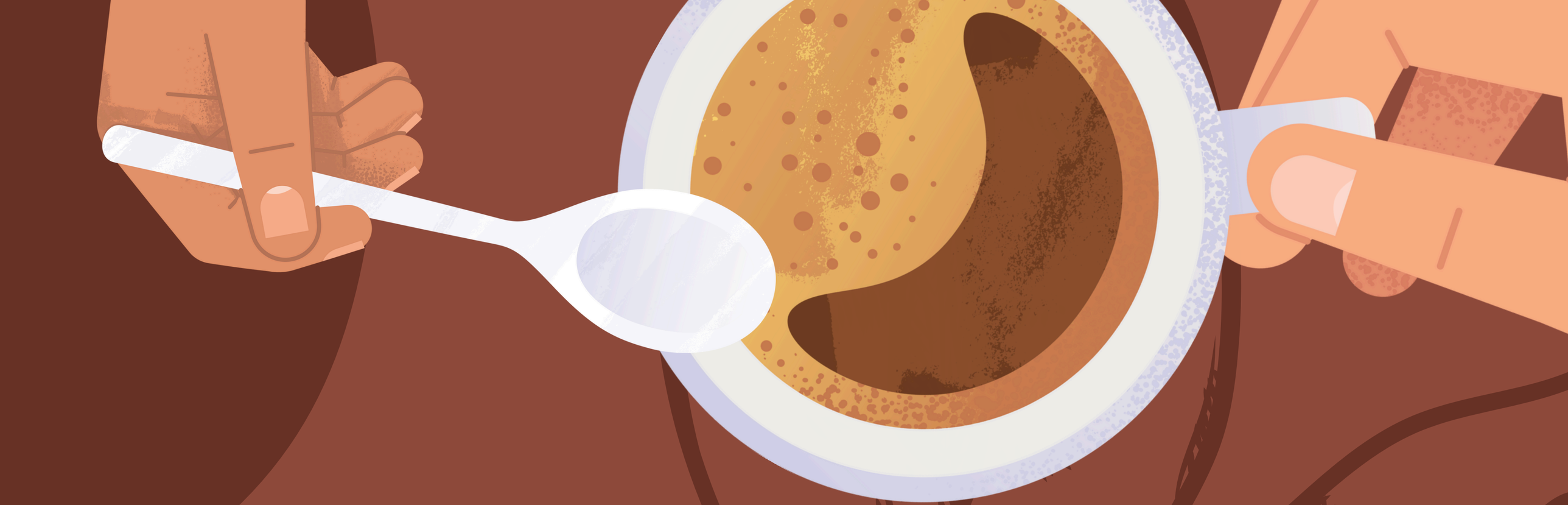

COMPREHENSIVE DATA MODELING FOR RIYADH COFFEE SHOPS

TEAM MEMBERS: LEEN ALQAHTANI, NORAH ALFAHEED, LAMYA ALNAHDI, HISSAH ALOTAIBI

AGENDA

1- Overview: Data Collection to Model Evaluation
2- Algorithms Selected and Rationale
3- Evaluation Metrics and Their Insights
4- Technical Challenges and Solutions
5- Future Work and Improvements
6- Conclusion

OVERVIEW: FROM DATA COLLECTION TO MODEL EVALUATION

Data Collection:

- Reviews and comments extracted from 26 Riyadh coffee shops using the Instant Data Scraper tool.
- Service options and atmosphere features also extracted using Instant Data Scraper.
- Busy hours scraped from Google Maps using Selenium automation.
- Locations, ratings, rating counts, and average prices were collected manually.

Data Preprocessing:

- Standardized and merged all datasets into a single structured dataset.
- Handled missing values and standardized field formats.

OVERVIEW: FROM DATA COLLECTION TO MODEL EVALUATION

Sentiment Analysis:

- Applied VADER Sentiment Analyzer to classify customer comments into Positive, Neutral, or Negative classes.

Model Development:

- Developed multiple machine learning models for predicting the sentiment of customer comments.

Model Evaluation:

- Models evaluated and compared using metrics such as Accuracy, Precision, Recall, and F1-Score.

SELECTED ALGORITHMS

Algorithms Applied:

- **Logistic Regression**

A baseline linear model for binary and multiclass classification; fast and interpretable.

- **Decision Tree Classifier**

Non-linear model capable of capturing complex patterns in sentiment data.

- **Random Forest Classifier**

Ensemble method combining multiple decision trees to improve generalization and reduce overfitting.

