# 1. Preamble

## 1.1 Introduction

As e-commerce has grown quickly, product reviews have emerged as a crucial component affecting what customers decide to buy. Product review reading has become common among online consumers, as 99.5% consult reviews before purchasing products. The extensive number of reviews in existence produces reliability issues. The existence of false product reviews damages consumer trust in online reviews because these deceptive ratings manipulate product ratings to misguide customers. Unrelated reviews entering the review systems create a visual noise that complicates user access to helpful customer feedback. The inconsistent or inadequate business responses to review sentiments create additional challenges regardless of the sentiment orientation.

The proposed solution builds an Automated Review Analysis and Response Generation system that resolves current market challenges. The system completes multiple key operations, including artificial review identification and product type recognition, with sentiment evaluation and automatic response automation. Four different operational elements tackle distinct obstacles in the review process by detecting fake content while categorizing and evaluating sentiments to produce automated business responses. E-commerce review management systems aim to achieve multiple purposes, which support both enhancing review trustworthiness and improving the consumer experience.

This research presents considerable potential to enhance the quality standards of internet-based product testimonials. Establishing this system enables platforms to reduce operational delays, deliver superior customer service, and uphold genuine review content. The system applies advanced Natural Language Processing (NLP) techniques based on large language models (LLMs).

The research examines the implementation of LLMs for evaluating fake reviews, automatic response generation, sentiment detection, and review classification in e-commerce product assessment systems. It necessitates the training and optimization process of multiple LLMs for task execution. The system works to boost review quality through its capabilities to identify and delete fake content, review classification, and detect sentiment with appropriate response generation. The research will help develop an automated system that enhances e-commerce operations.

## 1.2 Current Studies

### 1.2.1 Fake Review Detection

Modern fake review detection advances utilize LLMs to detect fraudulent reviews while considering factors like reviewer and merchant activities to achieve better accuracy rates. The paper described in [1] by Sun et al. combines **BERT,** which analyzes review text, with an analysis of reviewer conduct that includes rating trends and timing habits to establish superior detection capabilities compared to basic models. The paper [2] by Li et al. demonstrates how synthetic fake review generation with LLMs improves fake review detection capability by feeding them an enlarged training dataset. The application of LLMs for detecting fake news is explored in [3] by Papageorgiou et al., where the authors emphasize their ability to perform text classification and fact-checking tasks that can be applied to fake review detection. The research [4] by Wang et al. presents a

multimodal detection system that unites **LLMs** with visual elements for fake news recognition and this method could enable review analysis using product images. The research in [5] by Ma et al. adds high-level semantic analysis with graph-based feature propagation to address traditional LLM embedding problems. This leads to improved fake news detection abilities that can work for fake review detection by mapping between product features and topics. The combination of LLMs and multimodal methods proves effective in fake review detection per these collective research findings.

### 1.2.2 Sentiment Analysis

Implementing Large Language Models (LLMs) and new approaches has improved sentiment analysis in product reviews. The research paper [6] by Lan et al. investigates Chinese financial text sentiment analysis while studying a new dataset through comparisons between **LLaMA** and **ChatGPT** alongside specialized models that produce superior results compared to general LLMs. The authors in [7] by Miah et al. present work on cross-lingual sentiment analysis, which realizes better accuracy through machine translation integration with GPT-3 and **RoBERTa** ensemble models. The authors in [8] by Shaik et al. employ BERT and T5 to run aspect-based sentiment analysis on Amazon reviews, producing automated consumer insights for product design. Researchers in [9] by Ghatora et al. reported that GPT-4 performs superior to traditional machine learning models in processing lengthy reviews containing mixed sentiments because it provides improved explanation capabilities and transparent results. Multiple agents work together in [10] by Xing et al. to establish financial sentiment analysis through **GPT-3.5** and **BLOOMZ**, which target particular prediction errors and show effective collaboration for sentiment tasks. According to these recent studies, product review analysis benefits significantly from the emergence of LLMs and multimodal techniques, together with domain-specific adaptation technologies.

### 1.2.3 Category Classification and Relevancy Check

The classification of product categories represents a vital operational process within business analytics and e-commerce, which relies increasingly on Large Language Models (LLMs) for description categorization that supports recommender systems, customer profiling, and compliance tracking, as described in [11] by Park et al. The research study in [12] by Yamaguchi et al. illustrates that product feature reviews enable crucial consumer needs, enhancing new product development and promoting improved purchase choices. The paper [13] by Hou et al. explores how LLMs serve as zero-shot rankers inside a recommender system to fix various review rank deficiencies that arise from interaction histories and item bias, which boosts ranking accuracy. Consumer behavior changes are examined in a conceptual model analyzing the credibility and relevance of reviews per the research [14] by Mumuni et al., who show that review relevance strongly impacts purchase intentions and product recommendation behavior. Integrating LLMs with generative models, hand in hand with fine-grained analysis techniques, enables more profound and more accurate insights to transform ABSA and review relevancy checks in product review content.

