Coffee Rebrewer: Using Aspect-Based Sentiment Analysis to Find the Best Coffee in Town

Anhtuan Ho, Bo Wang, Jason Young, Donna Yung (Team 67)

Introduction: Customer review sites have transformed the way consumers decide on products or services and have become an integral part of the customer journey. Yelp is an example of a popular online review site used for local businesses, with 265 million reviews by the end of 2022.[1] Studies have shown a positive association between businesses' good online reputation based on Yelp reviews and consumers' patronage of the business (i.e., increases in intentions and behaviour),[2–3] marking its importance.

Problem: Yelp has a 5-star rating scale in which users rate their experience from 1–5, averaged to provide an overall rating of the business. However, with this approach, consumers may provide a rating based on different factors of their experience – such as service, ambiance, and others, beyond quality of the product. For consumers that are solely interested in quality, this system may not be accurate as the rating could incorporate a myriad of other contextual-dependent factors. We decided to create a guide for the best tasting coffee in Philadelphia by analyzing Yelp reviews to identify coffee shops serving the highest rating of coffee, using aspect-based sentiment analysis (ABSA) to disentangle consumer reviews and focus on coffee quality only. The guide will include a map of Philadelphia as well as useful visualizations that highlight features of each business.

Literature Survey: Many others have previously analyzed consumer reviews using different techniques and for different applications, such as sentiment analysis, predictive recommendation, and fake review detection.[4] Our methodology was motivated by learnings from previous studies in this area. For instance, researchers previously used SparkSQL to analyze the Yelp dataset, providing descriptions of how they loaded and queried data efficiently with different partition settings,[5] which provided guidance in setting up our computation platform. The Bag-of-Words method was previously used to analyze the most frequent words used in positive, negative, and neutral reviews,[6] which we considered using similarly to create our initial word list. Sentiment analysis has previously been used to improve restaurants' performance based on Yelp reviews[7] and other studies found that profanity in reviews and social relationships impacts the usefulness of the reviews,[8,9] motivating us to consider using different weights for different sets of reviews during sentiment calculation.

Given our proposed intent to utilize ABSA, we looked to find opportunities to extend its capabilities or fine-tune its accuracy. For instance, prior researchers have developed methods such as Latent Aspect Rating Analysis, which can both discover topical aspects and assign sentiment scores.[13] Similarly, in a comparison of aspect-based recommendation systems, Hernández-Rubio et al. (2019)[14] assessed the impact of different aspect-identification methods and user-specific information on recommendation accuracy. We decided that such extensions would complicate our project too much to justify their usage. Instead, we manually specified coffee-related aspects and ignored user-specific preferences to focus on interactive visualization – an approach justified by a previous model that assigned weights to various human-controllable aspects of a business review and showed that users who followed their aspect recommendations rated their experiences significantly higher than others.[15]

Methods: There's a distinct absence of accessible tools that allows consumers to assess coffee shops based on specific criteria like coffee quality or service. Our innovative solution aims to bridge this gap by introducing an interactive tool for users to evaluate and compare different coffee shops based on pre-defined criteria. While existing academic research explores sentiment analysis applied to online business reviews, no user-friendly tools currently empower consumers to make informed judgments about coffee shops according to their unique preferences. We employed the following methods in developing our innovation using the Yelp dataset obtained from their official website:

1. Algorithms

While there are various approaches to building an ASBA classifier, supervised methods are among the most popular. Because there are no training datasets labeled with coffee quality aspects and relevant sentiments, we looked to the PyABSA library, an open-source tool for sentiment analysis.[10] We selected yangheng/deberta-v3-base-absa-v1.1, a model that is trained based on the FAST-LCF-BERT model with DeBERTAv3 serving as its base architecture.[11]

Since we are solely focusing on coffee quality, and because there is more than one type of coffee drink, we explicitly identified a list of common coffee shop menu items to collectively serve as the aspect. This list includes but is not limited to: coffee, espresso, latte, americano, etc. For each review, our model outputs three sentiment scores which correspond to the probability that the sentiment label is 'negative', 'neutral', or 'positive'. If the 'positive' score is the highest, then the review is classified as having a positive sentiment for coffee quality. We computed these sentiment scores for all the coffee shops reviews in Philadelphia, and for each coffee shop, we devised two custom metrics:

