Configuration Manual: Dynamic ABSA for E-commerce Customer Feedback Optimisation

MSc Research Project
Data Analytics

Jenat Jiffna Rodrigues
Student ID: x23337451

School of Computing
National College of Ireland

Supervisor: Bharat Agarwal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	……… Jenat Jiffna Rodrigues…………………………………………………………………………
Student ID:	……… x23337451…………………………………………………………………………………………
Programme:	……… MSc Data Analytics……………… **Year:** …2025-26………
Module:	……… Research Project …………………………………………………………………………………
Supervisor:	……… Bharat Agarwal…………………………………………………………………………………
Submission Due Date:	………28/01/2026…………………………………………………………………………………………
Project Title:	……Dynamic ABSA for E-commerce Customer Feedback Optimisation…………………………………………………………………………………………
Word Count:	……………870……………… **Page Count**…4………………………………………………………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …………… Jenat Jiffna Rodrigue……………………………………………………

Date: ……………28/01/2026…………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual: Dynamic ABSA for E-commerce Customer Feedback Optimisation

Jenat Jiffna Rodrigues
x23337451

1 Project Overview

This project performs a comprehensive Aspect-Based Sentiment Analysis on product review data. It involves:

- **Data Preprocessing**: Cleaning and preparing raw text data.
- **Embedding Generation**: Converting text into numerical representations using DistilBERT.
- **Dimensionality Reduction**: Reducing the complexity of embeddings using PCA and UMAP.
- **Clustering**: Grouping similar reviews into aspects using MiniBatchKMeans.
- **Aspect Labeling**: Assigning meaningful names to clusters using KeyBERT and spaCy.
- **Sentiment Classification**: Training a DistilBERT model to classify review sentiment (positive, neutral, negative).
- **Business Cause Identification**: Pinpointing specific issues or praise within reviews using rule-based pattern matching.
- **Visualization**: Presenting sentiment distributions and business causes per aspect.

2 Prerequisites

To run this notebook, you will need:

- **Google Colab Environment**: The notebook is designed to run in Google Colab.
- **GPU Access**: A GPU runtime is highly recommended for faster embedding generation and model training (Runtime > Change runtime type > GPU).
- **Google Drive Access**: The project assumes the dataset is located in your Google Drive, mounted at /content/drive.

3 Setup Instructions

a. Mount Google Drive:

- Ensure your Google Drive is mounted to access the dataset. The relevant code is in the cell that executes drive.mount('/content/drive').

b. Install Libraries:

- All necessary Python libraries (keybert, wordsegment, sentence-transformers, spacy, etc.) are installed directly within the notebook using !pip install commands. These will run automatically when their respective cells are executed.

c. Download SpaCy Model:

- The English spaCy model (en_core_web_sm) is downloaded using !python -m spacy download en_core_web_sm.

4 Data Configuration

- **Dataset Path**: The project uses a CSV file located at /content/drive/MyDrive/Colab Notebooks/jenat /1429_1.csv. **To use a different dataset, update the path in the pd.read_csv() cell.**
- **Relevant Columns (keep_cols)**: The analysis focuses on reviews.text, reviews.title, reviews.rating, reviews.doRecommend, brand, categories, and asins. If your dataset has different column names for these types of information, you will need to adjust the keep_cols list accordingly.

5 Key Adjustable Parameters

Several parameters throughout the notebook can be adjusted to fine-tune the analysis:

a. Embedding Generation (Phase 1):

- batch_size (in get_embeddings_in_batches function): Controls the number of texts processed simultaneously during embedding generation. Default: 16. Adjust based on GPU memory.
- max_length (in tokenizer for embeddings): Maximum sequence length for DistilBERT. Default: 128. Longer sequences capture more context but require more memory.

b. Clustering and Dimensionality Reduction (Phase 2):

- PCA(n_components): Number of principal components to retain. Default: 30.
- UMAP(n_neighbors, n_components, n_epochs):
    - n_neighbors: Controls local vs. global structure preservation. Default: 15.
    - n_components: Final dimensionality after UMAP. Default: 25.
    - n_epochs: Number of iterations for UMAP optimization. Default: 400.
- K_values (for Silhouette Score evaluation): Range of cluster numbers to test for MiniBatchKMeans. Default: range(10, 35, 2).
- MiniBatchKMeans(batch_size, n_init, max_iter):
    - batch_size: Batch size for k-means. Default: 1024.
    - n_init: Number of times the k-means algorithm is run with different centroid seeds. Default: 30.
    - max_iter: Maximum number of iterations for a single run. Default: 1000.

c. Aspect Labeling (Phase 2):

- SAMPLE_PER_CLUSTER: Number of representative texts sampled from each cluster for keyword extraction. Default: 200.
- KEYBERT_TOP_N: Number of keywords to extract per sample using KeyBERT. Default: 10.
- MMR_DIVERSITY: Diversity parameter for KeyBERT's Max Marginal Relevance (MMR). Default: 0.6.
- w_tf, w_cent, w_spec: Weights for the combined scoring mechanism (TF-IDF, semantic centrality, corpus specificity) for aspect label selection. Defaults: 0.45, 0.40, 0.15 respectively.

d. Business Cause Identification (Phase 2):

- BUSINESS_RULES: This dictionary contains regular expression patterns used to identify specific issues. You can **add, modify, or remove entries** to tailor the cause detection to your specific needs.