### 1.2.4 Review Reply Generation

Rapid automated customer review reply generation development occurred when Large Language Models (LLMs) were deployed in systems. The two-phase framework described in [15] by Wang et al. enables LLM agents to extract substantial customer needs from UGC before creating product enhancement suggestions. The method showcases how LLMs effectively mine customer feedback with the same performance outcomes as deep learning models. The SCRABLE

system employs LLMs to produce and enhance customer review responses by using a self-improving iterative generation process according to [16] by Azov et al. SCRABLE uses Retrieval-Augmented Generation (RAG) and human-like evaluation to enhance response quality by more than 8.5% and proves that LLMs offer real-time optimization for review reply generation. An LLM-based Smart Reply (LSR) system that helps clients with contextualized and individualized responses during collaborative work is the subject of research in [17] by Bastola et al. Initially designed for workforce interaction, the system provides methods to minimize mental workload and boost workplace efficiency that user review response automation can utilize. The study in [18] by Kickler et al. demonstrates how open-source LLMs and RAG improve SME customer support through affordable information retrieval systems. The authors in [19] by Geng et al. studied how large language models (LLMs) function to produce multiple comments using in-context learning for different code snippet intents. Implementing few-shot learning allows the model to produce multiple suitable comments with diversity while requiring minimal fine-tuning, improving LLM adaptability. The studies present evidence of LLM versatility through high-quality context-aware customer review responses by demonstrating iterative development approaches with domain-specific adaptations and multi-agent systems for output optimization. In the research on [20] by Wang et al., Their findings demonstrate that LLM-based evaluation can match or even outperform traditional metrics and non-targeted human evaluators, reinforcing the role of LLMs not only in generating but also in reliably evaluating customer review replies. Emerging trends have demonstrated the versatility of large language models (LLMs) in applications. LLP4ServiceRobot [21] by Silva et al. illustrates how fine-tuned LLMs can improve human-robot interactions by generating contextually sensitive, natural language responses, thus improving communication in real-world service scenarios. Similarly, an automatic review generation method by Wu et al. [22] leverages LLMs to create scientific literature with high precision, encompassing a multi-layer quality control strategy that reduces hallucination rates to below 0.5%, thus providing reliability and scalability of automated reviews. Conversely, Zhou et al. [23] and Hu & Zhou [24] address LLM testing, with concerns such as benchmark leakage where there is overlap between training and testing sets that inappropriately increases performance, and the challenges of selecting strong, domain-relevant test metrics, particularly in sensitive domains such as biomedicine.

### 1.2.5 Zero Shot Classification & Labeling Dataset

Zero-shot learning (ZSL) has become a highly effective classification method when no labeled data is available. The work of Puri and Catanzaro [25] presented a generative method for zero-shot text classification through GPT-2 fine-tuning with natural language task prompts that enabled the model to execute numerous jobs without the need for task-specific instruction and outperform baseline models by a considerable margin. Building on this, Alhoshan et al. [26] applied ZSL to requirements classification using transformer-based models and contextual embeddings to achieve promising F1 scores (66%–80%) for several classification instances and thus demonstrate that ZSL can be realized in resource-scarce domains. Also advancing the generalizability of open-source LLMs, Alizadeh et al. [27] showed that fine-tuned models such as LLaMA-2 and FLAN-T5 can perform at par with GPT-3.5 in annotation tasks and even surpass crowd-sourced annotators and provide an open and cost-effective solution for computational social science applications. A survey documents how artificial intelligence has recently

improved review management through false review detection tools alongside various sentiment analysis methods, including relevance measurement and response simulation platforms. Large Language Models create significant changes by enabling swift evaluation of reviews with high accuracy and better contextual comprehension. Multi-agent technology and the Recovery-Based Algorithm (RAG) enhance review management processing efficiency. Integrating zero-shot learning approaches and generative models with detailed analysis provides evidence to develop automated systems that deliver customer sentiment insights.

## 1.3 Main Work of This Project

This research develops an AI-controlled review management solution for e-commerce websites. The evaluation process benefits from four fundamental tasks that streamline the review management operations:

- **Fake Review Detection**: Mistral-LLMs detect fraudulent reviews, which function as fine-tuned LLMs for identifying genuine versus fake reviews. The system employs sophisticated language detection methods for accurate fake content assessment to maintain proper review authenticity.
- **Category Classification**: Mistral-LLMs operates with a specific model that assigns each review to a particular product category by selecting automotive, fashion, home, electronics, or health. This classification method makes the review organization more efficient while allowing users to perform productive analysis based on product types.
- **Sentiment Analysis**: Review sentiment analysis depends on a sentiment model developers explicitly train for the review data. The model system differentiates reviews into three emotional groups: positive, neutral, and negative, allowing businesses to gain insights for better decision-making processes.
- **Automated Reply Generation**: The system uses its review content, sentiment analysis, and category information to generate appropriate responses through its review response generation model. The programmed responses utilize advanced algorithms to handle customer complaints, express gratitude for positive input, and boost the customer interaction process.

