The Food Delivery Dilemma: Classification and Evaluation of Real vs. Ghost Kitchens

Akhil Aggarwal
NYU Stern
aa7830@nyu.edu

Alan Chen
NYU CAS
azc242@nyu.edu

Mimi Chen
NYU CAS
xc2063@nyu.edu

Joshua Le
NYU CAS
joshua.le@nyu.edu

Abstract

The rise of virtual kitchens, or ghost kitchens, in the food delivery service industry has generated concerns about food quality and transparency in the virtual dining market. In this study, we develop a classification tool to distinguish between ghost kitchens and traditional restaurants using natural language processing (NLP) techniques. Leveraging word and sentence embeddings, we compare a Logistic Regression classifier to analyze linguistic patterns in menu item descriptions from online food delivery platforms. We evaluate the classifiers' performance using cosine similarity scores while providing insights into the prevalence and characteristics of ghost kitchens. Our findings demonstrate the effectiveness of NLP in identifying ghost kitchens, enabling food delivery platforms to enhance transparency and empower consumers to make informed choices, ultimately supporting local restaurants and ensuring a higher standard of food quality in the virtual dining industry.

1 Introduction

Following the COVID-19 pandemic, the food service industry has experienced a paradigm shift with the emergence of virtual kitchens (also referred to as ghost kitchens). These kitchens have become increasingly popular due to their low operational costs as well as their ability to adapt to changing market demands. Ghost kitchens, especially large establishments with many restaurants operating out of the same location, have begun to overpower the virtual dining industry, driving out local businesses. We will deploy NLP methods to build a classification tool that can identify whether a restaurant is a virtual/ghost kitchen or a real kitchen based on computational linguistic patterns. A prevalent way to compare similarities between two documents based on word embeddings (in this case, menu item descriptions of ghost kitchens versus real kitchens) is to apply the cosine similarity measure to the values obtained from the embedding vectors (Brück, Pouly 2019). Traditionally, document similarity estimation relied on deep semantic approaches or standard information retrieval techniques. However, in recent years, word and sentence embeddings have emerged as the state-of-the-art method (Brokos et al., 2016). We utilize OpenAI's API to access and interact with the text-embedding-ada-002 model. This will allow us to fine-tune the model, submit queries, and receive embeddings. In our contribution, we compare different approaches, including a Logistic Regression classifier using similarity features based on word embeddings. Additionally, we compare the performance of this method and further our analysis by evaluating the cosine similarity of menu item descriptions across restaurants.

2 Problem Statement

With the increasing popularity of food delivery apps such as Uber Eats, ghost kitchens have established a significant presence in the virtual dining industry. These off-premise consumption "restaurants" often operate multiple stores out of a single warehouse, serving generic items across various cuisines and potentially compromising food quality due to health code loopholes. As these ghost kitchens compete with local restaurants that invest in delivery apps, consumers and delivery platforms are becoming increasingly concerned about the transparency and quality of the food they consume or offer. Therefore, there is an urgent need for a system that can accurately identify and differentiate between ghost kitchens and traditional restaurants using computational language patterns found in their online presence, such as menu descriptions and restaurant profiles. This system will enable delivery apps to declutter their platforms by removing ghost kitchens and empower consumers to make informed choices about their orders. Ultimately, we hope to support reputable local restaurants and en-

sure a higher standard of food quality in the virtual dining industry.

3 Related Works

The emergence of ghost kitchens has gained significant attention due to their impact on the food delivery industry. However, much of the research on virtual kitchens focuses on their business model and industry impact. While the classification of ghost kitchens using natural language processing techniques has not been vastly explored, this section reviews some of the relevant literature that informed our approach to the problem.

One of the foundational works in the area of text similarity estimation using word embeddings is by Mikolov et al. (2013)(1), who introduced word2vec, a set of models for learning vector representations of words. Additionally, Pennington et al. (2014)(11) introduced GloVe, a global vectors model for word representation that combined the advantages of both count-based methods and prediction-based methods like word2vec. Their approach demonstrated state-of-the-art performance on various word analogy and similarity tasks. Their work paved the way for numerous embedding models available today, including ext-embedding-ada-002. Embeddings provide a facile way for us to quantify the similarity between bodies of text.

