TopicImpact: Improving Customer Opinion Analysis with LLM Preprocessing for Topic Modeling and Star-Rating Prediction

Emil Häglund
Department of Computing Science
Umeå University, Sweden
emilh@cs.umu.se

Johanna Björklund
Department of Computing Science
Umeå University, Sweden
johanna@cs.umu.se

Abstract

We improve the extraction of insights from customer reviews by adding an LLM preprocessing step to the traditional topic-modeling pipeline. This step segments reviews into opinion units — distinct opinions with relevant text excerpts and sentiment scores. The result is a heightened performance of the subsequent topic modeling, leading to coherent and interpretable topics while also capturing the sentiment associated with each topic. By correlating the topics and sentiments with business metrics, such as star ratings, we can gain insights on how specific customer concerns impact business outcomes. We present our system's implementation, use cases, and advantages over other topic modeling and classification solutions. We also evaluate its effectiveness in creating coherent topics and assess methods for integrating topic and sentiment modalities for accurate star-rating prediction.

1 Introduction

Understanding customer feedback is important for businesses aiming to strengthen their offerings, identify growth opportunities, and deliver superior customer experiences. Many companies accumulate extensive text-based feedback from customers, employees, and other stakeholders. The challenge, often is not a lack of feedback but the difficulty of extracting actionable insights from vast and unstructured data (Tavakoli et al., 2024).

We propose the analysis framework TopicImpact, which integrates neural topic modeling with an LLM-enabled preprocessing step that extracts and structures opinions. In the preprocessing step, raw reviews are transformed by an LLM into opinion units (Häglund and Björklund, 2025) — extracted phrases that encapsulate a customer's sentiment on specific aspects. Additionally, the LLM assigns a sentiment score on a scale from 1 to 10, where 1 indicates a very negative sentiment and 10

represents a very positive one. (See Part A of Figure 2 for examples of opinion units.) By clustering these opinion units instead of entire reviews, TopicImpact generates more coherent and interpretable topic clusters. A detailed outline of the system is provided in Section 2.

The improved topic clusters allow marketers to establish several important facts. By considering the themes of the generated clusters, they learn *(1) the topics discussed in the reviews*, and by considering the cluster sizes, also *(2) the frequency with which these topics occur*. Through the sentiment score of each opinion, they can infer *(3) the sentiments associated with each topic*, and by applying regression to the the combination of topics and sentiments, *(4) the contribution of each topic to the business metric, both in terms of polarity (positive/negative) and degree*. By providing these four pieces of information, our solution contributes to a complete and detailed understanding of customer opinions in unstructured review text.

The proposed framework offers clear advantages over traditional topic modeling methods. Unlike approaches that cluster entire reviews, TopicImpact generates more coherent topics by clustering aspect-delineated opinion units, this is an important strategy because individual reviews often address multiple aspects. Statistical methods like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) assume that documents are mixtures of words and, similar to TopicImpact, allow reviews to belong to multiple topics. However, LDA lacks the contextual understanding provided by embeddings created with pretrained language models. While LDA represents topics using keyword lists, it offers limited insight into which specific parts of a text prompted the assignment of a review to a topic. In contrast, TopicImpact enhances interpretability by using the opinion units' aspect-delineated label and excerpt as the clusterable document, offering clear context for understanding the topic assignment. Addition-

ally, metadata links back to the original review.

We can also relate topic modeling to classification, which is the appropriate approach when dealing with static and clearly defined categories. However, classification fails to capture emerging themes and niche customer concerns. Nor is classification suitable for exploratory analysis — situations where we do not know what opinions to expect in advance.

Another drawback with classification is the difficulty of adapting the level of topical abstraction. For example, shifting from analyzing general pricing concerns to examining a specific coupon code campaign would typically require retraining and relabeling datasets in traditional models. Even with LLMs that do not require extensive training data, this necessitates reclassification of all reviews, which incurs significant consumption of time, cost and compute. Topic modeling overcomes these limitations by enabling dynamic adjustments, such as setting the desired number of topics or providing seed words to refine the analysis toward specific topics. While our system also requires LLM-based preprocessing and embedding of opinion units—both of which are computationally demanding—these, are performed only once, enabling later rapid iteration and exploration of customer feedback across different levels of abstraction.

Beyond extracting insights from reviews, the applications of TopicImpact extend to other forms of opinionated free-text data, such as employee surveys, customer support interactions, course evaluations, and patient feedback. Instead of correlating topics with star ratings, we can analyse other business outcomes, such as employee or patient satisfaction, purchase likelihood or customer churn.