## 1.4 Organization and Structure

The body of this thesis is organized into six chapters:

**Part I: Preamble**: Provides the context of the research, its importance, and the overarching aims of the project, including automatic spurious review detection, sentiment analysis, and response generation. Current research is also discussed in the field.

**Part II: Background Knowledge**: This section includes the theories and primary methodologies employed in the project, such as large Language Models (LLMs), sentiment analysis, fake review detection, and category classification.

**Part III: The Analysis and Design of the System**: This section discusses the system's initial design choices, including model selection and architecture.

**Part IV: System Implementation**: Describes the step-by-step implementation of the system, from model fine-tuning to integration.

**Part V: System Test**: Reports the test procedure, strategy, and performance metrics, with results like accuracy, F1-score,

and other performance metrics. Describe the strengths and weaknesses of the system.

**Part VI: Work Summary and Reflection:** Summarizes the project's key findings, contributions, and problems. It also addresses the system's performance implications and recommends future research.

# 2. Background Knowledge

## 2.1 Overview of Transformer Architecture and Large Language Models

This thesis depends on Large Language Models (LLMs) as its fundamental technological foundation and uses **Mistral-LLMs** that operate with **transformer architecture**. The transformer models enable long-range dependency processing through **self-attention**, which evaluates input sequence segments, thus determining their prediction relevance. In contrast to older networks like RNNs and LSTMs, processors experienced difficulties working with effective long-term dependencies.

Modern Natural Language Processing (NLP) operates with the transformer architecture because this foundation combines high processing efficiency, scalability, and parallelization advantages. The Mistral-LLMs model operates with 7 billion parameters to work on intricate NLP projects such as text classification, generation, and reasoning systems. Through training across extensive language sources, the model acquired advanced language patterns it employs to deliver effective solutions in fake review detection, product category classification, sentiment analysis, and automated review reply generation.

Mistral-LLMs achieves efficient GPU execution by undergoing data-reduction training via the Low-Rank Adaptation (LoRA) and 4-bit quantization (QLoRA) protocols, lowering system memory requirements and processing intensity. These methods allow the model to perform effectively on particular tasks while using minimal hardware and training time, thus making it suitable for this project.

## 2.2 Low-Rank Adaptation (LoRA) and 4-bit Quantization (QLoRA)

LoRA and QLoRA are optimization tools that enhance Mistral-LLMs by improving efficiency and performance outcomes. LoRA's operation involves low-rank matrix additions to transformer layers, which reduce parameters to minimize resource use without affecting computational or storage needs. Such environments benefit from this method. QLoRA achieves memory optimization through a weight reduction process that converts 16-bit and 32-bit models into 4-bit precision for devices with reduced GPU capabilities.

The Unsloth library enhances optimization features that automatically integrate with PyTorch. The system enables deployment of big models through low-rank adaptation and 4-bit quantized models, which optimize performance in systems with GPU memory limitations.

## 2.3 NLP Tasks and Applied Algorithms

### 2.3.1 Fake Review Detection

Technical systems use Fake Review Detection to identify misled customers through fabricated reviews and reviews that other users have manipulated. Either sellers or external agencies create these reviews to enhance product scores and sanitize business rivals. The purpose of fake review detection systems is to protect review systems from fake information by identifying untruthful feedback. Review fraud detection applies NLP techniques and machine learning models to identify suspicious

behavioral patterns in review texts by spotting repetitive language, generic statements, and review-text/viewer-behavior inconsistencies. E-commerce systems depend heavily on review trust because consumer purchasing decisions rely on product reviews. The authentication of user reviews happens through fake review detection across different online marketplaces, including Amazon, eBay, and AliExpress. The technology supports both sites, which provide customer-based feedback, including hotel review services like TripAdvisor and restaurant review applications like Yelp, to fight deceptive reviews.

### 2.3.2 Category Classification

Category Classification is a task that categorizes reviews or products into predefined categories based on their content. This can be achieved using supervised learning techniques, where a model is trained to predict the most likely product category based on review text. The popular review categories for business analysis include automotive, fashion, and electronics. Systematic classification enables companies to handle numerous reviews by grouping feedback according to product categories, which helps evaluate customer insights per product type. E-commerce platforms, product recommendation systems, and content management systems utilize this task to categorize product reviews before analyzing them through customer sentiment analysis and marketing campaign generation.

### 2.3.3 Sentiment Analysis

Sentiment analysis systems successfully identify emotional sentiments and text-based emotional expressions in written material. The analysis technique expands customer text feedback into two classifications: positive and negative reviews, with optional neutral content determinations. Organizations need to evaluate customer emotional states through evaluations of their verbal product satisfaction and product dissatisfaction expressions and statements about product apathy. Machine learning models perform the analysis by assessing linguistic elements found in reviews based on word selections, sentence organization, and context evaluation. The sentiment analysis mechanism allows various economic sectors to measure public product opinions through feedback, market survey methods, and social media content inspection. Businesses can monitor customer satisfaction through this approach, which identifies specific issues by tracking public responses concerning their products and services. Facebook, Twitter, and Amazon's sentiment analysis software extracts user-submitted feedback in posts, tweets, and review content.

### 2.3.4 Review Reply Generation

A computer system produces well-suited feedback responses to reviews as part of Review Reply Generation. Programs generate responses after processing the review sentiment as well as the review content. The system develops purposefully generated empathetic responses that match particular customer conditions to increase interaction quality. Transformers and other NLP models conduct this operation through a process that involves fine-tuning on product review datasets with matching human response examples. The automatic response system used by e-commerce companies like Amazon and AliExpress, plus service vendors such as restaurants and hotels, saves customer support time and provides personalized interactions to customers. Systems powered by this technology frequently use it as part of automated customer service machines and chatbots.

### 2.3.5 Data Augmentation

Data Augmentation consists of transformation techniques that produce new dataset points by processing original entries to expand the dataset virtually. Text modification methods produce additional review data points by applying synonyms,

sentence transformations, and text-disturbing techniques to original materials. The method addresses problems with unbalanced datasets, which contain categories and sentiments with small numbers of reviews. Data augmentation is the standard approach for training machine learning models, particularly in NLP applications such as sentiment analysis, category classification, and fake review detection. Such representation in the training dataset enables better model generalization and reduces biases. Limited datasets receive performance improvements from this approach, allowing both image classification and spoken word identification, besides its medical applications.

### 2.3.6 Labeling Tasks

The process of assigning predefined labels to data through manual work or automated systems constitutes Labeling Tasks. An example would be designating reviews into fake or real classes or assigning them positive, neutral, and negative scores. The grounding truth required for training machine learning models comes from the critical labeling process in supervised learning. The thesis consists of labeling processes to determine fake reviews, analyze sentiment, and classify products while generating automated responses. Labeling tasks run directly through the development phase of labeled datasets to prepare machine learning models across different application domains spanning e-commerce, healthcare, social media, and finance sectors. Training models for sentiment analysis, fake review detection, and diagnostic models use labeled product reviews as input.

### 2.3.7 Zero-Shot Classification

The machine learning methodology called zero-shot classification enables models to perform classification of unknown categories that were not present during their training period. The main difference between standard supervised learning and zero-shot classification occurs when supervised learning depends on labeled samples per category. In contrast, zero-shot classification helps the model apply generalization from available knowledge acquired from reading the input text. Besides other widely used models, BERT and BART achieve this capability by extracting contextual information to predict categories outside their training scope.

Zero-shot classification grants the main advantage of adapting beyond training parameters. The model can classify reviews for categories beyond training some while learning without needing new examples or additional training instances. Zero-shot classification served as the selected method for product categorization throughout this project deployment. Through this method, the system gains the capacity to inspect an extensive assortment of product categories that exceed training parameters.

There are trade-offs, though. Zero-shot classification allows for speed and scalability in data labeling, but its accuracy is lower than supervised learning when used on specific categories. The classification system makes more errors in sorting complex product categories. Supervised models provide higher accuracy, yet their implementation demands expensive data labeling for every category.

The research adopted zero-shot classification for product categorization since it proved both affordable and expandable. This method proved beneficial when dealing with various categories while needing minimal model adjustments. The precision of supervised learning is higher than zero-shot classification, but product categories evolve so rapidly that zero-shot proved more practical.

## 2.4 Summary of This Chapter

The second chapter provided essential background understanding and fundamental algorithms required for designing the thesis accurately. This section explores how **Mistral-LLMs** function within various NLP functions, including fake review detection, category classification, sentiment analysis, and review reply generation. As detailed in the discussion, these models achieve their best operational characteristics through LoRA and QLoRA. The research described every method separately, from binary classification for fake review detection through multi-class classification and text generation used in sentiment analysis and reply generation. A workflow based on these approaches is the core foundation for the developed automated review management system to reach its performance peak.

# 3. The Analysis and Design of the Proposed System: Automated Review Analysis and Response Generation Using AI

## 3.1 Analysis and Modeling of System Requirements

### 3.1.1 Overview of System Requirements

The AI-driven review management system aims to automate and optimize the handling of customer reviews in e-commerce platforms. It is designed to address several critical aspects, such as detecting fake reviews, classifying the sentiment of reviews, ensuring the relevancy of reviews to the correct product category, and generating contextually relevant automated responses to customers. The system needs to handle enormous quantities of reviews spanning numerous product categories, with the ability to process thousands or millions of product ratings efficiently and accurately.

- **Functional Requirements**:
    - **Fake Review Detection**: The system must determine fake computer-generated reviews from legitimate human-written ones through double analysis of content data, plus user behavioral indicators such as abnormal language usage and sentiments that vary from review to review. The system provides text classification abilities through LLMs and identifies patterns associated with manipulative review actions, including excessively high ratings and repeated verbalization.
    - **Sentiment Analysis**: Advanced models, including LLMs, should automatically classify reviews into three categories: positive, neutral, and negative. This method allows businesses to evaluate customer emotions and gather insight, which helps them make future product development decisions.
    - **Review Relevancy Check**: The system requires functionality to validate that reviews fit within their corresponding product category and eliminate unrelated ones. It utilizes product categorization models to validate review relevance and block wrong information from altering the analytical results.
    - **Automated Reply Generation**: The system requires the capability to deliver responses that adapt to specific review sentiments, product categories, and review content. The system offers individualized computerized customer interactions, which improve user satisfaction.

- **Non-Functional Requirements**:
    - **Performance**: The system must operate quickly by processing reviews immediately while working through high data volumes. The system performs fast reactions through model quantization and batch processing without sacrificing performance standards.
    - **Scalability**: The solution must demonstrate data scale efficiency through its ability to grow easily while attaining distributed processing so review volumes can expand without affecting performance levels.
    - **Accuracy**: The system must keep its performance optimal when detecting fake reviews, analyzing sentiment, and generating suitable replies. The review management process achieves trustworthy accuracy by performing continuous model improvement cycles and training, which result in precise predictions.

### 3.1.2 Analysis of System Data

- **Data Collection**:
    - AliExpress, one of the biggest e-commerce sites, is the primary source of the review data for this study, using a **web scraping** method. The focus is collecting reviews for various products, including Crossbody Bags, Car Phone Holders, Dashcams, and Smartwatches.
    - The data scraped includes: reviewContent, username, userCountry, userStar, reviewTime, and language
    - The scraping is performed using the **ApifyClient** tool, which automates the extraction process and collects large volumes of reviews without manual intervention.
- **Preprocessing and Cleaning**: Once the data is collected, several preprocessing steps are carried out to prepare it for analysis and model training:
    - **Noise Removal**: Any irrelevant or incomplete reviews (e.g., reviews without text content or metadata) are removed.
    - **Handling Missing Values**: Missing values, such as empty fields or incomplete metadata (e.g., missing star rating or review time), are handled through imputation or removal, depending on the context.
    - **Text Normalization**: The review content is normalized by converting all text to lowercase, removing special characters, URLs, and user mentions (e.g., "@user").
    - **Text Structuring**: The cleaned data is stored in a structured format, primarily as a **CSV** file, to make it easily accessible for subsequent labeling and model training processes.
- **Data Labeling**:

  **(a) Fake and Real Labeling**: To train the fake review detection model, the dataset must be divided into real and fake reviews, which the review labeling procedure ensures. The process involves several stages to automate the classification of reviews.
    - **Data Loading**: The review dataset is loaded from a CSV file containing the raw reviews. The dataset includes multiple product reviews collected from AliExpress through web scraping.
    - **Review Classification**: A pre-trained fake review detection model (*'astrosbd/fake-reviews-distilbert-v3'*) from Hugging Face's transformers library is used for classifying reviews into either "real" or "fake." This

model leverages a DistilBERT architecture, fine-tuned on fake review detection tasks. The pipeline is used to classify each review based on its content.

- **Function for Review Classification**: A function classify_review() is defined, which takes a review as input, applies the fake review detection model, and returns a classification label:
  - **Real Reviews** are labeled as 1.
  - **Fake Reviews** are labeled as 0.
- **Output Dataset**: The modified dataset is saved into a new CSV file after the reviews have been categorized, containing the original reviews and their respective labels (real or fake). Among 17,607 reviews, 12,916 were real and 4,691 were fake.

**(b) Product Category Labeling Process:** Labeling each review's product category involves two main stages: first, classifying the reviews into 20 specific categories using a zero-shot classification model, and then mapping these 20 categories into five broader categories for more balanced representation. Below is a detailed explanation of this methodology:

- **Loading Dataset**: The process starts with loading the dataset containing real reviews that must be labeled with the appropriate product category. The dataset is imported into a Pandas DataFrame for easy processing and modification.
- **Zero-Shot Classification**: The Hugging Face zero-shot classification pipeline is used to classify the reviews into one of the 20 predefined product categories. This model (*'facebook/bart-large-mnli'*) can assign a review to a category without requiring additional training. The categories are based on product types such as "Crossbody Bags," "Car Phone Holders," "Smartwatches," etc. The model classifies each review into one of these categories and stores the results in the categoryLabel column of the DataFrame.
- **Mapping Categories to Broader Labels**: Since some categories have very few reviews, the next step is to map the 20 detailed categories into five main categories. This step helps balance the dataset by grouping similar categories into broader groups. The 20 categories are mapped into the following five main categories:
  - **Electronics**: This category includes pendrives, Bluetooth Earbuds, Smartwatches, Security Cameras, and Dashcams. It has 5606 reviews.
  - **Automotive**: Includes categories like Portable Car Vacuums, Automotive Accessories, and Car Phone Holders. The automotive category has 5062 reviews.
  - **Fashion**: This category includes Crossbody Bags, Dog Collars, Wallets, and Phone Cases. It has 1391 reviews.
  - **Home**: The Home category has 642 reviews and includes categories like Sofa Covers, Broom Holders, Kitchen Accessories, LED Lamps, and Fairy Lights.
  - **Health**: This category has 215 reviews, including categories like Electric Toothbrushes, Fitness Equipment, and Beauty and Health Products.

The reviews initially classified into the 20 specific categories are now re-labeled based on these broader categories.

After classification and mapping, the final dataset is saved as a new CSV file containing the product review, the initially predicted category, and the mapped broader category.

**(c) Sentiment Labeling Process:** The sentiment labeling process involves classifying product reviews into three categories: **positive**, **neutral**, and **hostile**. This process uses a pre-trained sentiment analysis model specifically fine-tuned for Twitter-style text, particularly suitable for reviews that may contain informal language and internet jargon. Here is a detailed breakdown of the procedure:

- **Data Loading**: Data Loading lies at the beginning of the process, where the dataset enters, which contains reviews for sentiment classification. The CSV file contains the reviews, which Pandas DataFrame converts into a format for processing and manipulation.
- **Preprocessing the Review Text**: The review text undergoes preliminary processing to achieve standardization, resulting in better sentiment processing outcomes. On this step, the system replaces Twitter handles (@user) and URLs (http://...) with placeholder markers to prevent such elements from distorting sentiment analysis results.
- **Sentiment Classification with Pre-trained Model**: The *'cardiffnlp/twitter-roberta-base-sentiment-latest'* model ensures sentiment analysis by utilizing its pre-trained function optimized for social media data sentiment classification. The chosen model succeeded in processing unstructured product review text similar to social media data because of its established performance on such informal content. The sentiment classification process operates on every review from the dataset by utilizing Pandas' progress_apply function to conduct sentiment analysis with an animated progress bar. The sentiment classification function evaluates each review by generating sentiment predictions alongside corresponding confidence scores within the 0 to 1 range.
- **Output**: Final dataset is saved as a new CSV file containing review content and sentiment.

**(d) Reply Generation Labeling Process:** The reply generation labeling process automatically generates context-dependent responses to product reviews based on their sentiment, product category, and content. The procedure uses a previously trained language model to produce relevant responses, which have been refined using a collection of manually constructed context-based responses. The following is a step-by-step description of how the reply generation labeling process works:

- **Data Preparation**: The process starts with a dataset of 100 manually created context-specific responses to fine-tune the language model. The manually created responses offer a comprehensive and varied set of examples for the model to learn from, encompassing sentiment variations, product categories, and content contexts. The dataset comprises reviews with sentiment labels (positive, neutral, negative) and their respective product categories (e.g., electronics, automotive, fashion, etc.), in addition to the manually generated responses.

- **Small Fine-Tuning the Model**: Use the 100 manually created context-based responses dataset to improve the language model that has already been trained (here, the Mistral-LLMs model). The model learns from the input's review content, sentiment, and product category to generate responses. Model refinement enables the model to capture the particular style, tone, and context of reactions, which is critical for creating high-quality, context-relevant responses.
- **Generating Replies for Each Review**: The model generates a reply using each review's context (sentiment, category, and review content). The model's output is a context-appropriate response that aligns with the sentiment and category of the review. The generated replies are stored alongside the original review content, sentiment, and category.
- **Storing the Results**: The generated replies are added to the dataset, and the updated dataset (with the generated replies) is saved to a CSV file. The final dataset now includes:
  - Review Content: Original customer review text.
  - Sentiment: Predicted review sentiment (positive, neutral, negative).
  - Category: Predicted product category for the review.
  - Generated Reply: The context-aware reply generated by the model.

