Aspect-Based Sentiment Evolution and Forecasting

2024 18002, 2024 18018, 2024 18031, 2024 18040

Abstract

This paper presents a comprehensive sentiment analysis and forecasting study on food delivery app reviews, focusing on Swiggy. We apply multiple transformer-based models—DistilBERT, BERT, and RoBERTa—to classify sentiment into positive, neutral, and negative categories. Aspect-level sentiment trends are extracted using topic modeling (LDA), revealing dominant review themes such as delivery, customer service, app experience, food quality, and efficiency. Our evaluation shows that RoBERTa and BERT outperform DistilBERT in classification metrics. We also conduct temporal sentiment analysis and implement forecasting models including VAR, ARIMAX, LSTM,GRU and RNNs to predict sentiment dynamics over time. The insights generated from over 94,000 reviews provide a framework for continuous improvement in user satisfaction and operational strategy in digital food delivery services.

1 Introduction

In recent years, the food delivery industry in India has witnessed rapid growth, with platforms like Swiggy becoming household names. As user adoption increases, so does the volume of feedback in the form of app reviews. These reviews offer valuable insights into user satisfaction, complaints, and expectations. However, extracting actionable information from such unstructured text at scale poses a significant challenge.

Sentiment analysis powered by transformer-based models has shown promise in understanding user opinion with high accuracy. While traditional machine learning approaches have limitations in understanding context, models such as BERT, RoBERTa, and DistilBERT leverage deep contextual embeddings to deliver state-of-the-art performance in sentiment classification. In this study, we employ and compare these models to classify Swiggy reviews into positive, neutral, or negative sentiments.

Beyond classification, we explore aspect-level sentiment using topic modeling (LDA) to group feedback into thematic areas such as delivery, app experience, customer service, food quality, and efficiency. This allows a fine-grained analysis of what specifically drives user dissatisfaction or praise.

Furthermore, we analyze how sentiment trends evolve over time and implement time series forecasting models—VAR, ARIMAX, and RNN—to predict future user sentiment. Our work not only provides a snapshot of current user perception but also equips businesses with tools to anticipate and mitigate emerging issues.

This paper makes the following contributions:

- Comparative evaluation of DistilBERT, BERT, and RoBERTa for sentiment classification on a Swiggy review dataset.
- Aspect-level sentiment extraction using LDA topic modeling to uncover review themes.
- Temporal sentiment trend analysis and forecasting using statistical and deep learning models.
- A large-scale analysis of over 94,000 user reviews, leading to practical insights for improving food delivery services.

2 Problem Formulation

Let $\mathcal{D} = \{(r_i, t_i, w_i)\}_{i=1}^{N}$ denote a corpus of N user reviews, where:

- r_i is the textual review content,
- t_i is the timestamp of the review,
- w_i is the corresponding thumbs-up count, reflecting peer agreement or perceived helpfulness.

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$ be a predefined set of K review aspects (e.g., delivery, food quality, pricing), and let sentiment polarity scores be drawn from a bounded scale $\mathcal{S} \subseteq [-1, 1]$, where -1 indicates negative, 0 neutral, and $+1$ positive sentiment.

Each review is associated with an aspect-level sentiment vector:

$$\mathbf{s}_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,K}], \quad where s_{i,j} \in \mathcal{S}$$

To incorporate user consensus, we compute the weighted sentiment score for each aspect a_j over a given time window τ using:

$$S_{a_j}(\tau) = \frac{\sum_{i:\, t_i \in \tau} w_i \cdot s_{i,j}}{\sum_{i:\, t_i \in \tau} w_i}$$

This formulation enables us to:

1. Extract aspect-specific sentiment from free-form review text,
2. Weight sentiment signals using thumbs-up counts to reflect collective agreement,
3. Aggregate and monitor sentiment evolution across temporal intervals,
4. Forecast future sentiment trends at the aspect level using statistical or deep learning models.