<u>Metric 1:</u>

$$Average\ positive\ sentiment\ score\ =\ \frac{sum\ of\ positive\ sentiment\ scores}{total\ number\ of\ reviews}$$

<u>Metric 2:</u>

$$Positive\ sentiment\ ratio\ =\ \frac{number\ of\ reviews\ classified\ as\ positive}{total\ number\ of\ reviews}$$

We then calculated a **weighted combination** of the two metrics to achieve a composite score out of 1:

$$Composite\ score\ =\ (0.6\ \times\ METRIC\ 1)\ +\ (0.4\ \times\ METRIC\ 2\ )$$

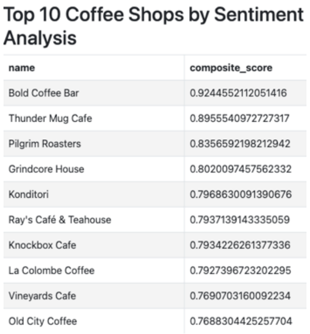
Finally, we select the 10 coffee shops in Philadelphia with the highest composite scores and use these for our interactive visualizations.

2. Interactive Visualizations

Although others have used aspect-based sentiment analysis of reviews to assess the sub-components of a product, we have gone a step further by developing a unique interface to make that information much more accessible. We implemented our features in an interactive, browser-based application which connects to our processed data in a MySQL database and renders interactive html with Python's Plotly and Dash frameworks.

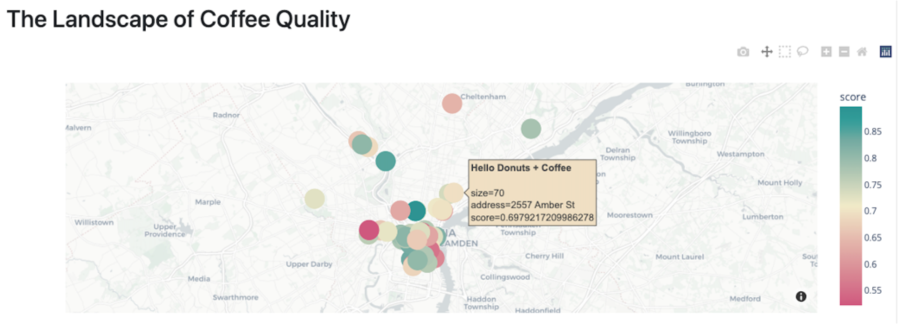
Our browser application consists of 3 main features. The first feature is a simple table of the top 10 coffee shops ranked by our unique composite score described above. While the table appears static to the user, it is sourced from our MySQL database, which means it will automatically refresh whenever new Yelp data is processed and added to the database.

The next feature is an interactive map, which allows users to explore all coffee shops in our dataset. Each café or restaurant is color coded according to its average positive sentiment score. When the user hovers over a location, they are shown the average score as well as the store's address.



Top 10 Coffee Shops by Sentiment Analysis

name	composite_score
Bold Coffee Bar	0.9244552112051416
Thunder Mug Cafe	0.8955540972727317
Pilgrim Roasters	0.8356592198212942
Grindcore House	0.8020097457562332
Konditori	0.7968630091390676
Ray's Café & Teahouse	0.7937139143335059
Knockbox Cafe	0.7934226261377336
La Colombe Coffee	0.7927396723202295
Vineyards Cafe	0.7690703160092234
Old City Coffee	0.7688304425257704

