



# Recalibrated Restaurant Star Ratings with LLMs

Sang-won Shim

# Online reviews and ratings can make or break a business

1	Project Overview
2	Data Overview
3	Analysis
4	Recommendation
5	Next Steps

**88%** of consumers **read online reviews to decide** whether to experience or purchase a business' product or service <sup>1</sup>

**94%** of consumers **will not do a business** with a company due to **negative online reviews** <sup>1</sup>

*Which restaurant do you want?*



OR



<sup>1</sup> Data Source: Digital Air Strike

# Business Problem

1	Project Overview
2	Data Overview
3	Analysis
4	Recommendation
5	Next Steps

## Current Challenge

→ Restaurant owners believe Google Maps reviews are **arbitrary and inconsistent**

## Project Scope

→ Investigate whether the **user complaints are valid** so that Google Maps PM can make a data-driven decision

# Data Overview

1	Project Overview
2	<b>Data Overview</b>
3	Analysis
4	Recommendation
5	Next Steps

## Source

→ Google Maps API

## Data

→ **~3K star ratings and text reviews** for ~570 unique restaurants

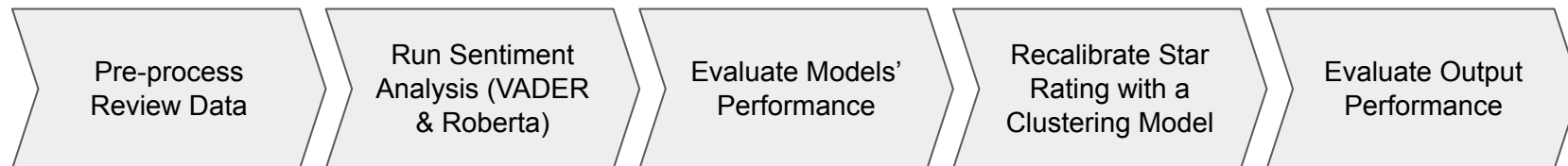
## Limitations

- Due to the limited computational power, a **small dataset** is used for the analysis
- The dataset does not have all/full text reviews for a given restaurant (i.e., limited to **5 reviews per restaurant**)
- The dataset exclusively examines reviews of **New York City restaurants**. Different geography may have quite different relationship between text review's sentiment score and star rating

# Analysis Overview

1	Project Overview
2	Data Overview
3	<b>Analysis</b>
4	Recommendation
5	Next Steps

## Analysis Process



### LLM / Sentiment Analysis Model Used

- VADER
- Roberta (latest model from 2022)