In the upcoming sections, we outline the implementation of TopicImpact and discuss its intended use cases. We then conduct experiments to address the following key research questions.

RQ1: Can topic modeling of opinion units generate coherent topics with regards to subject matter and sentiment? How does a sentiment-aware embedding model compare to a general-purpose embedding model in this context?

RQ2: To what extent can topics generated through topic modeling accurately predict star ratings? Furthermore, how can the integration of sentiment and topic modalities enhance prediction accuracy?

> **Value Proposition.** TopicImpact is a fast, interpretable and cost-effective solution for iteratively exploring customer opinions. It generates coherent topic clusters, provides supporting customer quotes, and identifies the sentiment associated with each topic, as well as its correlation to business metrics.

2 System description

The TopicImpact system is presented in Figure 1, through four steps (A-D). In step A, individual reviews are processed into opinion units using LLMs. These opinion units consist of an opinion label, a supporting excerpt, and a sentiment score (1–10), where 1 is very negative and 10 is very positive. An example review is processed in the figure. More details on the conceptualization of opinion units and an evaluation of LLM's ability to generate them can be found in (Häglund and Björklund, 2025). To isolate the contribution of specific topics, we instruct the LLM to tag overall sentiments towards the restaurant or experience (e.g., "we had a wonderful time") with an "overall experience" label. Since these statements are not informative for individual topics, we remove them from the analysis.

Step B starts by embedding the opinion units label and excerpt using a word embedding model, along with relevant metadata such as review ID and sentiment score. The opinion units are clustered through topic modeling, based on the widely used BertTopic (Grootendorst, 2022a), with the number of clusters as a key parameter. The final step in this phase is to inspect the clusters and select topics for further analysis.

In Step C, the relationship between the chosen topics and a business metric (here: star rating) is analyzed using multiple linear regression (MLR). Each review is a data point, with independent variables X representing clustered topics, each in its own column. For example, if an opinion unit linked to a review is clustered under a 'service' topic, the 'service' column for this review data point is populated with the sentiment score of the opinion unit. If there are multiple mentions of 'service' within the same review, an average sentiment score is calculated; if there are no mentions, the value is set to zero. The dependent variable y is the star rating which ranges from 1 to 5. This analysis provides coefficients for each topic, reflecting their strength of association with star ratings, along with p-values to assess statistical significance.
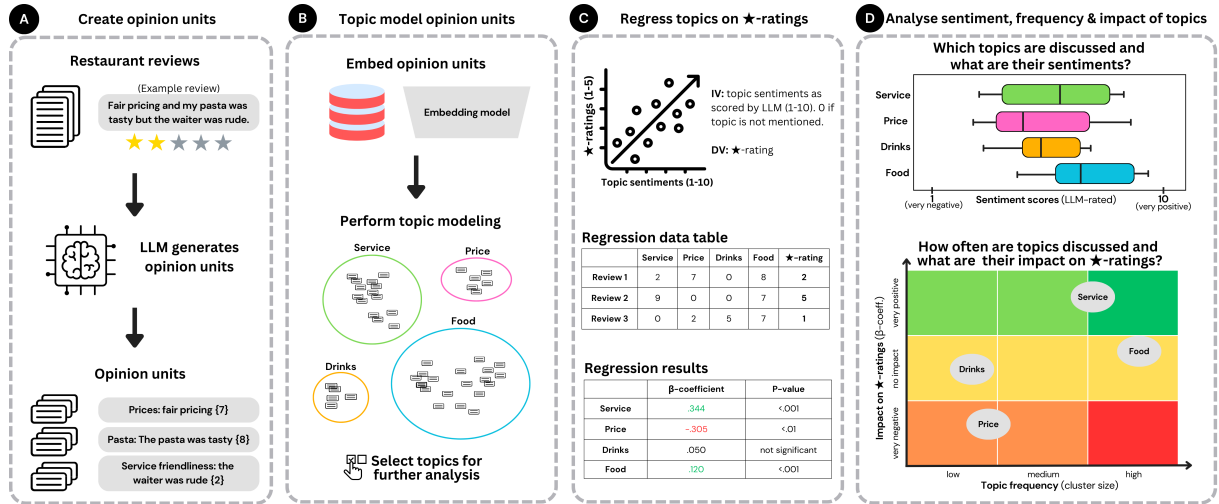
Figure 1: Overview of TopicImpact. **A:** LLM processes reviews to opinion units: *label: excerpt {sentiment 1-10}*. **B:** Embed and cluster opinion units through topic modeling. **C:** Regress topics and sentiments onto star ratings. **D:** Analyze sentiment, frequency, and topic impact.

In Step D we draw insights from the analysis. The topic clusters generated in Step B allow us to identify key topics discussed- their frequency, and sentiment. Regression analysis ranks topics by their relation to star-ratings. Combining this with frequency data highlights the topics most critical to customer satisfaction, enabling the creation of a priority matrix to rank issues by importance (Slack, 1994). For example, a frequently discussed topic that negatively affects satisfaction should be urgently addressed, while a positive but infrequent topic (e.g. sustainability) may present an opportunity for product promotion, such as highlighting locally-sourced ingredients.

Regression Model. Multiple linear regression is widely used in social science and marketing research for its interpretability and ability to assess the significance of relationships (Gordon, 2015). While non-linear models may improve prediction accuracy, they often compromise explainability. Despite criticisms of applying continous outcome regression to ordinal scales like star ratings or Likert scales (Binder et al., 2019), it remains the standard approach. Therefore, we adopt this method.

Word Embeddings. In our research questions, we ask whether coherent clusters of opinion units can be formed with respect to topic and sentiment. A key design consideration is the choice of word embedding. While general-purpose embeddings effectively capture semantic similarity, they often struggle to distinguish between opposing sentiments (Kim et al., 2024)—e.g., "the food is de-

licious" and "the food is disgusting" might be considered similar due to shared topical context. Recent research addresses this by fine-tuning embeddings to incorporate sentiment information (Tang et al., 2016; Yu et al., 2017; Fan et al., 2022; Kim et al., 2024). We compare a general-purpose embedding, all-mpnet-base-v2 (Transformers, 2024), with SentiCSE (Kim et al., 2024), a state-of-the-art, RoBERTa-based sentiment-aware model, to assess their impact on topic- and sentiment-cluster coherence. SentiCSE employs contrastive learning to separate positive and negative sentiment representations in embedding space. Through their LLM-rated sentiment metadata (1-10), opinion units offers an alternative approach to achieving sentiment-topic coherence by splitting data based on sentiment scores, then clustering using general-purpose embedding models. In our experiments, we compare these methods' ability to predict star ratings.

3 Applications

Our system offers distinct advantages for a range of use cases. We divide these into three classes:
1. *Exploratory Analysis of Customer Feedback.* Comparative analysis of product aspects or features: Understand which aspects resonate positively or negatively with customers. Evaluate how customer sentiments towards specific aspects of the business influence metrics like satisfaction, sales, churn, or brand loyalty. This customer knowledge could allow businesses to create more personalized promotions or experiences, tailored to specific customer segments' priorities and needs.

2. *Identification of Emerging Trends and Issues.* Real-time tracking of emerging trends in customer feedback. Businesses can quickly detect responses to new features or pinpoint emerging issues.

3. *Hypothesis-driven exploration* For targeted investigations, the system enables a non-compute-intensive, iterative approach. For example, analysts can evaluate sentiment toward a new coupon campaign, refining parameters such as cluster size and using keyword-seeded clustering to explore this predefined topic of inquiry—without the compute and time costs of classification methods.

TopicImpact is especially valuable for industries with large, diverse feedback data, such as e-commerce platforms and large retail chains, where customer preferences can vary widely and are often unpredictable. It is also beneficial for traditional service and product-based industries, including hotel and restaurant chains, fashion, automotive, or any business that gathers customer feedback.

The demand for customer opinion analysis, is also evident in academic research. Several studies, particularly in the hotel, restaurant, and consumer-tech sectors, have focused on: (i) clustering opinions in reviews to analyze topic frequency and sentiment (e.g. (Linshi, 2014; Debortoli et al., 2016; Hu et al., 2019; Krishnan, 2023)), and (ii) correlating topics with star ratings to evaluate their impact on overall business outcomes (e.g. (Debortoli et al., 2016; Linshi, 2014; Fu et al., 2013; Pappas and Popescu-Belis, 2014; Ganu et al., 2013; Radojevic et al., 2017; Binder et al., 2019)). These articles use methods such as LDA, keyword extraction, or available metadata, and therefore lack the advancements in explorability, cluster coherence and interpretability that LLM-preprocessing enables.

4 Experiments

We describe the experiments conducted to evaluate the proposed framework and research questions.

4.1 Dataset and Preprocessing

We base our experiments on the Yelp review dataset (Yelp, 2015). Due to the its vast size, we restrict ourselves to reviews from US restaurants, and focus on three gastronomic styles—Italian, Mexican, and Japanese. The filtering of reviews is based on Yelp's metadata. This results in three distinct datasets for topic modeling and star prediction analysis. Each dataset is subsampled to 5000 reviews.

Opinion units are generated using GPT-4, with a prompt in Appendix A1. We exclude opinion units related to the overall experience, focusing instead on specific aspects as described in Section 2. On average, each review generates 5.65 opinion units.

4.2 Topic Modeling

For our experiments, we adopt the BERTopic pipeline outlined by (Grootendorst, 2022b), employing hard HDBSCAN (Campello et al., 2013) for clustering and UMAP for dimensionality reduction. We consider two alternative sentence embedding models: a general-purpose model `all-mpnet-base-v2` (Transformers, 2024) and a state-of-the-art sentiment-aware embedding model (`SentiCSE`; see (Kim et al., 2024)). For our evaluation, we set the number of topics to 20 to ensure a manageable workload for human evaluation and the minimum topic size to 50 to provide sufficient data for statistical significance in regression analysis.

We then evaluate the coherence of the topic clusters and their predictive ability for star ratings. To accurately reflect system performance, we include all topics in the analysis without any sub-selection. However, subselection is often helpful in applications to focus on the topics that are of greatest interest to the user (discussed in Sections 2 and 3).

4.3 Evaluation of Topic Modeling

We evaluate the coherence of topics clusters with regards to topic and sentiment. For each cluster, three evaluators identify a dominant topical theme and determine which opinion units align with this theme (an inclusion-based approach (Eklund et al., 2024)). The precision of a cluster is defined as the proportion of opinion units that belong to its dominant theme. Evaluators were assigned 20 randomly sampled opinion units per topic. Further details on the evaluation are provided in Appendix A2.

To assess evaluator consistency, 5 opinion units per topic were reviewed by all evaluators, allowing us to calculate inter-rater agreement as (m/n), where m is the number of overlapping opinion units on which all evaluators agreed, and n is the number of total overlapping units.

Sentiment precision is calculated based on the distribution of LLM-assigned sentiment scores. We define it as the percentage of opinions with the dominant sentiment (positive or negative). For example, if 18/20 opinion units in a cluster have a sentiment score > 5 (positive), the precision is 90%. The same applies if 18/20 opinion units are negative (sentiment score ≤ 5).

4.4 Star Prediction: Regression analysis

After the topic modelling has been applied to the extracted opinion units, we can predict the star-rating of an individual review r through multiple linear regression (MLR). The basis for the regression is the opinion units extracted from r— their sentiment scores and their topic membership assigned through the topic modeling described in Section 4.2. The regression model is given in Equation 1, where K is the number of topics and $s_k(r)$ represents the average sentiment scores of the opinion units belonging to r that occur in topic cluster k (see Figure 1 and Section 2).

$$\star\text{-rating}(r) = \beta_0 + \sum_{k=1}^{K} \beta_k \cdot s_k(r) \qquad (1)$$

We implement three different methods for integrating topic and sentiment information to predict star ratings and compare their performance. To control for the effect of clustering granularity, we vary the number of clusters $K \in \{10, 20, 30\}$. For each method, we conduct evaluations both with sentiment scores (as defined in Equation 1) and without sentiment scores. It the latter case, we let $s_k(r) = 1$ if topic k is mentioned in review r and 0 otherwise.

M1. Cluster with a general embedding model, here `all-mpnet-base-v2` (Transformers, 2024).
M2. Cluster using a sentiment-aware embedding model, here SentiCSE (Kim et al., 2024).
M3. Split opinion units into negative and positive data based on LLM ratings (negative: ≤ 5, positive: > 5), then perform topic modeling separately using the general embedding model.

For Method 3, the sentiment scores (1-10) are adjusted to a 1-5 scale. For the cluster based on the positive data split, ratings of 1-5 correspond to the previous 6-10 range on the 1-10 scale, while for the negative cluster, the ratings of 1-5 remain. Splitting the data before clustering means that positive and negative influences are weighted separately. For instance, good service might not significantly impact the star rating because it is often expected, whereas poor service can be a decisive factor in lowering the overall experience.

We evaluate the predictive performance of the regression models using R^2 and RMSE on a holdout sample with 5-fold cross-validation.

Embedding	Dataset	% Outliers	Topic (P) Avg.	Topic (P) $\geq .9$	Sentiment (P) Avg.	Sentiment (P) $\geq .9$
General-Purpose	Italian	24.9	.905	79.0	.761	20.0
	Mexican	16.9	.917	79.0	.734	5.0
	Japanese	29.3	.863	63.2	.730	15.0
Sentiment-Aware	Italian	22.9	.843	55.3	.908	75.0
	Mexican	32.0	.832	60.5	.866	60.0
	Japanese	20.1	.821	52.6	.902	70.0

Table 1: Average topic and sentiment precision for cluster coherence, % of clusters with precision $\geq .9$ and % of outliers (opinion units not assigned to a cluster by BERTopic) for each embedding-dataset pair ($K = 20$).

5 Results

The alternative combinations of clustering approach and embedding model are evaluated based on (i) cluster coherence and (ii) effectiveness in predicting star ratings.

5.1 Topic and Sentiment Coherence

For the general-purpose embedding model (that is, `all-mpnet-base-v2`), the average topic precision over the clusters for each dataset fall in the range 86.3-91.7%, with 63.2-79.0% of topics achieving 90% precision (see Table 1), demonstrating a high topic coherence (Eklund and Forsman, 2022). The sentiment-aware model (`sentiCSE`) consistently performs worse across all three datasets. The results suggest that the general-purpose model largely ignores sentiment polarity when clustering, whereas the sentiment-aware model creates multiple clusters for the same topic with opposing sentiment (e.g., positive vs. negative price clusters). This focus on sentiment partly explains the lower topic precision, as it forms clusters that are cohesive in terms of sentiment but inconsistent in terms of topic. For instance, opinions about service, food quality, and cleanliness may be grouped together simply because they were all described as "okay".

Inter-rater agreement among evaluators was 90.3%. For topics with a clear theme (e.g., precision $\geq$80%), agreement was 96.1%. The percentage of outliers not assigned to a cluster ranges from 17-32%. While BERTopic can optionally force documents into clusters, we believe that having outliers is valid, as some opinions naturally fall outside the larger groups. Therefore, we allowed outliers.

To answer RQ1, our topic modeling results—comparing one general and one sentiment-aware embedding— achieve high topic cohesion but not high sentiment precision simultaneously. However, the metadata of opinion units can be

leveraged to create sentiment coherence. We implement this in the next section, where one of our star-rating prediction method splits the data based on sentiment-scores before topic modeling.

5.2 Star Prediction: Regression Analysis

In Table 2, we present the results of star-rating predictions using the three methods outlined in Section 4.4. The regression results show that sentiment-aware embeddings (Method 2) yield more accurate star rating predictions than general embeddings (Method 1). This improvement is expected, as topic clusters are formed based on both sentiment and topic, providing more relevant information about the customer's overall satisfaction. Method 3, which splits the dataset based on LLM-sentiment scores into positive and negative opinion units before clustering each split separately, achieves the highest accuracy with an R^2 value of 0.726, indicating a strong model fit. When clustering with general embeddings, incorporating sentiment scores of the individual opinions as described in Section 4.4, rather than using one-hot encoding for topic membership, enhances prediction accuracy.

The results are consistent across datasets and numbers of clusters. Table 2 shows regression outcomes averaged over the three datasets, with 5-fold cross-validation (K=20 clusters). Detailed results for each dataset and number of clusters ($K = 10, 20, 30$) are provided in Appendix A3.

To answer RQ2, our results show that TopicImpact accurately predicts star ratings. Topic modeling alone using general embeddings yields unsatisfactory results due to insufficient sentiment capture. However, combining topic assignment with the opinion units' sentiment scores—by segmenting data by sentiment before clustering and incorporating sentiment scores into regression (Eq. 1)—achieves high regression accuracy.

Method	s_k scores	# Sig. β	R^2	RMSE
M1. General Embeddings	without	14.5	.113	1.31
	with	14.6	.383	1.12
M2. Sentiment-Aware Embeddings	without	14.5	.487	1.01
	with	15.3	.393	1.10
M3. Sentiment Splitting	without	11.5	.652	.825
	with	15.1	**.726**	**.731**

Table 2: Star-prediction: No. of significant β-coefficients (out of 20), R^2 and RMSE ($K = 20$).

To illustrate the output from TopicImpact, we include results from the Japanese dataset using Method 3 with $K = 20$ clusters (10 negative, 10 positive). Table 3 lists the top 5 topics of each sentiment, ranked by impact on star ratings. The complete table is found in Appendix A4.

Topic	β	Size	Opinion Unit Example
Food safety	-1.21	56	Food poisoning: The next day was spent in the bathroom for several hours getting rid of the sushi
Service	-.886	1718	Staff friendliness: Our waiter had a very unfriendly standoffish personality
Food	-.852	2398	Food quality: My food was awful... overcooked
Order error	-.621	76	Order accuracy: I got the complete wrong order
Portions	-.511	73	Portion size: The serving was pathethically small
Food	.835	7182	Sushi: The sushi is always spot on delicious
Service	.350	1623	Service: the service is super good. The waitresses will check on you and do refills
Price	.347	1268	Food value: the food is very good for the price
Atmosphere	.346	182	Atmosphere: Casual and comfortable atmosphere!
Interior	.229	758	Décor: The decor is a simple and clean aesthetic

Table 3: *Topic* is a description of the topic, β is the regression coefficient in star rating prediction, *Size* is the number of opinion units in the topic, and the final column provides an example opinion unit from the topic.

6 Conclusion

TopicImpact enhances the extraction of actionable insights from customer reviews by integrating topic modeling with LLM-powered segmentation of reviews into distinct opinion units—individual, separated opinions supported by text excerpts. This provides both topic coherence and interpretability, while the metadata of opinion units enables sentiment-based segmentation. By incorporating sentiment scores and correlating topics with business metrics, TopicImpact effectively predicts star ratings and reveals how customer concerns influence business outcomes. This approach offers a fast, cost-effective solution for businesses to explore and act on customer feedback.

A promising avenue for future research is integrating keyword or example-guided topic seeding within our framework of topic modeling with LLM preprocessing. This approach could combine the predefined segmentation of classification with the exploratory, iterative, and cost-effective benefits of topic modeling. Furthermore, verifying the performance of TopicImpact in other domains, such as retail product reviews, course evaluations, or employee satisfaction surveys, could improve opinion analysis in these critical areas.

A limitation of LLM preprocessing is that it sometimes misses opinions in reviews or creates excerpts lacking full context, which hurts opinion analysis (Häglund and Björklund, 2025). However, preprocessing also addresses key issues in opinion clustering, such as removing non-opinionated text.

References

Markus Binder, Bernd Heinrich, Mathias Klier, Andreas Alexander Obermeier, and Alexander Schiller. 2019. Explaining the stars: Aspect-based sentiment analysis of online customer reviews. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden. Research Papers.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Stefan Debortoli, Oliver Müller, Iris Junglas, and Jan Vom Brocke. 2016. Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems (CAIS)*, 39(1):7.

Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Anton Eklund, Mona Forsman, and Frank Drewes. 2024. CIPHE: A framework for document cluster interpretation and precision from human exploration. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 536–548, Miami, USA. Association for Computational Linguistics.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, JIan Guo, and Nan Duan. 2022. Sentiment-aware word and sentence level pre-training for sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4994, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1276–1284.

Gayatree Ganu, Yogesh Kakodkar, and AméLie Marian. 2013. Improving the quality of predictions using textual information in online user reviews. *Information Systems*, 38(1):1–15.

Rachel A Gordon. 2015. *Regression analysis for the social sciences*. Routledge.

Maarten Grootendorst. 2022a. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Maarten Grootendorst. 2022b. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Nan Hu, Ting Zhang, Baojun Gao, and Indranil Bose. 2019. What do hotel customers complain about? text analysis using structural topic model. *Tourism Management*, 72:417–426.

Emil Häglund and Johanna Björklund. 2025. Opinion units: Concise and contextualized representations for aspect-based sentiment analysis. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*.

Jaemin Kim, Yohan Na, Kangmin Kim, Sang-Rak Lee, and Dong-Kyu Chae. 2024. SentiCSE: A sentiment-aware contrastive sentence embedding framework with sentiment-guided textual similarity. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14693–14704, Torino, Italia. ELRA and ICCL.

Anusuya Krishnan. 2023. Exploring the power of topic modeling techniques in analyzing customer reviews: a comparative analysis. *arXiv preprint arXiv:2308.11520*.

Jack Linshi. 2014. Personalizing yelp star ratings: a semantic topic modeling approach. *Yale University*.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pages 455–466.

Tijana Radojevic, Nemanja Stanisic, and Nenad Stanic. 2017. Inside the rating scores: A multilevel analysis of the factors influencing customer satisfaction in the hotel industry. *Cornell Hospitality Quarterly*, 58(2):134–164.

Nigel Slack. 1994. The importance-performance matrix as a determinant of improvementpriority. *International Journal of Operations & Production Management*, 14(5):59–75.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and M. Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28:496–509.

Asin Tavakoli, Holger Harreis, Kayvaun Rowshankish, and Michael Bogobowicz. 2024. Charting a path to the data and ai-driven enterprise of 2030. Accessed: 2025-02-06.

Sentence Transformers. 2024. Pretrained models. https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. Accessed: 2024-04-20.

Yelp. 2015. Yelp open dataset. Dataset.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.

A Appendix

A.1 Prompt template

Perform aspect-based sentiment analysis for the restaurant review provided as the input. Return each aspect-sentiment pair with a label and a corresponding excerpt from the text. Also rate the sentiment of each aspect on a scale from 1-10 where 1 is highly negative and 10 is highly positive.

Aspect-sentiment pairs should not mix opinions on different aspects. Make sure to include all aspects. An aspect should be independent and not have to rely on other aspects to be understood.

If an opinion in the review is about the restaurant or experience in general then label this aspect as "overall experience". Opinions not related to the restaurant should not be included.

Example input: I just left Mary's with my lovely wife. We had a very mixed experience. 3 out of 5 stars. The gorgeous outdoor patio seating was fantastic with a nice view of the ocean. First, we split a dozen oysters. They were the best I had in my life! FRESH! Delicious! The avocado toast was excellent as were the crab cakes. However, I absolutely hated the dessert we ordered and did not particularly like my cocktail. Also, the staff could have been a little friendlier.

Example output:
[["Overall experience","We had a very mixed experience. 3 out of 5 stars.",6],
["Outdoor patio seating","The gorgeous outdoor patio seating was fantastic with a nice view of the ocean",9],
["View","a nice view of the ocean",8],
["Oysters","we split a dozen oysters. They were the best I had in my life! FRESH! Delicious!",10],
["Avocado toast","the avocado toast was excellent",9],
["Crab cakes","the crab cakes were excellent",9],
["Dessert","I absolutely hated the dessert we ordered",1],
["Cocktail","I did not particularly like my cocktail",3],
["Staff friendliness","the staff could have been a little friendlier",4]]

Input: Review to be processed
Output:

Figure 2: Prompt template for generating opinion units.

A.2 Instructions for Evaluators Analyzing the Clustered Topics

1. Select the Topic: Click on a topic in the topic list.

2. Review the Excerpts: Carefully read through the 20 opinions/restaurant review excerpts belonging to the selected topic.

3. Identify a Common Theme: Look for a common theme among the review excerpts. Focus only on the topic being discussed (e.g., service, desserts, cleanliness, etc.). Ignore the sentiment of the opinions—whether they are positive or negative is not relevant.

4. Name the Topic: In the Excel file, write a descriptive name for the topic in the "Topic Name" column. If there is no clear majority theme, choose the most common or representative theme based on your judgment.

5. Flag Errors: Identify any review excerpts that DO NOT FIT the topic. In the Excel file, list the IDs of these excerpts under the "Error IDs" column. Separate the IDs with commas (e.g., 5,7,15).

A.3 Additional Star Rating Regression Results.

The following tables present the detailed star rating regression results as discussed in Section 5.2. These results are shown for each dataset (Italian, Mexican, and Japanese) and when varying number of clusters (K) used in the topic modeling. Results are averages over 5-fold cross validation on a hold-out sample.

Method	s_k-ratings	K	R2	RMSE	Sig. Coeffs. (β)
M1. General Embeddings	With	10	.482	1.02	6.2/10
	Without	10	.090	1.35	8.6/10
	With	20	.380	1.12	12.6/20
	Without	20	.136	1.32	16.0/20
	With	30	.335	1.16	20.2/30
	Without	30	.145	1.31	21.2/30
M2. Sentiment-Aware Embeddings	With	10	.366	1.13	8.8/10
	Without	10	.441	1.07	9.4/10
	With	20	.378	1.13	17.2/20
	Without	20	.463	1.05	17.0/20
	With	30	.350	1.16	25.2/30
	Without	30	.456	1.06	25.4/30
M3. Sentiment Splitting	With	10	.766	.687	8.0/10
	Without	10	.671	.814	8.6/10
	With	20	.750	712	14.4/20
	Without	20	.678	.808	15.4/20
	With	30	.728	.744	20.0/30
	Without	30	.661	.832	21.0/30
	With	40	.716	.763	31.0/40
	Without	40	.666	.827	31.0/40

Table 4: Italian restaurant dataset: Star prediction results (R^2 and RMSE) and no. of significant β-coefficients, varying the number of clusters (K).

Method	s_k-ratings	K	R2	RMSE	Sig. Coeffs. (β)
M1. General Embeddings	With	10	.544	.900	7.2/10
	Without	10	.089	1.27	8.2/10
	With	20	.364	1.06	14.8/20
	Without	20	.119	1.25	14.8/20
	With	30	.335	1.09	19.0/30
	Without	30	.136	1.25	17.8/30
M2.Sentiment-Aware Embeddings	With	10	.385	1.05	8.6/10
	Without	10	.503	.945	7.2/10
	With	20	.396	1.05	12.2/20
	Without	20	.531	.922	9.8/20
	With	30	.401	1.04	21.2/30
	Without	30	.531	.918	20.0/30
M3. Sentiment Splitting	With	10	.720	.705	8.8/10
	Without	10	.641	.797	8.8/10
	With	20	.705	.726	15.6/20
	Without	20	.635	.807	14.8/20
	With	30	.697	.736	23.4/30
	Without	30	.642	.802	21.8/30
	With	40	.695	.743	29.0/40
	Without	40	.641	.806	28.4/40

Table 5: Mexican restaurant dataset: Star prediction results (R^2 and RMSE) Star prediction results (R^2 and RMSE) and no. of significant β-coefficients, varying the number of clusters (K).

Method	s_k-ratings	K	R2	RMSE	Sig. Coeffs. (β)
M1. General Embeddings	With	10	.456	1.06	8.8/10
	Without	10	.049	1.40	8.0/10
	With	20	.326	1.18	16.4/20
	Without	20	.084	1.37	12.8/20
	With	30	.320	1.18	20.2/30
	Without	30	.125	1.34	20.6/30
M2.Sentiment-Aware Embeddings	With	10	.498	1.02	6.6/10
	Without	10	.533	.985	7.6/10
	With	20	.404	1.12	16.6/20
	Without	20	.466	1.05	16.8/20
	With	30	.422	1.1	27.6/30
	Without	30	.482	1.04	25.4/30
M3. Sentiment Splitting	With	10	.727	.750	9.6/10
	Without	10	.640	.861	10.0/10
	With	20	.724	.754	15.2/20
	Without	20	.643	.859	17.2/20
	With	30	.691	.800	22.8/30
	Without	30	.627	.880	22.2/30
	With	40	.688	.807	29.6/40
	Without	40	.629	.880	30.2/40

Table 6: Japanese restaurant dataset: Star prediction results (R^2 and RMSE) and no. of significant β-coefficients, varying the number of clusters (K).

A.4 Example Star Rating Regression Output

Below in Table 7 we present results from the japanese dataset using Method 3: Sentiment Splitting, with K = 20 clusters (10 negative and 10 positive). Below we summarize the topics ranked by their influence on star ratings (coefficients).

Topic	β	Size	Opinion Unit Example
Food safety	-1.21	56	Food poisoning: The next day was spent in the bathroom for several hours getting rid of the sushi
Service	-.886	1718	Staff friendliness: Our waiter had a very unfriendly standoffish personality
Food	-.852	2398	Food quality: My food was awful... overcooked
Order error	-.621	76	Order accuracy: I got the complete wrong order
Portions	-.511	73	Portion size: The serving was pathethically small
Hibachi	-.458	363	Hibachi grill entertainment: the chef was lacking in entertainment. I was borderline bored
Pricing	-.346	755	Sushi pricing: Pricing was too high for a conveyor belt style. These shops are popular in Japan, Taiwan, and China yet they were charging $11.00 a roll compared to other country's $1.50 a roll
Cleanliness	-.311	921	Cleanliness: the restaurant was filthy
Drinks	-.215	170	Sake selection: it tasted like water with a little vodka poured into it
Parking	-.161	121	Parking: A negative is the parking spots are narrow. It would deter me from going there when it's busy
Food	.835	7182	Sushi: The sushi is always spot on delicious
Service	.350	1623	Service: the service is super good. The waitresses will check on you and do refills
Price	.347	1268	Food value: the food is very good for the price
Atmosphere	.346	182	Atmosphere: Casual and comfortable atmosphere!
Interior	.229	758	Décor: The decor is a simple and clean aesthetic
Noodles	.137	72	Nabeyaki udon: The udon chicken tastes AMAZING. it hands down the best nabeyaki udon in town
Busyness	.137	248	Restaurant crowding: Whenever we've gone on a weekend evening, it's never been terribly crowded
Family friendly	n.s.	75	Suitability for families: great for families!
Location	n.s.	337	Location: The location of this place is actually beautiful
Views	n.s.	66	Views: The place has wonderful views, since it's right on the river

Table 7: *Topic* is a description of the topic, β. is the regression coefficient in star rating prediction, *Size* is the number of opinion units in the topic, and the final column provides an example opinion unit. "n.s." indicates that the β-coefficient was not statistically significant (i.e., p > .05).