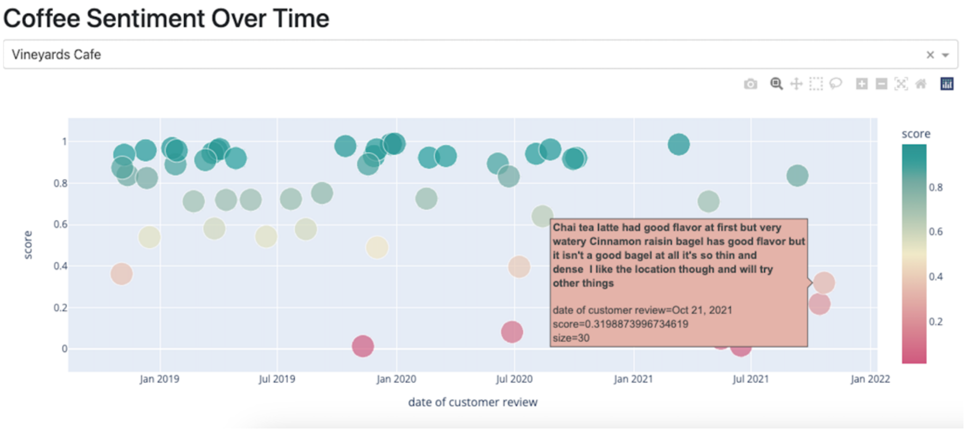


The Landscape of Coffee Quality

Hello Donuts + Coffee

size=70
address=2557 Amber St
score=0.6979217209986278

Our final feature allows users to view coffee sentiment scores over time for any business in our dataset. First, the user selects a business through the dropdown menu. The graphic then shows each review for the business and how the coffee sentiment has trended over time. When the user hovers over a review, the full text of the review is shown along with its sentiment score. This could allow users or business owners to track how the quality of coffee is changing over time, independent of other review factors.



Coffee Sentiment Over Time

Vineyards Cafe

Chai tea latte had good flavor at first but very watery Cinnamon raisin bagel has good flavor but it isn't a good bagel at all it's so thin and dense I like the location though and will try other things

date of customer review=Oct 21, 2021
score=0.3198873996734619
size=30

Experiments/Evaluation: In order to test and verify whether our sentiment score works as expected, we have designed several experiments to compare our results with other methods on a sample data set.

The Testing Data

Due to resource limitations, we selected a random sample dataset of 20 coffee shops and all their filtered reviews. All experiments were done on this testing dataset, which contains each coffee shops' overall average stars, the recent average stars, and the ranks based on these ratings.

The Definition of Model Error

We tested our results based on 2 different aspects: the data trend and the rank difference. The data trend was mainly verified by comparing them on the line charts, and the rank difference was measured using Spearman's rank correlation, which is defined as follows:

$$\rho = 1 - \frac{6\Sigma \, \mathrm{d}_i^2}{n(n^2 - 1)}$$

where:
$\boldsymbol{\rho}$ = Spearman’s rank correlation coefficient
d_i = difference between the two ranks of each observation
n = number of observations

The Spearman’s rank correlation can take a value from +1 to -1, where:

- A value of +1 means a perfect association of rank;
- A value of 0 means that there is no association between ranks; and
- A value of -1 means a perfect negative association of rank.[16]

To make the calculated scores and the average star ratings compatible, we converted the sentiment scores’ ranges from [0, 1] to [0, 5], so that we can compare them in one chart.

Experiment 1: **Comparing our results with the overall average ratings on Yelp**
In this experiment, we tried to understand if the sentiment scores are consistent with the overall star ratings, the recent star ratings, and the ranks. As depicted from the following Figure 1, although the calculated sentiment scores don’t match exactly, the trend is consistent. The Spearman’s rank coefficient between the rank from overall stars and the calculated score is around 0.639. This means that our predicted sentiment score was associated with the overall rates and the result is consistent. In addition, the coefficient between the rank from recent stars and the calculated score is 0.777. This is an even better score, and this result is likely because our sentiment scores were calculated based on the recent reviews. Both scores serve as indicators that our methods work as expected.

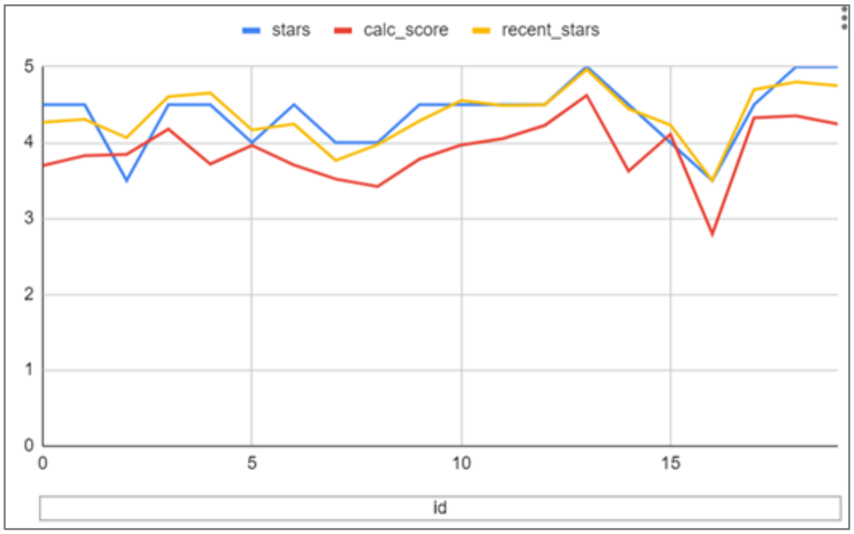


Figure 1: Coffee shop rating stars and sentiment scores

Experiment 2: Fitting a model to use different weights for *metric 1* and *metric 2*