- TOP_N_CAUSES: Number of top business causes to display in visualizations. Default: 5.

e. Sentiment Classification (Phase 3):

- test_size (for train/validation split): Proportion of data used for validation. Default: 0.15.
- num_train_epochs: Number of training epochs for the DistilBERT sentiment model. Default: 3.
- learning_rate: Learning rate for the sentiment model. Default: 2e-5.
- per_device_train_batch_size, per_device_eval_batch_size: Batch sizes for training and evaluation. Defaults: 16 and 32 respectively.

6 Workflow Phases

The project is structured into three main phases:

- **Phase 1 (25%)**: Data Loading, Initial Exploration, Preprocessing, and DistilBERT Embedding Generation.
- **Phase 2 (50%)**: Dimensionality Reduction, Clustering, Dynamic Aspect Labeling, and Business Cause Identification.
- **Phase 3 (50%-100%)**: Sentiment Model Training, Evaluation, Prediction, and Aspect-wise Sentiment Visualization.

Each phase builds upon the previous one, ensuring a logical progression through the analysis. It is recommended to execute cells sequentially.

7 Expected Outputs/Results

The successful execution of this notebook will produce a variety of analytical outputs and models:

1. **Transformed Data (df DataFrame)**: A cleaned and enriched DataFrame containing original review data, augmented with features like clean_text, brand_clean, combined_text, sentiment_label, embedding, aspect_cluster IDs, aspect_label, and predicted_sentiment.
2. **Embeddings (embeddings NumPy array)**: High-dimensional (768D) and reduced-dimensional (30D PCA, 25D UMAP) numerical representations of the reviews.
3. **Clustering Results (aspect_cluster column)**: Cluster assignments for each review, grouping semantically similar texts into 12 distinct aspects.
4. **Aspect Labels (aspect_label column)**: Descriptive, human-interpretable labels (e.g., 'computers', 'wireless speakers') assigned to each cluster.
5. **Fine-tuned Sentiment Classification Model**: A trained DistilBertForSequenceClassification model, saved to disk (./phase3_best_model), capable of predicting 'negative', 'neutral', or 'positive' sentiment.
6. **Sentiment Predictions (predicted_sentiment column)**: Granular sentiment predictions for every review, as determined by the fine-tuned model.
7. **Aggregated Sentiment Summaries (aspect_summary DataFrame)**: Tabular summaries detailing the distribution of predicted sentiments and the positive_ratio for each identified aspect.
8. **Business Cause Aggregations (bc_agg, aspect_tables DataFrames)**: Detailed tables showing the counts of specific business causes within each aspect, sentiment, and brand category.

9. **Data Visualizations (Matplotlib & Seaborn plots)**: Numerous plots illustrating: the overall distribution of aspects, their positive sentiment ratios, and the top business causes for negative, neutral, and positive sentiments within various key aspects.

References

Ahmad, W., Khan, H. U., Alarfaj, F. K., & Alreshoodi, M. (2025). *Aspect-based sentiment analysis: A comprehensive review and open research challenges*. IEEE Access, 13, 65138–65182. Available: https://ieeexplore.ieee.org/iel8/6287639/10820123/10945359.pdf

Alhadlaq, A., Amin, T., Ahmed, J., Alenazi, A., Alshahrani, A., & Al-Khateeb, H. (2024). *A new hybrid deep learning approach for aspect-based sentiment analysis*. PeerJ Computer Science, 10, e1877. Available: https://peerj.com/articles/cs-2267/

Giannakopoulos, T., Musat, C., Hossmann, A., & Baeriswyl, M. (2017). *Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets*. WASSA/EMNLP, 268–275. Available: https://arxiv.org/abs/1709.05094

He, R., Sun Lee, W., Ng, H. T., & Dahlmeier, D. (2017). *An unsupervised neural attention model for aspect extraction.* ACL, 388–397. Available: https://www.comp.nus.edu.sg/~leews/publications/acl17.pdf

He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. (2019). *An interactive multi-task learning network for end-to-end aspect-based sentiment analysis*. arXiv:1906.06906. Available: https://arxiv.org/abs/1906.06906

Henning, S., & Lenhardt, L. (2022). *Class imbalance in natural language processing: A comprehensive survey*. arXiv:2210.04675. Available: https://arxiv.org/pdf/2210.04675.pdf

Hozumi, Y., Kawashima, T., Sugimoto, M., Nakazato, T., & Murakami, T. (2021*). UMAP-assisted K-means clustering of large-scale biological sequence datasets with silhouette score validation*. Scientific Reports, 11, 4361. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7897976/

Islam, A., Hossen, M. R., Ahmed, A., & Haque, B. M. T. (2025*). BanglaASTE: A novel framework for aspect-sentiment-opinion extraction in Bangla e-commerce reviews*. arXiv:2511.21381. Available: https://arxiv.org/html/2511.21381v1

