Adeline Setiawan

NYC Supermarket Secrets:
Super humble insights from Google reviews of NYC grocery stores

—

ABSTRACT

To what extent do linguistic patterns in online reviews reveal demographic and spatial differences in everyday life? This paper analyses Google Maps reviews of New York City supermarkets to examine how shoppers describe and evaluate grocery stores across chains and neighbourhoods. Using a dataset of 1,331 reviews from 267 stores spanning 14 supermarket chains and three boroughs, the study combines TF–IDF and sentiment analysis to identify distinctive vocabulary, emotional framing, and evaluative structure.

The findings show that reviews encode more than general satisfaction: they reflect chain-specific shopping experiences, neighbourhood-level constraints, and culturally situated expectations. Distinctive vocabulary highlights differences in store format and branding, while sentiment patterns reveal systematic asymmetries between praise and complaint language. By comparing reviews across supermarket chains and locations, this study finds that reviews encode distinct chain-specific experiences, spatial constraints, and cultural identities. Rather than expressing generic sentiment, distinctive vocabulary reflects differences in store format, branding, and neighbourhood context.

KEY RESEARCH QUESTIONS

1. Vocabulary Differences Do different supermarkets attract different vocabulary?
2. Neighbourhood Patterns: Do neighbourhoods shape reviews?
3. Product Mentions: Do people mention different products at different chains?
4. Neighbour Inference: Can we infer socioeconomic factors from language patterns?
5. Complaint Patterns: Are unhappy shoppers unhappy for the same reasons?
6. Praise Patterns: Do happy shoppers value similar things?

—

1. CONTEXT AND MOTIVATION

Online reviews are culturally informed texts. Who leaves reviews, why they do so, and how experiences are described are shaped by varying motivations, including altruism, validation, social norms, incentives, digital affordances, and interactions with high emotional valence. Leaving reviews is one way of user-generated participation on digital public life. It turns private experiences into shared signals, building a network of trust.

In 2007, Google Reviews launched as a feature within Google Maps, distinguishing itself by anchoring online reviews to physical locations. Today, Google Maps hosts billions of reviews for businesses worldwide and accounts for more than 55% of the online review market[1].

Most consumers read reviews before engaging with the business. Sharing opinions publicly about one's experience do more than provide feedback to the business. They serve as social proof; a

[1] https://www.soci.ai/insights/state-of-google-reviews/

smoke signal for your village tribe. Other times, consumers have also taken to Google maps to publicly shame businesses and escalate feedback to headquarters. Since the COVID-19 pandemic, the volume of new reviews has declined and has not yet recovered[2]. At the same time, the proliferation of "fake" reviews written by paid contributors, bots, or AI-generated systems has further eroded trust in online review ecosystems.

The average business on Google holds a rating of 4.11 stars — an average that has steadily inflated alongside a decline in textual review content (from 76.2% in 2015 to 54.2% in 2022)[1]. Prior research indicates grocery stores in North America rank among the top five industry verticals in both review volume and star ratings, while ranking in the bottom five for review length and among the bottom ten for business response rates. Some claim that shoppers are "creatures of habit" and "brand-loyal when it comes to where they shop and what they prefer"[2]. It remains unclear what motivates individuals to leave reviews especially when grocery store businesses rarely respond.

This question is especially salient amid rising grocery prices. Survey evidence shows that nearly 9 in 10 Americans experience grocery stress due to recent inflation[3], prompting consumers to reevaluate shopping habits and increasingly cross-shop across multiple stores. Despite these shifts, grocery retailers continue to overlook online reviews as a channel for engagement and conversion.

This project situates supermarket reviews as a site where consumer experience, geographic context, and brand identity intersect. Supermarkets provide a valuable case for analysis because they are everyday institutions used by diverse populations with different shopping needs, brand identity associations, spatial constraints, and socioeconomic variation.

—

2. METHODS

2.1 DATA COLLECTION

Review data was collected from Google Maps using the Google Places (New) API and a custom script[4]. We searched for 2 values, "supermarket" or "grocery_store". The final dataset consists of 1,331 individual reviews from 267 supermarket locations [Table 1], representing 14 supermarket chains across 3 New York City boroughs and 16 neighbourhoods. Neighbourhoods were selected to capture socioeconomic and demographic diversity. We also included a mix of business and residential districts to capture a variety of shopping patterns. Each observation includes a store name, address, numeric rating, types, place id, review count, and review text, in JSON array.