### Clustering Model Used

- K-Means
- Agglomerative

### Evaluation Method

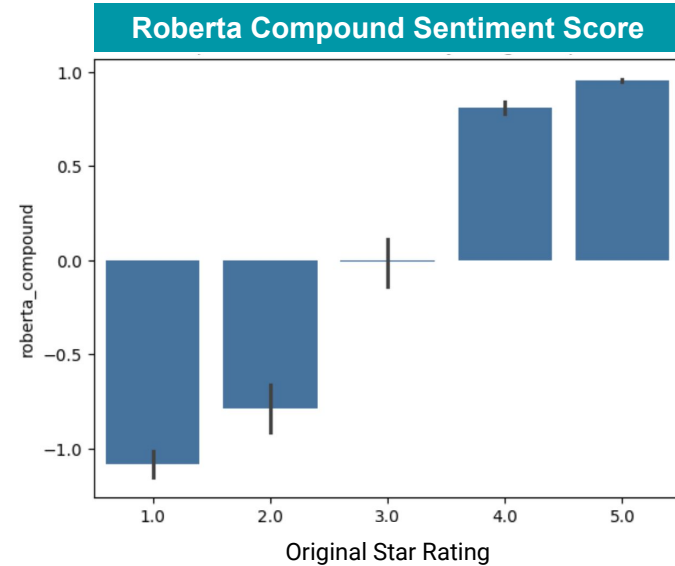
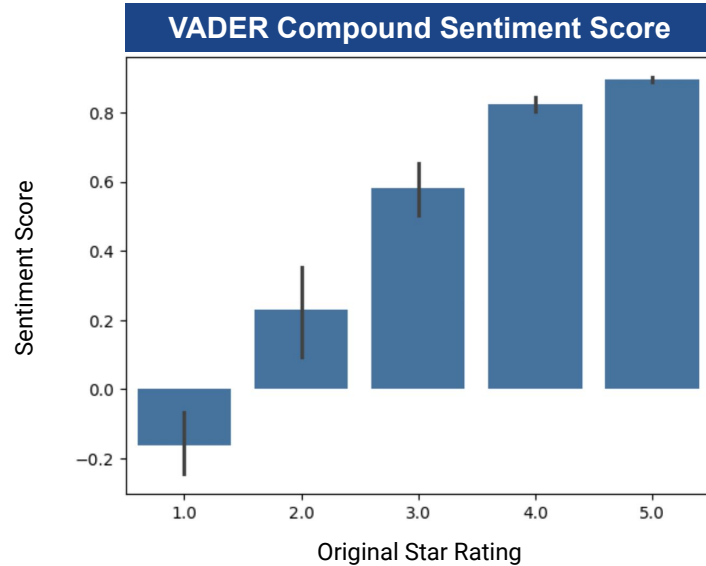
- Sentiment Scores
- Human Inspections

## Important Note

- Sentiment analysis model does not require usual NLP preprocessing steps (e.g., tokenizing, stemming, and lemmatizing) as the models rely on **pre-built lexicon** that contains sentiment scores

# Analysis - LLM Performance

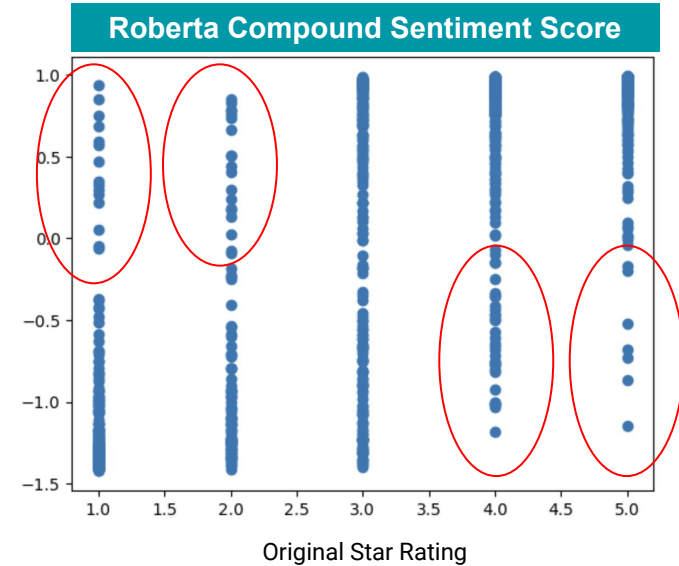
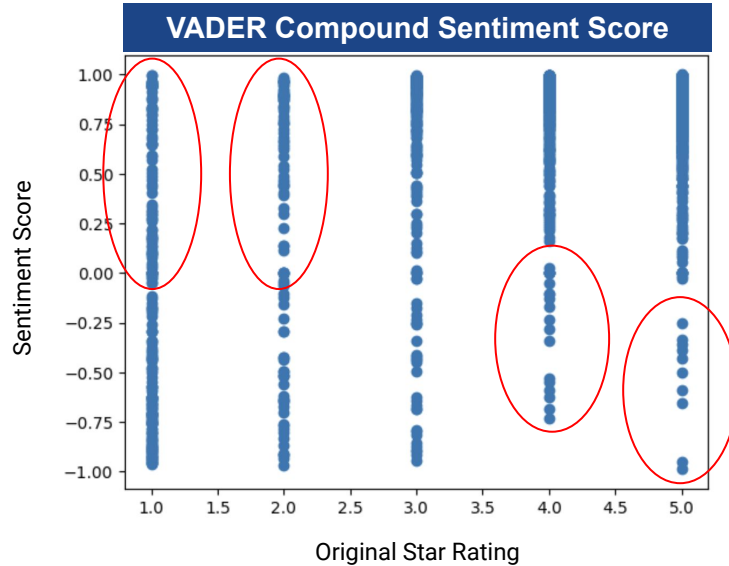
1	Project Overview
2	Data Overview
3	<b>Analysis</b>
4	Recommendation
5	Next Steps