Jayarao, V., Akram, M., & Agrawal, S. (2021*). Domain-specific contextual representations using DistilBERT for retail applications*. arXiv:2103.15737. Available: https://arxiv.org/pdf/2103.15737.pdf

Joshi, R. (2025). *TASCI: Transformers for aspect-based sentiment analysis with contextualized interactions.* Scientific Reports, 15(1), 12345. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12190667/

Kumar, A., Singh, A., Sharma, P., & Krishnamurthy, N. (2025*). Fine-tuning transformers for domain-specific sentiment analysis in e-commerce*. IJFMR, 1–15. Available: https://www.ijfmr.com/papers/2025/3/42227.pdf

Lawan, I., Sanda, I., et al. (2025*). Enhancing long-range dependency with state space models in aspect-based sentiment analysis.* COLING, 2173–2190. Available: https://aclanthology.org/2025.coling-main.148.pdf

Leung, C. (2023). *User-centred evaluation of keyword extraction algorithms for digital libraries*. arXiv:2504.21667. Available: https://arxiv.org/html/2504.21667v1

Li, X., Wang, X., Zhou, Y., & Liu, J. (2025). *Graph-enhanced implicit aspect-level sentiment analysis based on multi-prompt fusion*. Scientific Reports, 15, 10952. Available: https://www.nature.com/articles/s41598-025-02609-4

Luo, L., Ji, X., Zhang, M., Nie, Y., & Jiang, Z. (2019). *Unsupervised neural aspect extraction with sememes.* IJCAI, 5123–5129. Available: https://www.ijcai.org/proceedings/2019/0712.pdf

Maragheh, R. Y., Yaldaei, A., & Eftekhari, M. (2023). LLM-TAKE*: Theme-aware keyword extraction using large language models.* arXiv:2312.00909. Available: https://arxiv.org/pdf/2312.00909.pdf

Negi, G., Sarkar, R., Zayed, O., & Buitelaar, P. (2024). *A hybrid approach to aspect-based sentiment analysis using transfer learning*. arXiv:2403.17254. Available: https://arxiv.org/abs/2403.17254

Ngeyen, Y., & Yinkfu, N. (2025). *Improving QA efficiency with DistilBERT: Fine-tuning and inference on mobile Intel CPUs*. arXiv:2505.22937. Available: https://arxiv.org/abs/2505.22937

Putatunda, S. (2023). *A BERT based ensemble approach for sentiment classification of customer reviews*. arXiv:2311.10782. Available: https://arxiv.org/pdf/2311.10782.pdf

Qin, L., Zhang, X., & Wang, M. (2024*). Absolute evaluation measures for machine learning with imbalanced multiclass data*. arXiv:2507.03392. Available: https://arxiv.org/html/2507.03392v1

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv:1908.10084. Available: https://arxiv.org/abs/1908.10084

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.* arXiv:1910.01108. Available: https://arxiv.org/abs/1910.01108

Saxena, N., Pattanaik, P., & Verma, R. (2024*). Beyond architectures: Evaluating the role of contextual embeddings in transformer-based sentiment analysis.* arXiv:2507.14231. Available: https://arxiv.org/html/2507.1423

Scaria, K., Gupta, H., Goyal, S., Sawant, S. A., Mishra, S., & Baral, C. (2023). *InstructABSA: Instruction learning for aspect based sentiment analysis*. arXiv:2302.08624. Available: https://arxiv.org/abs/2302.08624

Schulz, M. (2022). *User-centred evaluation of KeyBERT and other keyword extraction techniques*. arXiv:2504.21667. Available: https://arxiv.org/html/2504.21667v1

Tan, B., & Shi, Y. (2024*). A rapid review of clustering algorithms*. arXiv:2401.07389. Available: https://arxiv.org/html/2401.07389v1

Tulkens, S., & van Cranenburgh, A. (2020). *Embarrassingly simple unsupervised aspect extraction*. ACL, 7132–7142. Available: https://arxiv.org/abs/2004.13580

Vysala, A., & Sinha, A. (2020). *Evaluating and validating cluster results*. arXiv:2007.08034. Available: https://arxiv.org/pdf/2007.08034.pdf

Wu, Q., Gajula, Y., et al. (2025). *AI-driven sentiment analytics: Unlocking business value from customer feedback in e-commerce.* arXiv:2504.08738. Available: https://arxiv.org/pdf/2504.08738.pdf

Zhang, L. (2022). *Scalable, stable, and interpretable text clustering via large language models*. arXiv:2204.09874. Available: https://arxiv.org/pdf/2204.09874.pdf

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). *A survey on aspect-based sentiment analysis: Tasks, methods, and challenges*. arXiv:2203.01054. Available: https://arxiv.org/abs/2203.01054

Zhao, Y., Zhang, X., Zhu, Q., & Yan, R. (2022). *A multi-task dual-tree network for aspect sentiment triplet extraction*. COLING, 6994–7005. Available: https://aclanthology.org/2022.coling-1.616.pdf