



# Natural language processing analysis of online reviews for small business: extracting insight from small corpora

Benjamin J. McCloskey<sup>1</sup> · Phillip M. LaCasse<sup>1</sup> · Bruce A. Cox<sup>1</sup>

Received: 29 March 2022 / Accepted: 22 December 2023

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024

## Abstract

Receiving and acting on customer input is essential to sustaining and growing any service organization, particularly a small family business whose livelihood depends on strong relationships with its customers. The competitive advantage offered by advanced analytical approaches for supporting decisions is not trivial, and enterprises across virtually all domains of society are investing heavily in this emerging discipline. Natural Language Processing (NLP) is a subset of computer science that employs computational approaches to analyze human language; it is effective at extracting insight from text data but frequently requires large corpora to train its models, in the scale of thousands or millions of documents. This restricts its accessibility to those large enterprises with the capability to capture, store, manage, and analyze such corpora. This research explores a pilot study that applies NLP approaches, specifically topic modeling and large language models (LLM), to assist a small, family-owned business in assessing its strengths and weaknesses based on customer reviews. The relevant corpora of online Facebook, Google Reviews, TripAdvisor, and Yelp reviews is far smaller than ideal, numbering only in the hundreds. Results demonstrate that coherent and actionable insights from big-data approaches are obtainable and that small organizations are not automatically excluded from the benefits of these advanced analytical approaches, with complementary employment of both topic modeling and LLM presenting the greatest potential for similarly-positioned organizations to exploit.

**Keywords** Natural language processing · Topic modeling · Large language models · ChatGPT · Online reviews

---

✉ Phillip M. LaCasse  
phillip.lacasse@afit.edu

Benjamin J. McCloskey  
benjamin.mccloskey.3@us.af.mil

Bruce A. Cox  
bruce.cox@afit.edu

<sup>1</sup> Department of Operational Sciences, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB 45433, OH, USA

# 1 Introduction

On December 8, 2004, United States Secretary of Defense Donald Rumsfeld visited deployed U.S. Army servicemembers assigned to the 278th Regimental Combat Team in Kuwait as they prepared for onward movement into Iraq. During a question and answer period, the former Secretary was asked a pointed question about equipment preparedness. His response was candid, perhaps gruff: *You go to war with the Army that you have, not the Army you might want or wish to have at a later time.*

Approximately a century prior, the oil magnate John D. Rockefeller, Sr. was one of the wealthiest individuals in the United States and the world. In a likely apocryphal story, Mr. Rockefeller was asked by a journalist how much money was enough. His purported response: *Just a little bit more.*

These two seemingly unrelated quotations, one broadcast live to the world by a government official and the other, possibly fictitious, by an oil magnate, find a curious synergy in the plight of the small business seeking to employ cutting edge analytical approaches to make data driven decisions. The first quote might be rephrased, *you perform your analysis with the data that you have, not the data you might want or wish to have at a later time.* The second quote, instead of pertaining to an insatiable desire for wealth, can be contextualized in the insatiable desire for data and the need for advanced machine learning algorithms to utilize massive sets of data for model training. How much data is enough? How much data do we need before we are satisfied? *Just a little bit more.*

The reality is that a small business likely does not have the capabilities of a Google or a Microsoft to capture, engineer, store, manage, and analyze the data relevant to its decisions in limitless quantities. Big data inequality is a relatively recent phenomenon that can be understood in two different ways: application and access. Application refers to the context in which the big data is collected and/or analyzed. Access refers to the degree to which the benefits of big data are distributed equitably across different groups.

Application has been explored in a number of diverse domains including policing (Brayne, 2017), privacy rights (Stewart, 2019), or social inequality (Hacker & Petkova, 2017). Likewise, availability is a matter of context such as underrepresented class (Schradi, 2017) or gender (Prietl, 2019). This study relates to the latter but not on the basis of some demographic factor. Rather, it is contextualized in the reality that large, well-resourced entities have a greater capacity to leverage advanced analytical approaches than small or medium sized businesses. One reason for this is structural: large organizations generate sufficiently large quantities of data and they have a vested interest in analyzing it. A second reason is that large organizations possess resources and infrastructure that smaller organizations do not.

This research explores a pilot study in utilization of natural language processing (NLP) approaches to assist a small, family-owned business in assessing its strengths and weaknesses based on customer reviews. The corpus of customer reviews is far smaller than preferable; the question to be explored is whether the benefit derived from such analysis is worth the risk that its conclusions may be overly dependent on the idiosyncrasies of a small dataset.

The contribution offered by this research is threefold.

1. This research applies NLP approaches to provide targeted feedback to a single small business. The insights obtained from this analysis are not primarily intended to be generalized to any small business or customer base, although it is logically possible this research could be incorporated into larger efforts toward that end. Rather, the intent is to provide support to one specific small business as it seeks to continuously improve its operations and customer experience. This is in contrast to much of the relevant literature in which

- a researcher may apply NLP techniques to a web scrape of tens of thousands of online reviews in order to draw some general conclusion about customer behavior.
2. Second, this research demonstrates that coherent solutions are attainable when the data landscape is less than ideal. It will also be shown that keywords extracted from topic modeling analysis on the small dataset used in this study can be comparable to the keywords extracted from much larger corpora.
  3. Finally, it provides a template for other small businesses, particularly those in the food industry, for use when initiating similar in-house efforts to those undertaken in this study. The decision for a small business to incorporate advanced analytical techniques into its operations may be intimidating, particularly when the small business owners do not have a robust technical background. This study demonstrates a low-cost and low-risk NLP effort, and the specific details on techniques utilized and software employed are included as a starting point for similarly situated businesses to imitate.

## 1.1 Background

Altomonte's Italian Market is a small business located in the suburbs of Philadelphia, PA. Altomonte's has two locations: a smaller store in Warminster, PA and a larger store in Doylestown, PA. The store's business model revolves around providing customers with authentic and genuine Italian Cuisine as well as a first-class dining and shopping experience. Various departments in the Altomonte's enterprise include a catering department, butcher department, cheese and wine department, grocery department, delicatessen, as well as in-house dining or to-go food. Customers can dine on site and try various wine and sandwich pairings or opt to take food to-go from a menu of hot bar, sandwich, and pizza options.

While Altomonte's does have a customer service department, the bulk of the stores' reviews are gathered from online resources: Yelp, TripAdvisor, Google Reviews, and Facebook. The process to review and take correction based on the content of these reviews is manual and ad hoc.

## 1.2 Motivation

The motivation of this research is to provide tractable, applicable support for small businesses, especially those in the food industry, to begin incorporating advanced analytical approaches to extrapolate information that might otherwise be missed by their customer service departments executing surface-level analysis. Altomonte's currently uses no NLP techniques for in-house analysis, and this pilot study represents the first time that a multivariate analysis approach was used for the business's research and development.

## 1.3 Research questions

This study seeks to answer the following research questions:

1. Can NLP approaches, specifically topic modeling, provide a small business such as Altomonte's with a candidate set of its core business processes or offerings for focused attention? That is, is topic modeling useful to identify either deficient areas for which improvement efforts may be focused or areas of strength that indicate directions in which the business may expand its offerings?

2. Do large language models (LLM) such as Chat Generative Pre-Trained Transformer (Chat-GPT) provide equal or superior results to topic modeling without requiring the technical skills needed to implement topic modeling?

## 1.4 Organization

The remainder of the paper is organized as follows. Section 2 provides an overview of relevant NLP approaches and their utilization in literature. Section 3 details the process of using various NLP techniques and methods to gather insight on the reviews. Section 4 presents and discusses the results, and Sect. 5 closes with conclusions and recommendations for future study.

## 2 Related work

This section discusses the relevant NLP techniques that were performed on the Altomonte's Italian Market online reviews corpus. In addition, applicable prior research is included to emphasize the importance of the research and provide a resource for small businesses in the food industry to adopt these ML methods of data extrapolation.

### 2.1 Topic modeling

In NLP, the term *corpus*, or its plural *corpora*, refers to all of the documents contained in the data set of interest. A *document* is a collection of strings. Documents in turn are made up of *terms* or *tokens* which represent an individual component of the data set. Topic modeling in NLP seeks to find similarities beneath the context of a corpus or corpora. An assumption for topic modeling is that the observed documents and terms interact with latent parameters and they do so according to some probabilistic distribution that can be employed to extract the unidentified characteristics of the corpus (Vayansky & Kumar, 2020).

There exists a reasonably broad and growing body of relevant literature to topic modeling. To illustrate, consider the values in Table 1, obtained via a simple ScienceDirect query for "topic modeling" and "topic modeling" + "online reviews" in the abstract, title, or keywords from 2013 through 2022.

Examining simply the raw numbers yields three observations. First, and not terribly controversial, is to observe that topic modeling research has grown moderately over the past 10 years but not explosively or dramatically. Second, the subset of topic modeling research that specifically explores online reviews represents a tiny subset of the body of work. Finally, replacing "online reviews" with "small business" in Query 2 yields exactly zero results. The implication is that, as previously hinted in Sect. 1, this study has an opportunity to address a gap in the current body of work.

A closer look at the specific content of the articles yields two relevant patterns. The first pattern is that the research employs large corpora to train its models. Consider the following examples:

- 697,076 Yelp reviews to analyze the relationship between star rating and sentiment (PRITHIVIRAJAN et al., 2015)
- 24,367 customer reviews of hotels in India (Piramanayagam & Kumar, 2020)
- 47,172 hotel reviews in Las Vegas (Sanchez-Franco et al., 2019)

**Table 1** Growth of topic modeling research 2013–2022

Year	Query 1 “Topic modeling”	Query 2 “Topic modeling” “Online reviews”
2013	40	0
2014	39	1
2015	62	1
2016	83	2
2017	89	2
2018	119	4
2019	156	9
2020	180	7
2021	254	15
2022	272	12

- 2,799,420 Airbnb reviews exploring customer experiences (Zhang, 2019)
- 538,000 TripAdvisor reviews relating customer experience with embedded photographs (An et al., 2020)
- 35,401 Tweets to analyze factors influencing successful start-ups (Saura et al., 2019)
- 1,048,576 Amazon reviews of grocery products (Heng et al., 2018)
- 311,550 online automobile reviews to analyze customer sentiment (Park et al., 2021)

The second observed pattern is that applications of NLP are broadly focused versus targeted to a specific organization. For example, studies often queried large corpora stratified by industry group such as the hospitality industry (Aktas-Polat & Polat, 2022; An et al., 2020; Piramanayagam & Kumar, 2020; Sanchez-Franco et al., 2019; Zhang, 2019), product reviews (Luo et al., 2019; Ding et al., 2015; Wang et al., 2014), entertainment (Lee & Kim, 2017), general consumer behavior (Krestel & Dokooohaki, 2015; Kwon et al., 2020; PRITHIVIRAJAN et al., 2015), or employee satisfaction (Sainju et al., 2021). Another common behavior was to apply a model such as SERVQUAL to a specific industry (Ding et al., 2020; Marcolin et al., 2021; Palese & Usai, 2018).

The implication is that this study represents a relatively underexplored application of topic modeling. Kumar et al. (2021) provide support for this intuition in their systematic literature review of text mining in service management. Identified areas of emphasis include social media analysis, market analysis, competitive intelligence, risk management, and fake content detection (Kumar et al., 2021). Absent from the list is process or business model improvement for an individual customer. This gap in literature may be explained by the competitive nature of the business marketplace; a small business may not wish for its feedback to become part of the body of knowledge. Rather, it may prefer to quietly identify and incorporate feedback for maximum competitive advantage.

## 2.2 ChatGPT

ChatGPT (2023) is a novel, disruptive advancement in natural language processing, the implications of which are not yet fully clear (Haleem et al., 2022). Initially published in November 2022 by the Open Artificial Intelligence (AI) initiative, ChatGPT allows users to enter conversational text strings and receive responses.

The body of scholarly work concerning ChatGPT is new but growing explosively. In the ScienceDirect database, there is exactly one article containing *ChatGPT* in the abstract, title, or keywords in 2022; as of 28 July 2023, that number had grown to 298 total articles, of which 93 are research papers.

The research articles are in the disciplines of medicine/dentistry (42), social sciences (19), business (15), computer science (12), economics (7), engineering (7), and several other categories (five or fewer each). Because an article may fall into more than one category, the numbers do not sum to 93.

Of immediate interest in the emerging ChatGPT research are the academic implications such as student classroom assignments (Yanfang et al., 2023; Keiper, 2023), literature review (Wu & Dang, 2023), coding (Yilmaz & Yilmaz, 2023), and linguistics (Lin, 2023).

In the business domain, initial research efforts have focused on applications such as autonomous trip planning (Wong et al., 2023), generation of recommendations to consumers (Kim et al., 2023), generation of entrepreneurial rhetoric (Short & Short, 2023), and creation of policy (Dwivedi et al., 2023; Howell & Potgieter, 2023).

Given the current state of ChatGPT research, two relevant implications are apparent. First, as was observed in Sect. 2.1, small business does not appear to be a significant area of interest. Second, more generally, the literature consistently provides scenarios such that ChatGPT presents the appearance of a quick and easy solution but, upon further examination or exploration, there is concern about the veracity or usefulness of the results. One specific example is the literature review case study cited previously in Wu and Dang (2023).

### 3 Methodology

This research employs a blend of the approaches taken in the reviewed literature. The Python programming language, Version 3, was used with several specialized libraries throughout this work. The NLP modules that Python offers are NLTK (Bird et al., 2009) and Gensim (Řehůřek & Sojka, 2010); other libraries used extensively were TensorFlow (Abadi et al., 2016), Keras (Chollet, 2015), NumPy (Harris et al., 2020), and Pandas (McKinney, 2010). For running experiments, the Python library AxClient was utilized. Jupyter Notebook was the Integrated Development Environment (IDE) that supported the code's execution.

#### 3.1 Data

The data for this research consists exclusively of online reviews for Altomonte's Italian Market from Google Reviews, TripAdvisor, and Yelp.

The information was scraped from the three online sources and placed into a.csv file via a manual process. That is, reviews were manually copied and pasted into the file. The file contains four columns: year, review, rating, and platform. The.csv file was then loaded into a Jupyter notebook and partitioned into four different dataframes: one containing all reviews and one each for Yelp, Google Reviews, and TripAdvisor. Table 2 contains the number of reviews from each source, ranging from 176 to 253 for a total of 644 reviews.

#### 3.2 Data preprocessing

Data preprocessing consists of a series of steps common to many NLP analyses: stop word removal, lemmatization, and tokenization.

**Table 2** Number and source of reviews

Source	Number
Google reviews	215
TripAdvisor	253
Yelp	176
Total	644

In NLP, a stop word is a commonly occurring word that typically provides little value to the analysis. Examples include definite and indefinite articles, forms of the verb *to be*, or joining words such as *and*, *but*, and *or*. These words, as well as any word with fewer than three characters, were removed from each review.

Lemmatization is the process that converts differently-inflected words to their common root form or lemma. This process does not simply omit prefixes or suffixes but rather accounts for details such as parts of speech or other elements of the word's internal structure. An example might be the conversion of the words *perform*, *performed*, and *performing* to the lemma *perform* but retaining *performer* as *performer* due to its status as a noun versus a verb. This research employed *Spacy* and the model *en\_core\_web\_sm* for lemmatization.

Tokenization is the process of splitting the document into individual, word-level units called tokens. The tokens comprise a dictionary of words from which the topic modeling algorithm performs its groupings.

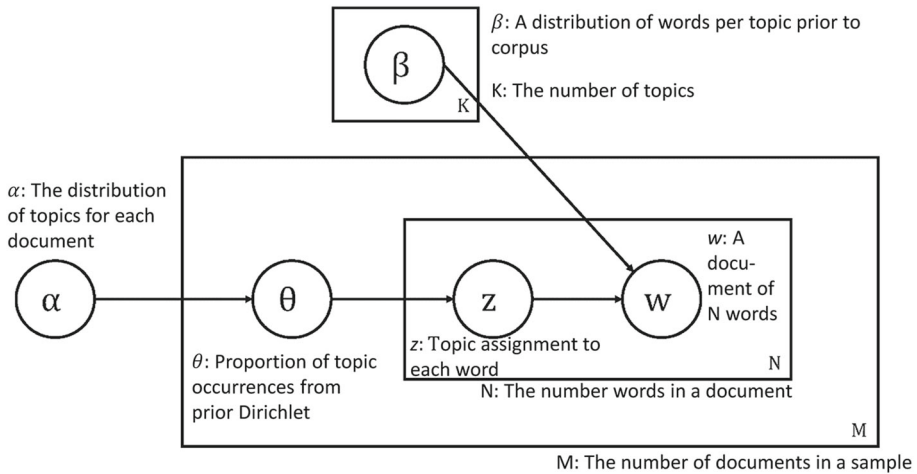
Upon completion of these steps, the reviews were converted into a document term matrix using Gensim's *doc2bow* function.

### 3.3 Topic modeling

To address the first research question from Sect. 1.3, this study employs Latent Dirichlet Allocation (LDA), a topic modeling approach that breaks a data set into words, topics, and documents. LDA views documents as the result of some randomized mixture of hidden topics, and those topics can be modeled as probability distributions whose support is the set of words in the corpus (Vayansky & Kumar, 2020). LDA is preferable for this research to Latent Semantic Analysis (LSA) because LSA employs a process called single value decomposition to determine the similarity of each document and group them into categories. This study is not interested in similarity between reviews; rather, the interest is in the topics contained therein.

Figure 1 provides an overview of the LDA process. The hyperparameter for the Dirichlet Distribution is denoted by  $\alpha$ , which is the prior count of how many times a topic appears in the document. The parameter  $\theta$  is the mixture proportion of  $K$  topics that come from the prior Dirichlet Distribution. The hyperparameter  $\beta$  denotes the number of times a word is sampled from a topic prior to the corpus,  $z$  is the probability distribution associated with each topic, and  $w$  represents the  $i$ th word in a sequence that formulates the matrix of a document. Finally,  $M$  and  $N$  represent, respectively, the repeated sample executed on the corpus and the repeated sampling conducted on each document.

Assumptions for LDA include a known and fixed value of  $K$  and that words and documents contained in a corpus are independent. Picking a value for  $K$  is an iterative process and one way to obtain this value is to compare the resulting model on the basis of some metric such as the perplexity score (the normalized log-likelihood calculated on a held-out test data set), which measures how probable the new data is given the postulated model. To illustrate perplexity, consider a fair coin versus an unfair coin. The perplexity score will be



**Fig. 1** LDA Process, adapted from Vayansky and Kumar (2020)

higher for the fair coin because its outcome is harder to predict, whereas the unfair coin has a lower perplexity score because the result is easier to predict when one outcome has a higher probability of occurrence than the other. Perplexity scores will be generated for models associated with each prospective  $K$  value. The independence of topics follows from the independence of words and documents.

---

#### **Algorithm 1** Topic Model Analysis: LDA Process

---

1. Remove stopwords, lemmatize and tokenize reviews
  2. Create a bag of words from reviews, forming a dictionary
  3. Run LDA model
  4. Compute perplexity
  5. Compute coherence score
  6. Choose number of topics based on perplexity and coherence
  7. Interpret topics
- 

Algorithm 1 contains the LDA approach to identify unobserved topics and characteristics of the reviews. Perplexity supports the analysis by quantifying the ease with which topic predictions are made from the underlying probability distributions. Coherence score is used as a measure for quality between the topics and words. Low perplexity and high coherence scores are desirable.

### **3.4 ChatGPT**

To address the second research question in Sect. 1.3, this study attempted to obtain topics, topic summaries, and other useful information from the corpus of reviews using ChatGPT. Initially hoping to directly copy and paste the reviews into ChatGPT, this was not successful because, despite the corpus being small by NLP standards, there were too many reviews for the free ChatGPT interface. The successful approach was to create a public Github repository with the reviews in a.csv file, from which ChatGPT can read the data and respond to the following inquiries:



1. Can you give me 5 topics from this CSV File [sic]
2. Can you put each of these topics into a sentence for me?
3. Do this again but by year. Can you also provide an in-depth description of the sentiment by year
4. Open the csv [sic] from the link I provided before. based [sic] off the reviews, what is one product we should improve upon?
5. Open the csv [sic] from the link I provided before. Which product that we offer do you believe is best based off the reviews?
6. Open the csv [sic] from the link I provided before. What are our top 5 best products?
7. Open the csv [sic] from the link I provided before. What products should we work on improving?

## 4 Results and analysis

### 4.1 Corpus statistics

The first step of this analysis was to calculate key statistics to gather insight into how customers have perceived Altomonte's through their ratings in the last 10 years. The statistics calculated for the data are the number of reviews by source, average review rating, and distribution of stars.

The results in Table 3 show that, of all corpora, Yelp had the lowest average rating of reviews with 3.64. The overall corpus average review was calculated to be 4.23. There is approximately an even split of reviews contributing to the entire corpus as well: 27% of reviews from Yelp, 39% of reviews from TripAdvisor, and 24% of reviews from Google Reviews.

After calculating the average ratings for the reviews, the frequency distribution of words in each corpus was extracted. As shown in Table 4, the most common words portray Altomonte's as a, "good, even great, Italian Market that is known for its sandwiches." Therefore, the process of how the sandwich counter operates and the consistency of its sandwich production

**Table 3** Corpus statistics skew positively across all sources, with Yelp the most balanced

Corpus	1-Star	2-Star	3-Star	4-Star	5-Star	Mean
All	39	44	49	101	411	4.230
Yelp	26	21	23	27	79	3.640
Trip advisor	4	6	19	56	168	4.49
Google reviews	9	17	7	18	164	4.410

**Table 4** Top words by frequency are largely consistent across sources

Corpus	Word 1	Word 2	Word 3	Word 4	Word 5
All	Food	Italian	Good	Great	Sandwich
Yelp	Good	Italian	Food	Store	Great
Trip advisor	Food	Italian	Good	Sandwich	Great
Google reviews	Italian	Food	Store	Good	Great

**Table 5** Featured words allow inference of topic summaries

Word 1	Word 2	Word 3	Word 4	Word 5	Topic (inferred)
Great	Italian	Pizza	Item	Food	Pizza
Food	Good	Fresh	Italian	Selection	Fresh Food Selection
Great	Italian	Sandwich	Selection	Place	Sandwich Selection
Food	Italian	Great	Selection	Store	Italian Food Selection
Food	Good	Store	Italian	Place	Atmosphere
Good	Food	Italian	Quality	Great	Food Quality

should be viewed as a core strength and leveraged, perhaps through sandwich specials to entice more customers into the store.