In this experiment, we tried to identify whether there is a better fit for the coefficients for *metric 1* and *metric 2* in our algorithm for calculating sentiment scores. Currently, the metrics are coded to 0.6 and 0.4, respectively, as discussed in our Methods section. We ran a linear regression using the average stars as the target and fit the parameters for *metric 1* and *metric 2*. As seen in Figure 2, the trends of the linear regression model result match the stars well, and Spearman's coefficient between average stars and recent average stars and the fitted result was 0.672 and 0.824, respectively. Both are slightly better than the scores obtained from Experiment 1, but not by a big difference. Considering the small size of our dataset, our conclusion is that it would be difficult to find reasonable parameters that fit most other cases and have good predictability, and future work on this is warranted.

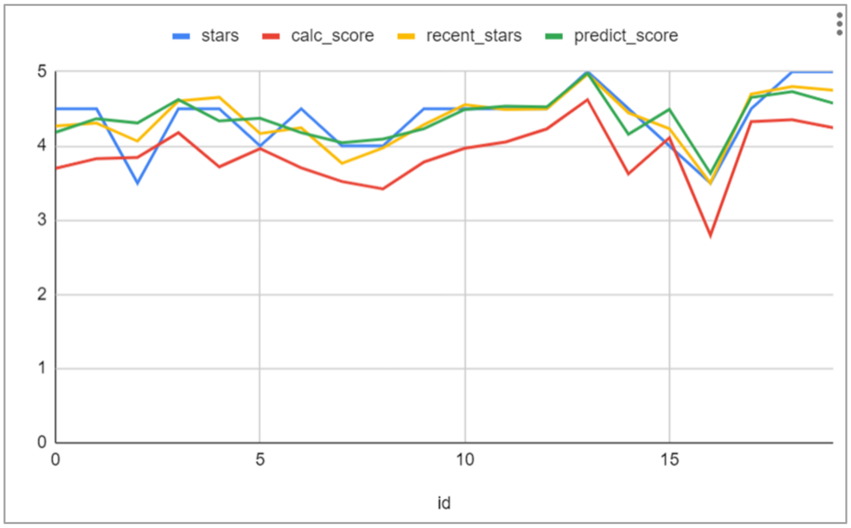


Figure 2: Coffee shop rating stars, sentiment scores and linear model predicated scores

Experiment 3: **Comparing our results with human-rated data based on the review**

One of the reasons for developing this new approach of calculating composite sentiment scores rather than simply using the average star ratings is that many different users contribute to the average star ratings, and each user may have different standards and preferences of value to them. We therefore decided to enlist one person to check all the recent reviews for a coffee shop and give a star rating, which would be more consistent given a single individual's influence. However, due to the massive number of reviews per shop and time/capacity constraints, we randomly selected 10 comments per shop and ranked the shops based on those as well as their recent average scores. The Spearman's coefficients between the human rank and the calculated sentiment score rank / linear regression fitted rank are 0.977 and 0.953, respectively – both results are much better than the results with the Yelp stars, providing evidence that our method's result is consistent with our assumptions.

Experiment 4: **Using different weights for the review on calculating *metric 1* and *metric 2***

In the experiment, we tried to understand if we need to treat reviews differently when we calculate *metric 1* and *metric 2*. The weight that we decided to use is the "useful" count for each comment, and we used that as the weight for each comment. The Spearman's coefficient between this weighted calculated rank and the human rank is 0.734, which is worse than the result we got in the last experiment. Our analysis is that the "useful" information may not be that useful when we manually rank the shops, and we don't see a visible benefit to consider this during calculation.