- **Dataset Augmentation**:

  **(a) Augmentation Process for Fake Review Dataset:** The augmentation process is performed by generating additional spurious reviews from existing ones using natural language processing techniques, in this instance, utilizing a **WordNet Augmenter** from the **TextAttack** library. This process is intended to balance the dataset by generating enough spurious reviews to match the number of actual reviews, thus making the dataset evenly balanced for training deep learning models. Some steps in the description of the augmentation process used on the dataset:

  - **Initial Dataset Analysis**: The dataset contains product reviews labeled "real" or "fake". Initially, the "real" reviews outnumber the "fake" reviews.
    - **Real reviews**: 12,916
    - **Fake reviews**: 4,691
  - **Targeted Augmentation for Fake Reviews**:
    - **Augmentation Goal**: The objective is to equilibrate the dataset by producing fake reviews. The system must produce 12,916 fake reviews in total, matching the quantity of real reviews, as the count of fake reviews is equivalent to that of real ones.
    - **TextAttack Library**: The **WordNetAugmenter** from the TextAttack library is used to generate new fake reviews. This augmenter replaces words in a sentence with their synonyms from the WordNet lexical database. This ensures that the meaning of the review is preserved while altering the text slightly, making it different from the original content.
  - **Generating Augmented Fake Reviews**:
    - The system randomly selects fake reviews from the dataset.

- For each selected fake review, the **WordNetAugmenter** generates multiple variations (augmented reviews).
- The procedure is repeated until the quantity of fabricated reviews equals the amount of authentic reviews (12,916).

  Augmentation Example:

  Original fake review: "This product is terrible, it broke after one use."

  Augmented version: "This product is awful, it shattered after one use."

- **Final Dataset**: After combining the augmented data, the final dataset now has 25,832 reviews, 12,916 of which are real and 12,916 of which are fake. The augmented data is balanced, and the final dataset is saved in a CSV format. Machine learning models will be developed utilizing this balanced dataset, guaranteeing that the model is not biased towards the real reviews.

Table 3.1 Fake review dataset analysis after applying data augmentation

**(b) Product Category Augmentation Process:** To ensure that the product category dataset is balanced, augmenting it involves generating additional reviews for categories with insufficient data. Below is a description of the augmentation process used for product category labeling.

- **Initial Dataset Analysis:** The dataset originally contained imbalanced data, with categories like "electronics" and "automotive" having significantly higher numbers of reviews than others like "fashion," "home," and "health." To create a balanced dataset, we need to augment the reviews of the categories with fewer reviews.

  Before Augmentation:

  - Electronics: 5606 reviews
  - Automotive: 5062 reviews
  - Fashion: 1391 reviews
  - Home: 642 reviews
  - Health: 215 reviews

- **Augmentation Strategy:** We used **Contextual Word Embeddings** via **BERT (*'bert-base-uncased'*)**, a model capable of understanding the context of the words in the sentence and providing suitable synonyms to augment the review texts. Specifically, **nlpaug's ContextualWordEmbsAug** was used for augmentation, which inserts contextually relevant words into the reviews.

  Augmentation Process:

  - For each category, Verify the quantity of reviews required to attain the desired volume (the maximum number of reviews from any category).
  - If the number of reviews for a category was insufficient, we used ContextualWordEmbsAug to generate additional reviews by modifying the words within the existing reviews.

- The augmentation continued until the category had enough reviews to match the maximum size.

- **Dataset Recomposition:** After the augmentation process, the augmented reviews for each category were combined with the original reviews. This step resulted in a balanced dataset where every product category had an equal number of reviews (5606 reviews per category). The balanced dataset was then saved, ensuring that all categories now had the same data for training machine learning models, thus preventing any category from being underrepresented.

Table 3.2 Analysis of the category classification data after augmentation

**(c) Sentiment Review Augmentation Process:** The sentiment analysis dataset's augmentation procedure aims to uniformly allocate the quantity of positive, neutral, and negative evaluations. Below is the detailed explanation of how the augmentation is done:

- **Initial Dataset Analysis:** The original dataset consists of imbalanced reviews, with more positive reviews than neutral or negative reviews. The initial counts were as follows:
  - **Positive:** 8801 reviews
  - **Neutral:** 2101 reviews
  - **Negative:** 2014 reviews

  This imbalance can affect the training process of sentiment analysis models. To address this, we need to augment the reviews of the **neutral** and **negative** categories to ensure that each sentiment category possesses an equal quantity of reviews, as many as the positive category (8801 reviews).
- **Augmentation Techniques:** To augment the reviews, we used **TextAttack**'s augmenters, specifically:
  - **WordNetAugmenter:** This augmenter uses WordNet to generate semantically similar words by replacing words in a sentence. It is particularly effective for augmenting negative reviews as it can preserve the negative sentiment while introducing some variations in wording. Negative reviews are augmented using the WordNetAugmenter.
  - **EmbeddingAugmenter:** This augmenter modifies reviews using word embeddings to introduce small variations while preserving the sentiment. It is used for neutral reviews. Neutral reviews are augmented using the EmbeddingAugmenter.

  Each review in the target category is augmented until the total count reaches the required size.
- **Combining Original and Augmented Reviews:** Once the augmentation process is complete, the augmented reviews are combined with the original reviews. After applying the augmentation process, the dataset is now balanced, and all three sentiment categories have an equal number of reviews (8801 reviews per sentiment). The augmented dataset is now ready to train a sentiment analysis model.

Table 3.3 Final analysis of the sentiment classification data after augmentation

### 3.1.3 Model Training and Evaluation Analysis

**(a) Data Splitting:** Preparing the dataset for model training and evaluation occurs subsequent to its labeling and enhancement. Training and test data are the categories into which each labeled dataset is separated. The test data is segmented to assess the model's efficacy, while the training data is employed to develop the models. In this case, 40% of the data is utilized for model testing, and 60% is used for training. We selected the random seed 42 as it is frequently utilized in machine learning research and practice due to its historical popularity and ease of reproducibility. Although the specific value of the seed does not affect the randomness itself, using a fixed value ensures that results remain consistent across multiple runs, facilitating debugging, comparison, and reproducibility of experiments.

- **Training Dataset**: The training set includes a subset of the labeled reviews, used to teach the models how to detect fake reviews, classify product categories, analyze sentiment, and generate automated replies.
    - For the fake review detection model, among all reviews, 7000 real and 7000 fake reviews are used for training.
    - For the category classification model, 17500 reviews are used for training, with an equal portion of automotive, electronics, fashion, home, and health-related categories.
    - For the sentiment analysis model, 16500 reviews are used for training, where positive, negative, and neutral sentiment labels have equal numbers of reviews.
- **Test Dataset**: The test set assesses the performance of each model post-training. This facilitates evaluating the model's generalizability to new, untested data.
    - Where the fake review detection model is tested with 10000 reviews, having an equal portion of fake and real reviews.
    - A category classification model is also used, with 10,500 reviews for testing the model, where each category has an equal portion.
    - The sentiment analysis model utilized 9,000 reviews, evenly distributed across positive, negative, and neutral categories.

The system obtained training and testing data by splitting modules in equal ratios of 60% for training and 40% for testing. When applicable to machine learning practice, this ratio provides adequate data points to train deep learning models alongside an unbiased testing sample. The class distribution received equal focus in training and testing sets by maintaining balanced enrollment of target categories (fake versus real and positive/neutral/negative values, together with five product types). The models required this approach for effective learning and generalization because it eliminated bias toward dominant classes. The evaluation method becomes more reliable and fairer through balanced splitting, following standard supervised learning best practices.

**(b) Model Training:** The models employed in this thesis are fine-tuned using the labeled training dataset. The following models are trained:

- **Fake Review Detection Model**: Fine-tuned to categorize reviews as fake or real based on the review content and metadata.

- **Category Classification Model**: Fine-tuned to predict the correct product category (automotive, fashion, home, electronics, health) based on the content of the review.
- **Sentiment Analysis Model**: Fine-tuned to classify reviews into sentiment categories (positive, neutral, negative) based on the language and tone used in the review.
- **Reply Generation Model**: Fine-tuned to generate context-specific replies based on the sentiment, product category, and review content.

**(c) Evaluation:** The test dataset assesses the models once trained. The performance metrics used for evaluation include:

- **Accuracy:** This denotes the ratio of correct predictions made by the model. It serves as a comprehensive metric that provides a general assessment of the model's efficacy in classification tasks.
- **Recall, Precision, and F1-Score:** These metrics are calculated to evaluate the performance of the models for each category (real/fake for fake review detection, category labels for product classification, sentiment labels for sentiment analysis, etc.).
    - **Precision** measures the ratio of true positive events correctly identified by the model.
    - The quantity of actual positive cases that the model correctly identified is referred to as **Recall**.
    - The **F1-score** reconciles precision and recall by calculating their harmonic mean.
- **Confusion Matrix:** The confusion matrix encapsulates true positives, false positives, true negatives, and false negatives for each classification task. The matrix helps visualize how the model handles each class and shows any issues related to misclassification.
- **Evaluation of Reply Generation Model:** The reply generation model's performance is being assessed by several natural language generation metrics:
    - **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Determines the n-gram overlap between the reference and generated answers. Greater ROUGE scores indicate more overlap and better-generated content.
    - **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Compares the quality of the generated response with the reference response, taking synonyms, stemming, and word order into consideration.
    - **BERTScore:** Compares the semantic similarity of the generated and reference responses based on pre-trained BERT embeddings. Higher score, indicating more semantic coherence and relevance of the generated response to the input review.