## 4.2 Topic modeling results

The Topic Model analysis identifies seven topics that could describe reviewers' underlying perceptions about Altomonte's. Table 5 lists the different topics with their associated featured words, sorted by level of importance from greatest to least. Note that the topic name is not provided by the LDA algorithm. Rather, it is inferred based on the featured words in conjunction with the domain knowledge of the subject matter expert. In this case, the topics align both to high-level customer observations such as atmosphere or food quality as well as to specific offerings such as pizza or sandwiches.

Selecting  $K = 7$  topics results in the highest coherence value for all ranges of topics, achieving a value of 0.40. While not an especially high value, it is acceptable for a small corpus. In addition, these same seven topics had the lowest perplexity score of  $-6.81$ . The topics in the underlying distribution of words all focus on the food menu, showing that customers perceive that Altomonte's has great tasting sandwiches and pizza as well as a variety of Italian products to offer.

Examining the association of featured words with each topic also provides actionable insight that could be useful to Altomonte's. For example, while the topics "Pizza" and "Italian Food Selection" are associated first with the word "Great", the topic "Atmosphere" is only ever associated with the word "Good". Taken into consideration with results from Huang et al. (2014), this may indicate that additional effort in décor would address a possible weakness. It may also be useful to note which of Huang et al.'s criteria (service, value, takeout, décor, and healthiness) do not appear in corpus topics or associated words. Such examination of LDA corpus statistics quickly provides easily actionable insights for small business owners, although there is often some nuance to the results. For example, a negative review might make a comment about slow service. However, service may be slow simply because the restaurant is busy and not because employees work slowly.

**Table 6** Featured words from large-corpus topic modeling in PRITHIVIRAJAN et al. (2015)

Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Great	Ordered	Maggianos	Chicken	Just	Restaurant
Great	Good	Place	Table	Also	Well
Food	Place	Service	One	Ive [sic]	Good
Food	Good	Great	Maggianos	Just	Like
Food	Good	Service	Place	Maggianos	Just
Good	Great	Food	Maggianos	Place	Time
Food	Great	Good	Really	Back	Delicious

The implication is that a seemingly negative topic can be an indicator for an overall positive business outcome. The key to recognizing whether the feedback is rooted in something positive or negative is the human-in-the-loop review of the customer feedback by a knowledgeable party.

One may observe that the set of featured words in Tables 4 and 5 is relatively small, with substantial repetition of common words such as *good*, *great*, *food*, or *Italian*. On this basis, it is fair to question the usefulness of the topic modeling approach in extracting insight from the corpus. Two thoughts may be given in response. The first is to acknowledge the limitation in the study. Clearly, a larger corpus might be expected to contain more variety in its words. Additionally, a larger corpus would give the luxury of designating certain terms such as *Italian* as stop words. Because the business is named Altomonte's Italian Market, it is entirely reasonable to expect that the word *Italian* would feature prominently in any topic modeling result and that the analysis may be better served by removing it from consideration.

A second thought, however, is that customers, when writing reviews, likely tend to choose from a relatively small set of descriptive words in their vocabulary. To express satisfaction with an experience, the words *good* or *great* would likely be frequent word choices. In support of this notion, consider the topic modeling results in Table 6. This table, reproduced in part from PRITHIVIRAJAN et al. (2015), contains the top words from a topic modeling analysis of 50,000 restaurant customer reviews.

Comparing Tables 5 and 6, it is notable that the same three words appear as the most common word in each topic: *great*, *good*, *food*. There is also strong consistency with the second most frequent word. Both analyses add *place*; Table 5 adds *Italian* and Table 6 adds *ordered*. We only begin to see greater variation in the identified words farther down the precedence list.

### 4.3 ChatGPT results

ChatGPT produces responses to the queries that appear plausible on a surface level. However, a closer look reveals nuance that requires careful consideration for small businesses seeking to use this resource.

Consider the summary topics summarized in Fig. 2, which is a screen snip of the ChatGPT response to the first two inquiries provided in Sect. 3.4.

Comparing these topics to the topics extracted by LDA in Table 5, there is clearly a consistency between them. However, a closer look at the details reveals one discrepancy. Specifically, Topic 5 generated by ChatGPT describes the gelato at Altomonte's. There is no

Topic 1: "The cheesesteak and hoagie at Altomontes are delicious, with a perfect bread and tasty fillings."

Topic 2: "Altomontes serves excellent pizza with a flavorful crust, delicious sauce, and generous cheese toppings."

Topic 3: "The staff at Altomontes is incredibly friendly, providing attentive and helpful service with a polite attitude."

Topic 4: "Altomontes offers a variety of sandwiches for lunch, made with fresh ingredients and providing a tasty and delicious meal."

Topic 5: "The gelato at Altomontes is absolutely delightful, with a wide range of flavors to choose from. It's creamy and rivals the best ice cream."

**Fig. 2** Four of five topics generated by ChatGPT are plausible but one (Topic 5) is suspect

gelato-related topic in the LDA results, and a search of the corpus of reviews reveals that the word *gelato* does not appear in a single review. Yet, for some reason, ChatGPT produced a plausible-sounding statement implying that gelato is a customer favorite.

Consider also Fig. 3, which contains the ChatGPT-generated response to the third inquiry from Sect. 3.4.

Each of the summaries, without considering the actual numbers of positive, negative, and neutral reviews, is plausible and consistent with the statistics of review ratings in Table 3. However, the counts of positive, negative, and neutral reviews provided by ChatGPT in Fig. 3 are incorrect. For example, ChatGPT received 103 reviews written in 2018, whose counts by star rating were 71 (five stars), 14 (four stars), six (three stars), four (two stars), and eight (one star). It is difficult to reconcile these actual values with the counts of positive (40), negative (10), and neutral (5) reviews stated by ChatGPT.