Conclusions and Discussion: All businesses must gain insights from user reviews to identify the positive and negative characteristics of their offerings. This enables them to enhance what's already working well as well as address key areas that need improvement.[12] Similarly, consumers rely on understanding product or service strengths and weaknesses to make informed purchase decisions. For example, when comparing phone models, a potential buyer interested in extended battery life may prioritize reviews about battery performance. ASBA serves a crucial role in these contexts. Similarly, our interactive tool provides coffee enthusiasts in Philadelphia with a filtering mechanism to select coffee shops aligning with their preferences, allowing them to make informed choices based on what they value most.

The success of this tool, as evidenced through our robust evaluations above, could impact and have positive implications for all customer review sites. By using ASBA to disentangle customer reviews, we were able to extract those on coffee quality only. Similarly, this could inform Yelp and/or other customer review sites to implement a similar approach to allow their users to filter averaged ratings based on factors that matter to them, and therefore making their site more useful and effective for users.

Our novel, objective tool will allow users to locate the highest-quality coffee in Philadelphia, yet it still has some limitations, such as only including coffee shops located in one city which was pre-defined and within expectations given our team's capacity and time constraints. Potential future extensions of this tool could include expansion of the geographic scope beyond Philadelphia, application to different aspects of business quality (e.g., ambiance vs. quality), and applications to different types of restaurants or businesses. Adaptation to be user-friendly on mobile devices would also be beneficial, given the substantial proportion of people who access services using mobile devices today.

All team members have contributed a similar amount of effort.

Reference List:

1. Yelp. (2023). *Fast Facts.* https://www.yelp-press.com/company/fast-facts/default.aspx
2. Banerjee, S., Bhattacharyya, S., & Bose, I (2017). Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. *Decision Support Systems, 96,* 17-26. https://doi.org/10.1016/j.dss.2017.01.006
3. Fogel, J. & Zachariah, S. (2016). Intentions to use the Yelp review website and purchase behavior after reading reviews. *Journal of Theoretical and Applied Electronic Commerce Research, 12,* 53-67. https://doi.org/10.4067/S0718-18762017000100005
4. Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review, 31,* 100413. https://doi.org/10.1016/j.cosrev.2021.100413
5. Kalariya, D., Kalariya, Vyas, S., Vyas, Savasni, D., Savasni, & Patel, S. P. (2022). Big data analysis on yelp user-generated reviews (Vols. 978-1-66542577-3/22). *2022 International Conference for Advancement in Technology.* https://doi.org/10.1109/ICONAT53423.2022.9726108
6. Alamoudi, E. S., & Azwari, S. A. (2021). Exploratory Data Analysis and Data Mining on Yelp Restaurant Review. *2021 National Computing Colleges Conference.* https://doi.org/10.1109/nccc49330.2021.9428850
7. Ching, M. R. D., & De Dios Bulos, R. (2019). Improving Restaurants' Business Performance Using Yelp Data Sets through Sentiment Analysis. *ACM International Conference Proceeding Series.* https://doi.org/10.1145/3340017.3340018
8. Hair, M. L., & Ozcan, T. (2018). How reviewers' use of profanity affects perceived usefulness of online reviews. Marketing Letters, 29(2), 151–163. https://doi.org/10.1007/s11002-018-9459-4
9. Corradini, E., Nocera, A., Ursino, D., & Virgili, L. (2021). Investigating negative reviews and detecting negative influencers in Yelp through a multi-dimensional social network based model. *International Journal of Information Management, 60*, 102377. https://doi.org/10.1016/j.ijinfomgt.2021.102377
10. Yang, H., & Li, K. (2022). PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis. *arXiv.* https://doi.org/10.48550/arXiv:2208.01368
11. He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv.* https://doi.org/10.48550/arXiv:2111.09543
12. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.
13. Wang, H., Lu, Y., & Zhai, C. X. (2011). Latent aspect rating analysis without aspect keyword supervision. *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11).* Association for Computing Machinery, New York, NY, USA, 618–626. https://doi.org/10.1145/2020408.2020505
14. Hernández-Rubio, M., Cantador, I. & Bellogín, A. (2019). A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews. *User Model User-Adap Inter 29,* 381–441. https://doi.org/10.1007/s11257-018-9214-9
15. Bauman, K., Liu, B., & Tuzhilin, A. (2017). Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. *In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 717–725. https://doi.org/10.1145/3097983.3098170
16. Gupta, A. (2023, February 2). *Spearman's Rank Correlation: The Definitive Guide to Understand*. Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation