



A FEEDBACK SYSTEM FOR E-COMMERCE BUSINESSES FROM CUSTOMER REVIEWS USING ASPECT-BASED SENTIMENT ANALYSIS

by

Durga Gurajala

Applied Research Project submitted in partial fulfilment of the requirements for
the degree of

MSc. in Data Analytics

at

Dublin Business School

Supervisor:

Salah Aberkane

May, 2024

DECLARATION

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of MSc in Data Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Name: Durga Gurajala

Student Number: 10612928

Date: 20-05-2024

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Salah Aberkane, for the continuous support of my research, for his patience, motivation, and immense knowledge. His guidance helped me during the time of research and writing of this thesis.

I would also like to thank my professors and fellow students for their or their invaluable tutoring and guidance, which greatly assisted me in the completion of this thesis. Their expertise and support were instrumental in shaping my research and ensuring its success.

I am also grateful to Dublin Business School for providing the necessary facilities and resources to conduct my research.

I would like to thank my family and friends for supporting me spiritually throughout writing this thesis and my life in general. Their love, patience, and unwavering support have been my greatest source of strength and inspiration.

Name: Durga Gurajala

Student Number: 10612928

Date: 20-05-2024

ABSTRACT

In the contemporary business landscape, understanding customer sentiment is paramount for the success of new enterprises. This research project aims to develop a sophisticated feedback system utilizing aspect-based sentiment analysis (ABSA) of customer reviews. By dissecting customer feedback into specific aspects and analysing the sentiment associated with each, businesses can gain granular insights into customer experiences. This system will empower new businesses to make data-driven decisions, improve products and services, and ultimately enhance customer satisfaction and loyalty. The study used the customer reviews extracted from Amazon for several products from different sub-categories of women's footwear which were rigorously pre-processed. The aspects were identified using Term Frequency Inverse Document Frequency. The reviews were labelled with sentiments for the aspects using OPE GPT-3.5 turbo. The pre-trained models of Microsoft's DeBertaV3 and Google's Flan-T5 were used to evaluate their performance against GPT 3.5. Both the models performed moderately.

Table of Contents

INTRODUCTION	6
Overview:	6
Literature Review:	6
Rationale:	9
Hypothesis:	9
RESEARCH PLAN, METHODOLOGY AND DESIGN:	10
Research Plan:	10
Research Design:	12
Research Process:	14
IMPLEMENTATION AND ANALYSIS:	25
Exploratory Data Analysis:	25
Data Pre-Processing:	26
Aspect- Extraction:	27
Annotation of the data:	27
Modelling and Evaluation:	28
CONCLUSIONS AND FUTURE WORK	33
Key Findings:	33
Future Work:	34
Conclusion:	35
REFERENCES	37
GLOSSARY	38

INTRODUCTION

Overview:

The growth of online platforms has greatly changed how consumers interact with businesses, especially in the digital age. In Ireland, this change is evident in the increasing number of customer reviews on various e-commerce and service platforms. According to a 2023 report by Statista, the e-commerce market in Ireland is expected to reach €3.5 billion by the end of the year, with an annual growth rate of 11.2% projected through 2025. This increase in online shopping has led to a significant rise in customer feedback, mostly in the form of online reviews.

For new businesses, extracting useful insights from these reviews can be challenging because the data is unstructured. Traditional sentiment analysis methods provide a general view of customer sentiment but lack the granularity needed to address specific customer concerns. This is where aspect-based sentiment analysis (ABSA) becomes valuable. ABSA breaks down customer reviews into specific components or aspects, such as product quality, customer service, and delivery experience, and evaluates the sentiment associated with each aspect. This method offers a more detailed understanding of customer feedback, helping businesses identify precise areas of strength and weakness. For new businesses in Ireland, which face the challenges of establishing a market presence and building a loyal customer base, ABSA can be highly beneficial.

The importance of effective feedback systems is highlighted by data from the Central Statistics Office (CSO) of Ireland, which shows that about 70% of new businesses fail within their first three years. One major reason for this high failure rate is the inability to adequately respond to customer needs and preferences. By using ABSA, new businesses can gain a competitive edge, ensuring they are not only aware of customer sentiments but also capable of acting on them in a timely and focused manner.

This research project aims to develop a sophisticated feedback system for new businesses using ABSA. The system will analyse customer reviews, extract relevant aspects, and classify the sentiment associated with each aspect. The goal is to provide business owners with a detailed, actionable understanding of customer feedback, enabling them to make informed decisions.

Literature Review:

The study from the paper titled “*A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis*” highlights the limitations of existing ABSA datasets containing only one aspect or multiple aspects with similar sentiment polarity which reduces the complexity of the ABSA task, making it almost identical to sentence-level sentiment analysis. The need for more sophisticated datasets that can challenge and improve ABSA models was made evident in the analysis. The analysis uses Multi-Aspect Multi-Sentiment (MAMS) dataset

that addresses this limitation by providing sentences with at least two different aspects that have different sentiment polarities pushing researchers to develop more nuanced and capable ABSA models. The MAMS dataset's complexity mirrors the real-world scenarios where customers often express varied sentiments towards different aspects within the same review.

Two versions of the MAMS datasets are created for two subtasks – aspect term sentiment analysis (ATSA) and aspect category sentiment analysis (ACSA). The ATSA dataset has been filtered out for reviews with multiple aspects of different sentiments by manual annotation. The ACSA dataset has been filtered out for reviews with multiple pre-defined categories that are manually annotated.

In their study, the authors introduce two models: CapsNet and CapsNet-BERT. These models are designed to leverage recent advancements in natural language processing (NLP) to improve ABSA performance.

- *CapsNet (Capsule Network)*: This model is noted for its ability to capture hierarchical relationships in data, which is beneficial for understanding the nuances in sentiment analysis.
- *CapsNet-BERT*: This model combines the strengths of CapsNet with BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art language representation model known for its deep contextual understanding of text.

The study's experiments conducted on the MAMS dataset demonstrate that both CapsNet and CapsNet-BERT significantly outperformed the existing baseline methods. The baseline methods adopted for the analysis were variants of LSTM and attention-based models. These results underscored the effectiveness of the proposed models in handling the complexities introduced by the MAMS dataset.

The paper titled “*Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis*” discusses the common challenges as below that hinder the effectiveness of ABSA:

- *Domain-Specificity*: ABSA models often struggle to generalize across different domains. A model trained on restaurant reviews, for example, may not perform well on hotel reviews due to variations in language and context.
- *Reliance on Labeled Data*: The performance of ABSA models heavily depends on the availability of labeled data, which can be costly and time-consuming to obtain.
- *Underutilization of Newer Large Language Models (LLMs)*: There is a lack of exploration into the potential of newer LLMs, such as GPT, PaLM, and T5, which have shown promising results in various NLP tasks.

To address these challenges, researchers have evaluated the performance of several prominent models on diverse datasets, including DOTSA, MAMS, and SemEval16. The

models evaluated include ATAE-LSTM, flan-t5-large-absa, Deberta, PaLM, and GPT-3.5-Turbo.

The findings from these evaluations reveal the strengths and weaknesses of the models across different domains:

- *Deberta*: Emerges as a consistently high-performing model, demonstrating robust performance in both aspect term sentiment analysis (ATSA) and aspect category sentiment analysis (ACSA) tasks.
- *PaLM*: Shows remarkable competitiveness, particularly for ATSA tasks, and performs well across various domains including restaurant, hotel, books, clothing, and movie reviews.

Understanding the strengths and weaknesses of different ABSA models across various domains can inform the selection of appropriate models for specific business needs. Additionally, exploring the potential of newer LLMs like GPT-3.5-Turbo and PaLM can lead to significant advancements in ABSA, providing businesses with more accurate and detailed insights from customer reviews.

The paper titled “*Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT*” showcases the performance of pre-trained models of BERT and T5 and their ability to capture contextual information and semantic nuances in aspect-based sentiment analysis. This study helps to address the challenges in extracting actionable information from unstructured reviews.

- *BERT (Bidirectional Encoder Representation from Transformers)*: A model that captures bidirectional context in text, making it highly effective for understanding the intricacies of language and sentiment expressed in customer reviews.
- *T5 (Text-to-Text Transfer Transformer)*: A versatile model that treats every NLP task as a text-to-text problem, offering flexibility and robustness in handling different types of text data.

The study utilized both synthetically generated and manually labelled datasets to train BERT and T5 models. The synthetic dataset is generated by prompt engineering using GPT 3.5. These datasets were designed to capture the specific features of review data related to eco-friendly products. The models were trained with synthetic data and performance was tested with manual analysis data. The training process involved fine-tuning the models for aspect detection and sentiment classification tasks, achieving high accuracy rates of 92% for BERT and 91% for T5.