A brief summary of the dataset. All 267 stores have complete ratings with no missing values. The review ratings range from 1.9 to 4.9 stars, with an average of 4.23 stars (out of 5.0). The median is 531 reviews, with a mean of 1132 reviews per grocery store, indicating some stores have significantly higher review counts. Due to Google Maps API limit of max 5 review counts per store, we were unable to perform analysis on the entire corpus of reviews.

---

[2] https://www.foodandwine.com/grocery-shopping-habits-inflation-lentingtree-survey-2025-11721837

[3] https://www.lendingtree.com/debt-consolidation/grocery-shopping-habits-survey/

[4] http://github.com/halfabluebanana/goo…

The dataset is unevenly distributed across chains, reflecting real-world differences in store prevalence and review activity amongst shoppers.

Trader Joe's leads with 3,454.9 average reviews despite having fewer stores than Whole Foods.
Whole Foods comes second with 3,029.1 average reviews and has the most locations (18 stores). We notice a significant drop-off after the top two chains
Target has notably fewer reviews (99.7) despite having 6 locations.

What makes Trader Joe's shoppers generally more motivated than Whole Foods shoppers to leave reviews, and Target shoppers the least so?

—

2.2 CLEANING + PREPROCESSING + STANDARDISATION

All review text was lowercased and stripped of punctuation, URLs, extra whitespace, and stopwords using the stopwords (NLTK English stopwords). Store locations and neighbourhood labels were standardised to ensure consistency. Supermarket chains were identified using pattern matching on full store names (e.g. grouping all "Trader Joe's" locations under a single chain label).

We retained reviews in their original textual form. To preserve consistency of word forms and interpretability, we conducted lemmatization with WordNetLemmatizer, with no stemming nor aggressive lemmatization (e.g. bought → buy, better → good). We also conducted minimum word length filtering with words that are more than 2 characters.

We also created a list of words to exclude after initial TF-IDF analyses. Some of the categories of words excluded:

- unique employee names (e.g. 'Denisse'),
- common grocery terms (e.g. foods),
- neighbourhood names (e.g. soho, Chelsea),
- street names (e.g. 57th),
- name of stores (e.g. Eataly), except in cases where the name of the store was related to unique food items (e.g. 'dashi' in 'Dashi Okume')
- misspelled words (e.g. 'survice')

—

3. METHODS

3.1 TF-IDF: INDIVIDUAL AND CHAIN - LEVEL

We apply term frequency–inverse document frequency (TF–IDF) analysis to highlight differences across supermarket stores / chains. The assumption is shoppers who perceive commonalities / differences between supermarkets / chains will use distinctive vocabularies in their reviews. In TF-IDF analysis,

- Distinctive words which appear frequently in that store / chain's reviews (high TF)
- Distinctive words which appear infrequently in other chains' reviews (high IDF)

TF-IDF Score = (Word Frequency in Store) / (Total Number of Stores with Word)

Terms with high TF – IDF scores are interpreted as distinctive vocabulary, capturing words and phrases that reviewers disproportionately associate with a particular supermarket. It can be interpreted as suppressing generic sentiment words while amplifying concrete nouns, anything related to products, incidents, or adjectives of experiences.

We analysed TF-IDF at two levels: by individual grocery stores, and by supermarket chains (e.g. Trader Joe's, Whole Foods). For chain-level TF-IDF analysis, we concatenated reviews and treated this as a single "document" to analyse the chain's linguistic profile. TF-IDF is computed for both unigrams and bigrams, and computed consistently across all chains to enable comparison.

—

3.2 SENTIMENT ANALYSIS

- Are TJ shoppers happier than WF shoppers? Are they unhappy for similar reasons?
- Are shoppers from wealthier neighbourhoods happier than shoppers from less affluent neighbourhoods?

We applied VADER sentiment analysis, a lexicon and rule based sentiment analysis tool, to quantify review polarity. VADER returns 4 sentiment scores: positive, negative, neutral and the main score, compound. Compound is a normalised, weighted composite score. We used VADER's standard threshold (0.05) for general sentiment classification, where sentiments are classified as

Positive: compound ≥ 0.05
Neutral: $-0.05 <$ compound < 0.05
Negative: compound ≤ -0.05

In order to assess reviews that are clearly praise or complaints, we first applied VADER analysis with stricter thresholds (0.1) to filter out weak sentiment (0.05 to 0.1 range). We then classified filtered keywords by manually putting them into dictionaries based on categories contextual to grocery shopping. (e.g. Keywords such as 'rude', 'slow', 'unhelpful', 'staff' can be categorised under 'service'; keywords such as 'dirty', 'clean', 'mess', 'hygiene', 'filthy can be categorised under 'cleanliness'].