→ On an aggregated level, there is a clear distinction between each star rating and average compound sentiment score for both models

# Analysis - LLM Performance (contd.)

1	Project Overview
2	Data Overview
3	<b>Analysis</b>
4	Recommendation
5	Next Steps



- On an individual review level, there is a **wide variation in sentiment score** for a given star rating
- There must be some **mismatch** between star ratings and text reviews

# Analysis - LLM Performance (contd.)

1	Project Overview
2	Data Overview
3	<b>Analysis</b>
4	Recommendation
5	Next Steps

## Selected 1-star Review

"You're paying \$9 for the fancy truck to get ripped off with mediocre food, barely any chicken and a ton of raw onions after on top to create an illusion. The guy also shouting talking on phone preparing your food spit flying out his mouth. So many better other options near. Never again"

	Negative	Neutral	Positive
VADER	0.000	0.874	0.126
Roberta	0.946	0.048	0.005

## Another 1-star Review

"Delicious dumplings were only surpassed by the warm beer. Limit your ordering to that."

	Negative	Neutral	Positive
VADER	0.073	0.615	0.313
Roberta	0.013	0.053	0.932

- Based on several spot checks, **Roberta model appears to have better performance** than VADER model
- However, neither LLMs were good at picking up on **context or nuances**



# Analysis - Clustering Output Evaluation

1	Project Overview
2	Data Overview
3	<b>Analysis</b>
4	Recommendation
5	Next Steps

## **Example of Good Re-classification (Disagree with user's 3 star rating, and it was re-calibrated as 5 star review)**

"Loved their pizza and calzones! We loved how the pizza bread were chewy and had some garlicky flavor, slices were big and affordable too. Their menu options is wide and freshly made! Will definitely go back when I'm in the city."

## **Example of Mis-classification (Agree with user's 3 star rating, but it was re-calibrated as 5 star review)**

'First off this place is smaller than it looks. The tables are tiny. The pizza is OK. It is not amazing but not awful. The ingredients are fresh and tasty. The service is great here. Prices are good. The place is pretty and stylish.'

→ Since the LLM models were not great at picking up on contexts, nuanced reviews, and sarcasms, clustering performance was **not always optimal**

# Recommendations

1	Project Overview
2	Data Overview
3	Analysis
4	<b>Recommendation</b>
5	Next Steps

1

## Provide Guidance

Post **example reviews** per different star rating buckets to create a general criteria/guideline for what each star rating means (to prevent very harsh or loose star ratings)

2

## LLM Based Star Rating

Beta-test **LLM-based restaurant star rating system** (i.e., based on sentiment analysis model) in addition to the traditional star rating system and get user's feedback

3

## Assistant Feature

Beta-test **star rating assistant feature**: when there appears to be a big mismatch in text review's sentiment score and user's star rating, assistant feature can provide a suggested star rating

---

**With the recommendations, the PM should be able to address concerns from the restaurant owners**

# Next Steps

1	Project Overview
2	Data Overview
3	Analysis
4	Recommendation
5	<b>Next Steps</b>

- With more computational power, re-run the analysis with a bigger dataset
- Research into more advanced sentiment analysis model that can better pick up nuances (e.g., ChatGPT 4.0 API)



**Thank you!**  
**Any Questions?**