- **Support Vector Machine (SVM)**

Effective in high-dimensional spaces, particularly suitable for text-based features and small-to-medium datasets.

RATIONALE BEHIND ALGORITHM CHOICES

Goal:

- To classify customer reviews into Positive, Neutral, and Negative categories.
- Multiple classifiers were tested to select the model with the best overall performance

Why these algorithms?

- Logistic Regression:
  - Acts as a strong baseline for multiclass classification.
  - Interpretable and quick to train.
- Decision Tree Classifier:
  - Captures non-linear feature interactions.
  - Easy to visualize for initial analysis.
- Random Forest Classifier:
  - Reduces overfitting compared to a single decision tree.
  - Provides feature importance insights.
- Support Vector Machine (SVM):
  - Effective with high-dimensional data (text features).
  - Strong margin separation improves classification generalization.

EVALUATION METRICS AND THEIR PURPOSE

Metrics Used for Model Evaluation:

- **Accuracy**: Measures the percentage of total correct predictions.
- **Precision**: Evaluates the proportion of positive identifications that were actually correct.
- **Recall**: Measures the proportion of actual positives that were correctly identified.
- **F1-Score**: Harmonic mean of Precision and Recall; balances false positives and false negatives.

EVALUATION METRICS AND THEIR PURPOSE

Why multiple metrics?

- To handle potential class imbalance.

- To evaluate models not just based on overall accuracy, but also on their ability to correctly classify each sentiment class.

WHY THESE METRICS

- **Accuracy** alone may be misleading if classes are imbalanced (e.g., more positive reviews than negative).
- **Precision** is critical to reduce false positives (wrongly classifying neutral/negative reviews as positive).
- **Recall** ensures capturing as many real positive/negative reviews as possible.
- **F1-Score** provides a balanced view when there's a trade-off between Precision and Recall.

TECHNICAL CHALLENGES

Data Collection Challenges:

- Inconsistent Data Structures:
  - Different formats and missing fields across coffee shop data.

Selenium Automation Issues:

  - Busy hours scraping sometimes failed due to dynamic loading delays or missing elements.

Modeling Challenges:

- Class Imbalance:
  - Positive reviews significantly outnumbered neutral and negative ones.
- Text Noise and Variability:
  - Reviews included non-English text, emojis, abbreviations, and irrelevant content.

Evaluation Challenges:

- Small Dataset Size:
  - Limited data per class impacted model generalization and robustness.

SOLUTIONS TO CHALLENGES

Data Handling Solutions:

- Unified all data sources under a standardized format.
- Manual collection of missing fields (location, ratings, rating counts, average prices).

Selenium Scraping Solutions:

- Added explicit wait times and fallback mechanisms.
- Marked missing busy hours as "No data available" to avoid scraping failures stopping the pipeline.

Modeling Solutions:

- Chose classifiers that perform relatively well under imbalance (Random Forest, SVM).
- Focused on using robust features like sentiment scores and busy hour data instead of raw text alone.

Evaluation Solutions:

- Evaluated models with Precision, Recall, and F1-Score, not just Accuracy, to account for imbalance.

FUTURE WORK AND IMPROVEMENTS

Data Collection Improvements:

- Expand the dataset to include more coffee shops and time periods (seasonal analysis).
- Automate data extraction for additional features like customer demographics.

Modeling Enhancements:

- Implement text preprocessing techniques like TF-IDF or word embeddings (Word2Vec, GloVe).
- Perform hyperparameter tuning using GridSearchCV or RandomizedSearchCV.

FUTURE WORK AND IMPROVEMENTS

Advanced Modeling:

- Experiment with Deep Learning models (e.g., BERT-based classifiers) for sentiment analysis.
- Explore ensemble stacking to combine multiple models for improved accuracy.

Business Insights Expansion:

- Predict future rating trends based on review sentiment and busy hour patterns.
- Develop recommendation systems based on customer feedback sentiment.

CONCLUSION

This project demonstrated a complete machine learning pipeline:

1. Data collection from multiple sources (web scraping, manual extraction).
2. Sentiment analysis using VADER.
3. Busy hours integration through Selenium-based automation.
4. Building and evaluating multiple classification models.
5. Comparative model evaluation revealed the strengths and trade-offs of different algorithms.
6. Despite challenges such as data imbalance and small dataset size, robust solutions were implemented to ensure reliable results.
7. Future improvements can further enhance predictive capabilities and business insights through deeper feature engineering and advanced modeling techniques.


THANK YOU