Logistic regression has proved very effective in binary classification of texts, as with this paper in 2022 which used the methodology of POS tagging probability to determine whether short bodies of text belonged to works of fiction or nonfiction (3). The success of logistic regression models can be attributed to their ability to capture local and long-range dependencies within a text, making them particularly suitable for tasks such as identifying linguistic patterns associated with ghost kitchens and traditional restaurants.

In the context of the food delivery industry, a study by Cai et al. (2022) (4) examined the role of ghost kitchens in the digital transformation of the restaurant industry. Their work highlighted the need for tools that can help consumers differentiate between ghost kitchens and traditional restaurants, as these two types of establishments often have different business models, service quality, and food standards. Such tools pave the way for increased trust and clarity between consumers and virtual restaurants.

In summary, our approach to classifying ghost kitchens and traditional restaurants is informed by the existing literature on word embeddings, logistic regression, and their applications to text classification and similarity estimation. By leveraging these NLP techniques, we aim to develop a robust classification tool that can accurately identify ghost kitchens and support the delivery platforms and consumers in making informed choices.

4 Data

4.1 Data Collection

In the process of data collection, we employed web-scraping techniques to extract relevant information from the widely-used food delivery platform, Uber Eats. To ensure a diverse and representative sample of ghost kitchens and traditional restaurants, we targeted seven major metropolitan areas across the United States: Los Angeles, Austin, Seattle, San Jose, Orlando, New York City, and Chicago.

For each city, the first part of our custom-built web-scraping program scrapes over 200 of the listed restaurants' unique URLS for the given location. Thereafter our program then iterates through each URL and scrapes essential data such as the establishment's name, geographical location, and a comprehensive list of menu items along with their respective descriptions.

After obtaining data from more than 1,000 restaurants, we employed a data-cleaning process to eliminate unwanted or redundant information, such as repetitive terms and item prices. This step was accomplished with the aid of carefully crafted regular expressions (regex) tailored to our specific requirements.

For a more detailed examination of our data collection methodology, the corresponding repository containing the code and documentation can be accessed here.

4.2 Limitations

One significant limitation of our study is that we had to restrict our data to Uber Eats, as we encountered web scraper blockers on other popular services such as GrubHub and Doordash. This restriction may have resulted in a less comprehensive dataset, potentially affecting the generalizability of our findings to other food delivery services.

Furthermore, we encountered numerous web scraping deterrents on Uber Eats, such as dynamically changing CSS class names and frequent popups, which made data collection challenging.

While we took measures to overcome these obstacles, it is possible that some data was missed or inaccurately collected.

Another limitation is the difficulty of formatting and cleaning the data once it was scraped. Additionally, the need to merge data collected by multiple team members and remove duplicates added an additional layer of complexity.

These limitations may have impacted the quality and reliability of our data and subsequent analysis. However, we believe that our findings still provide valuable insights into the food delivery industry, and we suggest that future studies consider these limitations when conducting similar research.

Another limitation of our study was the difficulty in identifying ghost kitchens on Uber Eats. Uber Eats has publicly stated that they are taking measures to remove ghost kitchens from their platform, which may result in algorithmic down-ranking of these establishments. Additionally, ghost kitchens are often newer establishments which may not have a significant online presence. As a result, it was challenging for us to find ghost kitchens on Uber Eats. This limited our ability to collect data on them, despite our efforts to scrape additional sites. Consequently, some ghost kitchens may have been missed, which could have affected the comprehensiveness of our dataset.

Additionally, we experienced challenges when balancing the trade-off between an over-representation of ghost kitchens versus producing a meaningful classification model. Our initial concern was having too many real kitchens and creating a system that yielded too many false negatives. To combat this, we aimed for a 45/65 split between ghost kitchens and real kitchens. However, while we made extra efforts to scrape for ghost kitchens, we are unaware of the underlying distribution between the two in the real world.

A technical limitation of our study is that the results, particularly those of the logistic regression model, maybe less explainable due to the limited transparency of the word embedding model we used, OpenAI's text-embedding-ada-002 word embedding. The model incorporates a variety of complex linguistic features and abstractions, and many details of its training data and methodology are unavailable to us. As a result, it may be challenging to fully interpret and explain the underlying factors driving the word embeddings generated and, thus, the logistic regression model, which uses the embeddings as feature vectors.

Despite these limitations, our study provides valuable insights into the prevalence and characteristics of ghost kitchens on Uber Eats. We suggest that future studies take into account these limitations when examining the topic.