The performance of the models was assessed using standard evaluation metrics such as precision, recall, F1-score, and computational efficiency. The BERT model outperformed T5 in these evaluations, demonstrating superior capability in detecting aspects and classifying

sentiments in customer reviews. This led to the selection of BERT as the classifier for the prediction pipeline.

Rationale:

The key insights from the literature survey that define our research's rationale are:

The traditional sentiment analysis methods perform sentiment classification at the document or sentence level which fails to capture opinions about specific aspects of products or services. This limitation restricts the ability of businesses to gain detailed insights necessary for targeted improvements. In the dynamic landscape of product design, particularly for new businesses, understanding detailed customer sentiments and preferences is critical. Aspect-based sentiment analysis (ABSA) addresses this need by linking sentiments to specific aspects mentioned in reviews, providing more actionable insights for product designers and business analysts.

High performance in ABSA models often depends on extensive labelled datasets, which are costly and time-consuming to obtain. There is a lack of comprehensive exploration into the potential of newer large language models (LLMs) such as GPT, PaLM, and T5, which have demonstrated significant advancements in various NLP tasks. Many ABSA models struggle to generalize across different domains, resulting in reduced accuracy when applied to varying types of reviews.

Pre-trained models like DeBertaV3, Flan-T5 and GPT 3.5 have shown substantial improvements in various NLP tasks due to their ability to capture complex contextual information and semantic nuances. These models have demonstrated high accuracy in detecting aspects and classifying sentiments, making them suitable candidates for advancing ABSA research.

The rationale for this research is grounded in the identified limitations of traditional sentiment analysis, leveraging the advanced capabilities of pre-trained models like DeBertaV3, Flan-T5 and GPT 3.5, and addressing domain-specific issues, this research aims to develop a sophisticated feedback system that provides detailed, actionable insights from customer reviews.

Hypothesis:

Implementing an aspect-based sentiment prediction pipeline using the pretrained DeBertaV3, Flan-T5 and GPT 3.5 to develop a feedback system that helps new businesses improve their product quality and customer satisfaction and evaluating the best models to use with unlabelled data.

RESEARCH PLAN, METHODOLOGY AND DESIGN:

Research Plan:

The research plan is formulated by adopting the CRISP – DM process. The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a structured approach to data mining projects. The following tasks are included in the research plan for developing an aspect-based sentiment analysis (ABSA) system for customer reviews using the CRISP-DM process:

1. *Business Understanding:*

Objective:

A clear understanding of the business problem must be obtained, and objectives have to be formulated to ensure the research aligns with the needs of new businesses. The primary goal must be defined from a business perspective and translate them into data mining goals. A detailed plan of the project must be developed including the timeline, resources and tasks.

Task:

The main business goal is to develop an efficient system that provides aspect level sentiments for new businesses. The idea is to develop a model that can efficiently analyse sentiments of aspects and provide feedback to the ecommerce businesses to improve customer satisfaction and product quality. The key stakeholders would be the product designers, business analysts and marketing teams.

2. *Data Understanding:*

Objective:

Data has needs to be gathered from relevant sources and understand the structure and characteristics of the data by conducting exploratory data analysis. The quality of the data must be assessed, and any issues need to be addressed.

Tasks:

Customer reviews have to be gathered from relevant e-commerce platforms like Amazon, Flipkart or E-bay using web scraping. The gathered data needs to be analysed for the data structure, content and distribution. Issues such as missing values, duplicates and any inconsistencies have to be identified.

3. *Data Preparation:*

Objective:

The collected data has to be prepared for analysis and modelling. Any irrelevant data needs to be ignored from analysis. Issues identified during the data understanding have to be addressed.

Tasks:

The duplicates have to be removed from the data and the missing values have to be handled. The text data has to be normalized and emojis and irrelevant special characters have to be removed. Stemming or Lemmatization has to be performed. Data from multiple platforms is integrated into one dataset for analyse. The relevant features have to be identified for the analysis. Aspects have to be identified for implicit extraction.

4. Modelling:

Objective:

Appropriate models for data analysis have to be chosen and assessed for the data available. Test the model for efficiency and fine-tune if necessary to fit for the data.

Tasks:

NLP Models and algorithms have to be identified for the aspect extraction and sentiment analysis. Transformer and LLM models like Bert, GPT, T5 etc, that are pre-trained on similar datasets can be considered for modelling. These models have to be trained on the prepared data and fine-tune, if necessary, with appropriate parameters. Evaluation metrics such as accuracy, precision and recall and F1-score to assess model performance.

5. Evaluation:

Objective:

The models need to be evaluated to ensure they meet business objectives and assess if they achieve the data mining goals and are ready for deployment. If not, the model has to be iterated again until its ready for deployment.

Tasks:

The models' performance has evaluated and compared to determine the best performing model. Assess if the model provides the actionable insights for feedback. The precision, recall, F1-Score and accuracy are considered to ensure the model is performing well.

This structured research plan ensures a thorough and systematic approach to developing an advanced aspect-based sentiment analysis system, aligning with business objectives and leveraging state-of-the-art models.

Research Design:

1. Overview:

This study utilizes a mixed design approach, combining both correlational and quasi-experimental methods to analyse the relationship between the identified aspects of customer reviews and the corresponding sentiment classifications. The goal is to develop and validate an aspect-based sentiment analysis (ABSA) system for customer reviews, particularly focusing on the footwear category from Amazon.

2. Independent/Predictor Variables:

- **Aspect Terms:** The specific features or attributes of the footwear products identified from the reviews (e.g., comfort, fit, style).
- **Sentiment Labels:** The sentiment classification (positive, negative, neutral) assigned to each aspect term using GPT-3.5 Turbo.

3. Dependent/Criterion Variables:

- **Sentiment Analysis Accuracy:** The accuracy of the sentiment classification for each aspect.
- **Model Performance Metrics:** Evaluation metrics for the NLP models (DeBERTaV3 and Flan-T5), including precision, recall, F1-score, and accuracy.

4. Experimental and Control Groups:

- **Experimental Group:** The dataset of customer reviews processed and analysed using advanced NLP models (DeBERTaV3 and Flan-T5).
- **Control Group:** Not explicitly used in this design.

5. Participant Assignment to Groups:

"Participants" refer to the data points (customer reviews). The reviews are naturally divided based on the brands and subcategories of footwear. Reviews are split into training and test datasets for model evaluation.

6. Between-Subjects and Within-Subjects Variables:

- **Between-Subjects Variables:**
Brand and Subcategory: The different brands and subcategories of footwear. Each review belongs to a specific brand and subcategory.

- *Within-Subjects Variables:*
Aspect Terms and Sentiments: Each review can contain multiple aspects with associated sentiments, analyzed within the context of the review.

7. Actual Design

Correlational Design:

Objective: To explore the relationship between aspect terms identified in the reviews and the sentiment labels assigned.

Predictor Variables: Aspect terms and sentiment labels.

Criterion Variables: Sentiment analysis accuracy and model performance metrics.

Quasi-Experimental Design:

- *Objective:* To evaluate the effectiveness of the ABSA models (DeBERTaV3 and Flan-T5) in accurately classifying sentiments for each aspect.
- *Independent Variables:* Aspect terms and sentiment labels.
- *Dependent Variables:* Model performance metrics such as accuracy, precision, recall, F1-score, and computational efficiency.

Detailed Methodology

1. Data Collection:

- Tool: Review extraction tool by Symanto.
- Sources: Amazon.co.uk and Amazon.com.
- Focus: Six brands and four subcategories of footwear.

2. Data Preprocessing

Steps:

- Remove duplicates, null reviews, URLs, and emojis.
- Normalize text using NFKD and apply lemmatization.

3. Aspect Extraction

Techniques:

- TF-IDF: Identify high-frequency terms.
- LDA: Validate identified aspects through topic modelling.

4. Aspect Labelling

- Tool: GPT-3.5 Turbo.
- Process: Develop prompts to classify the sentiment of each identified aspect as positive, negative, or neutral.

5. Modelling and Evaluation

- Models: DeBERTaV3 and Flan-T5.
- Evaluation: Use accuracy, precision, recall, F1-score, and computational efficiency to assess model performance.

Research Process:

The below flow diagram summarizes the entire process of the research.

