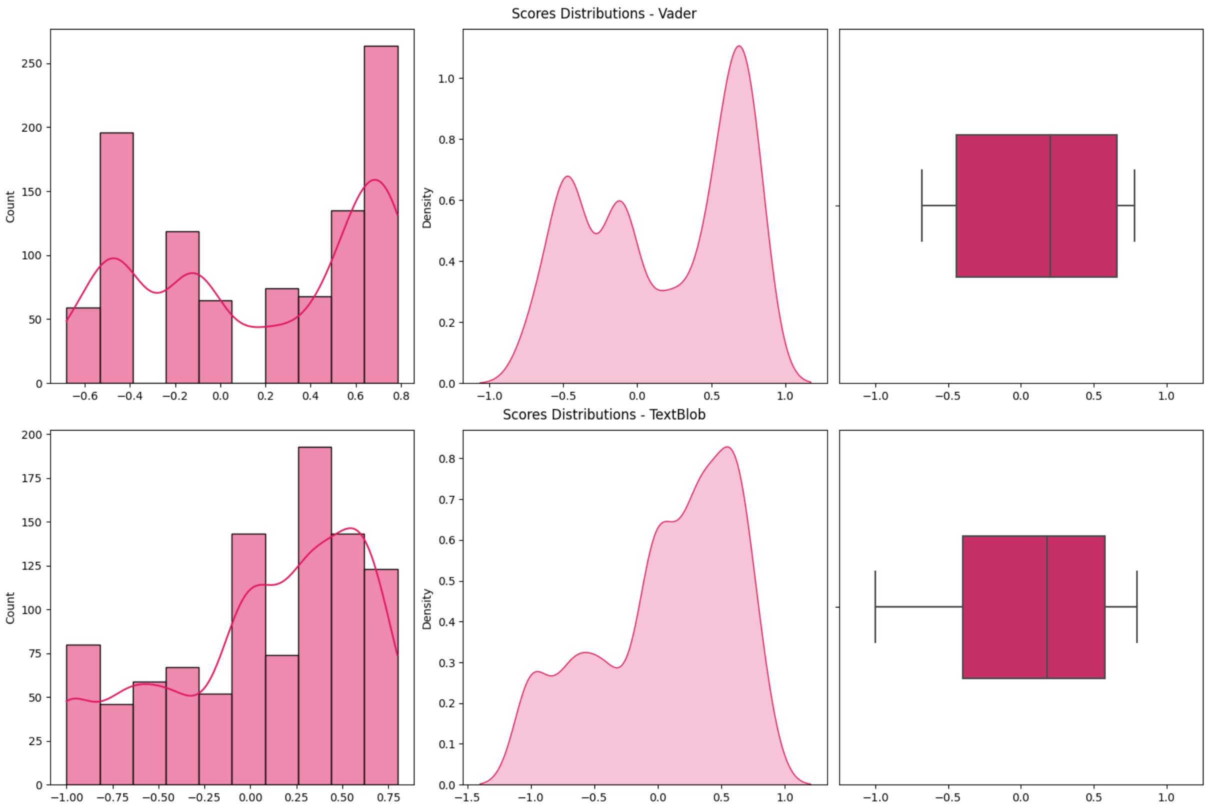
- Vader classifies reviews 2 and 58 as neutral
- Textblob has two reviews (6, 9) with score 0. With three categories, review 6 is assigned a neutral sentiment what is clearly incorrect.
- The BERT model I applied is specifically trained for binary sentiment analysis, so it would be better to use another model trained for multiclass classification.

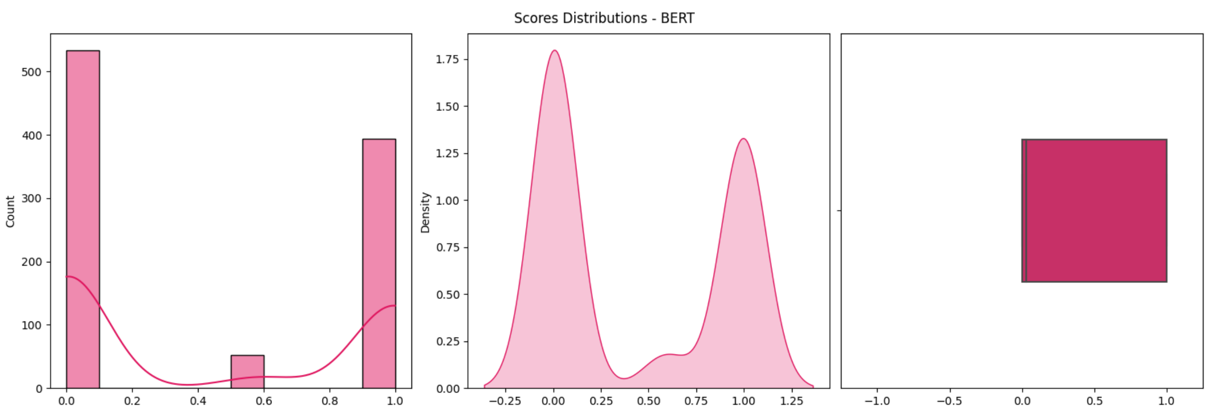
As future work:

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

Visualizations of sentiment distribution.







VADER

Positive Reviews

Negative Reviews

BERT

Positive Reviews

Negative Reviews