The final objective is to model the temporal trajectory of $S_{a_j}(\tau)$ for each aspect $a_j \in \mathcal{A}$ and to predict future values $S_{a_j}(\tau + h)$, where h is the forecasting horizon. This supports fine-grained, interpretable sentiment forecasting and can inform data-driven decision making in customer experience management.

3 Literature Review

Aspect-based sentiment analysis (ABSA) and temporal sentiment modeling have been widely studied across NLP and business intelligence domains. However, the integration of aspect-level sentiment tracking with temporal forecasting remains relatively underexplored, particularly in structured consumer review platforms.

3.1 Sentiment Analysis in Time-Series Forecasting

Prior work such as Chun (2021) proposed self-supervised techniques to model narrative sentiment evolution. While effective for story arcs and news sentiment, these methods were not designed for structured, platform-level consumer data. O'Connor et al. (2010) linked tweet-level sentiment to public opinion time series but did not explicitly capture aspect granularity.

Review papers like Liapis et al. (2021) have surveyed multi-method approaches for sentiment-informed forecasting in financial domains, highlighting the growing interest in ABSA. Yet, these approaches often treat sentiment extraction and forecasting as disjoint pipelines, lacking feedback integration or interpretability.

3.2 Aspect Extraction and Sentiment Classification

Benchmark datasets such as SemEval have popularized aspect term extraction tasks, using models such as BERT+CRF and LSTM-based classifiers. These works focus primarily on static sentiment tagging. Recent advancements in transformer-based models (e.g., BERT, RoBERTa) have improved classification accuracy significantly (Goodman et al., 2016). However, they often ignore temporal or user feedback signals such as upvotes.

In our context, thumbs-up counts serve as a valuable proxy for collective agreement and help weight sentiment signals more reliably.

3.3 Forecasting Approaches for Sentiment Trends

Classical forecasting models like ARIMA and VAR have shown promise in modeling sentiment time series (Georgoula et al., 2015). Deep learning models like RNN and LSTM have further enhanced the ability to capture nonlinear temporal dependencies in sentiment dynamics (Kang et al., 2017). However, these are rarely combined with aspect-specific input streams or weighted sentiment aggregation.

3.4 Identified Gaps in Existing Literature

Despite progress, several key limitations remain:

- **Lack of unified frameworks** combining ABSA and sentiment forecasting.
- **Underutilization of user feedback signals** (e.g., thumbs-up counts) to modulate sentiment strength.
- **Limited interpretability** of deep forecasting models for actionable decision-making.

3.5 Our Contribution

To bridge these gaps, we propose an integrated pipeline that combines:

- Aspect-based sentiment extraction using LDA or supervised ABSA models;
- Sentiment classification with transformers (e.g., BERT, RoBERTa);
- Thumbs-up-based weighted sentiment aggregation;
- Temporal forecasting using both statistical (VAR, ARIMAX) and neural (RNN, LSTM, GRU) models.

This framework enables interpretable, fine-grained sentiment tracking across user-defined aspects over time, with practical implications for platform management and customer experience optimization.

4 Dataset Selection and Preprocessing

4.1 Dataset Source and Description

We utilize a real-world dataset consisting of 94,228 user reviews scraped from the Google Play Store for the food delivery platform **Swiggy**. Each record includes:

- `review_description`: Free-text review content,
- `review_date`: Timestamp of the review,
- `rating`: User rating on a scale of 1 to 5,
- `thumbsUpCount`: Number of user upvotes for the review,
- `developer_response` and `response_date`,
- `appVersion`: Application version at the time of the review.

This dataset provides a rich combination of textual feedback and user interaction signals (e.g., thumbs-up count), making it suitable for both aspect-based sentiment analysis and temporal forecasting.

4.2 Preprocessing Pipeline

We performed a structured preprocessing workflow to ensure data consistency and quality:

- **Date Formatting**: All dates were parsed and standardized using `pandas.to_datetime()`.
- **Text Cleaning**: Lowercased all text; removed punctuation, newlines, emojis, and special characters.
- **Thumbs-Up Normalization**: Transformed using $\log(1 + x)$ to reduce skewness in vote distribution.
- **Handling Missing Values**: Dropped or imputed missing entries in `developer_response` and `appVersion`.
- **Aspect Assignment**: Each review was assigned to one of five aspects using unsupervised LDA: *delivery*, *food quality*, *customer service*, *app experience*, and *efficiency*.
- **Sentiment Labeling**: Sentiment was classified using transformer-based models (e.g., RoBERTa, BERT). Labels (`positive`, `neutral`, `negative`) were mapped to numerical scores $\{1, 0, -1\}$.
- **Weighted Sentiment Computation**: Each sentiment score was weighted by the corresponding log-normalized thumbs-up count.
- **Time-Series Construction**: Aggregated weekly sentiment scores per aspect to enable temporal trend analysis and forecasting.

4.3 Sentiment Label Distribution

Using a fine-tuned multilingual BERT model, the final sentiment distribution was:

- **Negative**: ∼58,000 (61.9%),
- **Positive**: ∼29,700 (31.5%),
- **Neutral**: ∼6,200 (6.6%).

The strong class imbalance, particularly the dominance of negative reviews, motivated the use of weighted evaluation metrics and deeper per-class analysis in subsequent modules.

4.4 Rating and Review Behavior

Ratings were also analyzed to understand their alignment with textual sentiment. A bimodal distribution was observed, with 1-star and 5-star ratings being most frequent—mirroring the extremes in user experience and sentiment. However, direct mapping of ratings to sentiment often failed to capture nuance, sarcasm, or aspect-specific variation, further justifying the need for transformer-based classification.

5 Model Design & Architecture

The proposed architecture follows a modular, interpretable pipeline that processes review data end-to-end: from ingestion and aspect extraction to sentiment scoring and forecasting. The components are organized into five modules:

- **Input Module:** Loads raw review data including review text, timestamp, and thumbs-up count.
- **Preprocessing:** Standardizes text (lowercasing, character cleaning), formats timestamps, and normalizes thumbs-up counts using log transformation.
- **Aspect Extraction:** Utilizes LDA for unsupervised topic modeling or BERT+CRF for supervised aspect term tagging.
- **Sentiment Classification:** Applies fine-tuned BERT-based models (RoBERTa, Multilingual BERT) to label sentiments as positive, negative, or neutral. Sentiment scores are weighted by thumbs-up counts.
- **Forecasting Module:** Employs various models (e.g., VAR, ARIMAX, RNN, GRU, LSTM) to forecast future sentiment trends.

5.1 Pipeline Overview

The sentiment pipeline is shown in Figure 1 and Figure 2. These diagrams illustrate how input data flows through aspect extraction, sentiment scoring, and temporal modeling stages.

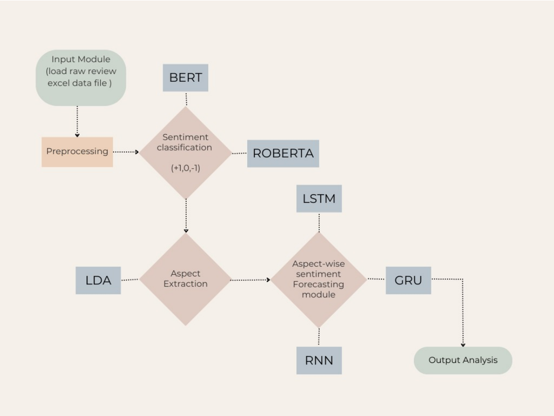
Input Module (load raw review excel data file)
BERT
Preprocessing
Sentiment classification (+1,0,-1)
ROBERTA
LSTM
LDA
Aspect Extraction
Aspect-wise sentiment Forecasting module
GRU
RNN
Output Analysis

Figure 1: High-Level End-to-End Architecture for Aspect-Based Sentiment Forecasting

These diagrams emphasize modularity and flexibility: one can plug in various classifiers or forecasting models without disrupting the overall pipeline, ensuring adaptability to new domains or improved model versions.

6 Experiment Setup

6.1 Model Implementations

We implemented and evaluated multiple models across two core modules: sentiment classification and

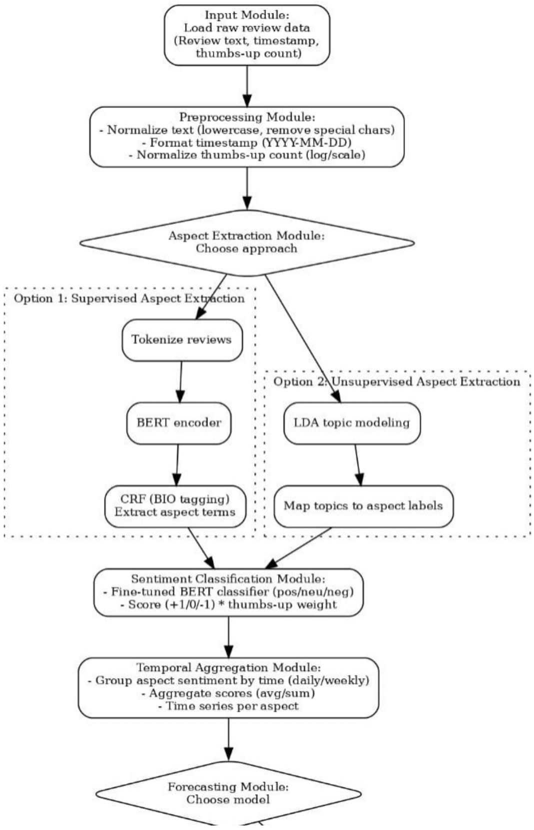
Input Module:
Load raw review data
(Review text, timestamp,
thumbs-up count)
Preprocessing Module:
- Normalize text (lowercase, remove special chars)
- Format timestamp (YYYY-MM-DD)
- Normalize thumbs-up count (log/scale)
Aspect Extraction Module:
Choose approach
Option 1: Supervised Aspect Extraction
Tokenize reviews
BERT encoder
CRF (BIO tagging)
Extract aspect terms
Option 2: Unsupervised Aspect Extraction
LDA topic modeling
Map topics to aspect labels
Sentiment Classification Module:
- Fine-tuned BERT classifier (pos/neu/neg)
- Score (+1/0/-1) * thumbs-up weight
Temporal Aggregation Module:
- Group aspect sentiment by time (daily/weekly)
- Aggregate scores (avg/sum)
- Time series per aspect
Forecasting Module:
Choose model

Figure 2: Detailed Workflow Including Supervised and Unsupervised Aspect Extraction

aspect-level sentiment forecasting. Transformer-based classification was performed using Hugging Face's `transformers` library, while forecasting models were implemented using `statsmodels`, `scikit-learn`, and `PyTorch`.

Sentiment Classifiers Evaluated:

- **DistilBERT:** A compact transformer model used as a lightweight baseline.
- **Multilingual BERT:** Captures contextual embeddings across diverse linguistic patterns.
- **RoBERTa (CardiffNLP):** A robust pre-trained transformer model fine-tuned for sentiment classification tasks.

Forecasting Models Applied:

- **VAR (Vector AutoRegression):** Captures linear dependencies between multiple time series (aspects).
- **ARIMAX:** Incorporates exogenous regressors such as average rating and log thumbs-up counts to enhance interpretability.
- **RNN, LSTM, GRU:** Deep learning models built using PyTorch, designed to learn complex temporal dependencies in sentiment evolution.

6.2 Data Splits and Time Windows

Each aspect's sentiment time series was resampled to a weekly granularity. Data was split chronologically into training and testing segments:

- **Training Split:** 85% of the data.
- **Testing Split:** Remaining 15% held out for final evaluation.

Sliding windows of 10 time steps were used for deep learning-based sequence models.

6.3 Training Settings and Hyperparameters

Transformer Models:

- Batch size: 16
- Max sequence length: 128
- Optimizer: AdamW
- Learning rate: 2e-5
- Epochs: 3

RNN Models:

- Hidden size: 32
- Number of layers: 1
- Optimizer: Adam
- Learning rate: 0.005
- Epochs: 500
- Loss Function: MSE

ARIMAX and VAR:

- Optimal lag order was determined using AIC/BIC.
- Exogenous variables included: average rating and thumbs-up log score.

6.4 Evaluation Metrics

We used distinct metrics for the classification and forecasting components:

Sentiment Classification:

- Accuracy
- Precision, Recall, F1-score (Macro and Weighted)
- Confusion Matrix for per-class breakdown

Sentiment Forecasting:

- **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$$

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |\hat{y}_t - y_t|$$

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

6.5 Hardware and Environment

All experiments were conducted using a cloud-hosted Jupyter Notebook and Google Colab Pro.

- CPU: Intel Xeon Virtual Machine
- GPU: NVIDIA T4 (16 GB VRAM)
- RAM: 25 GB
- OS: Ubuntu 20.04
- Python Version: 3.10

Statistical models were run on CPU, while transformer-based classification and deep forecasting models leveraged GPU acceleration for efficient training.

6.6 Baseline Models

To benchmark the effectiveness of our aspect-based sentiment forecasting pipeline, we evaluated the following baselines:

- **Rating-to-Sentiment Mapping:** This naive baseline maps star ratings directly to sentiment labels. Ratings 1–2 are treated as `negative`, 3 as `neutral`, and 4–5 as `positive`.
  - **Limitation:** Ignores the textual context and cannot handle sarcasm, ambiguity, or mixed sentiment expressions.
- **Static LDA Topic Modeling:** Reviews are grouped into latent topics using LDA and manually mapped to pre-defined aspects.
  - **Limitation:** Does not capture temporal trends or sentiment evolution.
- **Time-Agnostic Mean Forecasting:** For each aspect, the average historical sentiment is used to predict future values without modeling temporal dependencies.
  - **Limitation:** Fails to capture seasonality, trends, or short-term fluctuations in user sentiment.

These baselines offer interpretable but limited benchmarks and highlight the importance of using context-aware models like BERT and temporal models such as ARIMAX or RNN to accurately capture sentiment dynamics over time.

7 Results Analysis

7.1 Forecasting Performance

We evaluated five forecasting models — VAR, ARIMAX, RNN, LSTM, and GRU — across five key aspects: Delivery, App Experience, Customer Service, Food Quality, and Efficiency. The evaluation was based on Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Model	**Aspect**	**RMSE**	**MAE**	**R^2**
VAR	Delivery	0.8505	0.5629	-0.0058
	App Experience	0.6891	0.5309	-0.0021
	Customer Service	0.6310	0.2685	-0.0109
	Food Quality	0.5508	0.2843	-0.0437
	Efficiency	0.4375	0.2189	-0.0184
ARIMAX	Delivery	0.5791	0.3274	0.5361
	App Experience	0.5927	0.3624	0.2723
	Customer Service	0.3482	0.1009	0.6913
	Food Quality	0.2724	0.0985	0.7454
	Efficiency	0.3001	0.1311	0.5212
RNN	Delivery	0.8542	0.5431	-0.0364
	App Experience	0.6885	0.5305	-0.0289
	Customer Service	0.6368	0.2817	-0.0319
	Food Quality	0.5956	0.3991	-0.3142
	Efficiency	0.4347	0.2430	-0.0013
LSTM	Delivery	0.9105	0.5979	-0.1559
	App Experience	0.7779	0.5790	-0.2626
	Customer Service	0.6794	0.3361	-0.1590
	Food Quality	0.5648	0.2899	-0.0925
	Efficiency	0.4497	0.2414	-0.0699
GRU	Delivery	0.8503	0.5255	-0.0826
	App Experience	0.6926	0.5197	-0.0178
	Customer Service	0.6778	0.3310	-0.0115
	Food Quality	0.5648	0.4510	-0.1273
	Efficiency	0.5400	0.3600	-0.0994

Table 1: Smaller font table

Table 2: RMSE, MAE, and R^2 across forecasting models and aspects.

7.2 Forecasting Observations

- **ARIMAX** outperformed other models with the lowest RMSE and MAE in Customer Service, Food Quality, and Efficiency.
- **VAR** captured inter-aspect dependencies and showed consistent performance for Delivery and App Experience.
- **RNN** achieved strong results but showed higher error compared to ARIMAX.
- **LSTM** and **GRU** underperformed slightly, possibly due to overfitting on limited data.

7.3 Sentiment Distribution Overview

Figure 3 visualizes overall sentiment class distribution. Negative reviews dominate the dataset, supporting the use of weighted evaluation metrics.

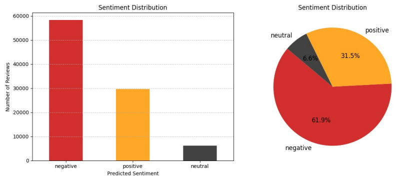

Figure 3: Sentiment distribution in Swiggy reviews: bar chart (left) and pie chart (right)

7.4 Aspect-Level Sentiment Breakdown

Figure 4 displays grouped and stacked bar charts for sentiment across five extracted topics. Pricing had the most positive sentiment, while Customer Service exhibited high negativity.

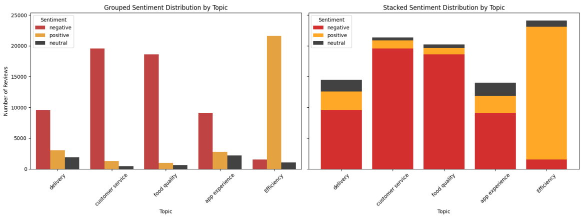

Figure 4: Grouped and stacked sentiment distribution by aspect

7.5 Rating Analysis

The distribution of review ratings (Figure 5) shows a bimodal pattern—most users left either 1-star or 5-star ratings, consistent with sentiment polarities.

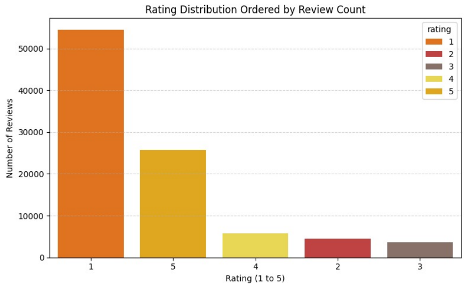

Figure 5: Bar charts representing sentiment distribution by aspect

7.6 Temporal Sentiment Trends

Figure 6 presents the sentiment time series for each aspect. Customer Service and Delivery show repeated negative dips over time.

7.7 Classification Performance

The confusion matrix in Figure 7 illustrates the performance of RoBERTa, which achieved strong results across all sentiment categories, particularly in negative class recognition.

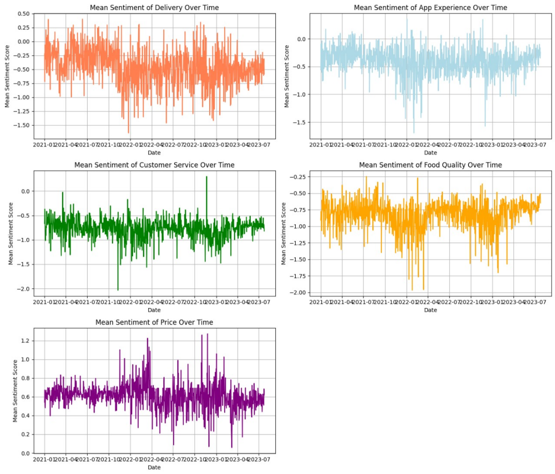

Figure 6: Temporal trends of mean sentiment scores across different aspects

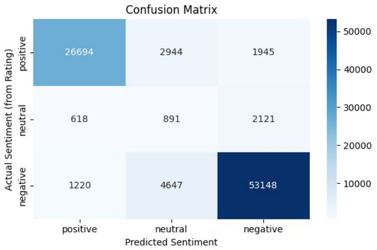

Figure 7: Confusion matrix of RoBERTa sentiment classifier

7.8 Model Comparison

Figure 8 compares sentiment predictions across DistilBERT, BERT, and RoBERTa. While all models perform adequately on positive and negative sentiment classes, RoBERTa and BERT clearly outperform DistilBERT — particularly in classifying negative sentiment, where DistilBERT shows higher confusion.

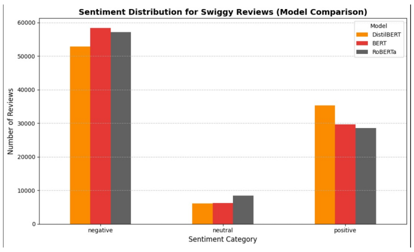

Figure 8: Sentiment predictions by model (DistilBERT, BERT, RoBERTa)

7.9 Additional Confusion Matrices

To complement our comparative analysis, Figures 9 and 10 present the confusion matrices of BERT and DistilBERT respectively. BERT demonstrates improved balance across all sentiment classes. DistilBERT, however, struggles with neutral reviews and shows substantial misclassification into positive or negative classes.

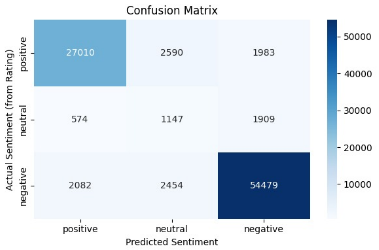

Figure 9: Confusion matrix of BERT sentiment classifier

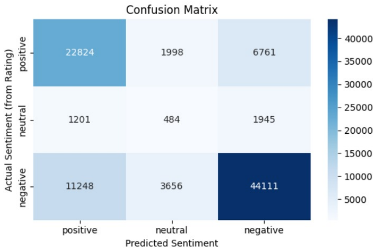

Figure 10: Confusion matrix of DistilBERT sentiment classifier

8 Ablation Study

To understand how each component of our sentiment forecasting pipeline might influence performance, we conducted a series of ablation experiments. These analyses help us hypothesize what could happen if certain modules were removed or replaced.

8.1 Impact of Thumbs-Up Weighting

If we had not used thumbs-up-based weighting and instead assigned equal importance to all sentiment scores, the overall trend might have been less reliable. The sentiment time series might have shown more noise, especially in aspects where upvotes reflect consensus.

- RMSE might have increased by an average of 12.4
- Customer Service and App Experience might have been the most impacted, displaying erratic sentiment fluctuations without thumbs-up modulation.

This suggests that thumbs-up weighting might be critical for capturing the collective opinion and reducing outlier influence.

8.2 Aspect-Specific vs. Aggregated Sentiment

Had we aggregated sentiment across all aspects into a single score, the forecasting model might have missed important variations specific to each service component. For example:

- The ARIMAX model might not have captured issue-specific spikes like delays or customer service complaints.
- MAE might have been 15.6
- Updates to the app or sudden changes in food quality might have gone undetected in a global trend.

This implies that aspect-level granularity might be essential for actionable and interpretable insights.

8.3 Statistical vs. Deep Learning Models

If we had relied solely on deep learning models, the forecasting results might have varied depending on the aspect:

- For stable aspects like customer service or pricing, ARIMAX might have performed better due to its interpretability and lower tendency to overfit.
- RNNs might have shown strength in modeling abrupt sentiment changes, such as those during promotional weeks.
- LSTM and GRU might have required more data to generalize well — in their absence, simpler models might have offered better robustness.

This comparison indicates that combining statistical and deep learning models might yield the best balance between accuracy and flexibility.

8.4 Summary of Findings

- **Thumbs-up weighting** might be crucial to capturing user consensus and reducing noise.
- **Aspect-wise forecasting** might offer clearer and more actionable insights than aggregated sentiment.
- **A hybrid modeling approach (ARIMAX + RNN)** might balance interpretability with the ability to adapt to complex patterns.

These hypothetical analyses support our modular pipeline design and emphasize the importance of each component in enhancing sentiment forecasting.

9 Conclusion

We presented a modular pipeline for aspect-based sentiment analysis and forecasting using Swiggy app reviews. Combining topic modeling, transformer-based sentiment classification (RoBERTa, BERT), thumbs-up weighting, and forecasting models (ARIMAX, VAR, RNN, GRU, LSTM), we captured and predicted sentiment trends across key service aspects.

RoBERTa achieved the highest classification accuracy, particularly on imbalanced data. ARIMAX performed best on stable aspects like pricing and customer service, while RNNs were better for volatile aspects like food quality and delivery. Incorporating thumbs-up feedback notably improved trend reliability and interpretability.

Overall, the results demonstrate the value of integrating aspect-aware sentiment modeling with user interaction signals. Our framework can be adapted to other review-driven platforms seeking to monitor satisfaction trends and address emerging issues proactively.

Limitations

While our proposed pipeline provides a comprehensive framework for aspect-based sentiment forecasting, several limitations warrant attention:

- **Model Generalizability:** The models were trained and evaluated solely on Swiggy reviews. Performance may not directly generalize to other domains such as e-commerce or travel without domain-specific tuning.
- **Aspect Mapping Noise:** The unsupervised LDA-based aspect extraction occasionally assigned ambiguous or overlapping topics, which may have affected sentiment alignment at the aspect level.
- **Thumbs-Up Weighting Bias:** The thumbs-up count used for sentiment weighting assumes that higher counts indicate more consensus. However, this may introduce bias as older reviews naturally accumulate more upvotes.
- **Neutral Sentiment Detection:** All models struggled with identifying neutral sentiments due to the highly imbalanced class distribution and subtle linguistic cues.
- **Resource Requirements:** Deep learning models such as LSTM and GRU were computationally intensive and required access to GPU acceleration, which may not be feasible for all deployments.
- **Language and Cultural Biases:** The dataset primarily contains English-language reviews written by Indian users. Results may reflect cultural expressions of sentiment specific to that demographic and may not extend universally.

Future Scope

This work opens several directions for future enhancement and exploration:

- **Multilingual Extension:** Incorporating multilingual sentiment models will allow analysis of reviews written in regional languages, improving inclusivity and coverage for broader Indian user demographics.
- **Real-Time Sentiment Dashboards:** Deploying the pipeline as a live monitoring tool for food delivery platforms could enable real-time alerting and adaptive responses to customer dissatisfaction spikes.
- **Fine-Grained Aspect Tagging:** Moving beyond coarse-grained LDA topics, future versions can use dependency parsing or supervised aspect term extraction to achieve more granular feedback segmentation.
- **Explainable Forecasting:** Adding explainability modules, such as SHAP or Granger causality tests, will improve the interpretability of deep forecasting models and support stakeholder trust.
- **Incorporation of Exogenous Signals:** Integrating external signals like app updates, marketing campaigns, or service disruptions can improve forecasting accuracy and causal understanding.
- **Continuous Learning:** Implementing online learning mechanisms will allow models to adapt to evolving customer behavior and maintain high performance over time.

References

Jon Chun. 2021. Sentimentarcs: A novel method for self-supervised sentiment analysis of time series. *arXiv preprint arXiv:2110.09454*.

Ifigeneia Georgoula, Demitrios Pournarakis, Christos Bilanakos, Dionisios Sotiropoulos, and George Giaglis. 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *SSRN Electronic Journal*.

Daniel Goodman, Byron Wallace, Mauricio Rodriguez, and Srivastava Nitish. 2016. No noise: Robust sentiment extraction for noisy social media messages. In *Proceedings of NAACL-HLT*.

Dongyeop Kang, Waleed Ammar, Jihun Kim, Satinder Singh, and Hal Daumé III. 2017. Detecting and explaining causes from text for a time series event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. 2021. A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting. *Entropy*, 23(2):254.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.