—

3.3 Stylistic and Complexity Measures

Language complexity, specificity versus vagueness, and lexical diversity are analysed to capture stylistic variation across chains and neighbourhoods.

—

4. FINDINGS + ANALYSIS

4.1 TF-IDF BASELINE FINDINGS

This section examines how distinctive vocabulary reflects perceived supermarket identity, differentiation, and store format.

TF–IDF analysis reveals systematic differences in the vocabulary shoppers use to describe different supermarkets and supermarket chains. Reviews encode distinctive experiential markers tied to store format, branding, and neighbourhood context.

TF–IDF scores cluster near 1.0 across stores, reflecting the routine nature of grocery shopping, with meaningful differentiation emerging primarily in the distribution tails (Fig. 2). Meaningful variations however emerge on the distribution tails. Stores categorised in 'Other', a catch-all construct created for small specialty shops, independent / local stores, or ethnic markets, have a markedly higher variance of TF-IDF scores than large chains. This suggests that shoppers are likely noticing the store' distinctive features and brand differentiation.

Top distinctive terms (e.g. "dashi", "pickle", "vegan") are associated with specialty stores in the 'Other' category. This suggests that shoppers value signature products rather than generalised shopping experiences. Manual inspection confirmed that these terms reflected substantive discussion rather than mere name repetition. We manually inspected review texts (Fig 3) to make sure the high TF score were not only referential to the store name, but substantive discussion of specific offerings. In the case of the exceptionally high scoring "Dashi Okume" (TF-IDF = 13.0), reviewers describe the product quality and cultural authenticity (*"…offers an exceptional level of dashi that even Japanese people would be deeply impressed by"),* preparation practices and experiences (*"they have a wide selection of Japanese ingredients, and you can even make your own dashi").*

In some instances, branding strategies of aligning store names with product lines further reinforce memorability and salience. E.g. "*I finally got to try one of their Dashi Omakase Series", or "a medley of fall vegetables cooked in Tokyo dashi (not sure what that means)"*.

While specialty stores dominate the extreme ends of TF–IDF rankings, keywords from a small number of large chains—e.g. Whole Foods, C-Town, and Trader Joe's— appear among the top distinctive terms. Unlike specialty retailers, these chains are not defined by niche products alone. How do we make sense of how these 3 chains outranked other specialty shops? Whole Foods (TF-IDF score = 6.0), C Town (TF-IDF score = 5.0), and Trader Joe's (TF-IDF) are known to be big chain stores, with their own in-house brand products.

We propose that it's not particularly due to their unique products. Linguistic distinctiveness appears to emerge when shopper experiences subvert expectations associated with large corporate retailers.

For example, a particular reviewer who went to Whole Foods describes an "unforgettable service" encounter that "turned a simple errand into a meaningful encounter," highlighting how emotional salience emerges when corporate retail norms are subverted.

Reviews describing unusually attentive service, unexpected personalisation, or emotionally salient interactions stand out precisely because they contrast with standardised retail norms.

Across chains, TF–IDF patterns suggest two modes of reviews. The heat map for top 20 distinctive words by all major chains (Fig 4) appear to capture words with higher emotional valence (e.g "overweight", "snarled", "jealous") in contrast to the concrete language of other non-chain supermarkets. These linguistic differences motivate a sentiment-based analysis of how distinctive experiences are emotionally framed. We'll use sentiment analysis to interpret how those experiences are emotionally framed.

4.2 SENTIMENT DISTRIBUTIONS AS INTERPRETATION OF TF-IDF PATTERNS

We conducted sentiment analysis across all 267 supermarket stores and chains (1331 reviews) to contextualise vocabulary differences, lexical diversity, and narrative structure. Sentiment scores are positively correlated with average star ratings (Fig. 6). While sentiment scores correlate with star ratings, several chains exhibit distributions that diverge from ratings alone, underscoring the limits of star ratings as summaries of shopper experience (Fig. 6). Some stores / chains (e.g. Food Bazaar, Fairway, C Town) display sentiment distributions that diverge from what ratings alone suggest. This hints at limits of star ratings as summaries of shopper experience.

We observe these clear differences in the mean and deviation of sentiment scores across chains (Fig. 7). For example,Trader Joe's reviews cluster tightly at the positive extreme, indicating consistent emotional tone, whereas Whole Foods exhibits greater variance with a pronounced negative tail, reflecting polarised experiences (Figs. 7 – 8). This compression suggests not only high positivity in sentiments, but also a striking consistency in emotional tone. In contrast, Whole Foods reviews how a more moderate mean alongside a larger variance with a negative tail, indicating polarised experiences. Shoppers of Trader Joe's generally express positive sentiments consistently. In Whole Foods, however, we can interpret this as highly positive narratives when expectations are met, but sharply negative responses when they are not.