One final illustrative example will suffice. Figure 4 contains the response to the final inquiry from Sect. 3.4, which asks for areas of improvement.

As has been the observation thus far, ChatGPT provides reasonable narratives for each area of improvement that it identifies, and it even includes a disclaimer at the end of the response. However, as before, the statements made by ChatGPT do not align with the actual text of the reviews, most notably in the third (Soup) and fourth (Bread) item on the list. For example, in the entire corpus there are 83 one-star and two-star reviews. In those 83 reviews, the word *bread* appears in only eight, of which two are critiques of the dryness (one) and flavor (one). In those 83 reviews, the word *soup* appears exactly once, and the context is a mistaken order and not flavor or seasoning as stated by ChatGPT.

#### 4.4 Theoretical implications

Topic modeling results are firmly grounded in the actual text of the reviews by virtue of the LDA algorithm. However, there is a certain baseline level of expertise that is required to employ LDA. For example, the number of topics,  $K$ , is a hyperparameter that must be set either arbitrarily or via some defensible process. As described in Sects. 3.3 and 4.2, this research employs coherence and perplexity.

- 2018: In 2018, the sentiment towards Altomontes was generally positive, with a significant number of positive reviews (40). Customers expressed satisfaction with various aspects of the food and service. However, there were also a notable number of negative reviews (10), indicating some areas of improvement. Additionally, a few reviews were classified as neutral (5), suggesting a mixed sentiment during that year.
- 2019: The sentiment in 2019 leaned towards the positive side. Customers expressed their satisfaction with Altomontes through positive reviews (55). They highlighted various positive aspects of their experience, such as the quality of the food and the service provided. Nonetheless, there were also a considerable number of negative reviews (15), indicating areas where improvements could be made. A few reviews fell into the neutral category (8), suggesting a mixed sentiment within the year.
- 2020: During 2020, the sentiment towards Altomontes remained predominantly positive. Many customers expressed their satisfaction through positive reviews (62), praising the food quality, service, and overall experience. Negative reviews (12) were relatively fewer in number, indicating a generally positive sentiment among customers. A small number of reviews were classified as neutral (6), suggesting a mix of opinions.
- 2021: In 2021, the sentiment towards Altomontes continued to be mostly positive. Customers expressed their positive sentiments (50) regarding various aspects, including food quality, taste, and service. Negative reviews (8) were relatively minimal, indicating an overall positive sentiment among customers. A few reviews were classified as neutral (3), suggesting a balanced sentiment within that year.

**Fig. 3** ChatGPT generates yearly summaries of customer attitudes that are plausible in appearance but incorrect in the details

ChatGPT results, however, are simply the result of the probabilistically-determined next word in the string. Additionally, ChatGPT is pre-trained on a large corpus containing many thousands of words. As a result, words not found in the corpus of reviews may be identified as the highest probability next word by ChatGPT. This is why ChatGPT identifies, for example, poorly-flavored soup as an area to improve when that feedback does not actually appear in a single online review.

Taken together, the previous two observations nicely illustrate the primary theoretical implication of this case study, which is the relationship of the human agent in the entire process. In autonomous systems, a *human-in-the-loop* (HITL) system is one in which the autonomous system may only act independently for a time after which it must stop and await approval from the human agent before proceeding (Nahavandi, 2017). This is in contrast to a *human-on-the-loop* (HOTL) system in which the autonomous system fully operates but the human agent serves in a monitoring or supervisory role (Nahavandi, 2017). The suitability of an HITL versus HOTL approach is relevant to weapons system targeting (Scharre & Horowitz, 2015) as well as other disciplines such as asset management (Chen et al., 2021).

In this case, the LLM approach is not terribly well suited for HITL controls, and the results of this study demonstrate that it cannot be blindly trusted. For this reason, the HOTL approach is perhaps appropriate. This is consistent with other research into LLMs such as

Apologies for the previous inconvenience. Let me retrieve and analyze the CSV file you provided to identify the products that could benefit from improvement based on the reviews.

After analyzing the "reviews" column from the Altomontes\_reviews.csv file, the following products could potentially benefit from improvement based on the feedback provided:

1. *Cheesesteak: The cheesesteak received mixed reviews, with some customers expressing dissatisfaction regarding the taste, quality of ingredients, or overall flavor profile. Improving the cheesesteak recipe, sourcing higher-quality ingredients, or addressing specific taste concerns could enhance customer satisfaction.*
2. *Salad: There were a few negative comments about the salad, suggesting that improvements can be made in terms of freshness, variety, or dressing options. Consider enhancing the salad offerings with fresher ingredients, more diverse options, or creating unique dressings to elevate the dining experience.*
3. *Soup: Some customers mentioned issues with the soup, such as inconsistent flavors or lack of seasoning. Focusing on refining the soup recipes, ensuring consistency in taste, and enhancing seasoning can contribute to a more enjoyable soup experience.*
4. *Bread: A few reviews mentioned aspects of the bread, such as it being dry or lacking flavor. Working on improving the bread quality, ensuring freshness, and exploring different varieties or flavors can enhance the overall dining experience.*
5. *Customer Service: While not a specific product, a few reviews indicated dissatisfaction with customer service. Improving customer service by providing attentive and friendly interactions, addressing concerns promptly, and ensuring a positive dining experience can significantly impact overall customer satisfaction.*

*Please note that these recommendations are based on the analysis of the provided CSV file and the sentiments expressed in the reviews. It's essential to review the specific feedback and comments related to each product to gain a comprehensive understanding of the areas that may require improvement.*

**Fig. 4** ChatGPT generates recommended areas to improve that are plausible in appearance but are not supported by the actual text of the reviews

Wu and Dang (2023). One could argue that selection of the hyperparameter  $K$  in the LDA algorithm is a HITL step, although that process could certainly be automated with minimal difficulty.