4.3 Processing

To process our data, we performed the following steps:

1. Remove all listings of the top 50 fast-food restaurants in the United States except for one instance, sourced from a Kaggle dataset. Our motivation for this was to ensure that our cosine similarity evaluation would not be skewed by multiple restaurant listings with identical menus.

2. Drop all duplicates of restaurant listings that may have occurred in the scraping process

3. Remove all stopwords from the menu item descriptions

4. Utilize OpenAI's API to obtain embeddings for the menu items in each restaurant. The model we use is OpenAI's newest embedding model, text-embedding-ada-002. We also use OpenAI's tiktoken, a byte pair encoding tokenizer.

5 Methodology

5.1 Initial Evaluation

We randomly selected 236 samples from both ghost kitchens and real kitchens, then computed the cosine similarity matrix for both the ghost kitchen samples and the real kitchen samples. Cosine similarity allows for comparison of text passages of different sizes by representing each word in a vector format. The text documents are then plotted as vectors in an n-dimensional space. The mathematical formula for cosine similarity measures the cosine of the angle between two n-dimensional vectors projected in a multi-dimensional space. The score ranges from 0 to 1, with a score of 1 indicating that the two vectors have the same orientation, and therefore are identical. Conversely, a score closer to 0 suggests that the two documents have little similarity:

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

OpenAI embeddings are normalized to length 1, which means cosine similarity and Euclidean distance will result in identical rankings. We calculated the similarity for both matrices by summing the similarity of the upper triangular matrix (excluding the diagonal) and dividing by the total number of unique pairs of samples.

In addition, we utilized two measures of central tendency, mean and median. Since the mean is sensitive to extreme values and outliers, we also calculated the median as it is a more robust measure when dealing with non-normally distributed data.

5.2 Same-Location Evaluation

Our initial evaluation plan consisted of marking all locations with 15+ restaurants as ghost kitchens. However, a significant limitation to this would be a false positive of ghost kitchens at locations where multiple real kitchens operate, such as malls and food halls. To evaluate if we can distinguish between real and ghost kitchens when there is a cluster of restaurants at the same location, we performed the process outlined in the initial evaluation on two locations in our dataset, one containing a cluster of real kitchens and the other all ghost kitchens. Both of these locations contained 15 restaurants to ensure equal representation. While using a case study of only two locations does not entirely represent our data as a whole, it can help us determine underlying text patterns of real and ghost kitchens at the same location.

5.3 Classification

We trained our classification model on 20 percent of the data, with kitchen type as the target variable. The remaining 80 percent was used for training and developing the classification model. We chose logistic regression for our first model because it provides a robust framework for binary classification tasks by modeling the probability of class membership. Then, we calculated the accuracy of the model as well as the precision and recall for both classes.

6 Results

6.1 Cosine Similarity Measures

1. **Initial Evaluation**
   Our initial evaluation yielded an average of 0.80243 for cosine similarity between ghost kitchens and a median of 0.800896. This indicates a relatively high degree of similarity between the vectors. In addition, the cosine similarity between ghost kitchens and normal kitchens yielded an average of 0.77992, with a median of 0.79460. Although there was a difference, the central tendencies are almost the same. This indicates that even though ghost kitchens overall are more similar to each other compared to normal kitchens, the difference is minute. cosine similarity evaluates the similarity between two vectors of numerical values that represent the text being compared. It takes into account the frequency of words in the text and their relative importance, but it does not directly evaluate the sentence structure or syntax of the text.

To visualize the cosine similarity scores in a 2-dimensional space, we use the t-distributed Stochastic Neighbor Embedding (t-SNE) to transform the data into two dimensions.t-SNE is beneficial for visualizing cosine similarity because it can effectively capture non-linear structures, handle high-dimensional data, reveal clusters, and provide a robust visualization despite the noise. The t-SNE model has previously been used to display high-dimensional embeddings of words in a 2-dimensional space (Lee & Mimno, 2014).

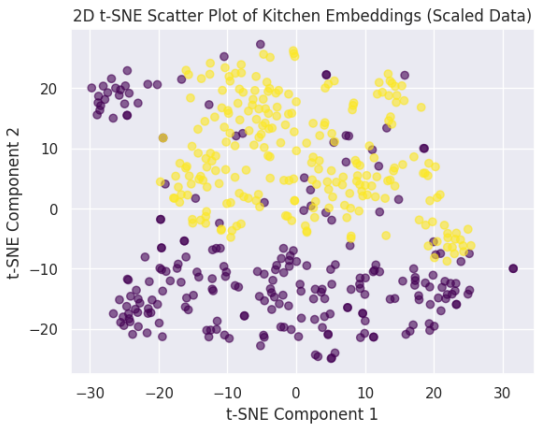

Figure 1: t-SNE Scatter Plot

The yellow data points indicate ghost kitchens and the purple datapoints indicate real kitchens. Looking at the plot, there are some distinct clusters of data, but the two groups are quite close to each other. There is also a cluster of real kitchens that are located relatively further away from each other. If data points are located far away from each other in a t-SNE plot, it typically means that they are a large distance away in the original high-dimensional space. This cluster of real kitchens could be contributing to the high cosine similarity between our real kitchens and ghost kitchens. This could have been caused by the structure of our data. However, given the fact that the t-SNE algorithm is built with stochastic gradient descent, it is difficult to interpret the underlying mechanisms that are resulting in this cluster.

	Average	Median
Ghost to Ghost	0.80243	0.80096
Normal to Ghost	0.79544	0.79460

Table 1: Initial Evaluation

2. **Same-Location Evaluation**
   We evaluated the cosine similarity scores of two clusters of restaurants at the same location, one belonging to the class of ghost kitchens and the other normal. Our results in this evaluation yielded a larger difference, with the average of the in-group ghost kitchen average cosine similarity being 0.81265 and the median being 0.81637. In comparison, the two locations of the normal kitchen and ghost kitchen yielded an average of 0.77992 and a median of 0.78870. This shows that there is a slightly larger similarity between ghost kitchens located in the same establishment in comparison to all the ghost kitchens in our data. In addition, normal kitchens located at the same location, such as a mall or food hall, are less similar to ghost kitchens than in the entire dataset. This difference could be explained by the fact that food halls and malls offer a large variety of cuisines, meaning the menu items would not be as similar to each other. However, in a larger dataset, more restaurants belonging to the same cuisine exist, resulting in more similarities. In addition, ghost kitchens located at the same location are often found to serve identical menu items across different cuisines, which could explain why they are more similar.

	Average	Median
Ghost to Ghost	0.81265	0.81637
Normal to Ghost	0.77992	0.78870

Table 2: Same-Location Evaluation

One observation is that our central tendencies were quite similar to each other across the board, meaning that our calculations were relatively robust to outliers.

6.2 Classification System

Our classification system using logistic regression performed well, achieving an accuracy of 0.9404. The precision and recall were 1.00 and 0.8364, respectively.

One possible explanation for why the logistic regression could classify ghost kitchens better than other methods such as cosine similarity could be the way the words in the embeddings are distributed. Word embeddings represent words as dense vectors in a high-dimensional space, where similar words are closer together. However, the relationship between words in the embedding space is not necessarily linear, which means that cosine similarity may not capture all the nuances of the relationship. Logistic regression, on the other hand, can learn non-linear decision boundaries that can better separate the categories.

Another possible explanation for our success when using logistic regression is that word embeddings are learned from large amounts of text data, which means that they may contain noise and spurious associations. Because cosine similarity measures the similarity between two vectors based on the angle between them, it is sensitive to the presence of noise in the vector space. In contrast, logistic regression is robust to noise in the input data due to assigning less weight to noisy data points in training.

Based on the evaluation metrics, our classification system appears to be more conservative when predicting ghost kitchens. Each time it classifies a restaurant as a ghost kitchen, it is correct. However, in actual cases where restaurants are ghost kitchens, it tends to misclassify them as regular restaurants. The most probable reason for this is because of the uneven balance between ghost kitchens and

Metric	Value
Accuracy	0.9404
Precision for ghost kitchen (g)	1.0000
Recall for ghost kitchen (g)	0.8364
Precision for regular kitchen (r)	0.9143
Recall for regular kitchen (r)	1.0000

Table 3: Performance Metric of Logistic Regression Classification

regular restaurants in the dataset, and thus an uneven balance in the training dataset. In our overall dataset, approximately 31.34% of data points were ghost kitchens, so the logistic regression model likely learned more from the majority class (regular restaurants in this case) and thus had a difficult time generalizing to ghost kitchens.

7 Conclusion

Our study aimed to develop a classification system capable of differentiating ghost kitchens from traditional restaurants using natural language processing techniques. By employing logistic regression and cosine similarity upon restaurant listing and menu text embeddings, our system was able to classify each type of listing with compelling results.

Our findings demonstrate the potential of using NLP-based methods for tackling the growing concern about transparency and quality in the food delivery industry. With relatively simple NLP data analysis techniques, food delivery platforms can enhance their service offerings and provide consumers with better-informed choices, ultimately supporting local businesses and maintaining higher food quality standards.

Despite the limitations and challenges faced in this study, such as data imbalance, web scraping deterrents, and the difficulties in obtaining a comprehensive dataset, our findings and system present compelling methods to help the food delivery industry address the rise of ghost kitchens. The results from our study could serve as a foundation for future research aimed at refining and expanding upon the methodologies presented here.

8 Future Work

While our current methodologies have demonstrated promising results, we believe there is potential for further refinement and exploration in the differentiation of ghost kitchens and traditional restaurants on food delivery platforms. In future research, we propose employing two additional approaches to bolster the validity of our findings and to provide a more comprehensive understanding of the underlying patterns.

Firstly, we plan to investigate the degree of similarity between probability distributions of menu offerings from both ghost kitchens and traditional restaurants. By quantifying these differences, we aim to uncover whether there are any unique characteristics or trends in the food options provided by the two types of establishments.

Secondly, we intend to utilize text analysis techniques, such as the computation of Term Frequency-Inverse Document Frequency (TF-IDF) vectors, for the evaluation of restaurant reviews sourced from Yelp. This will enable us to identify patterns and sentiment differences in the language used by customers when reviewing ghost kitchens versus their brick-and-mortar counterparts. For instance, we hypothesize that ghost kitchen reviews might predominantly focus on aspects related to "delivery", while traditional restaurant reviews may emphasize elements such as "service" quality, "ambiance", and in-person dining experiences.

Lastly, since the text-embedding-ada-002 word embedding model was released in December of 2022, we were unable to reference earlier works using this model. We plan to implement different types of embedding models to evaluate whether our results will vary between embedding methods.

By incorporating these additional methods, we aim to further enhance the robustness of our research and contribute to the growing body of knowledge on the emerging phenomenon of ghost kitchens within the food delivery ecosystem.

9 Acknowledgement

We would like to express our deepest gratitude to Professor Adam Meyers for his guidance and support throughout the course of this research project. His insightful comments, patient guidance, and unwavering encouragement were instrumental in our success.

We are also extremely grateful for the mentorship of Norihito Naka, whose expertise in guiding our methodology and providing technical guidance was critical to the success of this project. Norihito's exceptional mentorship helped us to navigate the challenges of the research process and provided us with invaluable insights into the field.

References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

[2] Levy, O., Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177-2185).

[3] Kazmi, A., Ranjan, S., Sharma, A., Rajkumar, R. (2022). Linguistically Motivated Features for Classifying Shorter Text into Fiction and Non-Fiction Genre. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 922–937).

[4] Cai, L., Luo, X., Zhang, M. (2022). Ghost kitchens: A new business model in the digital transformation of the restaurant industry. *Tourism Management*, 89, 104429.

[5] Thongtan, T., Phienthrakul, T. (2019). Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 407–414).

[6] vor der Brück, T., Pouly, M. (2019). Text Similarity Estimation Based on Word Embeddings and Matrix Norms for Targeted Marketing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1827–1836).

[7] Mihaylov, T., Frank, A. (2016). Discourse Relation Sense Classification Using Cross-argument Semantic Similarity Based on Word Embeddings. In *Proceedings of the CoNLL-16 shared task* (pp. 100–107).

[8] Zhelezniak, V., Shen, A., Busbridge, D., Savkov, A., Hammerla, N. (2019). Correlations between Word Vector Sets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 77–87).

[9] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*

[10] Garcia-Silva, A., Berrio, C., Gómez-Pérez, J. M. (2019). An Empirical Study on Pre-trained Embeddings and Language Models for Bot Detection. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 148–155), Florence, Italy. Association for Computational Linguistics.

[11] Pennington, J., Socher, R., Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP* (Vol. 14, pp. 1532–1543).

[12] Lee, Moonae, and David Mimno. Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1319–1328, Low-Dimensional Embeddings for Interpretable Anchor-Based Topic Inference.