Meanwhile in Fig 9, smaller or legacy chains such as Gristedes, Fairway, and C-Town display lower average sentiment and relatively higher proportions of negative reviews. It's remarkable to note how reviews for these stores seem more complaint-driven. We could perhaps infer that customers are disproportionately motivated to leave feedback following negative encounters or interactions with negative valence. Food Bazaar and H Mart, meanwhile, show comparatively high sentiment scores and lower negative proportions. We can interpret this as stronger community identification, cultural familiarity, or localised loyalty effects due to its perceived centrality within the neighbourhood. One review notes how one employee helped "download the Food Bazaar app". Another notes that Food Bazaar is "Best supermarket in Harlem. It's HUGE with a great variety of produce, low prices for the area… I've even taken a few of my friends here to shop and they love it too!" This suggests that further spatial - competitor analysis needs to be done to get a more complete interpretation of sentiment scores.

In some ways, the sentiment patterns surfaced with VADER sentiment analysis closely parallel the TF–IDF findings above in Section 4.1. Chains with highly positive *and* compressed sentiment distributions (e.g. Trader Joe's) seem to also exhibit smaller, more repetitive vocabularies [Fig 5] populated by abstract language (e.g. "amazing", "great"). These often mention no products, services, or events. Meanwhile chains with broader sentiment distributions (e.g. Whole Foods) display greater vocabulary diversity and more detailed experiential language (e.g. "staff helped me find", "manager apologised").

Across all chains, neutral sentiment makes up majority of reviews (Fig. 10), but proportion of positive and negative sentiment varies. Chains with more praises tend to elicit brief but affective endorsements. Complaint-heavy chains attract longer, more concrete reviews about specific incidents, service failures, or perceived injustices.

Finally, sentiment intensity does not always follow star ratings closely. Some chains receive high ratings accompanied by emotionally muted reviews, while others elicit emotionally charged reviews disproportionate to their numeric scores (Fig. 6).

4.3. PRAISE / COMPLAINT ASYMMETRY

Our results suggest that shoppers who leave reviews tend to praise and criticise grocery stores for systematically different reasons. To examine these asymmetries more rigorously, praise and complaint categories were manually curated across all chains. Keywords were grouped into higher-level dimensions based on their narrative function rather than sentiment polarity alone.

Importantly, overlapping terms (e.g. “fresh,” “price,” “staff”) were allowed to map onto different categories depending on context. This reflects the fact that the same attribute may be mobilised either as praise or as criticism, depending on whether expectations were met or violated.
The final praise categories were: quality, selection, service, price, convenience, cleanliness, and atmosphere. Complaint categories included: service, quality, price, availability, cleanliness, checkout, parking, and location.

Across all chains, we observe limited symmetry between praise and complaint categories. While certain dimensions (notably, service and quality) appear in both positive and negative reviews, they are mobilised differently.

Positive reviews are characterised by brief, abstract, affective terms such as “good,” “great,” and “fresh,” (Fig. 11). Negative reviews by contrast, are more likely to reference specific aspects, such as staff behaviour, service quality, pricing, and checkout processes (Fig. 12). Other dimensions such as checkout logistics, product availability, and price are also disproportionately represented in complaints while rarely appearing in praise. We infer that these features primarily function as baseline expectations, easily turning into sources of friction rather than drivers of satisfaction.

We compiled store-specific profiles to summarise what shoppers praise and complain at each chain (Table 3).

4.4 NEIGHBOURHOOD CONTEXT

To what extent does one’s geolocation and neighbourhood shape online reviews? While not geographically representative, the data allow for exploratory analysis of neighbourhood-level variation.

Sentiment scores vary by neighbourhood (Fig. 15). Reviews from neighbourhoods such as the East Village, Lower East Side, Greenwich Village, and Upper West Side exhibit higher average sentiment, while neighbourhoods such as Long Island City, Harlem, and Astoria display lower sentiment. These patterns are consistent with earlier chain-level findings (e.g. Food Bazaar in Harlem), suggesting that local retail conditions and competitor availability within the neighbourhood may shape how shoppers experience satisfaction or dissatisfaction.

Beyond sentiment, neighbourhoods differ markedly in how complaints are distributed across categories (Fig. 16). In Long Island City, the majority of complaints are concentrated in two categories (service and quality), suggesting specific, recurrent points of friction. By contrast, neighbourhoods such as Astoria, Park Slope, and the Upper East Side exhibit complaints spread across a wider range of categories (> 8), indicating more fragmented dissatisfaction that may reflect broader systemic issues rather than isolated failures.