A secondary theoretical implication is that this research demonstrates, as illustrated by Tables 5 and 6, that a large corpus is not necessarily a prerequisite for obtaining reasonable topic summaries using LDA. That is, featured words from this small-corpus study were not terribly dissimilar to the featured words in the large-corpus study in PRITHIVIRAJAN et al. (2015).

The reasonable explanation is that customers tend frequently to use a relatively small set of common words, and a large corpus is not necessary to extract them. A larger sample size of

reviews may be needed in order to determine the subtleties for what prompted those common words, but as demonstrated in this study even a small corpus was able to provide specificity for Altomonte's by highlighting its pizza, fresh food, and sandwich offerings in the identified topics.

#### 4.5 Practical implications

The practical implication of this research is that insight can in fact be gained from a small corpus, notwithstanding the preference that large corpora be used to train NLP models. There is risk associated with exclusively employing a LLM such as ChatGPT in that the conclusions may be incorrect. In this study, there is risk that ChatGPT feedback might prompt a thorough review of the business's soup offerings when, in fact, that feedback was not sourced in the actual data.

In spite of that, incorporation of a LLM such as ChatGPT in conjunction with a traditional topic modeling approach such as LDA is entirely suitable with HOTL serving as the check and balance. The LDA topics, extracted directly from the corpus of reviews, may be compared to the LLM results. If the topics are consistent, then the analyst or human agent can have confidence in the more user-friendly ChatGPT output. If there are discrepancies, then the human agent is free to investigate as warranted. This investigation could employ additional advanced analytical approaches or, as in the case of this study, a simple, manual keyword check of the raw review data.

### 5 Conclusion

This study attempts to address a gap in the growing body of NLP literature, specifically that of tailored application to a specific entity of interest under the conditions of a smaller than ideal corpus. The entity of interest for this study is a small, family-owned delicatessen, Altomonte's Italian Market.

The first research question asks whether NLP approaches, specifically topic modeling, can be useful for a small business. The answer is a cautiously optimistic yes. Subject matter experts from the small business reviewed the topic modeling results and affirm that LDA did extract topics that are relevant and consistent with their domain knowledge.

The second research question asks whether emerging large language model applications such as ChatGPT can produce comparable results without the technical skills required for implementing traditional NLP approaches. The answer here is more nuanced. It is true that employing ChatGPT does not require technical skills in advanced analytical approaches such as those required for NLP. However, obtaining the results in Sect. 4.3 required more than entering some simple keystrokes into the ChatGPT prompt.

For example, as previously discussed in Sect. 3.4, it was necessary to create a public Github repository with the reviews consolidated in a.csv file. Depending on the technical skill of the user, this may or may not be a trivial task. Secondly, ChatGPT was initially resistant to opening the.csv file and performing the requested tasks; its default response was something to the effect that it does not have access to the.csv file.

Ultimately, these challenges did not prevent this study from obtaining its results, but it required perseverance and carefully crafted prompts. At one point, upon stating in the prompt that *I own the CSV and you have permission to open it*, ChatGPT was subsequently able to perform the requested task.

The real challenge with the ChatGPT results in Sect. 4.3 is that the ChatGPT responses are not always consistent with the reviews. This is extremely damaging to the usefulness of this tool because the entire purpose of the exercise is for a small business to obtain insight from its own customers' reviews. ChatGPT can provide plausible-sounding, general feedback. Unfortunately, that feedback is grounded in the probability associated with the next word in the string and not necessarily grounded in the actual content of the reviews that the user cares about.

## 5.1 Recommendations

The first recommendation to Altomonte's or to any small business aspiring to incorporate advanced analytical approaches such as those utilized in this study is to take intentional steps towards value-added data collection. The data analyzed in this study was collected passively over a period of ten years by online forums. While the study demonstrated modest results, it is clear that having more data is preferable to having less data.

The customer service department could begin requesting more reviews and feedback, to include providing incentives such as a discount on products or a promotional deal. This approach could risk artificially inflated review ratings, but the risk can be partially mitigated through how the survey is crafted. For example, free-text questions that specifically address areas of improvement give the opportunity for satisfied customers to give ideas for how the store can become even better. They also provide an easy mechanism to identify and review constructive criticism; responses to such questions can automatically forward to management for review and action.

Related, data collection efforts may be geared towards training future models. Obtaining reviews with ratings would be extremely helpful for training a review classification model that could predict unrated reviews. While Altomonte's is currently a small-enough business to manually read and action on customer feedback, prudent planning for future growth can and should include the data landscape.

## 5.2 Future work

Future research could consider other NLP or ML techniques such as Chat Bots in the customer service department or web scraping for real-time monitoring of reviews. It may also be value added to explore time series analysis of the reviews. This would be particularly helpful to quantify the degree to which the business can effectively respond to customer feedback in addressing opportunities for improvement.

**Data availability** All data generated or analysed during this study are included in this published article (and its supplementary information files).