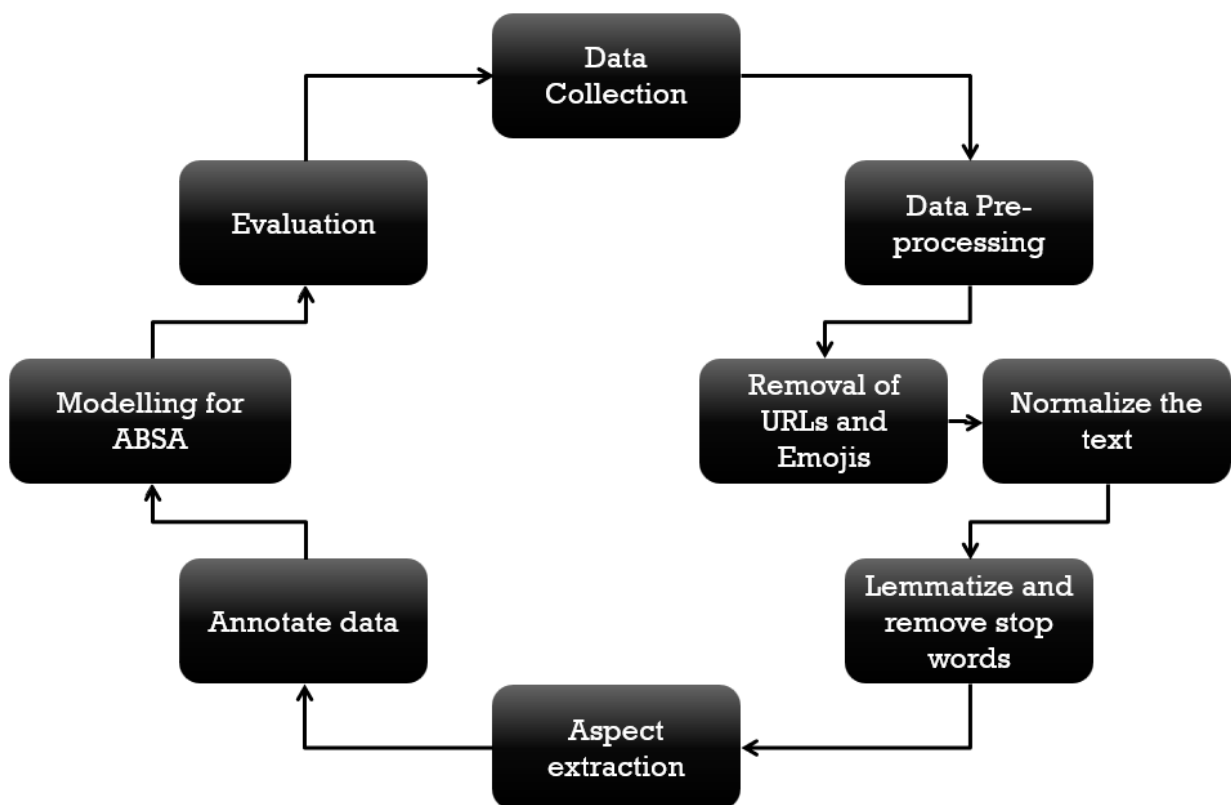


Fig 1: Flow diagram depicting the research process.

Data Collection:

The dataset for the research has been formulated by web scraping customer reviews from e-commerce websites. The reviews were extracted using the chrome plugin “Review

Extractor” developed by Symanto. The tool facilitates an automated process of extracting reviews from several e-commerce platforms and allows to save them in an excel sheet.

Since this research focuses on the analysis of products from a single category, the reviews extracted were that of Women’s footwear from Amazon UK and US platforms for six different brands under 4 sub-categories. The brands considered were:

```
df['Brand'].unique()
array(['clarks', 'crocs', 'dreampairs', 'drscholls', 'geox', 'skechers'],
      dtype=object)
```

1. Clarks
2. Crocs
3. Dream Pairs
4. Dr Scholls
5. Geox
6. Skechers

The 4 sub-categories are as shown below.

```
df['Sub-Category'].unique()
array(['Boots', 'Flats', 'Sandals', 'Sneakers'], dtype=object)
```

1. Boots
2. Flats
3. Sandals
4. Sneakers

For each sub-category of a brand, 100 reviews of each for 4 different products were extracted from the platforms. Each brand has 1600 reviews for 16 products. This accounted for a total of 9600 reviews extracted for the analysis. The data extracted contained the following attributes.

```
[5] df.head()
```

	postId	pageName	text	title	subtitle	stars	source	date	latitude	longitude	Brand	Sub-Category
0	R3K2FD3TRN61W	Clarks Women's Orinoco Club Short Shaft Boots	Nice boots but should state they are a very na...	3.0 out of 5 stars	Very narrow	NaN	3 Amazon Review	2024-02-12T00:00:00.000Z	NaN	NaN	clarks	Boots
1	R1DDKC5KOS9RGN	Clarks Women's Orinoco Club Short Shaft Boots	The boots came quicker than suggested. Slightl...	4.0 out of 5 stars	Comfy boots	NaN	4 Amazon Review	2024-01-26T00:00:00.000Z	NaN	NaN	clarks	Boots
2	R2LMJ8SR6YA9RL	Clarks Women's Orinoco Club Short Shaft Boots	The boots look really nice but unless you have...	3.0 out of 5 stars	Ankle boots	NaN	3 Amazon Review	2024-01-02T00:00:00.000Z	NaN	NaN	clarks	Boots
3	R2VASGJHF147EV	Clarks Women's Orinoco Club Short Shaft Boots	I had a pair of these boots before. Wanted exa...	5.0 out of 5 stars	Comfy boots	NaN	5 Amazon Review	2023-12-28T00:00:00.000Z	NaN	NaN	clarks	Boots
4	R3K165QI8VYLFE	Clarks Women's Orinoco Club Short Shaft Boots	These are great boots very stylish and very co...	5.0 out of 5 stars	Boots	NaN	5 Amazon Review	2023-12-27T00:00:00.000Z	NaN	NaN	clarks	Boots

postId: A unique alphanumeric ID assigned to each review.

pageName: refers to the product name.

text: input variable - is the text body of the review.

title: input variable - the title of the review.

subtitle: subtitle to the review

stars: rating given for the product out of 5 stars.

source: the platform from where the review was extracted.

date: date on which the review was posted.

latitude: latitude of the location from where the review was posted.

longitude: longitude of the location from where the review was posted.

Brand: Brand of the product

Sub-Category: Sub-category of the product.

Data Pre-processing:

The data is pre-processed to remove any duplicates or null values. Any URLs or emojis present in the text are also removed. The text is then normalized using the Normalization Form KD (compatibility decomposition) NFKD form to remove any inconsistencies in the text. The NFKD form converts any Unicode text into canonical and consistent format. This means that characters that may look different but are essentially equivalent in meaning or usage are treated as the same.

The text is then lemmatized and stop words are removed using the *Natural Language Processing Toolkit* library. Lemmatization has been chosen instead of stemming to retain the context which is very important for the sentiment analysis.

This completes the preprocessing of the text data and is now ready for aspect extraction.

Aspect Term Extraction:

Aspect Term Extraction is a crucial component of aspect-based sentiment analysis (ABSA). It involves identifying specific terms or phrases in a piece of text that refer to aspects or features of an entity being reviewed.

TF-IDF (Term Frequency-Inverse Document Frequency) is employed to identify high frequency terms to extract aspects. It is a numerical statistic that reflects the importance of a

word in a document relative to a collection of documents (corpus). It is widely used in information retrieval, text mining, and natural language processing (NLP) as a weighting factor. The components of TF-IDF are:

1. *Term Frequency (TF)*:

- Definition: Measures how frequently a term appears in a document. The assumption is that the more frequently a term appears in a document, the more important it is.
- Formula:

$$\text{TF}(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

2. *Inverse Document Frequency (IDF)*:

- Definition: Measures the importance of a term by considering how common or rare it is across all documents in the corpus. The assumption is that terms that appear in many documents are less informative than terms that appear in fewer documents.
- Formula:

$$\text{IDF}(t, D) = \log \left(\frac{\text{Total number of documents } N}{\text{Number of documents containing term } t} + 1 \right)$$

Here, N is the total number of documents, and the "+1" is added to avoid division by zero.

3. *TF-IDF Score*:

- Definition: The TF-IDF score combines the local importance (TF) and the global importance (IDF) of a term. A high TF-IDF score indicates that the term is significant in the document and not common across other documents.
- Formula:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Scikit Learn's TF-IDF vectorizer converts a collection of raw documents into a matrix of TF-IDF features. In this analysis only unigrams were considered for vectorization and the top frequency terms are then manually assessed and aspects are finalized. The synonyms and antonyms are also considered to cross check in the reviews for similar terms.

Latent Dirichlet Allocation (LDA) is a widely used topic modeling algorithm in natural language processing (NLP). Topic modeling is a method for identifying hidden topics within a collection of documents or texts. LDA is a probabilistic model that generates topics, each characterized by a distribution of words, for a given corpus of documents.

LDA seeks to uncover the underlying topics in the corpus and the corresponding proportions of each topic in each document. As an unsupervised learning technique, LDA does not require labelled data and is suitable for the dataset at hand as it is also unlabelled.

The corpus and dictionary are created from the pre-processed data. The LDA model is built using the corpus and dictionary. This model is used to get the top words for each topic. The coherence score is calculated to evaluate the degree of semantic similarity between the words in a topic. The coherence score provides the topic quality, degree of interpretability, and how well the model can capture the relevant topics.

Annotation of the data:

The review data extracted from Amazon is labeled using prompt engineering with GPT-3.5 Turbo for subsequent evaluation with other models. GPT, or Generative Pre-trained Transformer, is a large language model developed by OpenAI, capable of various applications including chatbots, question-answering systems, and text completion. GPT-3.5 Turbo excels in tasks that require a high level of natural language understanding and generation.

OpenAI offers an API for GPT-3.5 Turbo, enabling developers to integrate the model into their applications. This API supports multiple programming languages such as JavaScript, Python, C++, and C#, allowing developers to use the language they are most comfortable with for seamless integration.

The OpenAI API is fed with prompts using one-shot classification to extract the sentiments of the aspects. The prompt is included with a sample text and its output is provided followed by the text that needs to be analysed. OpenAI API charges by number of tokens to process the prompt and is relatively expensive when it comes to processing large datasets.

The labelled data is then used to model with transformer models.

Modelling:

The annotated data is then verified and processed for any inconsistencies or duplicates or null values again. This data is then modelled using DeBERTaV3 and Flan-T5 to evaluate their performance. The models used are from Hugging Face.

DeBERTaV3:

DeBERTaV3 (Decoding-enhanced BERT with disentangled attention) is an advanced variant of the BERT model developed by Microsoft. It aims to improve upon the original BERT

architecture by addressing some of its limitations and incorporating novel techniques to enhance performance on various natural language processing (NLP) tasks.

Key Features and Innovations

1. Disentangled Attention Mechanism:

- Traditional BERT uses a single attention mechanism that entangles content and position information. DeBERTaV3 separates these two, allowing the model to focus on the content of the words and their positional relationships independently. This disentanglement helps the model better understand the structure and semantics of the text.

2. Enhanced Positional Encoding:

- Instead of using absolute positional embeddings, DeBERTaV3 employs a more sophisticated relative positional encoding. This allows the model to capture positional information more effectively, particularly in longer sequences.

3. Masked Language Model (MLM) Training:

- Similar to BERT, DeBERTaV3 is trained using the MLM objective, where some tokens in the input are randomly masked, and the model is trained to predict these masked tokens. This encourages the model to develop a deep understanding of the language context.

4. Improved Training Techniques:

- DeBERTaV3 incorporates optimizations in the training process, such as better initialization, regularization techniques, and more efficient use of computational resources. These improvements contribute to faster convergence and higher performance.

5. Model Size and Variants:

- DeBERTaV3 is available in various sizes, like BERT (base, large, etc.), allowing for flexibility in deployment depending on the computational resources and specific use case requirements.

6. Benefits of Using DeBERTaV3:

- *Enhanced Performance:* The disentangled attention mechanism and improved positional encoding contribute to better performance on a variety of NLP tasks.
- *Better Generalization:* DeBERTaV3's ability to capture content and positional information separately allows it to generalize better, particularly in scenarios with longer text sequences.
- *Flexibility and Scalability:* Available in multiple sizes, DeBERTaV3 can be scaled up or down depending on the specific requirements and available computational resources.

- *State-of-the-Art Results:* DeBERTaV3 has shown to achieve state-of-the-art results on several benchmark datasets, making it a strong candidate for a wide range of NLP applications.

The model used in this research is pre-trained version of DeBERTaV3 – “*yangheng/deberta-v3-large-absa-v1.1*”. This model has been pre-trained on 30k+ ABSA samples and English datasets based on the FAST-LCF-BERT model with “*microsoft/deberta-v3-large*”. This model is used with “*autotokenizer*” and “*automodelsequenceclassification*” utilities from Hugging Face’s transformers library.

AutoTokenizer automatically load the appropriate tokenizer for a given pre-trained model. Tokenizers are responsible for converting raw text into a format that can be processed by the model (usually token IDs).

Key Features:

- *Automatic Configuration:* Automatically selects the correct tokenizer class for a given model.
- *Preprocessing:* Handles tokenization, padding, and truncation of input text.
- *Special Tokens:* Manages special tokens (e.g., [CLS], [SEP]) required by certain models.

AutoModelForSequenceClassification automatically loads a pre-trained model specifically designed for sequence classification tasks. Sequence classification tasks include sentiment analysis, spam detection, and more.

Key Features:

- *Automatic Configuration:* Automatically selects the correct model class for a given model.
- *Pre-trained Weights:* Loads pre-trained weights that can be fine-tuned on your specific dataset.
- *Versatility:* Supports a wide range of pre-trained models from BERT, RoBERTa, GPT-3, etc.

Flan-T5:

FLAN-T5 (Fine-Tuned Language Agnostic T5) is an advanced version of the T5 (Text-to-Text Transfer Transformer) model developed by Google. FLAN-T5 is specifically fine-tuned to improve performance across various natural language processing (NLP) tasks by leveraging instruction fine-tuning, which involves training on a wide array of tasks using human-readable instructions.

Key Features and Innovations:

- *Instruction Fine-Tuning*: FLAN-T5 is fine-tuned using task-specific instructions, which helps the model understand and perform a wide range of tasks more effectively. This method improves the model's ability to generalize across different types of tasks.
- *Text-to-Text Framework*: Like the original T5, FLAN-T5 uses a text-to-text framework where all NLP tasks are framed as text generation tasks. This simplifies the model architecture and makes it versatile for various applications.
- *Enhanced Generalization*: By training on a diverse set of tasks and instructions, FLAN-T5 demonstrates improved generalization capabilities, making it effective even on tasks it hasn't explicitly seen during training.
- *Multilingual Support*: FLAN-T5 is designed to work well across multiple languages, enhancing its utility for global applications.

Benefits of Using FLAN-T5:

- *Versatility*: The text-to-text framework allows FLAN-T5 to be applied to a wide range of tasks without needing task-specific architectures.
- *Improved Generalization*: Instruction fine-tuning enhances the model's ability to generalize across tasks, improving performance even on tasks not seen during training.
- *Multilingual Capabilities*: Supports multiple languages, making it useful for global applications.
- *Ease of Use*: The text-to-text approach simplifies the model usage, as the same model can be used for various tasks with minimal adjustments.

The model used in this research “*shorthillsai/flan-t5-large-absa*” is a pretrained version of “*google/flan-t5-base*” on custom dataset prepared by GPT 4 and can be useful for ABSA specific tasks. This model is also used with autotokenizer and automodelsequenceclassification utilities.

Evaluation:

When evaluating the performance of machine learning models, especially in classification tasks, several key metrics are commonly used: accuracy, precision, recall, and F1 score. These metrics help to understand different aspects of the model's performance, particularly how well it distinguishes between different classes.

Accuracy:

Accuracy measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances.

Formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

TP (True Positives): Correctly predicted positive instances.

TN (True Negatives): Correctly predicted negative instances.

FP (False Positives): Incorrectly predicted positive instances (Type I error).

FN (False Negatives): Incorrectly predicted negative instances (Type II error).

Pros: Simple to understand and compute, useful when the classes are balanced.

Cons: Can be misleading in imbalanced datasets, where the majority class dominates the prediction.

Precision:

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive.

Formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

High precision indicates a low false positive rate, meaning the model is reliable when it predicts a positive instance.

Pros: Useful when the cost of false positives is high.

Cons: Does not account for false negatives.

Recall (Sensitivity or True Positive Rate):

Recall measures the proportion of correctly predicted positive instances out of all actual positive instances.

Formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

High recall indicates a low false negative rate, meaning the model is good at identifying all positive instances.

Pros: Useful when the cost of false negatives is high.

Cons: Does not account for false positives.

F1 Score:

The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns.

Formula:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score considers both precision and recall, making it useful for evaluating models on imbalanced datasets where a balance between false positives and false negatives is needed.

Pros: Provides a single metric to evaluate both precision and recall.

Cons: Can be less interpretable than individual precision and recall scores.

Each performance metric offers unique insights into the model's behavior and suitability for specific tasks. Accuracy is suitable for balanced datasets, while precision is crucial when false positives are costly. Recall is vital when missing positive instances is critical, and the F1 score provides a balanced measure of precision and recall, making it useful for imbalanced datasets. Selecting the appropriate metric depends on the specific application and the cost associated with different types of classification errors.

Fine-Tune Bert:

BERT (Bidirectional Encoder Representations from Transformers) is a powerful pre-trained language model developed by Google. It has been trained on a large corpus of text data using unsupervised learning objectives, such as masked language modelling and next sentence prediction. While BERT's pre-trained model captures a vast amount of general language knowledge, fine-tuning it for specific tasks is crucial for several reasons:

1. Task-Specific Adaptation:

BERT's pre-trained model is generic and not tailored to any specific task. To leverage its capabilities for applications such as sentiment analysis, named entity recognition, or question answering, the model needs to be fine-tuned on task-specific data.

Fine-tuning adapts the model to understand and predict the nuances and requirements of the specific task, leading to improved performance and accuracy.

2. Improved Performance:

Directly using the pre-trained BERT model without fine-tuning often yields suboptimal results because the model is not specialized for the task at hand.

Fine-tuning significantly enhances performance metrics (e.g., accuracy, precision, recall, F1-score) as the model learns to apply its general language understanding to specific task-related patterns and data distributions.

3. Domain Adaptation:

Different domains (e.g., legal, medical, financial) have unique vocabularies, styles, and terminologies that are not adequately represented in BERT's pre-training data.

Fine-tuning on domain-specific data helps the model adapt to these specialized contexts, making it more effective in understanding and generating domain-specific text.

4. Handling Imbalanced Data:

Many real-world datasets are imbalanced, with certain classes being underrepresented. BERT's pre-training does not address these imbalances.

Fine-tuning with techniques such as data augmentation, resampling, or using weighted loss functions helps the model learn to handle imbalanced classes more effectively, leading to better generalization.

5. Efficiency and Practicality:

Necessity: Fine-tuning BERT on a smaller, task-specific dataset is computationally more efficient than training a model from scratch.

Importance: Fine-tuning requires fewer computational resources and time while still leveraging the extensive knowledge encoded in the pre-trained BERT model, making it practical for a wide range of applications.

IMPLEMENTATION AND ANALYSIS:

Exploratory Data Analysis:

The dataset consists of roughly 9600 reviews from the Amazon platforms for Women's footwear. The attributes of interest for the analysis would be the review text, title brand and stars to analyse brand-wise aspects, their sentiments and overall sentiments of the reviews.

The ratings are distributed unevenly across the dataset with most of the reviews being positive with a 5-star rating as shown below.

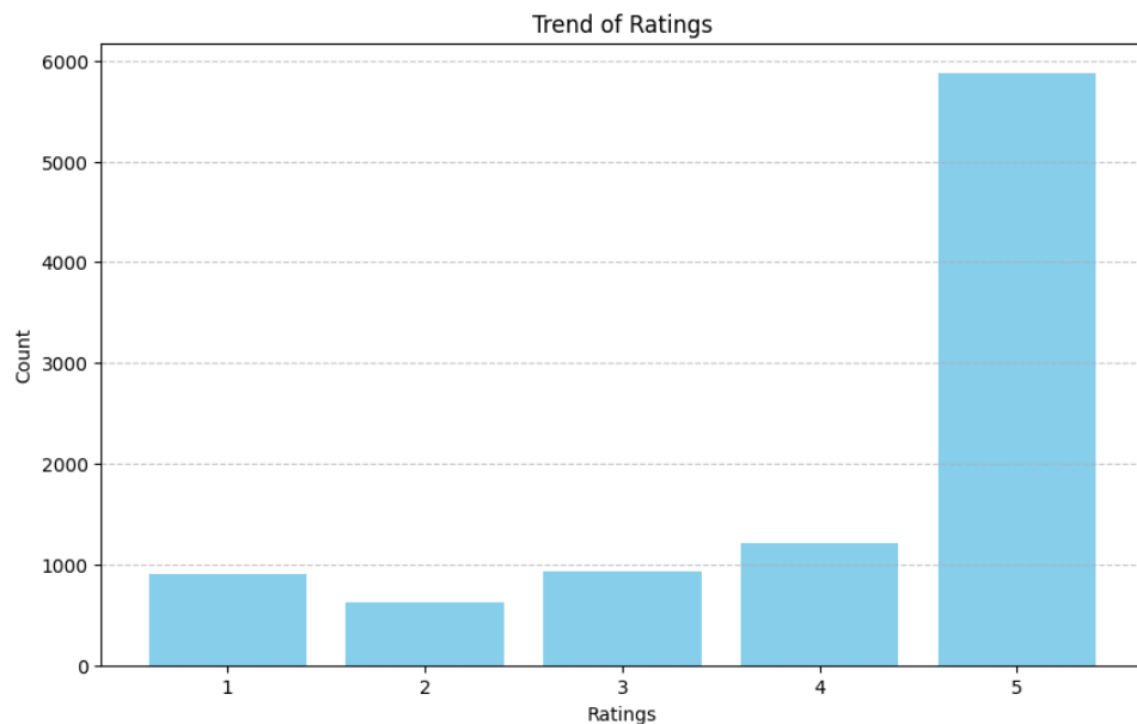


Fig: Bar plot showing the trend of ratings across the dataset

The reviews are distributed similarly even across the brands with most reviews being 5-star.

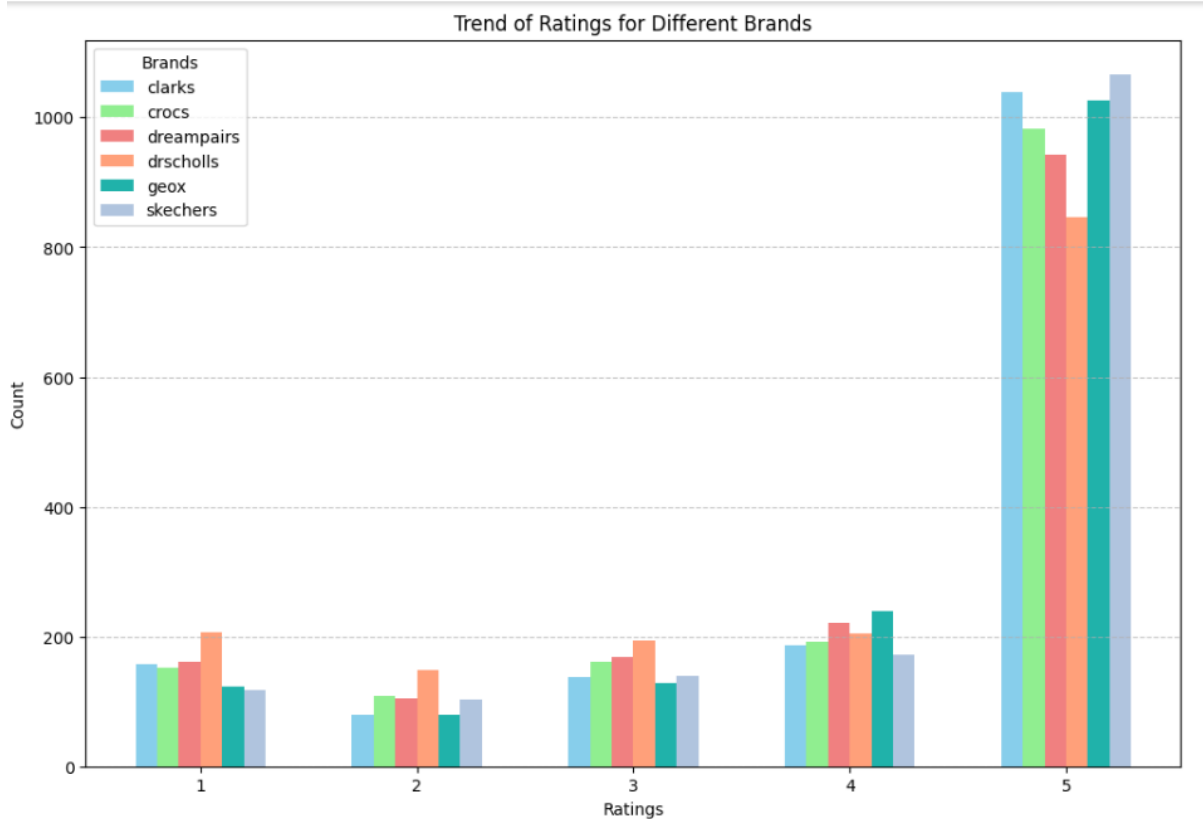


Fig: Bar plot showing the trends of ratings across the brands

This clearly indicates a class imbalance. Retaining this imbalance is crucial to accurately reflect real-world scenarios and prevent overfitting. Additionally, it helps assess the model's robustness. Moreover, it's important to identify relatively minor negative aspects, even for a highly reputable brand.

Data Pre-Processing:

The data had a couple of duplicate and null records and were removed during the pre-processing.

The attributes subtitle, latitude and longitude had no data on them and were irrelevant to the analysis. So, these were dropped from the dataset during the preparation.

The title column had relevant information to the review and could not be ignored. Hence it was added to the review text and analysed together.

The text was normalized with NFKD format to convert any Unicode text to canonical format. Any URLs in the review text do not provide any sentiment information and would become noise to the analysis. So, any URLs present have been removed from the text. Similarly, emojis and special characters would lead to misclassification and were removed from the text. This preprocess data has been exported to CSV for future use.

The text was lemmatized and stop words were removed before processing for aspect extraction.

Aspect- Extraction:

The TF-IDF vectorization was applied to the processed data to extract the high frequency terms to analyse for aspects. The top 50 TF-IDF terms were analysed, and the terms were narrowed down to ['comfort', 'fit', 'quality', 'style', 'color', 'size', 'price', 'look', 'color', 'support']

These terms were then brought down to their closest general aspects and the list was ['comfort', 'fit', 'quality', 'appearance', 'value'] and verified synonyms and antonyms against all the reviews to check if the relevant aspects popped up.

The text is further analysed with LDA for topic modelling to identify the words in each topic and compare with the above aspects. The coherence score was low at a 0.44 indicating the poor interpretability of the topics and that manual intervention is necessary to identify the terms in the topics. There were also a lot of short reviews in the dataset that could have affected the coherence score.

The LDA analysis further used TF-IDF weights with the topic modelling and there was no improvement in the coherence score which stood at .45.

Annotation of the data:

The dataset at hand is unlabelled which makes evaluating the models difficult. This created the necessity to label aspect sentiments using a reliable method. The records are 9600, which would consume a lot of man hours to manually annotate. The GPT 3.5 Turbo has been used as a baseline model to annotate the data with the sentiment labels for each of the aspects.

The GPT 3.5 Turbo used with OpenAI API was fed with a prompt that used a sample review and sample output for one-shot classification along with the text to be annotated. The annotated data was manually verified for a small batch of randomly selected 100 reviews and the annotation was 95% close to human annotation for all aspects.

The distribution of sentiments across the aspects is important to the feedback system. Below plot shows the sentiment distribution across the aspects for each brand.

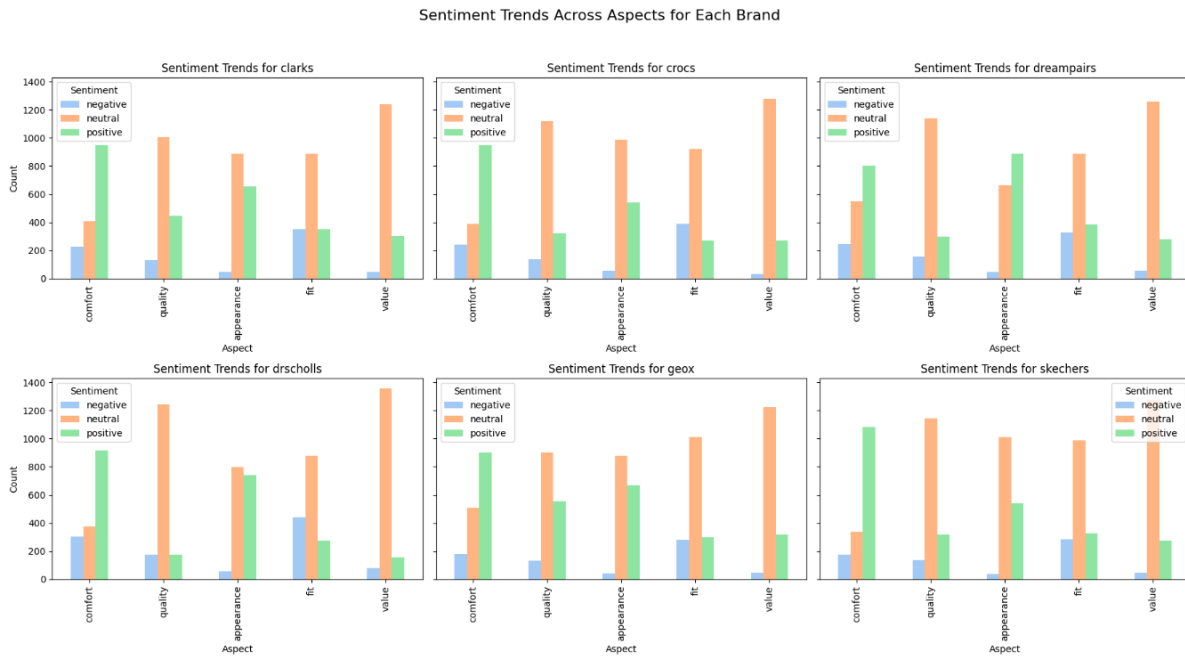


Fig: Distribution of aspects for each brand

This feedback enables brands to identify and enhance underperforming product aspects. The plots generally indicate positive sentiments, though some aspects are not performing well. The "fit" aspect appears to be relatively poor across all brands and needs attention. Conversely, comfort and appearance are consistently performing well across all brands.

Modelling and Evaluation:

DeBertaV3:

The data has been modelled using the pre-trained ABSA DeBertaV3-large - *yangheng/deberta-v3-large-absa-v1.1*. This model has been trained on 30,000 ABSA datasets and has demonstrated strong performance in aspect-based sentiment analysis for laptop and restaurant review dataset.

The modelling approach uses zero-shot classification, where the model is not specifically trained on the dataset but is utilized as-is to classify the sentiments of the aspects in the reviews. An overall sentiment analysis of the reviews was done along with the aspect-based sentiment analysis to obtain the sentiment polarity of each review.

The model's performance has been evaluated using the metrics of F1-score, accuracy, recall and precision. The evaluation is carried out assuming the annotated values as true values.

The model did not perform well with the dataset particularly for the aspect "comfort" which has the highest amount of positive sentiments failing which the model's overall performance has dropped down significantly. For the remaining aspects the model performed moderately with the metrics between 50-65%.

Evaluating aspect: comfort

Accuracy: 0.2588518751309449
 Precision: 0.3859827753197408
 Recall: 0.2588518751309449
 F1 Score: 0.1530825578666785

	precision	recall	f1-score	support
0	0.31	0.48	0.37	1374
1	0.24	0.70	0.36	2571
2	0.47	0.00	0.00	5601
accuracy			0.26	9546
macro avg	0.34	0.39	0.25	9546
weighted avg	0.39	0.26	0.15	9546

Evaluating aspect: appearance

Accuracy: 0.4056149172428242
 Precision: 0.4377124569087009
 Recall: 0.4056149172428242
 F1 Score: 0.3354421184358032

	precision	recall	f1-score	support
0	0.06	0.50	0.11	288
1	0.52	0.71	0.60	5221
2	0.36	0.01	0.01	4037
accuracy			0.41	9546
macro avg	0.31	0.41	0.24	9546
weighted avg	0.44	0.41	0.34	9546

Evaluating aspect: value

Accuracy: 0.6303163628745024
 Precision: 0.661520909776894
 Recall: 0.6303163628745024
 F1 Score: 0.6285407342854606

	precision	recall	f1-score	support
0	0.07	0.46	0.12	311
1	0.79	0.77	0.78	7635
2	0.16	0.00	0.01	1600
accuracy			0.63	9546
macro avg	0.34	0.41	0.30	9546
weighted avg	0.66	0.63	0.63	9546

Evaluating aspect: quality

Accuracy: 0.5613869683637126
 Precision: 0.5196814852731492
 Recall: 0.5613869683637126
 F1 Score: 0.5195741577118668

	precision	recall	f1-score	support
0	0.19	0.51	0.28	873
1	0.69	0.75	0.72	6554
2	0.13	0.00	0.01	2119
accuracy			0.56	9546
macro avg	0.34	0.42	0.33	9546
weighted avg	0.52	0.56	0.52	9546

Evaluating aspect: fit

Accuracy: 0.5569872197779174
 Precision: 0.4661485755513888
 Recall: 0.5569872197779174
 F1 Score: 0.4950250664928052

	precision	recall	f1-score	support
0	0.39	0.46	0.42	2072
1	0.62	0.78	0.69	5570
2	0.11	0.00	0.00	1904
accuracy			0.56	9546
macro avg	0.37	0.42	0.37	9546
weighted avg	0.47	0.56	0.50	9546

Fig: Performance metrics of all sentiments for DeBertaV3

The aspect sentiment plot for all brands is as shown below.