Comparing praise and complaint categories further reveals systematic neighbourhood-level asymmetries. Quality emerges as the most salient evaluative dimension overall, appearing as both the most frequent praise category (25.8% of all praise) and the most frequent complaint category (27.7% of all complaints). Service functions as a volatile dimension: neighbourhoods that strongly praise service also tend to complain about it, indicating high variability in interpersonal encounters rather than uniformly positive or negative experiences.

Two neighbourhoods illustrate these dynamics clearly. Astoria emerges as a negative outlier, combining low average sentiment with high complaint volume concentrated in quality and service, while the Upper West Side exhibits high sentiment, minimal complaints, and price as the sole recurring concern. These cases show that sentiment, review volume, and complaint structure together provide a more nuanced picture of neighbourhood experience than sentiment alone.

Linguistic measures reinforce these patterns. Neighbourhoods with higher mean review length and larger vocabulary sizes exhibit higher linguistic complexity scores and greater variance in TF–IDF values (Fig. 17), indicating more descriptive and varied experiential language. In contrast, neighbourhoods with lower vocabulary size and complexity produce shorter, more compressed reviews dominated by high-frequency evaluative terms and logistics-focused complaints, particularly around availability, checkout, and price (Fig. 16).
These linguistic and evaluative patterns align with neighbourhood-level socioeconomic markers (Table 4). Higher-income neighbourhoods and areas with greater retail density disproportionately praise experiential dimensions such as quality, selection, and atmosphere, while complaints in these areas tend to be selective and expectation-driven, commonly referencing price or service. By contrast, convenience-oriented neighbourhoods praise accessibility and price, while complaints concentrate on availability, quality consistency, and operational friction.

Importantly, these differences persist within the same supermarket chains This suggests neighbourhood context mediates how shoppers articulate experience rather than simply *which* stores they review.

—

5. CONCLUSION

This study demonstrates that Google maps reviews for grocery stories encode more than customer satisfaction or dissatisfaction. These findings show that review language encodes expectations and constraints that are flattened by star ratings or sentiment polarity alone.

This paper examined the extent to which linguistic patterns in online grocery reviews reveal demographic, spatial, and institutional differences in real-world shopping contexts. Using a dataset of 1,331 Google Maps reviews across 267 supermarkets in New York City, the analysis shows that review language encodes meaningful variation in how shoppers experience, evaluate, and narrate grocery stores—variation that is not captured by star ratings or aggregate sentiment alone.

Across supermarket chains, distinctive vocabulary reflects differences in store format, branding, and institutional expectations. TF–IDF analysis highlights how specialty and independent stores are associated with product-specific language, while large chains become linguistically distinctive primarily when shopper experiences deviate from standardised retail norms. Sentiment analysis further shows that highly positive sentiment does not necessarily correspond to rich description: chains with compressed sentiment distributions tend to exhibit repetitive, abstract praise, whereas chains with more polarised sentiment elicit longer, more concrete narratives.

The asymmetry between praise and complaint language provides additional insight into how reviews function as evaluative texts. Praise typically operates as a low-effort endorsement, relying on affective summaries of experience, while complaints are more diagnostic, documenting specific failures, interpersonal conflicts, or structural frictions. These asymmetries vary systematically by neighbourhood context. Spatial analysis demonstrates that the same

supermarket chains are evaluated differently across neighbourhoods, with socioeconomic context shaping whether stores are assessed primarily along experiential dimensions (e.g. quality, selection, atmosphere) or convenience-oriented dimensions (e.g. price, availability, checkout). Linguistic complexity and vocabulary diversity serve as measurable indicators of these locally situated expectations.

Several limitations should be noted. The dataset is constrained by Google Maps API limits, which restrict the number of retrievable reviews per store and preclude analysis of the full review corpus. Neighbourhoods were selected purposively rather than sampled representatively, and the analysis does not support causal inference. In addition, the manual construction of praise and complaint categories involves interpretive judgment, though this approach was necessary to preserve contextual meaning.

Despite these limitations, the findings demonstrate the value of combining interpretable NLP methods with spatial and institutional context. Rather than treating online reviews as homogeneous sentiment data, this approach reveals how language reflects expectations, constraints, and social positioning. Future work could extend this framework to larger datasets, additional cities, temporal comparisons, review authenticity analysis, and more personalised reviews and recommendations. More broadly, the study underscores that online reviews are not merely consequences of consumption, but socially and spatially embedded narratives that offer insight into redesigning everyday experiences — including the mundane act of leaving a review, to participate on the Internet, and perhaps one day, the neighbourhood.