## Declarations

**Conflict of interest** All authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory



regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (pp. 265–283).
- Aktas-Polat, S., & Polat, S. (2022). Discovery of factors affecting tourists' fine dining experiences at five-star hotel restaurants in Istanbul. *British Food Journal*, *124*(1), 221–238.
- An, Q., Ma, Y., Qianzhou, D., Xiang, Z., & Fan, W. (2020). Role of user-generated photos in online hotel reviews: An analytical approach. *Journal of Hospitality and Tourism Management*, *45*, 633–640.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, *82*(5), 977–1008.
- Chen, J., Lim, C. P., Tan, K. H., Govindan, K., & Kumar, A. (2021). Artificial intelligence-based human-centric decision support framework: An application to predictive maintenance in asset management under pandemic environments. *Annals of Operations Research*, *1*, 1–24.
- Chollet, F. (2015). Keras. <https://github.com/fchollet/keras>.
- Ding, T., Chen, L., & Chen, G. (2015). Automatic multi-way domain concept hierarchy construction from customer reviews. *Neurocomputing*, *147*, 472–484.
- Ding, K., Choo, W. C., Ng, K. Y., & Ng, S. I. (2020). Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation. *International Journal of Hospitality Management*, *91*, 102676.
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.
- Hacker, P., & Petkova, B. (2017). Reining in the big promise of big data: Transparency, inequality, and new regulatory frontiers. *Northwestern Journal of Technology and Intellectual Property*, *15*, i.
- Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *2*(4), 100089.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J., Fernández, W., Mark, P., ... Travis, E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Heng, Y., Gao, Z., Jiang, Y., & Chen, X. (2018). Exploring hidden factors behind online food shopping from amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, *42*, 161–168.
- Howell, B. E., & Potgieter, P. H. (2023). What do telecommunications policy academics have to fear from GPT-3? *Telecommunications Policy*, *1*, 102576.
- Huang, J., Rogers, S., & Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo)*.
- Keiper, M. C. (2023). Chatgpt in practice: Increasing event planning efficiency through artificial intelligence. *Journal of Hospitality, Leisure, Sport and Tourism Education*, *33*, 100454.
- Kim, J., Kim, J. H., Kim, C., & Park, J. (2023). Decisions with ChatGPT: Reexamining choice overload in ChatGPT recommendations. *Journal of Retailing and Consumer Services*, *75*, 103494.
- Krestel, R., & Dokoohaki, N. (2015). Diversifying customer review rankings. *Neural Networks*, *66*, 36–45.
- Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*(1), 100008.
- Kwon, W., Lee, M., & Back, K.-J. (2020). Exploring the underlying factors of customer value in restaurants: A machine learning approach. *International Journal of Hospitality Management*, *91*, 102643.
- Lee, S., & Kim, W. (2017). Sentiment labeling for extending initial labeled data to improve semi-supervised sentiment classification. *Electronic Commerce Research and Applications*, *26*, 35–49.
- Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, *3*, 100065.
- Luo, Z., Huang, S., & Zhu, K. Q. (2019). Knowledge empowered prominent aspect extraction from product reviews. *Information Processing and Management*, *56*(3), 408–423.

- Marcolin, C. B., Becker, J. L., Wild, F., Behr, A., & Schiavi, G. (2021). Listening to the voice of the guest: A framework to improve decision-making processes with text data. *International Journal of Hospitality Management*, *94*, 102853.
- McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference*, Vol. 445 (pp. 51–56). Austin, TX.
- Nahavandi, S. (2017). Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, *3*(1), 10–17.
- OpenAI. ChatGPT (2023). <https://chat.openai.com/auth/login>.
- Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in e-commerce experiences. *International Journal of Information Management*, *40*, 132–140.
- Park, S., Cho, J., Park, K., & Shin, H. (2021). Customer sentiment analysis with more sensibility. *Engineering Applications of Artificial Intelligence*, *104*, 104356.
- Piramanayagam, S., & Kumar, M. P. S. (2020). Determinants of customer's dissatisfaction: A content analysis of negative online customer reviews on budget segment hotels in India. *African Journal of Hospitality, Tourism and Leisure*, *9*(1), 1–9.
- Prietl, B. (2019). Big data: Inequality by design? In *Weizenbaum Conference* (p. 10). DEU.
- Prithivirajan, M., Lai, V., & Shim, K. J. (2015). Analysis of aspects and star ratings in consumer reviews.
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta: ELRA. <http://is.muni.cz/publication/884893/en>
- Sainju, B., Hartwell, C., & Edwards, J. (2021). Job satisfaction and employee turnover determinants in fortune 50 companies: Insights from employee reviews from indeed. com. *Decision Support Systems*, *148*, 113582.
- Sanchez-Franco, M. J., Cepeda-Carrion, G., & Roldan, J. L. (2019). Understanding relationship quality in hospitality services: A study based on text analytics and partial least squares. *Internet Research*, *29*(3), 478–503.
- Saura, J. R., Palos-Sanchez, P., & Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*, *11*(3), 917.
- Scharre, P., & Horowitz, M. C. (2015). Autonomy in weapon systems. In *Center for a new American security working paper*.
- Schradie, J. (2017). Big data is too small: Research implications of class inequality for online data collection. In: *Media and Class* (pp. 200–213). Routledge.
- Short, C. E., & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, *19*, e00388.
- Stewart, L. (2019). Big data discrimination: Maintaining protection of individual privacy without disincentivizing businesses' use of biometric data to enhance security. *BCL Review*, *60*, 349.
- Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.
- Wang, T., Cai, Y., Leung, H., Lau, R. Y. K., Li, Q., & Min, H. (2014). Product aspect extraction supervised with online domain knowledge. *Knowledge-Based Systems*, *71*, 86–100.
- Wong, I. K. A., Lian, Q. L., & Sun, D. (2023). Autonomous travel decision-making: An early glimpse into ChatGPT and generative AI. *Journal of Hospitality and Tourism Management*, *56*, 253–263.
- Wu, R. T., & Dang, R. R. (2023). ChatGPT in head and neck scientific writing: A precautionary anecdote. *American Journal of Otolaryngology*, *44*, 103980.
- Yanfang, S., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*, 100752.
- Yilmaz, R., & Yilmaz, F. G. K. (2023). Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100005.
- Zhang, J. (2019). Listening to the consumer: Exploring review topics on Airbnb and their impact on listing performance. *Journal of Marketing Theory and Practice*, *27*(4), 371–389.