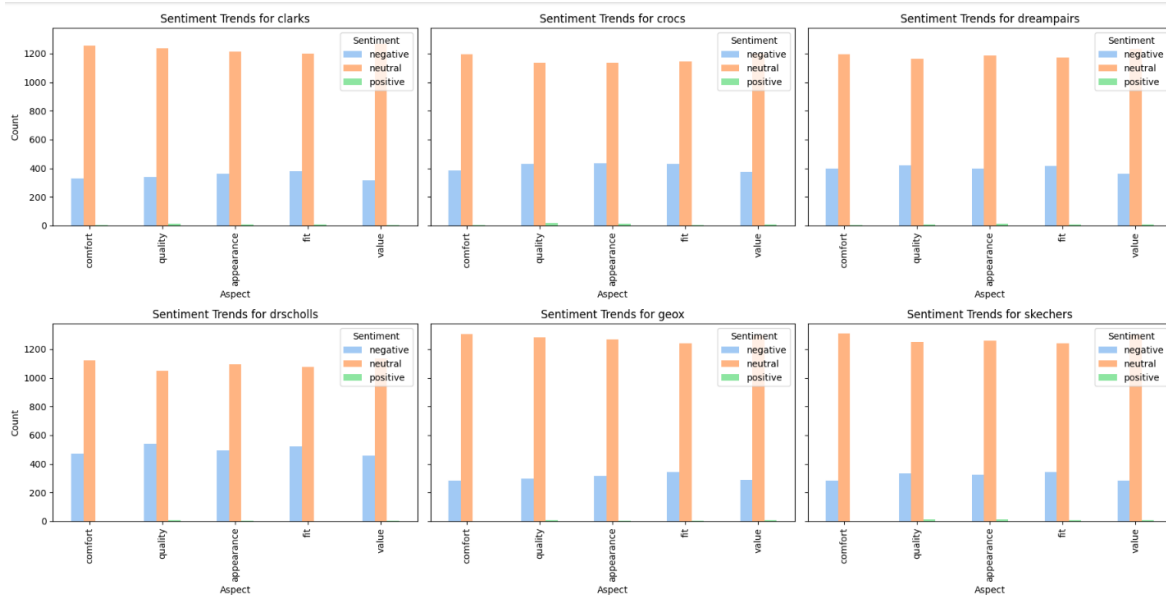


Fig: bar plot of aspect-based sentiments across all brands

The plot clearly demonstrates the model’s inability to capture any positive sentiments, and the uniform distribution across all brands indicates that the model's weights are inadequate for accurate classification. As it stands, the model is not suitable for the feedback system, as it could provide misleading information.

Flan-T5:

The data has been modelled using the pre-trained ABSA Flan-T5-large – “shorthillsai/flan-t5-large-absa”. This model has been trained on a custom dataset generated by GPT-4.

The modelling approach uses zero-shot classification, where the model is not specifically trained on the dataset but is utilized as-is to classify the sentiments of the aspects in the reviews. An overall sentiment analysis of the reviews was done along with the aspect-based sentiment analysis to obtain the sentiment polarity of each review.

The model’s performance has been evaluated using the metrics of F1-score, accuracy, recall and precision. The evaluation is carried out assuming the annotated values as true values.

This model also did not perform well with the dataset particularly for the aspect “comfort”. For the remaining aspects the model performed moderately with the metrics between 48-63%.

Evaluating aspect: comfort

Accuracy: 0.3274670018856065
Precision: 0.40275065819163963
Recall: 0.3274670018856065
F1 Score: 0.28552003501340917

	precision	recall	f1-score	support
0	0.11	0.00	0.01	1374
1	0.26	0.75	0.39	2571
2	0.54	0.21	0.31	5601
accuracy			0.33	9546
macro avg	0.30	0.32	0.23	9546
weighted avg	0.40	0.33	0.29	9546

Evaluating aspect: quality

Accuracy: 0.6274879530693485
Precision: 0.5235201020828607
Recall: 0.6274879530693485
F1 Score: 0.5620665494026141

	precision	recall	f1-score	support
0	0.04	0.01	0.01	873
1	0.69	0.88	0.77	6554
2	0.22	0.11	0.14	2119
accuracy			0.63	9546
macro avg	0.32	0.33	0.31	9546
weighted avg	0.52	0.63	0.56	9546

Evaluating aspect: appearance

Accuracy: 0.509009009009009
Precision: 0.502976043321435
Recall: 0.509009009009009
F1 Score: 0.5036630125168033

	precision	recall	f1-score	support
0	0.00	0.00	0.00	288
1	0.58	0.52	0.55	5221
2	0.44	0.53	0.48	4037
accuracy			0.51	9546
macro avg	0.34	0.35	0.34	9546
weighted avg	0.50	0.51	0.50	9546

Evaluating aspect: fit

Accuracy: 0.481981981981982
Precision: 0.3862450711020711
Recall: 0.481981981981982
F1 Score: 0.42865442275744187

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2072
1	0.60	0.74	0.66	5570
2	0.18	0.26	0.21	1904
accuracy			0.48	9546
macro avg	0.26	0.33	0.29	9546
weighted avg	0.39	0.48	0.43	9546

Evaluating aspect: value

Accuracy: 0.53886444584119
Precision: 0.6600369274230279
Recall: 0.53886444584119
F1 Score: 0.5813112215404148

	precision	recall	f1-score	support
0	0.04	0.00	0.01	311
1	0.79	0.60	0.68	7635
2	0.15	0.36	0.21	1600
accuracy			0.54	9546
macro avg	0.33	0.32	0.30	9546
weighted avg	0.66	0.54	0.58	9546

Fig: performance metrics of the model Flan-T5

The aspect sentiment bar plot across the brands is as shown below:

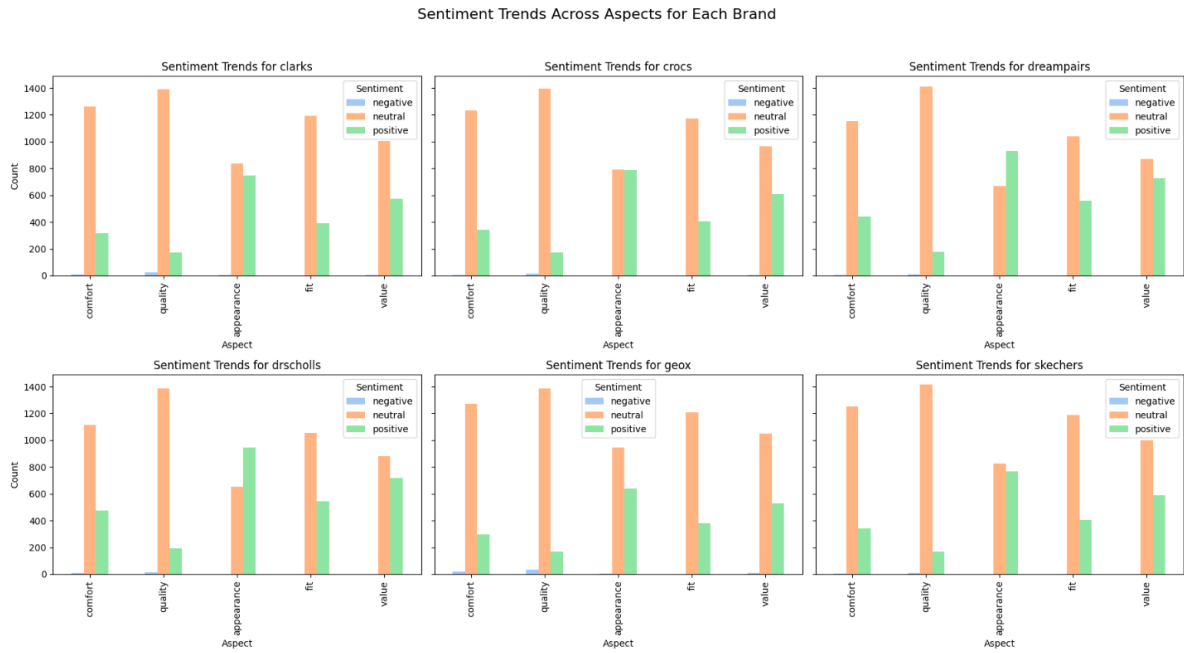


Fig: bar plot of aspect-based sentiments across all brands

This model failed to capture most of the negative sentiments. Majority of the negative sentiments have been classified either into neutral or positive sentiments resulting in a poor performance of the model. This model also is not a recommended one for the feedback system as it fails to capture the negative sentiments which are very important for improving customer satisfaction.

CONCLUSIONS AND FUTURE WORK

The rapid expansion of e-commerce platforms has dramatically reshaped consumer interactions, emphasizing the importance of understanding customer sentiments for new businesses. This research focused on developing an advanced feedback system using aspect-based sentiment analysis (ABSA) to analyze customer reviews of women's footwear from Amazon. The aim was to provide detailed insights into customer experiences by breaking down reviews into specific aspects and evaluating the sentiment associated with each aspect.

Key Findings:

The following are the key takeaways from the research analysis.

Data:

One of the primary issues identified was the imbalance in the dataset, with fewer examples of negative sentiments. This imbalance can cause the model to skew towards more frequent sentiment classes, such as neutral or positive, leading to poor performance in identifying negative sentiments. Analysing the class distribution depicted a significant disparity among the sentiment classes, necessitating the implementation of strategies to balance the dataset. Techniques such as oversampling the minority class, under sampling the majority class, or using class weights during training could be adopted as effective solutions.

Aspect Extraction and Annotation:

The study utilized Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) to extract relevant aspects from the reviews. There was notable low coherence in the words of the topics when modelled using LDA and LDA with TF-IDF. The aspects had to be manually identified from the TF-IDF frequent terms and LDA topic words. This did help much with the dynamic extraction of the aspects.

Sentiment labelling was performed using GPT-3.5 Turbo, which provided reliable annotations with an accuracy of 95% in a manual verification of a sample set. The model was easy to use with and did not need much intervention with respect to modelling and performed very well with one shot classification. However, there were a very few instances where the model assumed additional sentiment labels like “mixed feelings” or “unknown” and assigned to the aspects which had to be manually analysed. These misnomers were negligible compared to the size of the dataset. Although GPT-3.5 Turbo is a little expensive, it offers a trade-off in terms of improved performance and evaluation.

Model Performance:

DeBertaV3 achieved an accuracy range of 50-65% for most aspects, but it struggled significantly with positive sentiments.

Flan-T5 showed similar performance issues, particularly in correctly identifying negative sentiments, leading to an accuracy range of 48-63%. The models.

The models were not specifically trained on any footwear reviews, which could explain their poor performance in analysing them. Additionally, there is a possibility that they were not initiated with proper weights that could have contributed to the inaccurate analysis.

Evaluation Metrics:

The models were evaluated based on standard metrics, including accuracy, precision, recall, and F1-score. The results indicated that neither model was fully adequate for providing actionable insights for the feedback system, as they failed to accurately classify sentiments across all aspects.

Implications for New Businesses:

ABSA provides detailed insights into specific aspects of customer feedback, which is crucial for new businesses striving to improve customer satisfaction and product quality.

Challenges with Pre-Trained Models:

The moderate performance of DeBertaV3 and Flan-T5 suggests that pre-trained models need further fine-tuning and adaptation to specific datasets to enhance their effectiveness in aspect-based sentiment analysis.

Importance of Negative Feedback:

Accurate identification of negative sentiments is critical for businesses to address customer pain points. The failure of models to capture these sentiments adequately highlights the need for more robust training data and model optimization.

Future Work:

The research revealed several areas for further investigation and improvement to enhance the efficiency of aspect-based sentiment analysis in providing valuable feedback for new businesses. Future work should focus on addressing the limitations encountered and exploring new methodologies to improve model performance.

Key Areas for Future Research:

Efficient Hybrid Approaches for aspect extraction:

Combining multiple aspect extraction methods, such as TF-IDF, LDA, and neural network-based approaches, may yield better results. Hybrid models can leverage the strengths of different techniques to improve aspect identification.

Dynamic Aspect Identification:

Implementing dynamic aspect identification systems that can adapt to emerging trends and new aspects in customer reviews will keep the feedback system relevant and up to date.

Model Fine-Tuning and Customization:

Task-Specific Fine-Tuning:

Future research should involve fine-tuning pre-trained models like DeBertaV3 and Flan-T5 on task-specific datasets. This process should include training on a larger, class balanced set of reviews to improve the models' performance.

Custom Model Development:

Developing custom models tailored specifically for aspect-based sentiment analysis in the footwear domain may enhance accuracy and reliability. These models can be trained from scratch using a comprehensive dataset that covers a wide range of aspects and sentiment polarities.

Enhanced Data Annotation Techniques:

Increasing the volume of manually annotated data can improve the training process. Implementing rigorous quality checks on annotated data will ensure higher accuracy and reliability of sentiment labels.

Integration with Business Intelligence Systems:

Real-Time Feedback Systems:

Developing real-time feedback systems that integrate ABSA models with business intelligence platforms will enable businesses to respond promptly to customer feedback and make data-driven decisions.

Actionable Insights and Reporting:

Creating dashboards and reporting tools that translate model outputs into actionable insights for product designers, business analysts, and marketing teams will maximize the utility of the feedback system.

Conclusion:

This research has demonstrated the potential of aspect-based sentiment analysis in providing detailed insights into customer feedback. While the pre-trained models DeBertaV3 and Flan-T5 showed moderate performance, their limitations highlight the need for further fine-tuning, data augmentation, and model customization. Future work should focus on enhancing

model accuracy, reliability, and interpretability to develop a robust feedback system that empowers new businesses to improve their products and services based on customer sentiment. By addressing these challenges and exploring advanced methodologies, aspect-based sentiment analysis can significantly contribute to the success and growth of new enterprises in the competitive e-commerce landscape.

REFERENCES

1. *A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis*
Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.
2. *Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis*
N. Mughal, G. Mujtaba, S. Shaikh, A. Kumar and S. M. Daudpota, "Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis," in IEEE Access, vol. 12, pp. 60943-60959, 2024, doi: 10.1109/ACCESS.2024.3386969. keywords: {Sentiment analysis;Analytical models;Task analysis;Reviews;Computational modeling;Transformers;Biological system modeling;Large language models;Aspect-based sentiment analysis (ABSA);large language model (LLM);GPT;PaLM;BERT}
3. *Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT*
Vadla, Mahammad & Suresh, Mahima & Viswanathan, Vimal. (2024). Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT. Algorithms. 17. 59. 10.3390/a17020059.
4. *Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning*
Yang, Heng & Zeng, Biqing & Xu, Mayi & Wang, Tianxing. (2021). Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning.
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
6. <https://huggingface.co/yangheng/deberta-v3-large-absa-v1.1>
7. <https://huggingface.co/shorthillsai/flan-t5-large-absa>
8. <https://www.amazon.co.uk/>
9. <https://www.amazon.com/>
10. Review Extraction utility by <https://www.symanto.net/>

GLOSSARY

ABSA (Aspect-Based Sentiment Analysis): A technique in sentiment analysis that breaks down customer reviews into specific components or aspects (e.g., product quality, customer service) and evaluates the sentiment associated with each aspect.

Amazon: An online e-commerce platform where customers can buy a wide range of products and leave reviews based on their experiences.

AutoTokenizer: A utility from Hugging Face's transformers library that automatically loads the appropriate tokenizer for a given pre-trained model. It handles the conversion of raw text into token IDs.

CapsNet (Capsule Network): A neural network architecture known for capturing hierarchical relationships in data, beneficial for understanding nuances in sentiment analysis.

CRISP-DM (Cross-Industry Standard Process for Data Mining): A structured approach to data mining projects that includes phases like business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Customer Sentiment: The overall feeling or attitude expressed by customers in their reviews, which can be positive, negative, or neutral.

DeBERTaV3 (Decoding-enhanced BERT with Disentangled Attention): An advanced variant of the BERT model developed by Microsoft, designed to improve performance on various natural language processing tasks by using disentangled attention mechanisms and enhanced positional encoding.

Evaluation Metrics: Standard metrics used to assess the performance of machine learning models, including accuracy, precision, recall, and F1-score.

Flan-T5 (Fine-Tuned Language Agnostic T5): An advanced version of the T5 model developed by Google, fine-tuned using task-specific instructions to improve performance across various NLP tasks.

GPT-3.5 Turbo: A large language model developed by OpenAI, capable of various applications including sentiment analysis, text completion, and more. It excels in tasks requiring a high level of natural language understanding.

Hugging Face: A company that provides tools and libraries for natural language processing, including the transformers library which hosts a variety of pre-trained models.

LDA (Latent Dirichlet Allocation): A topic modeling algorithm that identifies hidden topics within a collection of documents or texts, used in this study to validate the aspects identified through TF-IDF.

LLM (Large Language Model): A type of machine learning model trained on large amounts of text data to understand and generate human language. Examples include GPT, PaLM, and T5.

NFKD (Normalization Form KD): A form of text normalization that converts Unicode text into a consistent format, helping in the preprocessing of textual data.

NLP (Natural Language Processing): A field of artificial intelligence focused on the interaction between computers and humans through natural language, encompassing tasks such as sentiment analysis, text generation, and more.

OpenAI API: An interface provided by OpenAI that allows developers to integrate models like GPT-3.5 Turbo into their applications for various NLP tasks.

Precision: A performance metric that measures the proportion of correctly predicted positive instances out of all instances predicted as positive.

Recall: A performance metric that measures the proportion of correctly predicted positive instances out of all actual positive instances.

TF-IDF (Term Frequency-Inverse Document Frequency): A numerical statistic used to reflect the importance of a word in a document relative to a collection of documents. It helps identify high-frequency terms relevant to specific aspects in reviews.

Topic Modeling: A method used in natural language processing to discover abstract topics within a collection of documents. LDA is a common algorithm used for this purpose.

Transformer Models: A type of deep learning model that uses attention mechanisms to process and generate text. Examples include BERT, DeBERTa, T5, and GPT.

Zero-Shot Classification: A machine learning approach where the model is not specifically trained on a particular dataset but is used as-is to classify data based on its general understanding from pre-training.