ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis – Data Science and Marketing Analytics

# Decoding Customer Sentiments: Utilising advancements in text analytics to extract strategic insights from E-Commerce reviews.

Name: Aris Niamonitakis

Student ID number: 557699

Supervisor: Carlo Cavicchia

Second assessor: Radek Karpienko

First Draft Version: 01/08/2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam

# Abstract

This thesis investigates the analysis of customer reviews using text analytics and machine learning modelling within the e-commerce sector of the clothing industry. As digital consumer feedback becomes more prominent, businesses face the challenge of guiding their strategic decisions using large amounts of unstructured data. The study addresses this challenge by developing a predictive model that combines deep learning and machine learning techniques to extract actionable insights from customer reviews. The aim is to improve customer satisfaction by utilising direct consumer feedback to guide business strategy.

The research employs sentiment analysis and topic modelling to interpret review data. The data is initially prepared by being translated into English. The text is then assigned a sentiment score to record the emotion portrayed within it numerically. Similarly, various commonly discussed topics are recorded in binary form. Finally, the presence of topics, customer demographic data and purchase data are used as predictor variables in a regression to predict the sentiment score. The finalised regression reveals how different product features and demographic data interact with the satisfaction displayed within the reviews, showing customer preferences and potential areas to improve upon.

The model was trained using a combination of customer review data from various online retailers and synthetic review data generated via large language modelling. This approach was selected to ensure the model could handle the diverse reviews that emerge, enhancing its effectiveness across different situations, products, and retailers.

The trained model was tested on consumer product reviews scraped from Zara, a website with a wide selection of products that serve a different purpose than those used to train the model, as to avoid overfitting. The testing showed high accuracy in the model’s ability to correctly translate reviews, apply a suitable sentiment score to them, and correctly identify the presence of topics within them, proving that the model can be deployed successfully in the environment.

Finally, the model analysed a dataset of 7500 customer reviews discussing clothing they purchased from the athletic apparel brand adidas and used its results to identify pathways to guide company strategy based on direct consumer feedback. The analysis showed that the model’s results can be interpreted to guide product, advertising, and pricing strategies and how to segment and target the market optimally.

The research acknowledges limitations related to the static nature of dataset used and suggests directions for future work to build upon its findings. Regardless, this study reached its goal of advancing the methodological landscape in e-commerce analytics and providing a strategic framework for businesses aiming to leverage customer feedback for competitive advantage.

# Contents

# Chapter 1: Introduction

This thesis aims to produce a solution that addresses the necessity for companies to utilise reviews to guide their product strategy to survive within today’s e-commerce industry whilst tackling the issue of big data producing more information than it is humanly possible to interpret manually.

## 1.1 Introduction to the E-Commerce Industry

Since the inception of the Internet in the last two decades, the e-commerce market has grown immensely competitive, due to to multiple vital elements, most of which result from globalisation and technological advancements.

These factors include lower barriers to entry. The availability of e-commerce platforms such as Shopify has made it easy for new businesses to enter the market and make enough sales to profit and grow. As the presence of such platforms has caused the initial investment cost necessary to start a business to decrease significantly, small businesses and even individuals can set up their online shops quickly without taking on significant financial risks; hence, the e-commerce market is more saturated and competitive than ever. Furthermore, as the internet is becoming more accessible to people each year, the e-commerce market is seeing steady growth, making the e-commerce industry an attractive endeavour for businesses to diversify into. Finally, globalisation efforts made by governments have caused global shipping to become accessible and cheap, meaning that e-commerce websites can advertise globally and have the whole world as their target market, multiplying the number of potential sales compared to local-based businesses.

These e-commerce industry innovations combined have transformed the market into a competitive environment where businesses must consistently innovate and differentiate themselves to remain relevant. As the industry gets more crowded, the two types of companies that will survive will be those that produce unique products not found elsewhere and provide consumers with the highest-quality products in the market. Companies must keep up with consumer demand and global market trends to grow efficiently and turn a profit.

## 1.2 Introduction to Reviews

The e-commerce industry has shown that understanding what consumers want in a product and keeping up with their changing behaviour is essential for a business’s survival. As e-commerce platforms evolve and market saturation increases, customer reviews have become critical for companies to understand their consumers. Consumer reviews reflect the consumer's voice and serve as feedback for companies looking to differentiate themselves from the crowded market. Understanding and responding to customer reviews is essential in the ongoing adaptation and innovation required to grow in the competitive e-commerce landscape.

Each coming year further cements the norm that reviewing a product has become part of the customer experience journey. Whereas previously, one would purchase a product, repurchase it if they were satisfied, and otherwise return it, company-driven incentives and the rising popularity of online discussion have resulted in customers leaving a short description of their experience using the product.

Companies incentivise their customers to leave reviews on the products they buy because management sees reviews as a powerful tool in guiding product strategy. A well-crafted review will discuss the product's qualities and what features need to be improved upon; it conveys customers' emotions at different points of the purchase experience and what changes they would have preferred.

Using mass consumer feedback to design the product will lead to the target market responding positively. Qi et al. (2016) focus on the usefulness of reviews in product improvement processes and extracting product quality-oriented information. The study cites a positive relationship between product improvement based on customer feedback and brand awareness. Gensler et al. (2015) conducted a case study on McDonald’s that revealed the strategic significance of using customer feedback, particularly negative feedback, as a tool for brand improvements.

Although companies know that customer reviews offer invaluable insights into product quality and customer satisfaction, the sheer volume of data generated daily presents a formidable challenge in analysing it.

## 1.3 Introduction to Data

Companies have been able to get ahead of their competitors by leveraging data-driven insights to shape their marketing and product strategies. However, the large volume of data generated daily have created a challenge for companies as it becomes increasingly difficult for humans to analyse them manually to reach meaningful conclusions.

It is estimated that millions of user reviews are constantly being generated across various platforms. Yelp accumulated 229 million reviews in the first quarter of 2021. It is unrealistic for companies to go over the user review data manually; companies will need to shift their priorities to utilising data science to produce the appropriate models to effectively analyse large amounts of data automatically. Doing so will allow businesses to understand the underlying product qualities of discussion and emotions that drive the production of the reviews.

## 1.4 Thesis Goal

This study proposes creating and applying a machine learning model dedicated to text analytics to extract insights from review data quickly. This approach enables businesses to analyse and understand extensive textual data effectively, promptly uncovering customer insights and market directions. The goal is to allow organisations to make quick, informed decisions, improve customer engagement, and secure a competitive stance in the ever-evolving e-commerce market with the click of one button. As the features that determine product quality vary significantly between product types, the research will be based on the most sold product family within the e-commerce space, clothing.

To ensure that this thesis applies to real-life cases, customer product reviews extracted from the various online domains of adidas will be used to assess the reliability and credibility of the methods used in this thesis.

## 1.5 Relevance of Study

In the current era of exponential data generation, a deep-dive study into extracting information from customer review data and improving upon its drawbacks is relevant from multiple perspectives—business, societal, and academic.

### 1.5.1 Scientific Relevance

Regarding academia, this thesis aims to add to the research by deep-diving into previous research regarding text analytics in customer reviews and improving upon their drawbacks and limitations. As technology evolves at a rapid pace, it is possible that past research limitations can be solved by using advancements made in technology and newly produced statistical methods.

### 1.5.2 Social Relevance

From a social perspective, quickly and effectively understanding customer reviews can significantly improve product quality and the shopping experience for everyone. Businesses can now instantly find out how customers feel and what they are talking about, allowing them to fix any issues immediately. This results in better clothing products and happier customers. Consumers will receive higher-quality products and services, and the market will be shaped by clothes created based on consumer-centric feedback. Over time, these improvements can lead to a better standard of living and more power for consumers as companies become more focused on meeting their customer's needs and preferences.

### 1.5.3 Business Relevance

For businesses, the impact of this study is significant. By using automated systems to analyse customer reviews, companies can improve their image and make their products more appealing to consumers. Getting quick updates on customers' thoughts and noticing new trends helps businesses make smart decisions fast. This way, they can create clothing products and marketing plans that meet customers' wants. Quickly responding keeps current customers happy and draws in new ones by showing that the company cares about customer satisfaction and is always working to improve. Also, being one of the first to act on customer feedback and innovate based on what customers say can give a company an edge over others in busy markets, ensuring the business stays relevant and competitive as things change online.

## 1.6 Research Questions

### 1.6.1 The Main Research Question

In the changing world of data science and its use in business, figuring out the best mix of large-scale text analysis methods to pull useful information from customer reviews is becoming increasingly important. The fast progress in text analysis technology has dramatically improved our understanding of unstructured text data, opening new doors for companies to use customer feedback for a competitive edge on the market. However, the challenge is still to find the right combination of techniques that can provide clear business insights from the correct information in customer reviews, a difficult task considering the wide variety of information found within them. This study aims to fill that gap by blending text analysis tools to get the most relevant business insights.
Therefore, the following main research question has been formulated:

**What is the optimal mixture of text analysis techniques to extract the most effective business-focused insights from customer reviews for an e-commerce-based retailer?**

### 1.6.2 Research Sub-Questions

The main research question has been broken down into several sub-questions to explore the main research question thoroughly. These questions help understand different aspects of analysing text, like customer reviews, for businesses:

The research aims to find and assess the most practical ways to understand and extract the feelings and topics in customer reviews for further analysis. Therefore, the first sub-research question is:

**What are the most effective methods for analysing unstructured text data to accurately determine the emotions and topics in consumer opinions?**

The research aims to uncover the challenges in pulling out feelings and topics from customer feedback and determine whether these challenges can be overcome with newly released technology or applications that have not been attempted yet; the next sub-research question is:

**What are the roadblocks behind extracting sentiment and topics from customer reviews, and can they be overcome?**

To understand how best to use automated insights tools in a business setting, the difficulties and limits of using instant analysis to help shape business strategies must be identified; therefore, the next sub-research question is:

**What are the challenges and limitations of analysing online consumer opinions for business strategy development within the clothing e-commerce sector?**

This thesis aims to simplify the process of analysing reviews. Therefore, another sub-question to research would be based on how the process of analysing reviews can be simplified for someone not specialised in data science to utilise the techniques shown within this paper to extract insights from their business's reviews. Therefore, the final sub-research question is:

**How can a user-friendly tool be developed to enable non-specialists to analyse customer reviews and extract critical business insights effectively?**

## 1.7 Ethical Issues

Studies in data science tend to contain a range of ethical issues as they use large amounts of real-world data, which must be collected and used ethically and with the individual's privacy in mind. As such, the data utilised will be masked to respect the confidentiality of the adidas consumers whose purchases are found within the datasets.

## 1.8 Research Limitations

Due to its exploratory and practical nature, the research is expected to run into various limitations. These potential obstacles must be considered and addressed to ensure the credibility and integrity of the research outcomes.

The limitations of this thesis primarily arise from the characteristics of the review dataset utilised for analysis. As an ad hoc dataset, it consists of a static collection of reviews and does not update in real-time. This static nature may not accurately capture the evolving dynamics of customer feedback over time.

Additionally, the dataset's scope is limited to clothing e-commerce reviews. This limitation restricts the findings' applicability across different sectors, such as hospitality, services, or local businesses, where customer reviews are also crucial. The specific nuances and ways language, sentiment, and topics are expressed and deemed relevant in these sectors may not be well-represented by a dataset solely focused on clothing, thus limiting the broader applicability of the research outcomes.

Moreover, only public text analysis methods are explored in the thesis due to the unavailability of private models for replication and analysis. Private models, which corporations and research institutions often develop, could employ more advanced techniques or proprietary algorithms, potentially enhancing the accuracy and depth of text analysis. The exclusion of these models means that the research will not incorporate the latest advancements in text analytics, which could offer improved performance or new insights. The reliance on publicly available methods limits the exploration of techniques accessible for academic research, possibly excluding cutting-edge approaches that are not publicly shared.

# Chapter 2: Literature review

The literature review aims to gain insight into the process of preparing review data for analysis and the various techniques that can be used to analyse and extract interpretable information from the said review data to guide strategy.

## 2.1 Extracting Insights from Text

A combination of text analysis techniques should be utilised to effectively guide product strategy using customer reviews to extract how customers react to different product features and the consumer journey (Liu, 2012). This approach involves dissecting how customers feel about various aspects of the product and their overall experience. It allows businesses to make informed decisions based on comprehensive customer feedback by producing a model that predicts how happy a customer feels regarding a product based on the features they mention when discussing it.

The topic analysis should be deployed to identify the aspects of the product discussed in different reviews (Blei et al., 2003). By placing the specific aspects of the product that customers are discussing, this technique categorises the content of reviews into distinct topics. These topics can range from product features to service aspects. By categorising the data into a binary format, one if a particular subject is mentioned and 0 if otherwise, the product elements present within the reviews can be identified on a mass scale.

To extract the emotions a customer expresses within their review, sentiment analysis must be utilised (Pang & Lee, 2008). Sentiment analysis identifies the feelings expressed in the reviews. It extracts them as a numerical outcome that can be easily interpreted to understand whether the customer is speaking positively or negatively within a review.

To understand how customers feel about specific product features, the sentiment score of a review can be predicted based on the presence of topics within the text through predictive modelling such as regression analysis (Hastie et al., 2009).

The sentiment score is the dependent variable (Y) in this approach. At the same time, the presence of specific product aspects in the text is treated as independent variables (X), coded as binary indicators (1 if the element is mentioned, 0 otherwise). Through this method, if the regression shows a positive significant correlation between mentioning a product aspect and sentiment scores, it suggests that the aspect positively impacts customer satisfaction. Conversely, a negative correlation would indicate areas where the product fails to meet customer expectations.

Businesses can transform raw review data into actionable insights using these analytical techniques. By understanding both what customers are talking about (topic analysis) and how they feel about it (sentiment analysis) and predicting the impact of specific product features on customer satisfaction (regression analysis), companies can strategically address the strengths and weaknesses of their offerings, thereby enhancing the overall customer experience and boosting product strategy.

For the final regression to be produced, the ideal models for topic and sentiment analysis must be chosen based on previous research and context, and the textual data must be prepared as necessary to ensure the most robust results are extracted (Manning et al., 2008).

## 2.2 Preparing the Data for Analysis

### 2.2.1 Data Cleaning

Text data preparation is the practice of cleaning and processing raw text to convert it into a format that can be analysed. Data cleaning is a necessary stage in the preparation of text for analysis. (Rahm & Do, 2000).

The removal of duplicates is a primary step in the data-cleaning process. Duplicate data entries can hurt the analysis by incorrectly overrepresenting certain information because they appear multiple times, leading to biased outcomes. Moreover, duplicates may cause analytical models to overfit, causing the model to overperform in predicting the data it is trained on but failing to accurately predict unseen data, a fatal flaw as the goal of this research is to produce a model that can be applied to various clothing-related reviews from a variety of retailers.

Handling missing values is another critical part of the cleaning process. Many datasets inherently contain gaps or missing entries, which, if unaddressed, can lead to incomplete analyses and incorrect conclusions. Analytical techniques and machine learning algorithms typically require complete datasets to operate effectively; hence, addressing missing values by removing them from the dataset entirely if the data size allows it or replacing the value with a statistically average value is necessary to maximise the predicting capability of an analytical model (Davenport & Harris, 2007).

Standardising text ensures consistency across the dataset by converting all text to the same case and removing redundant spaces, punctuation, and non-relevant characters. Standardisation used to be necessary for accurate data comparison. For example, variations in the representation of the same word (e.g., “Email,” “email,” “E-mail”) are treated as identical only after standardisation, thereby avoiding the fragmentation of data. However, in today’s age, text analytics models are built to operate without the necessity for standardisation, and therefore, whether standardisation is applicable depends on the model(s) used to analyse the text (Kandel et al., 2011).

Finally, text data can be cleaned by separating it into smaller, more concise sections. A large block of text tends to contain information regarding various topics and may even contradict itself, making it difficult for machines to correctly dissect it. Through separating large blocks of texts into paragraphs or sentences, machines will be able to read each section separately and interpret it more effectively whilst also providing superior insights (Davenport & Harris, 2007).

### 2.2.2 Spelling and Grammatical Mistakes

Spelling and grammatical errors can lead to the misclassification of data and inaccurate results from subsequent analyses. Correcting grammatical errors ensures that the data is accurately represented and interpreted. Grammatical errors tend to result from human-made mistakes and although minor, they can change the context and meaning of phrases written. Models, such as SymSpell, have been produced to detect grammatical errors within text data and fix them, and text data should be run through such models before being analysed (Socher et al., 2013).

Finally, it should be noted that most text analytics models are monolingual. They are built and can only be used to analyse the text of a single language - as a result, text data must be translated into one common language before analysis is conducted. Choosing the language of translation can vary based on context. Still, most text analytics tools were developed within the USA and the UK, where the most spoken language is English. Thus, the models are specialised for that language- as a result, translating text into a language that is not English may limit the analysis from most of the analytical tools (Agarwal et al., 2011).

These data-cleaning techniques form the foundation of reliable and valid text data analysis, ensuring the final dataset is well-prepared for detailed examination and modelling.

### 2.2.3 Data Preparation

Once the data is cleaned, it must be transformed into a format suitable for analysis. Tokenisation is a process in which a text is broken down and divided into words and phrases that can then be analysed separately; it allows for the structured analysis of text by transforming the text so that subsequent Natural Language Processing tasks can be applied to the original raw text (Manning et al., 2008).

Lemmatisation is a technique used to reduce words to their base form, converting varying words that possess the same meaning into one universal format. It involves comparing each word against a dictionary to transform each word into the base form shown in the dictionary. Lemmatisation reduces textual data's complexity, resulting in improved performance of NLP tasks (Hastie et al., 2009).

Stop word removal involves eliminating common words from text data that may appear frequently but offer little to no value in understanding the meaning of the content. Such words include ‘and’, ‘the’, and ‘a’; by removing these words, subsequent data analysis will not be clouded by their inclusion and will thus focus on words that carry more significant meaning in understanding the text (Manning et al., 2008).

Another data preparation technique is Vectorisation. It is a process for converting text into a numerical format, making it possible for traditional mathematical models to extract information from text. The most utilised vectorisation technique is the Term Frequency-Inverse Document Frequency approach, where the weight of terms depends mainly on how often the said term appears within the different documents of text. A term that appears in most documents would not provide insight into what differentiates a document from the others and thus is applied to a lower weight. TF-IDF transforms text into a form suitable for various machine-learning algorithms, facilitating classification, clustering, and sentiment analysis (Rahm & Do, 2000).

It should be noted that these data preparation techniques vary depending on the models utilised. Some modern machine-learning models do not need the data to be prepared; therefore, the preparation techniques rely mainly on the deployed models (LeCun, Bengio, & Hinton, 2015).

## 2.3 Sentiment & Topic Analytics

### 2.3.1 Sentiment Analysis

Sentiment analysis is a field of text analytics focused on extracting the emotion portrayed within a text. It interprets the text to understand whether the subject(s) discussed within it are mentioned in a positive, neutral, or negative light (Pang & Lee, 2008).

Although initially, sentiment analysis took a simple approach of labelling words with a numerical value from -1 to 1 that represents whether a word carries positive (1), negative (-1), or neutral (0), and afterwards adding the scores within a text up to extract the overall sentiment, the complexities of human language held back this method; irony, metaphors, and context are essential to understand an emotion being portrayed within a text, and rule-based approaches are unable to capture them correctly. For example, in the sentence 'Dutch weather is amazing; I love heavy winds and frequent storms', the usage of sarcasm will cause the dictionary approach to incorrectly label the negative sentiment due to the words 'amazing' and 'love'. Similarly, describing a coat as heavy is positive, but describing a laptop as heavy tends to be negative; however, both statements would produce the same sentiment score when analysed using the dictionary approach (Socher et al., 2013).

Furthermore, the existence of negations, amplifiers, and double negatives leads to a phrase's sentiment being different from that of the dictionary labelling. The phrase "I am happy" has a positive sentiment score because the word happy is positive; however, "I am not happy" is negative despite using the word ‘happy’. This limitation can lead to inaccurate sentiment assessments, affecting the overall analysis quality (Pang & Lee, 2008).

Machine learning methods account for the complexities of human language. The current state-of-the-art machine learning model in text analytics is RoBERTa, designed to improve how computers interpret and generate human-like text (Devlin et al., 2018).

RoBERTa is an advanced language processing tool that META developed to enhance how a machine can understand written text. Unlike previous models that linearly go through the text, RoBERTa is built on the foundation that it should read complete text documents at once to understand the relationship between different words and phrases and, as a result, capture context more accurately than previously assumed possible. RoBERTa has been pre-trained on a large, diverse corpus of internet text and further fine-tuned for improved general language understanding, being, therefore, exceptionally proficient at tasks demanding deep text comprehension—sentiment analysis, content summarisation, and question answering (Devlin et al., 2018).

RoBERTa can be fine-tuned to perform specific tasks after its initial pre-training whilst retaining the previously rigorous training it received. As a result, an extra processing layer can be inserted into the model to specialise it and conduct a variety of text analytics-related tasks such as sentiment analysis. The final layer can be trained on labelled datasets for further specialisation. Therefore, when context is necessary for producing correct predictions, RoBERTa is a powerful solution (Devlin et al., 2018).

### 2.3.2 Machine Learning Sentiment Analysis

Machine learning-based sentiment analysis utilises pre-labelled text data with sentiment to predict the sentiment of new texts. It is a method of supervised learning; a model is fed a large dataset of text data with predefined sentiment labels, leading to the model identifying sentiment patterns and accurately determining the sentiment of the text regardless of the complexities of human language (Pang & Lee, 2008).

One of the most robust models available that can be used to label reviews with their corresponding sentiment is the ‘cardiffnlp/twitter-RoBERTa-base-sentiment’ model. This model is fine-tuned to perform sentiment analysis, categorise text into positive, negative, and neutral labels, and produce a ‘sentiment score’ for each text analysed. The model was trained on 10,000 unique English-written tweets that the founders of the model manually labelled. The rigorous training done to the model so that it can precisely extract sentiment from published online content dramatically enhances its ability to detect subtle expressions of sentiment within that field, making it suitable for application within other types of online content, such as reviews (Pang & Lee, 2008).

### 2.3.3 Type of Topic Analytics

Topic analysis is used in text mining to identify and categorise a text's main themes or topics. Depending on the specific goals of the analysis, this analysis can be conducted with either predetermined or non-predetermined topics (Blei et al., 2003).

When topics are predetermined, the process involves classifying text into predefined categories based on the content included within the text. Pre-determined topic analysis often uses a supervised machine learning approach, where the model is trained on a labelled dataset to understand the difference between text found within the said topic category and text not. The pre-determined topic analysis is utilised in scenarios where the topics of interest are known ahead of time, such as monitoring for specific issues in customer feedback (Blei et al., 2003).

On the other hand, when topics of interest are not known ahead of time, unsupervised learning techniques such as Latent Dirichlet Allocation are used to discover hidden themes within extensive text data. Such topic analysis interprets the words in the documents to infer topic probabilities. This approach is precious for exploring large datasets with unknown themes, providing insights that can guide further research or business strategy (Blei et al., 2003).

Although both methods offer powerful ways to extract actionable insights from text by understanding the prominent themes that emerge from unstructured data sources, in the context of producing a specialised tool that can predict the sentiment score of a sentence based on the topics within it to determine the product's strengths and weaknesses, pre-determined topic analysis seems superior.

Pre-determined topic analysis allows for a more focused examination of specific areas of interest related to the product. In the clothing e-commerce sector, these areas are related to product quality and customer experience. By defining topics relevant to the product's features or aspects of the customer experience journey, the analysis can directly address these areas, making the insights more applicable and actionable for product development and marketing strategies (Blei et al., 2003).

Furthermore, pre-determined topic analysis allows for more straightforward comparability of the results across different items and datasets. The consistency in topics due to the pre-determined nature of the study allows for comparative analysis, such as tracking changes in sentiment over time or comparing sentiment across different products or services. It provides a standardised basis for measurement, making it easier to draw reliable conclusions to guide strategy (Blei et al., 2003).

## 2.4 Pre-Determined Topic Analysis

### 2.4.1 Deep Learning Modelling

A robust model type for pre-determined topic analysis is the deep learning model. Deep learning models are advanced algorithms within machine learning that use structured layers of artificial neural networks to process data and make predictions. They can also be used for classification tasks, such as labelling text based on whether it discusses a specific topic (LeCun et al., 2015).

The data is inserted into the deep learning model through the input layer and then processed through a pre-determined number of subsequent layers. Each layer comprises several nodes, all used to process and analyse the data before passing it onto the nodes in the next layer for further analysis. The data is continuously processed from layer to layer until it reaches the output layer, where the data is transformed into a form suitable for the task the deep learning model is built to solve. This structured approach allows the models to handle complex data transformations, making them particularly effective for analysing datasets regardless of form and size.

Deep learning models mimic how the human brain learns; they improve autonomously by identifying prediction errors and adjusting their parameters to improve accuracy in their next predictive attempt. The layer setup and the ability to learn from its own mistakes to self-optimise make deep learning models effective for highly complex tasks such as text analysis, where they can detect subtle differences in language. Furthermore, the flexibility of deep learning models enables them to adapt to new data and changing conditions, maintaining their predictive ability over time in situations where they need to adjust, such as analysing trends and consumer preferences over time (LeCun et al., 2015).

### 2.4.2 Deep Learning Model Optimisation Techniques

Deep learning models can be fine-tuned through adjusting their parameters to optimise their ability to analyse data and increase their predictive accuracy (Smith, 2021).

The learning rate determines how much the model's processing weighs in response to errors- an appropriate learning rate ensures that the model learns efficiently without taking too long or settling for a less optimal solution (Johnson, 2020).

An epoch is one complete cycle of the training data through the model. This parameter must be fine-tuned as increasing the number of epochs improves predictive accuracy due to the model having more opportunities to assess and learn from the data. However, it can also lead to overfitting if the model is trained on too many cycles, causing the model to predict new data with low accuracy (Lee, 2019).

The batch size is the number of data examples the model processes before updating its weights. A smaller batch size can lead to better performance and lower error rates. However, minimising the batch size below a certain point will negatively impact the model's predictive accuracy as it will cause instability in the training (Brown, 2018).

The correct settings for learning rate, epochs, and batch size are essential for the model to learn effectively and make accurate predictions (White, 2022).

### 2.4.3 Deep Learning Model Training

For a deep learning model to have high predicting accuracy, especially when dealing with complex tasks like language analysis, it is necessary to train the model on a balanced and diverse dataset (Buda et al., 2018).

A balanced dataset helps prevent the model from becoming biased and ensures it can correctly detect minority classes. A model trained on an unbalanced dataset tends to be biased towards the most commonly appearing class within the training data and misclassifies minority classes as a result (Japkowicz & Stephen, 2002). By having a balanced dataset, the model learns to recognize and accurately classify all classes, resulting in higher predictive accuracy (He & Garcia, 2009).

A balanced dataset also improves the model's ability to generalise; the model learns the patterns and features of all classes, allowing it to identify them better in data that it has not been trained on and label them correctly (Zhou & Liu, 2005). As a result, the model will perform with higher accuracy when analysing unseen data (Sun et al., 2009).

Synthetic data can help balance datasets and allow models to be trained on a large and diverse datasets. Creating synthetic data increases the number of training examples, giving the model more information to learn from (Esteban et al., 2017). This is useful for underrepresented classes and rare cases that are hard to come by in an authentic dataset. By generating synthetic samples of underrepresented classes, the dataset becomes more balanced, ensuring that the model is better prepared for the analysis of diverse situations (Douzas & Bacao, 2018). Furthermore, training a model for deployment in text analytics requires a large volume of specific scenarios for the model to be trained on to effectively operate, as there are various distinct ways to express a class within text, making synthetic data a strong addition to training deep learning models for application in text analytics. Through utilising advanced large language models, such as GPT-4, generating synthetic data that closely resembles authentic data is possible through prompt engineering (Brown et al., 2020).

## 2.5 Predicting Customer Satisfaction

Regression analysis is used to identify the relationship between a dependent variable and one or more independent variables. Predicting the customer satisfaction level, in the form of a numeric sentiment score, based on the presence of topics will reveal which topics can significantly impact the sentiment score, and their coefficient value will show whether the said impact is positive or negative. Furthermore, by comparing the coefficients, it can be interpreted which variable has the most significant impact on sentiment score and, thus, customer satisfaction (Hastie et al., 2009).

## 2.6 Use Cases to Predict Market Strategy

The research's modelling phase will gather insights to guide the company's strategic direction. The final part of the literature review will assess how text analytics can influence strategy. These methods will be applied to adidas's data to show how the model can use text analytics and reveal its potential to provide valuable insights to clothing retailers (Davenport & Harris, 2007).

### 2.6.1 Product Augmentation and Innovation

By analysing the contents of raw, unfiltered reviews left by consumers engaging with the brand, companies can understand what features of their products and overall shopping experience are generally liked and disliked by the public. This type of consumer-direct feedback allows companies to produce new product versions that align with the customers' wishes, leading to higher customer satisfaction and loyalty. Furthermore, companies can use consumer feedback on feature-overlapping products to brainstorm new product ideas that the masses will receive positively upon their initial release to the market (Davenport & Harris, 2007).

### 2.6.2 Competitive Advantage Analysis

Combining sentiment and topic analysis insights allows businesses to understand their products and competitors' products. With the most publicly available review data and easy-to-web-scrape, businesses can access and analyse competitor reviews to understand what drives customer value in competing products. Competitive advantage analysis using text analytics can reveal gaps in the market and superior features of competing products to improve upon, allowing businesses to position themselves to innovate and enhance competitor products to steal market share and drive competitors out of the market (Davenport & Harris, 2007).

### 2.6.3 Market Entry Decisions

Before you go into a new market, you need to understand it. Sentiment analysis gives insights into how the potential customers of competitors see the products. This understanding helps assess if there is room for a new entrant and what unique value proposition this entrant should offer to gain market share effectively (Davenport & Harris, 2007).

### 2.6.4 Customer Segmentation

Filtering reviews by demographic qualities of the customers leaving them and analysing them separately allows for producing a general profile of the demographic, how they react to a product, and what topics drive a positive response from them. Such information allows for targeted marketing and product development strategies, allowing businesses to apply a tailored approach that suits different customer segments, maximising the engagement and conversion rates of advertising campaigns and increasing sales within that demographic (Davenport & Harris, 2007).

### 2.6.5 Monitoring the Brand

By monitoring the customer response to the products and customer experience by measuring their sentiment over time, businesses can interpret the trend to assess whether the business brand has increased or decreased and act accordingly. A sentiment score that negatively trends over time suggests that the business is being perceived worse by consumers than previously, and the marketing strategy should be altered to prevent the public perception from worsening. To understand the root cause of the change in sentiment score, the sentiment score of different product features over time can be interpreted to identify changes in how the public perceives and values them, providing the business with a specific product issue to improve upon or bring to the forefront depending on the trend (Davenport & Harris, 2007).

### 2.6.6 Strategic Business Decisions

Overall, such insights derived from review analysis empower businesses to make decisions in alignment with consumer expectations and market demands. Whether changing their product lines, reshaping their brand strategies, or crafting impactful marketing campaigns, this data provides a robust foundational approach to better business outcomes. These use cases have demonstrated the power of advanced text analytics and how it moves beyond reacting to market conditions to helping anticipate them and plan strategically in response (Davenport & Harris, 2007).

## 2.7 Literature Review Conclusion

This literature review thoroughly examines how to prepare and analyse customer review data to help make strategic business decisions. It outlines essential methods such as topic, sentiment, and regression analysis, showing how text analytics can turn raw data into valuable insights.

It highlights the crucial step of preparing data: cleaning and organising raw text. This preparation is necessary to ensure that further analyses, such as topic and sentiment analysis, are accurate. Techniques like tokenisation, lemmatisation, and vectorisation are essential to prepare the data for natural language processing tasks.

Sentiment analysis, especially with advanced models like RoBERTa, helps identify the emotional content in customer reviews. It lets businesses understand the type of emotions present and their context and intensity, which are crucial to understanding overall customer satisfaction. Topic analysis is vital as it helps categorise what customers are talking about in their reviews, whether product features or service quality. This allows businesses to pinpoint what matters most to their customers. Furthermore, by combining insights from topic and sentiment analysis through regression modelling, companies can predict how changes in product features might affect customer satisfaction. This predictive ability is crucial for planning and making informed decisions.

These techniques are helpful for various business strategies, such as developing new products, entering new markets, segmenting customers, and analysing competition. Understanding the positives and negatives of customer feedback allows businesses to meet consumer needs better and stand out in the market.

In conclusion, this review emphasises the significant role of text analytics in using customer reviews to guide business strategies. By systematically analysing these reviews, businesses can better understand what their customers value, leading to more innovative and effective business decisions. This review sets a foundation for further exploration and improvement in using text analytics for strategic purposes. It provides a practical guideline for the models being built with specific goals, shaping the outline for the research methodology and beyond.

## 2.8 Conceptual Research Model

The existing literature describes how to turn unstructured text data into numerical form and predict its value, making it possible to derive business insights from it to shape strategy on a large scale. By converting the emotions displayed within a review into a numerical score that displays both the polarity and intensity of the text, the aspects of the purchase experience and the review that drive those emotions can be identified through predictive modelling. The research model will aim to identify the drivers of the sentiment score. Doing so will showcase whether aspects of the purchase experience negatively or positively impact the emotions displayed by the customer in their review.

The product aspects discussed in a review are drivers for the emotions displayed within it, and therefore, they must be used as predictors in the regression modelling. The product aspects can be extracted through applying topic modelling to the reviews. The product features tend to be the same when discussing a particular family of products, such as clothing. Therefore, pre-determined topic modelling is more appropriate based on the context than topic discovery models such as Latent Dirichlet Allocation. The presence of the topics will be binary labelled and used as a variable for the regression to predict the sentiment score.

Furthermore, data on the purchase and the customer's demographic will influence the emotions displayed within the review. Therefore, the regression modelling will control various aspects of the purchase and demographics displayed to avoid omitted variable bias. The finalised regression's significant scores will reveal which attributes impact the sentiment score and, thus, consumer opinion, and the coefficients will reveal the intensity of each topic.

The regression results will reveal how product aspects and purchase information influence the emotions displayed in reviews and how customers of different demographic backgrounds react to their shopping experience with the company. These insights will be applied, theoretically, to investigate how they can be used to guide company strategy by allowing customers to refine their product offerings, marketing approaches and market penetration tactics based on direct feedback from consumers.

# Chapter 3: Research Methodology

## 3.1 Data

Although pre-trained machine learning models will be used to translate the reviews, the deep learning model used to identify the presence of topics within the reviews will need to be trained and validated before deployment- hence, two separate review datasets will be gathered to build it. Finally, a third review dataset with information regarding the demographic of the customer and the purchase itself will be used to showcase the combination of the machine learning and deep learning models used in this research to extract insights for business cases.

The deep learning topic-labelling model will be trained and validated using review data collected specifically for use during this research. The research model's capabilities will be validated using a separate dataset that consists of the review text, as well as customer and purchase information.

### 3.1.1 Training Data for Deep Learning Modelling

A combination of real and synthetic data sources will be used to train the deep learning model to accurately identify topics within reviews. This hybrid approach is intended to create a versatile dataset that aligns with the specific objectives of the research, creating a model that can be used to assist various e-commerce clothing retailers in setting their strategy.

The authentic data component will consist of reviews from various online retailers. Data from the very extensively diverse retailers, Amazon & Zalando, will be used to comprehensively represent different product types, price points, and opinions of customer demographics. The selection of these reviews will carefully ensure a balanced representation of sentiments, product types, and consumer demographics. Such diversity is crucial for training a model capable of interpreting and categorising a wide range of real-world consumer feedback accurately.

In addition to genuine reviews, synthetic review data will be generated through advanced language model, GPT-4. Synthetic data offers the advantage of quickly producing large volumes of text data, which is essential when the available real data is insufficient to cover all potential scenarios the model might encounter, as well as producing a balanced dataset to train the model with.

The synthetic review data will be tailored to reflect the specific scenarios the model must be trained to understand. This data will address gaps in the real dataset, particularly regarding rare that might not be represented in the collected reviews.

By integrating real reviews with high-quality synthetic data, the training dataset will be large, varied and aligned to tackle the complexities of text analysis. This approach will ensure the model is well-prepared to handle diverse reviews and deliver reliable insights.

### 3.1.2 Testing Data for Deep Learning Modelling

The model's testing will be conducted using a dataset consisting exclusively of real customer reviews collected from the retailer Zara. Zara is chosen due to its broad assortment of products and diverse price points, making it an ideal source for testing the model's ability to handle varied real-world data effectively. This approach will ensure that the model's performance is assessed based on its response to genuine consumer feedback rather than synthetic approximations.

Zara was chosen as the source for the review data used to test the model to mitigate the risk of overfitting. Since the testing data will be separate from that used in the training phase, the model testing will be done on an unseen dataset, validating the model’s generalisation capabilities. The model will be tested on its ability to apply learned patterns to new, unseen data rather than repeating memorised responses.

Furthermore, synthetic data will not be included in the testing phase. Using only real data will ensure that the testing phase measures how well the model performs in a real-world scenario, which is necessary for the achieving the research of goal of insight extraction from practical applications. The focus on real data will help establish the model's effectiveness and reliability in analysing and interpreting customer discussions from actual review texts.

### 3.1.3 Research Model Validation Data

To validate that the combination of methods used to extract insights is effective, a dataset comprised of customer reviews regarding adidas products, extracted from the brand’s various European domains, will be used. This dataset allows for a detailed assessment of the model across a wide variety of demographics, cultures and languages, as well as across adidas’s extensive product catalogue – allowing for a comprehensive evaluation of the research model developed in the literature review.

Utilising adidas reviews introduces diversity within the reviews from a cultural and language perspective, hence allowing for the assessment of the model's ability to translate and interpret customer feedback that may include regional slang, varying sentence structures, and culturally specific reference an essential feature for model's applicability to global markets.

Table 1 of the appendix describes the finalised dataset and its columns. It includes the review text, data regarding when the review was left on the website, the website itself and financial information regarding the product being reviewed.

## 3.2 adidas Data Table Descriptive Statistics

The adidas review data was chosen for the research model validation section because it represents accurate, unaltered data. Applying the research model to a perfectly balanced dataset will not accurately reflect whether it is effective to use with the imperfect data businesses deal with daily. Below is a breakdown of the adidas review data, showing how a real dataset naturally has imbalances, unlike a 'perfect' dataset typically found in controlled experiments.

Figure 1 in the appendix shows that the monthly distribution of the review data is not constant. There is clear seasonality, shown by the increase of reviews left in the Summer and Winter months – a statistic consistent with the popularity of Summer and Winter sports (Jimenez et al.,2021). Furthermore, figure 1 shows a sharp decrease in the reviews left in the second half of 2022, followed by a steady recovery as the company enters Q2 of 2024, which reflects the company's poor performance in late 2022 because of brand ambassador Kanye West's public anti-Semitic rants demolishing adidas's public image and sales (Morrow, 2023).

Figure 2 shows the rating distribution of the reviews. The distribution indicates a clear bias towards the more extreme scores, with only 9% of the reviews having a lukewarm rating of 3 stars, while the majority gravitate towards extremely high or low ratings. This distribution pattern is common in the field of customer reviews, where moderate opinions are less likely to be shared online compared to extreme ones (Hu et al., 2009).

Figure 3 shows the domain in which the reviews were left. Adidas operates across 19 markets within the EU, each with their own domain and exclusive marketing and product lineups. Most reviews were extracted from the German website (.DE), while the majority were extracted from the English, French, and Spanish websites. Less than 20% of reviews were extracted from other domains. This statistic correlates with the sales distribution of adidas, which shows its largest markets within Europe being Germany, England, France and Spain (Adidas Annual Report, 2023).

The graphs show that the data is not perfectly balanced but rather a result of circumstances impacting company sales, brand image, and seasonal effects, reflecting real-world data's typical complexities and irregularities. The model's ability to turn such unbalanced data into insights will showcase its ability to bring value to businesses operating in realistic circumstances.

## 3.3 Summarised Goal of the Model

This model aims to identify different aspects of the online purchase experience discussed in customer reviews. The model will ideally be able to identify topics within reviews as well as the overall sentiments within them; once the reviews are labelled with a sentiment score and the issues that form the said review, the model will predict the sentiment score based on the presence and absence of issues within the reviews and other product data. By analysing the correlations between sentiment score and the presence of topics and other features of the review and product purchased, the research model should reveal what topics customers are satisfied with or need to be improved upon, guiding the company strategy.

## 3.4 Building the Sentiment Prediction Model Using Reviews

### 3.4.1 Translation

The first step in analysing reviews is ensuring they are all in the same language. Adidas is an international selling firm that operates in more than 160 different markets and has an e-commerce infrastructure in 67. As a result, its product reviews are discussed in various languages; thus, they must all be translated into one before being analysed. As shown in the literature review, ensuring that all languages are written in one language allows for easier and more effective specialisation of the model, yielding more accurate results.

The language chosen for all the text to be translated is English. English is one of the most widely spoken languages in the world, and it is used as a lingua franca in international business, which this review analysis tool is being built to cater to. English is also the most spoken language on the World Wide Web, resulting in most online data that can be used to train or test the model's validity to be written in English. Finally, because of data science being initially pioneered by American and English institutions such as M.I.T. and Oxford, the first natural language processing tools and the resources built upon it are built in English and have been trained on English text; therefore, the best analysis results will be extracted from English reviews.

The process of translating all the reviews to English began by first detecting the language of each review. Langdetect, a tool specialised in detecting the language used to write a piece of text, was utilised to identify the language in which each review was written accurately. This tool automatically assigned a language code, such as 'EN' for English, 'NL' for Dutch, or 'ES' for Spanish, to each review based on its content.

Following the detection of the language, the appropriate translation model was selected to convert the text into English. This was achieved using the Helsinki-NLP models available on the Hugging Face platform, specifically the `opus-mt-XX-en` series. By replacing 'XX' in the model's name with the detected language code, the system dynamically selected the translation model tailored for each specific language to English translation. The process is explained in table 2 below.

*Table 2 – ML Translation Demonstration*

| Review | Langdetect result | opus-mt-XX-en name | Helsinki-NLP Translated text |
|---|---|---|---|
| 'El dia esta muy nublado' | ES | opus-mt-ES-en | The day is cloudy. |

The translated texts were stored in a new column titled 'review_text_en' within the original database. This column served as the unified basis for subsequent data analysis, ensuring that all reviews were analysed in the same language, thereby enhancing the reliability and accuracy of the insights derived from the data.

### 3.4.2 Sentiment Analysis

According to the literature studied, the `cardiffnlp/twitter-RoBERTa-base-sentiment` model was utilised to extract the sentiment expressed within the customer reviews. This model is state-of-the-art in extracting sentiment.

Sentiment analysis was applied to the review data by first separating the text into sentences, and then tokenising the text data and converting it to a numerical ID format suitable for a machine learning model to read and analyse. Once the tokenisation of the data was done, the `cardiffnlp/twitter-RoBERTa-base-sentiment model was employed to classify the sentiment of each review and label them with a score from -1 to 1, where -1 indicates a strongly negative sentiment, 1 indicates a strongly positive sentiment, and scores around 0 represent neutrality.

The results were scored in two newly created columns within the initial data table. 'Sentiment score' is the numerical score, whilst 'Sentiment' categorises each review as 'positive', 'negative' or 'neutral' based on the score.

Figure 4 and figure 5 in the appendix show that majority of the reviews trend towards more positive scores; according to the sample, reviews of adidas products tend to be positive, indicating that customers reflect positively on their experiences with the company.

## 3.5 Extracting Topics out of Reviews

### 3.5.1 Topic Analysis Model Motivation

Initially, LDA (Latent et al.) was the preferred model for extracting topics from reviews due to its ability to uncover and combine underlying discussion points within text into coherent issues that are easy to summarise and interpret. However, in the specific context of clothing within the e-commerce space, topics in reviews tend to remain consistent across different products. Consequently, LDA's approach, which generates varying issues for each product requiring manual interpretation and labelling, is not ideal.

Since the nature of reviews in this domain tends to be similar, LDA might continuously generate similar or overlapping topics for different products. This redundancy does not add value and leads to inefficiency. Furthermore, LDA generates issues based on the distribution of words without prior knowledge of the domain. In a specialised field like clothing reviews, where the meaning of a word varies based on the context, LDA might not always prioritise or clearly distinguish these crucial topics without specific tuning.

Due to the drawbacks of LDA mentioned above, it was determined that an optimal model that can identify pre-determined discussion topics within reviews and label their presence was chosen. The labelling will be done by producing a specialised deep learning model built upon the RoBERTa framework. The deep learning model was determined as the topic analysis tool due to its ability to learn complex patterns, such as those of large volumes of text data. Deep learning models can understand nuances in language and context, which enables them to accurately identify and classify predefined topics within text based on subtle variations in how things are discussed in reviews.

### 3.5.2 The Goal of the Topic Analysis Deep Learning Model

The deep learning model will identify whether a topic is present within a review by analysing its contents and labelling each review with a binary value based on whether the topic is present.

### 3.5.3 Building the Deep Learning Model

The deep learning model will be built off the RoBERTa tool and specialised in predicting and labelling the presence of topics within a piece of text.  RoBERTa's output layer will be adjusted to label a piece of text with a binary value based on the presence of its contents. For RoBERTa to label text, it must be trained and optimised for the task.

### 3.5.4 Training the Model

For the RoBERTa tool to correctly classify text based on the presence of topics, the topics will first need to be defined, and then a training data set will need to be produced. Through conversations with personnel working within adidas's Customer Support division and manual inspection of review samples, a list of the five most prominent topics influencing consumer sentiment was determined, which can be found in table 3 below.

*Table 3 – Pre-Determined Topic Names and Definitions*

| Topic name | Definition |
|---|---|
| Quality | Evaluates the durability, material, and overall build of the product. It often includes comments on whether the product meets the expectations set by the brand or retailer. |
| Sizing | Discuss a product's fit, whether it runs large, small, or true to size. This may include specific mentions of dimensions or comparisons to standard sizing charts. |
| Style | It focuses on a product's aesthetic and design aspects, including trends, appearance, and how well the product matches the online description or images. |
| Customer Service Support | Covers interactions with the company's customer service team, including responsiveness, helpfulness, and resolution of issues. |
| Delivery Experience | Describes the shipping process, delivery timeliness, product condition upon arrival, and how the reality matches the promised delivery expectations. |
| Pricing | Describes the price of the product |

These five topics are believed to be the most effective in swaying one's perception of the product and shopping experience. Therefore, they are essential in correctly identifying insights to be extracted from the review to guide business strategy accurately and correctly.

### 3.5.5 Producing the Training Data Set

Producing the training data for classification tasks is done by producing a sample of text that fits the appropriate context for the task and manually labelling them so that the model can analyse them and understand the differences between differently labelled text; in this scenario, the model will be able to understand and predict with accuracy whether a topic is mentioned within a review by comparing reviews that discuss the topic with reviews that do not.

The training data must be balanced for the model to be trained appropriately and to ensure that the model learns to recognise and respond to all categories equally, without bias towards the more frequently discussed ones. If the data is unbalanced, the model may become biased towards the overrepresented class, leading to poor performance on underrepresented review topics, omitting valuable data from the insight generation process.

Furthermore, the dataset should contain a balanced representation across different clothing types to prevent the model from being biased toward more frequently reviewed items. table 4 below defines four categories of clothing.

*Table 4 – Clothing Category Names, Definitions and Examples*

| **Category** | **Definition** | **Examples** |
|---|---|---|
| Outerwear | Clothing worn over regular garments for warmth or protection from weather. | Coats, jackets, windbreakers |
| Inner tops | Shirts or tops are worn directly on the upper body, usually under outerwear. | T-shirts, blouses, tank tops |
| Bottoms | Clothing is worn on the lower part of the body. | Jeans, trousers, skirts |
| Shoes | Footwear is designed for comfort, protection, and style. | Sneakers, boots, sandals |

It is also necessary to have an equal distribution of positive and negative reviews to ensure that the model can identify the various ways a topic can be discussed and correctly label the topic's presence within each one.

Finally, having a large volume of training data is crucial to ensure that the model can effectively learn the complex patterns and nuances in the text. Due to the complexity of human language, many ways exist to discuss a topic. Therefore, the model must be trained on sufficiently large amounts of text data and example reviews to identify the presence of a topic within different, highly varied contexts. Hence, synthetic review data was used to generate a sufficient number of reviews that discuss and omit the topics positively and negatively for the different clothing categories, which was used to train the model.

This comprehensive training approach enables the model to perform reliably in real-world scenarios, making robust predictions across diverse situations. These steps help create a robust model that performs well across various data dimensions.

The model was trained to identify each of the five topics individually, meaning that it was broken down into five smaller deep-learning models specialised for each topic. Each model was trained on 1600 synthetic reviews and 400 real reviews, and each provided one binary label based on whether the topic was present. The 2000 reviews comprised 500 for each clothing type, 250 positive and 250 negative. 1000 of the 2000 reviews discussed the topic, and the rest did not; hence, the training data contained various potential review scenarios and was trained to identify whether the topic was present in them to achieve high predictive performance in unseen data; the reviews used can be found in the GitHub repository in the appendix below.

### 3.5.6 Producing the Testing Data Set

To test the trained model, a dataset consisting of 250 manually labelled reviews from the clothing retailer, Zara was created. Zara sells a variety of clothing items at a range of price points, providing a diverse collection of review text to test the deep learning model’s labelling capabilities with.

The choice to use a separate dataset from the one in the training process was due to assuring that overfitting was avoided. The testing was done with the goal of confirming that the model can accurately predict the labelling of unseen data under varying circumstances, which would have been a difficult to verify task if the training dataset was used for testing.

### 3.5.7 Optimising the Model

Optuna was employed as the optimisation framework to enhance each model's predictive performance. The strategic use of Optuna aimed to refine each model’s settings and improve their overall effectiveness in accurately classifying text. The optimal parameters were identified for each model, and the models were tested again for their predictive accuracy on the testing data. The optimal parameters are found in Table 5 below.

Table 5 – Optimal parameters

| Topic | Optimal Batch Size | Optimal Learning Rate (lr) | Optimal Epochs | Predictive accuracy |
|---|---|---|---|---|
| Quality | 8 | 9.31e-05 | 5 | 87.6% |
| Customer Service | 16 | 4.33e-05 | 1 | 82.7% |
| Shipping | 8 | 9.97e-05 | 2 | 97.8% |
| Size | 32 | 1.30e-05 | 3 | 95.7% |
| Style | 16 | 7.30e-05 | 7 | 91.5% |
| Price | 8 | 2.21e-05 | 5 | 84.0% |

The predicting accuracy of each separate model is significantly high due to the separation of each topic to be handled by a smaller, specialised deep learning model, the diverse training dataset and the optimisation of its model's parameters with Optuna; the predicting accuracy of each model is satisfying to finalise each deep learning model and apply it to review data to produce a regression to interpret.

## 3.6 Utilising the Review Data for Predictive Modelling

### 3.6.1 Extracting Review Data for Regression

The finalised regression aims to predict the sentiment score based on the presence of topics within the review texts, as well as other data regarding the review and the purchase; this is done with the goal of enhancing the regression's predictive accuracy by incorporating multiple predictor variables that control for the diverse contextual settings the reviews were produced within. By accounting for the factors that influence the customer to leave a review and the emotions they felt when writing them, the accuracy of the coefficients and significant values produced by the regression will be higher. For the raw review data to be used as predicting variables within the finalised regression equation, they must be selected and transformed into a numerical format.

One of the key variables in the analysis is the website domain, as indicated by the columns '.DE', '. CO.UK', '.ES', and '.FR'. The domain value is based on the adidas website from which the review was extracted; a customer's taste and expectations vary greatly based on cultural differences, and the domain variable will allow to track indirectly the customer's nationality. Hence, implementing the website domain as a predicting variable will allow the regression to reveal how customers making purchases from different countries react to the products and shopping experience at adidas; revealing the brand's appeal in different geological markets and potential nationality demographics to target the company strategy around. The selected domain variables were based on their share of visits within the overall adidas brand's domain visits, with the base variable being the remaining EU domains clustered under 'Other.'

The 'Word count' variable presents the number of words written within a review. Whilst longer reviews provide deeper insight into the customer's opinion, they also tend to be written by more passionate consumers who correlate with having more extreme opinions and hence sentiments. Furthermore, a more detailed review is like to discuss more topics and hence may be overrepresented in the finalised results if not controlled for its length. Controlling for the word count is necessary to ensure that the more detailed reviews do not sway the coefficients of other predicting variables.

The 'Product Price' variable represents the base price of the product purchased. Customer expectations regarding a product are higher when they paid a large sum of money for it, as supported by the utility theorem (Lancaster, 1966). As a result, two products that offer the same utility at different price points will be responded to with varying sentiments –product price will be controlled through this variable to account for that. Furthermore, the 'product price' variable will reveal the differing opinions of customers between adidas luxury product lines and budget-friendly products, which will be used to guide product strategy.

Finally, to account for seasonality, the review season is recorded in the form of 'Summer', 'Spring', &'Winter' ('Fall' being the base variable of comparison). The review season reflects the season of when the review was published on the adidas website and aims to understand the performance of seasonal product lines – such as the adidas Winter Sports Collection. The coefficients and p-values of the review seasons will indicate which seasonal product types outperform the others, assisting in guiding product and marketing strategies of the brand in the process.

By integrating the above variables, the regression model will gain a nuanced understanding of the factors affecting the sentiment of the customer expressed within their review – ensuring that the analysis results are not skewed and accurately reflect the opinions of the customer, whilst also revealing valuable information to guide strategy with.

### 3.6.2 Building the Regression Model

With the data prepared and labelled, the regression model's predicted and predictor variables are ready for analysis. The final linear regression model formula is seen below.

$$\begin{aligned} SentimentScore \\ &= \beta_0 + \beta_1 Quality + \beta_2 CustomerService + \beta_3 Shipping + \beta_4 Size + \beta_5 Style \\ &+ \beta_6 Price + \ \beta_7 DomainDE + \beta_8 DomainES + \beta_9 DomainUK + \beta_{10} DomainFR \\ &+ \beta_{11} WordCount + \beta_{12} ProductPrice + \beta_{13} Summer + \beta_{14} Winter + \beta_{15} Spring + \varepsilon \end{aligned}$$

This model predicts the sentiment score of a review, which quantitatively reflects the emotional tone of the feedback, ranging between very negative to very positive. As the predicted variable is numeric, ranging from -1 to 1, the base coefficient will reveal not only the overall sentiment customers express in their reviews but also the intensity of it; the coefficients and p-values of the predicting variables will reveal which ones have the highest, significant impact on the sentiment expressed and whether that impact is positive or negative.

## 3.7 Combining the code into a Smart Analysis Model

The steps used to analyse the data were all combined into one python script that can be found in the GitHub repository of this thesis. The script goes through the process of tokenising and translating the data, extracting the sentiment score, word count and applying the six deep learning models to identify whether the topics were discussed within the reviews. The script finally combines all the data within the reviews to conduct a regression analysis to predict the sentiment score. Note that due to the differing nature of reviews based on platforms, the script only performs the regression using the sentiment score, topics and word count, but it can be adjusted to fit the demographic data that is wished to be included with the specific reviews being analysed.

### 3.8 Research Methodology Conclusion

The research methodology provides a detailed framework for analysing customer reviews and transaction data for clothing e-commerce retailers to gain deeper insights into the online shopping experience of their consumers. By utilising advanced machine learning techniques and comprehensive transactional and review data, this approach allows businesses to understand consumer satisfaction levels and what drives them.

The methodology revealed that it is possible to take reviews from an international retailer and translate them into one universal language to analyse their contents. Applying context-aware sentiment analysis to reviews once translated also provided the desired results. Finally, through the usage of advanced machine learning frameworks and carefully balanced training dataset, building a deep learning model that can detect the presence of a topic within text of various contexts with high accuracy is realistic.

Converting the customer demographic, the details of the product they purchase and contents of their review into numerical format allows for the predictive modelling of consumer feedback to guide company strategy based on the real-time consumer response to products and shopping experience.

Overall, the methodology section provides a foundation for the analysis and application of review data for online retailers to utilise. It highlights the combination of advanced techniques that will help derive meaningful conclusions for business strategy adjustments.

The following section will evaluate the model by applying it to actual adidas transactions and review data from across the EU market. This testing will assess the model's predictive capabilities and effectiveness in a real-world scenario, providing strategic guidance to enhance marketing strategies and operational adjustments based on empirical data.

# Chapter 4: Results

### 4.1 Analysing the adidas Review Data

The adidas review data was fed into the model. Based on 7500 rows of reviews, table 6 was generated.

*Table 6 – Regression Results*

| | Coefficient | Std. Error | Pr(>\|t\|) |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Intercept | 0.61 | 0.06 | < 0.001 *** |
| Quality | 0.14 | 0.02 | 0.03 * |
| Customer Service | -0.21 | 0.05 | < 0.001 *** |
| Shipping | 0.05 | 0.04 | 0.43 |
| Size | 0.03 | 0.03 | 0.005 ** |
| Style | -0.07 | 0.01 | < 0.001 *** |
| Price | -0.13 | 0.04 | < 0.001 *** |
| .DE domain | 0.05 | 0.03 | 0.04 ** |
| . CO.UK domain | 0.03 | 0.01 | < 0.001 *** |
| .ES domain | 0.04 | 0.05 | 0.23 |
| .FR domain | -0.09 | 0.03 | 0.04 ** |
| Product Price | -0.03 | 0.02 | < 0.001 *** |
| Summer | 0.11 | 0.05 | 0.02 ** |
| Winter | -0.05 | 0.04 | 0.03 ** |
| Spring | 0.01 | 0.03 | 0.12 |
| Word Count | 0.001 | 0.005 | 0.56 |

### 4.2 Interpretation of adidas Review Data Analysis

The analysis of adidas's review data indicates that customers generally exhibit a high level of satisfaction with their purchases, as evidenced by a significant intercept value of 0.61. This high value is consistent with the high amount of 4 to 5-star ratings shown in figure 2.

/The statistical model reveals several insights into factors affecting customer sentiment. Notably, mentions of customer service in reviews negatively impact sentiment, with a significant value of -0.21. This suggests that customers are dissatisfied with their interactions with adidas customer support, highlighting an area where adidas should prioritise improvements to enhance overall satisfaction.

Product quality, on the other hand, has a significant positive impact on customer sentiment when discussed, with a significant value of 0.14. This indicates that customers greatly appreciate the quality of adidas products, making it a strong driver of positive satisfaction expression. Maintaining its high product quality is crucial for adidas to ensure customer contentment. Similarly, size has a positive, significant coefficient of 0.03, similarly suggesting that customers tend to be satisfied with the sizing of the adidas products they purchase.

Shipping, though having a positive impact of +0.05 on sentiment, is not a significant factor, suggesting that it evokes mixed reactions that are neither consistent nor strong enough to significantly influence customer emotions. This indicates that adidas should try to improve its shipping experience, but not with the same priority as topics that evoke negative emotions, such as Customer Service.

Price, however, negatively affects sentiment with a significant value of -0.13. This suggests that customers find adidas products overpriced, highlighting the need for the company to address pricing strategies or emphasize the value proposition of its products to improve customer perceptions. Similarly, style has a significant, negative 0.07, suggesting that although adidas products tend to have high quality and correct sizing, the aesthetics of the brand is discussed negatively.

The analysis also highlights demographic variations in sentiment. Customers in Germany, as recorded by the .DE domain, show a slightly higher sentiment score (+0.05), although this is not significant, indicating no strong difference in satisfaction compared to other geographical locations. English customers (. CO.UK) tend to have more positive sentiment with a significant value of +0.03, which shows that the brand resonates with the British market. Spanish customers (.ES) do not have a significant difference in sentiment compared to other markets across the EU. Conversely, French customers (.FR) display lower satisfaction with a significant negative impact of -0.09, highlighting a potential area for targeted improvements for the particular market.

Furthermore, the analysis suggests that the length of reviews, indicated by word count, does not significantly impact sentiment. Regarding seasonality, reviews left in the Winter and Summer months show a significant increase in sentiment compared to reviews left in Fall. Additionally, the price of products is a critical factor, with higher prices correlating with lower satisfaction, as evidenced by a significant negative impact of -0.03; suggesting that customers do not find the value provided by the higher priced adidas products to be sufficient.

While adidas customers are generally satisfied, there are clear areas for improvement, particularly in customer service and pricing. Enhancing service quality and reassessing pricing strategies, while continuing to focus on product quality and style, can further elevate customer satisfaction. Additionally, addressing specific linguistic and demographic needs can help tailor a more personalised and positive customer experience

## 4.3 Utilisation of Results to Guide Strategy

The coefficients from adidas's review analysis can be interpreted within the business context to guide various strategic decisions. Understanding how different attributes of the product and shopping experience impact customer sentiment allows companies to pick which parts of their production to improve upon and focus on to guide strategy.

### 4.3.1 Product Augmentation and Innovation

The analysis revealed that adidas' customers generally respond negatively to the Style of the company's clothing. Adidas can utilise this feedback to refine its clothing designs to keep up with modern fashion trends and meet customer expectations better. By enhancing the looks within its catalogue and introducing new designs and silhouettes to the market, adidas will see an increase in sales as previous lost-on sales due to style dissatisfaction will now be redundant. Furthermore, adidas has the opportunity to launch a marketing campaign alongside it to advertise its updated catalogue that is up to date on stylistic trends that the public will react positively to, as it addresses their feedback directly to improve upon an issue they have been vocal about within their reviews.

### 4.3.2 Marketing

Adidas can leverage the insights from the review data to address the negative feedback on customer service. To enhance its market position, adidas should conduct an overhaul of its customer service approach, aiming to transform itself into a customer-focused company. Doing so will significantly elevate the company brand and its competitive edge by emulating successful customer models like that of Amazon, which can help adidas become synonymous with top-tier customer service, making customers more willing to pay the premium to shop at adidas due to its enhanced customer-focused approach. This shift could positively impact public perception of the company's pricing, which the analysis currently shows to impact sentiment negatively.

To support this customer service-focused transformation, adidas should launch a robust marketing campaign to communicate its new business approach. Through traditional and social media storytelling, adidas can convey the message that it is a brand deeply committed to customer satisfaction. This campaign can attract both returning and new customers to its stores while also differentiating the adidas brand within the market. By highlighting the revamped customer service, adidas can rebuild customer trust and loyalty, ultimately leading to a more positive overall sentiment and improved market performance.

### 4.3.3 Competitive Advantage Analysis

Understanding that customer sentiment regarding price and style is less favourable provides adidas with a transparent, competitive vector. By analysing the catalogues of competing brands such as Nike and Puma, adidas can modify its product lineup and introduce new products incorporating these desirable features, increasing the company's market share and sales in the process.

### 4.3.4 Local Market Analysis

The regression analysis reveals differing customer sentiments across various demographic groups, with customers in Germany and England expressing more positive sentiments compared to those in France. Based on this insight, adidas should consider reallocating resources from the French market to other markets where the overall customer sentiment is more favourable. In the case that adidas sees diminishing returns in the reallocation of resources to the said markets, adidas can instead conduct a local analysis of the French market to understand why their sentiment is overall lower than other markets in an attempt to increase the French customers' satisfaction level and sales in the process.

Regardless of which market adidas chooses to invest its resources into, the process should involve enhancing physical & online store presence, improving localised marketing efforts, and tailoring product offerings to better meet the preferences and expectations of the geographical customer groups. This strategic shift would allow adidas to build a stronger brand image and customer loyalty in specific regions with the goal of increasing market share and sales within them.

### 4.3.5 Pricing Strategy

The analysis reveals that more expensive adidas products are reviewed more negatively, with price being a frequent point of dissatisfaction among customers. The fact that the topic of price is discussed negatively and the price of the product itself correlates negatively with the satisfaction level of the consumer, it can be concluded that adidas has a misalignment between the cost and consumer value of its high-priced products.

To address this issue, adidas can invest in better materials for its high-priced products to enhance their quality further – helping to justify the higher prices and meet the customer expectations in the process. Although this will result in a lower profit margin for the products, it will cause for the perceived satisfaction level to increase and present the opportunity to profit through increased sales.

### 4.3.6 Market Entry Analysis

Adidas should continuously evaluate its competitors' consumer feedback to its product offerings and shopping experience in markets it is looking to enter. Through applying the same methods used to analyse its own reviews to reviews found in its competitors' websites, adidas can understand its competitors' position within the market and anticipate challenges upon entering the market, allowing adidas to shape its own competitive position in the process.

For example, if a new retailer looking to enter the European market analysed adidas's product reviews, they would view its customers' poor reception to the company's customer services and choose to capitalise upon that flaw by marketing themselves as a consumer service focused company to attract consumers that are dissatisfied with their shopping experience at adidas.

Assessing its competitors' consumer feedback will provide adidas with insights that can be crucial in refining the brand's strategy for expanding into culturally diverse markets.

### 4.3.7 Monitoring the Brand

Monitoring the brand perception over time is crucial for brands like adidas to ensure sustained customer satisfaction and to maintain a strong brand reputation. By analysing the average sentiment score over time (measured by the review publish date), adidas can monitor the average customer perception towards their brand and products over time, and by correctly filtering, they can also monitor the consumer perception in relation to topics discussed. Doing so will allow adidas to effectively monitor its brand appeal and identify product strengths and weaknesses.

This method enables management to address underperforming areas and capitalise on successful ones quickly, tackling potential incoming issues before they get out of hand and cause damage to the brand perception in the long term. Continuously tracking changes in sentiment allows adidas to manage its brand reputation proactively. If a particular product feature begins to register declining satisfaction levels, the company can make immediate enhancements to reverse these issues' negative impact. Conversely, features that consistently receive positive feedback can be highlighted in marketing campaigns, bolstering the brand's public image.

Figure 6 illustrates a concerning trend in the sentiment scores of adidas reviews that discuss customer service from March 2023 to January 2024. Over this period, the sentiment score has declined noticeably, from roughly 0.55 to 0.4. This downward trajectory indicates growing dissatisfaction among customers regarding adidas' customer service and warrants and investigation into changes implemented to its CS practices over the course of 2023, to tackle what could result in long-term harm to adidas' reputation and customer loyalty.

Possible responsive actions include enhancing customer service representatives' training, improving response times by increasing the customer service workforce headcount and ensuring that customer concerns are resolved efficiently and effectively through creating case studies based on previously solved customer service cases. By proactively addressing these issues, adidas can prevent further deterioration in customer satisfaction, mitigate potential crises and restore positive sentiment among customers.

## 4.4 Results Conclusion

In conclusion, adidas's application of data-driven insights from customer feedback and sentiment analysis can ask as a guide for refining its product selection, customer journey and expanding its market presence. By tailoring its strategies to address specific customer preferences, adidas can strengthen its position within the market.

Through these focused efforts, adidas can optimise its operational and marketing strategies to achieve sustained growth in various markets and improve its brand’s relationship with its consumers. Through refining its strategic positioning based on the results of detailed review analysis, adidas can maintain a positive brand image and encourage repeat business, ensuring long-term success in a competitive market landscape.

# Chapter 5: Conclusions

## 5.1 Literature Review Summary

The literature review explored the advancements, integration, and implications of text analytics within the e-commerce sector by focusing on how customer reviews can be leveraged to guide business strategy. The review discussed the recent yet significant advancements in text analysis methods, contexts where these produce improved results, and how they must be prepared and utilised to yield actionable insights. The literature review further discussed how integrating advanced text analytics techniques gives e-commerce businesses critical insights into consumer behaviour and preferences, enabling them to make informed strategic decisions.

The literature describes the shift from simple lexical and rule-based approaches, to analysing text to advanced machine learning techniques. Models such as the RoBERTa framework have enhanced computers' ability to interpret text accurately by understanding and accounting for complex language features such as context, irony, and emotional depth—all of which are crucial for correctly understanding the message the customer is trying to convey through their writing.

Machine learning sentiment analysis built off RoBERTa's framework understands context by interpreting the text surrounding the word rather than only the text that came before it, giving the model a 360-degree view of the text. This allows the model to gather a better understanding of the text's contextual background and more accurately apply the correct sentiment score, indicating whether the sentiment is positive or negative and the intensity of the emotions expressed. This provides deeper insights for businesses to guide strategy.

Topic modelling techniques have also advanced by creating new and improved machine learning models. Techniques such as Latent Dirichlet Allocation (LDA) and deep learning models have allowed for known and unknown topics to be extracted from text and have proven effective in extracting themes from large volumes of text; and synthetic data will assist in creating superior datasets to train the models, leading to more accurate analytical results. Topic modelling techniques have allowed private businesses and public institutions to analyse text in mass amounts to understand the main topics of discussion and provide essential insights to understand the author's motivation without necessarily reading through the entire text, saving time and resources whilst also making it possible to accurately understand the large amount of text generated for different purposes daily.

The conversion of text to numerical format, such as transforming emotion portrayed to a sentiment score or marketing the presence of a topic by a binary number, has allowed for a text's contents to be predicted or used as predictors using regression modelling, which can facilitate predictive analytics for improve decision-making of those looking to use text to understand and influence the author of it. In the e-commerce clothing industry context, predictive text analytics can significantly impact business strategies by providing rapid and accurate assessments of customer opinions. Such ability allows businesses to alter their catalogue and marketing to better align with consumer preferences. The ability to process and analyse customer feedback efficiently helps businesses maintain a competitive edge by quickly responding to changing market trends while identifying weaknesses in their competitors' business models.

The review also notes the challenges of using text analytics in a business setting, such as the necessity for high-quality, balanced training datasets that prepare the model for the many different types of text it will likely have to analyse. Depending on the situation, the analyst must also understand which models to implement to get the best results. These challenges underscore the importance of continual model evaluation and adjustment.

Further deep dives into existing literature regarding the topic of text analytics in the e-commerce sector should include looking at specific e-commerce companies that have successfully implemented text analytics to guide their business strategies – doing so will provide examples of how the theory translates into practice, including the methodologies used, challenges faced, and results achieved. Furthermore, as the machine learning market increases, it will be necessary to investigate different text analytics tools and platforms to highlight their features, integration capabilities and effectiveness in various e-commerce-related scenarios. Finally, a further literature review should explore studies that have tracked the long-term impact of text analytics on business outcomes in e-commerce to identify trends, measure ROI over time, and determine the sustainability of text analytics strategies compared to others.

### 5.2 Research Methodology and Results Summary

The research methodology was designed to analyse the review data from adidas to gain insight into consumer preferences and behaviours, ultimately guiding the business strategy.

The data was standardised into English using the 'Helsinki-NLP' translation models; English was chosen to be the language the records were translated into as the majority of the current state-of-the-art text analytics-based machine learning models are built to analyse English text, and therefore translating all the reviews to English would allow these models to be utilised to reach the goals of this thesis.

Sentiment is the emotion attempting to be conveyed by the writer through the text that was produced; it provides insight into the author's feelings and their motivations. The sentiment within the translated data was extracted using the ‘cardiffnlp/twitter-RoBERTa-base-sentiment’ model in the form of a score ranging from -1 to 1. Values below zero represent negative sentiment, whilst values above represent positive - values closer to the limit represent a higher intensity in the expressed emotion; scores close to zero represent neutral emotions. The machine learning model was chosen due to its precision due to being trained on over 10,000 manually labelled online texts whilst also having an advanced understanding of context due to its ability to read the whole text before interpreting the meaning of a word, unlike traditional models that only assess the meaning of a word based on the words that came before it.

The contents of the reviews were further dived into to extract the discussion points through topic analysis. A deep learning model based on the RoBERTa framework was developed to identify predetermined topics within customer reviews. Quality, Sizing, Style, Customer Service, and Delivery Experience were manually labelled as topics of discussion across various reviews in combination with synthetic review data to train the model to identify whether a text discusses the mentioned topics correctly. The model was tested on reviews from the retailer, Zara. Different review data was chosen for the testing compared to the training to ensure that overfitting was avoided.

A regression model was developed to predict the sentiment score using topics within a review, customer demographic data, and product pricing data. The regression model aimed to reveal how different aspects of the product and shopping experience impacted the emotions portrayed within a customer’s review in combination with the customer’s purchase and geographical background. A case study was produced to showcase the model's abilities to guide business strategy by applying it to the adidas review data and interpreting it to produce actionable insights.

The model was applied to adidas’s review data, and the correlations, coefficients, and significant values were assessed, which theoretically provided adidas with actionable insights – which would allow the company to adjust its product offerings and marketing strategies to meet customer needs better. It was found during the case study that the insights derived from the project can have significant implications for strategic decision-making in various business operations areas. The analysis of predicting customer sentiment through product features allows for pinpointing areas needing improvement or innovation. Understanding which aspects of the products sold resonate well with customers also helps tailor marketing campaigns to highlight the product/brand strengths or sway attention away from underperforming aspects. Furthermore, the interpretation of the analysis revealed how consumers perceive the value of products at different price points, which can assist adidas in adjusting pricing strategies to match market expectations.

The methodological advancements made through this research demonstrate significant progress in applying text analytics and machine learning to understand customer reviews. These techniques effectively decode complex customer sentiments and preferences and enable businesses to anticipate customer reactions and proactively adapt their strategies.

Despite the goals of the research being met, several limitations and challenges were faced. For one, the final analysis was limited to data from a single retailer. Hence, it may not represent the broader e-commerce industry; this drawback was combated by training the model on data from various e-commerce retailers. Furthermore, the analysis on adidas's product selection was conducted on 7500 datapoints, a small sample in comparison to adidas's overall fiscal year sales, indicating that most of the customer information is missing. To add to that, online reviews by nature tend to be uploaded by customers with solid feelings (hence why adidas's reviews were either overwhelmingly positive or negative rather than lukewarm); therefore, the milder opinions are missing from the analysis and the shaping of the business strategy. Finally, the tool developed is hyperspecialised in analysing e-commerce clothing reviews, and it would not be as efficient in other industries.

Future research should address these limitations by exploring customer reviews across product categories and employing flexible machine-learning algorithms with the goal of producing a model that can be applied to various e-commerce contexts. Integrating real-time data analysis could also enhance the ability to capture dynamic changes in consumer behaviour and market conditions and provide real-time strategic decisions. Further studies should aim to verify that the approach introduced within this research is applicable across various products and services, as that would enhance the robustness and applicability of the research findings, ensuring its effectiveness across various market conditions.

## 5.3 Conclusions

The research has demonstrated that the tool developed is a powerful asset for e-commerce businesses, providing substantial utility in guiding strategic decisions. Its capacity to analyse customer reviews through regression models offers deep insights into customer sentiment and does so efficiently, requiring relatively minimal computing power. This efficiency ensures the models deliver fast results, enabling businesses to respond quickly to emerging trends and customer feedback. As the e-commerce industry continues to expand and the volume of customer reviews increases, the importance of such tools is expected to grow. The ability to quickly and effectively analyse vast amounts of data will become increasingly crucial, making this text analytics model an invaluable resource for businesses aiming to survive and grow in the competitive e-commerce landscape. This ongoing relevance underscores the need for continuous refinement and adaptation of the tool to meet the market's evolving demands and leverage the growing body of customer feedback effectively.

## 5.4 Recommendations for Businesses

In the fast-evolving landscape of e-commerce, businesses must prioritise integrating customer feedback into their product development and marketing strategies. Understanding the product properties that impact the satisfaction of the customer and the extent to which they do, as well as how the impact differs between demographics, provides businesses with a powerful tool that allows for the product strategy to be guided by the customer. The benefit of guiding strategy using such feedback is getting ahead of competitive marketing by enhancing overall customer satisfaction through building the products to the customers' vision, addressing specific customer concerns and identifying holes in the market to target.

To make the most of customer reviews, businesses should encourage more customers to leave feedback by offering incentives. This approach increases the volume of data available for analysis. It encourages groups who usually do not leave reviews, such as those with neutral feelings or older customers, to share their experiences. By incentivising customers to leave reviews, companies can gain access to a broader range of opinions and a fuller understanding of customer satisfaction to identify opportunities for further improvement.

## 5.5 Recommendations for Future Research

Future research should expand the scope and depth of variables considered in analyses to understand customer preferences and behaviours better. While the current study focused on six product qualities to avoid overfitting, incorporating additional variables could provide richer insights into the factors that influence customer satisfaction. This expansion will allow for a more comprehensive analysis that could uncover less obvious but equally significant consumer choice and satisfaction drivers.

Developing tools that transcend the specific context of clothing e-commerce is another valuable direction for future research. By creating analytical models that can be adapted to various retail sectors, researchers can enable businesses in different industries to leverage customer reviews effectively. This versatility would make the tool more useful across a broader spectrum of the retail sector, enhancing its applicability and impact.

The integration of Generative AI into review analysis tools offers exciting possibilities. Generative AI can be used to simulate customer responses or generate synthetic review data that can help test and refine the analytical models. This approach could provide businesses with predictive insights about how changes in products or services might be received, even before implementing those changes in the market.

Lastly, a deeper demographic analysis and more refined customer segmentation could significantly enhance the precision of customer insights derived from reviews. Future studies should focus on identifying and analysing demographic factors influencing review content and sentiment, such as age, geographic location, and purchasing power. This enhanced segmentation can help businesses tailor their marketing and product development strategies more effectively to meet different customer groups' specific needs and preferences.

By addressing these areas, future research can build on the current findings and offer more robust tools for businesses that aim to use customer reviews to shape their strategic decisions.

### 5.6 Research Limitations

The research faced several challenges that could be addressed in future studies to enhance the findings' reliability and practical application. A significant limitation was the overrepresentation of 1-star and 5-star reviews in the dataset, with fewer reviews and moderate ratings of 3 stars. This skewed distribution might have led to an analysis that captured more extreme sentiments than the lukewarm opinions of moderate reviews. Additionally, the study relied on static data, which does not capture the ongoing changes in consumer behaviour and market conditions over time, which may limit the relevance of the findings over time.

Another significant challenge was the extensive data preparation involved, including cleaning, training, and testing, which required substantial resources, time and knowledge. This process could be particularly daunting for smaller organisations or those with limited technical capabilities. The model developed was also specifically tailored for the clothing e-commerce sector. Its application to other industries would require considerable adjustments, including comprehensive retraining and testing to ensure accuracy.

Moreover, the dataset used in the study contained limited demographic information, with only the geographical location included. This restricted the scope for detailed demographic analysis, which could have provided more profound insights into how different groups of customers perceive and interact with products. Future research should consider incorporating a more comprehensive range of demographic variables to allow for more precise customer segmentation and targeted marketing strategies.

Addressing these issues in future studies could improve the models' adaptability and accuracy, making them more useful for businesses seeking to leverage customer feedback for strategic decision-making.

## 5.7 Closing Thoughts

The research showcases the many implications of using text analytics and machine learning methods to extract data from reviews within the e-commerce sector. As the industry expands, extracting actionable insights from customer-generated data becomes essential for maintaining a competitive advantage within the market.

Businesses' potential to leverage customer-centric data extends beyond mere product adaptation or marketing refinement. It allows for continuous improvement and innovation, where customer interaction contributes to strategic goals. This approach to business strategy is likely to set the benchmarks for success in the digital marketplace.

In the appendix below the GitHub repository of this research is linked. It contains all the code produced during the research for those interested in expanding upon this paper.

# Appendix

## References

Adidas Group. (2023). Global Sales - Adidas Annual Report 2023. Adidas. Retrieved from https://report.adidas-group.com

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media (pp. 30-38).

Blei, D. M., & Lafferty, J. D. (2006). Correlated topic models. In Advances in Neural Information Processing Systems (Vol. 18).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Brown, T. (2018). Batch size effects on neural network training: Dynamics and stability. Neural Network Engineering.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision-making. American Economic Review, 106(5), 133-139.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in Neural Information Processing Systems (Vol. 22).

Chui, M., Manyika, J., & Miremadi, M. (2016). Where machines could replace humans—and where they can’t (yet). McKinsey Quarterly.

Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: The new science of winning. Harvard Business Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Systems with Applications, 91, 464-471.

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. arXiv preprint arXiv:1706.02633.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl 1), 5228-5235.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Hu, N., Pavlou, P. A., & Zhang, J. (2009). Overcoming the J-shaped distribution of product reviews. Communications of the ACM, 52(10), 144-147.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-449.

Jimenez, A. (2021, March 31). Summer sports vs winter sports: Which one is better? The Jetstream Journal. Retrieved from https://thejetstreamjournal.com

Johnson, D. (2020). Learning rates and their impact on deep learning model performance. International Journal of Machine Learning Systems.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3363-3372).

Lancaster, K. J. (1966). A new approach to consumer theory. Journal of Political Economy, 74(2), 132-157.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Lee, A. (2019). Epochs and overfitting: Balancing accuracy and generalisation in neural networks. Computational Intelligence Review.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data (pp. 415-463). Springer.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: The management revolution. Harvard Business Review, 90(10), 60-68.

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 490-499).

Morrow, B. (2023, February 10). Adidas could lose over $1 billion after terminating Kanye West partnership. The Week. Retrieved from https://theweek.com

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

Poria, S., Cambria, E., Howard, N., Huang, G. B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 174, 50-59.

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

Smith, J. (2021). Optimising deep learning models: A case study on learning rates. Journal of AI Research.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1631-1642).

Srivastava, A. N., & Sahami, M. (Eds.). (2009). Text mining: Classification, clustering, and applications. Chapman and Hall/CRC.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1555-1565).

White, R. (2022). Parameter tuning in deep learning: Practices for enhanced model predictions. AI and Learning Systems Journal.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems (Vol. 28).

Zhou, Z. H., & Liu, X. Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 18(1), 63-77.

## Figures

Table 1 – adidas Review dataset

| Column name | Description of column | Data type |
|---|---|---|
| Review Title | Title of the review | Character |
| Rating of the review | Rating score given to product based on customer experience (out of 5) | Numerical |
| Review text | The body of text that is the review itself | Character |
| Product name | Name of product that is being reviewed, shown in the raw form it is stored within the adidas databases | Character |
| Product Family | Type of clothing item family the product falls into | Character |
| Domain | Adidas website domain where review was left | Character |
| Product Price | Monetary value of the product reviewed (In Euros) | Numerical |
| Review Date | Date when the review was initially left on the adidas platform | DateTime |

Figure 1

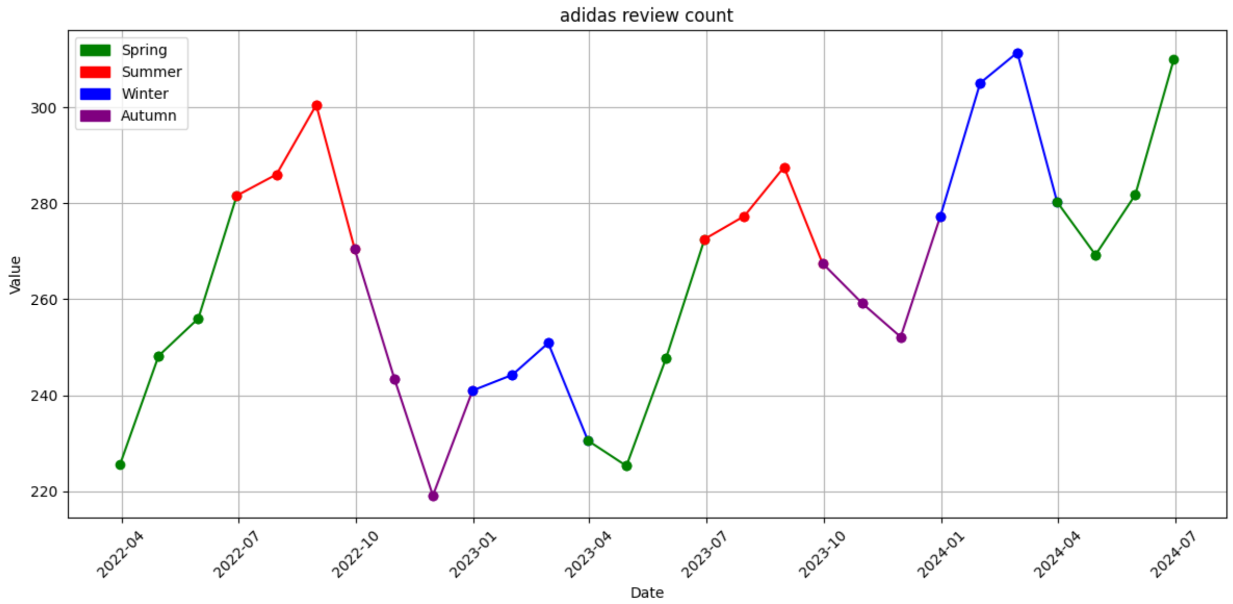


Figure 2

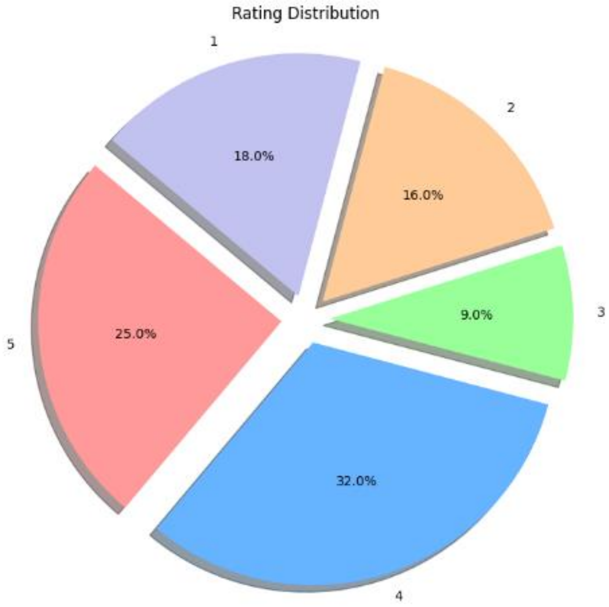


Figure 3

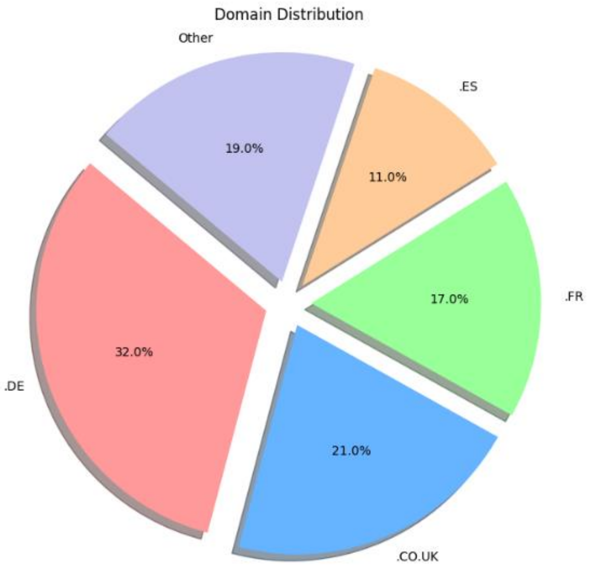


Figure 4

Sentiment distribution by review

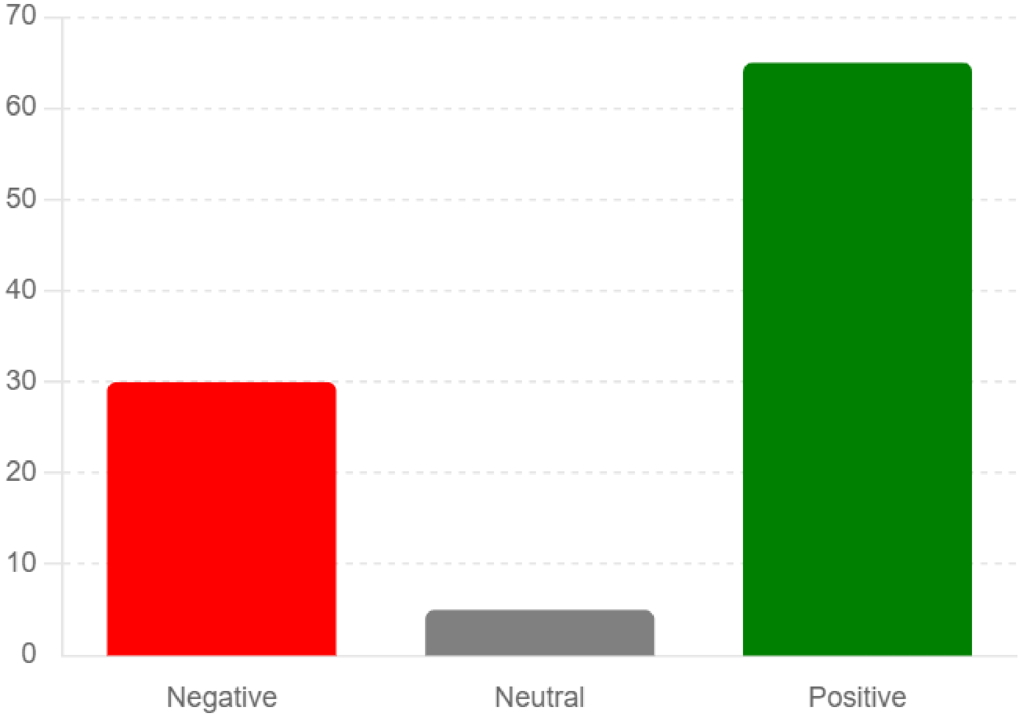


Figure 5

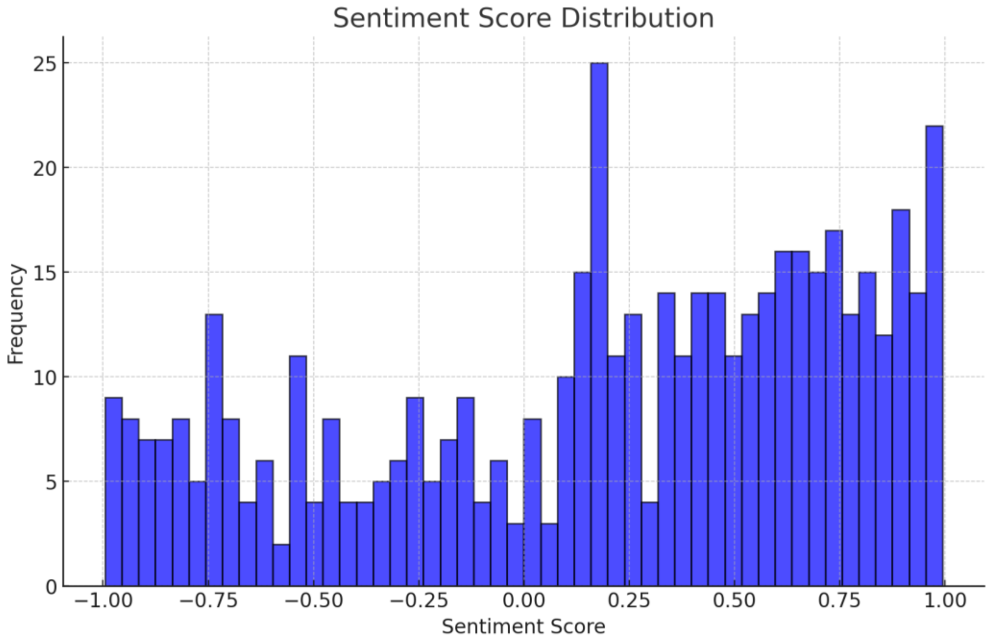
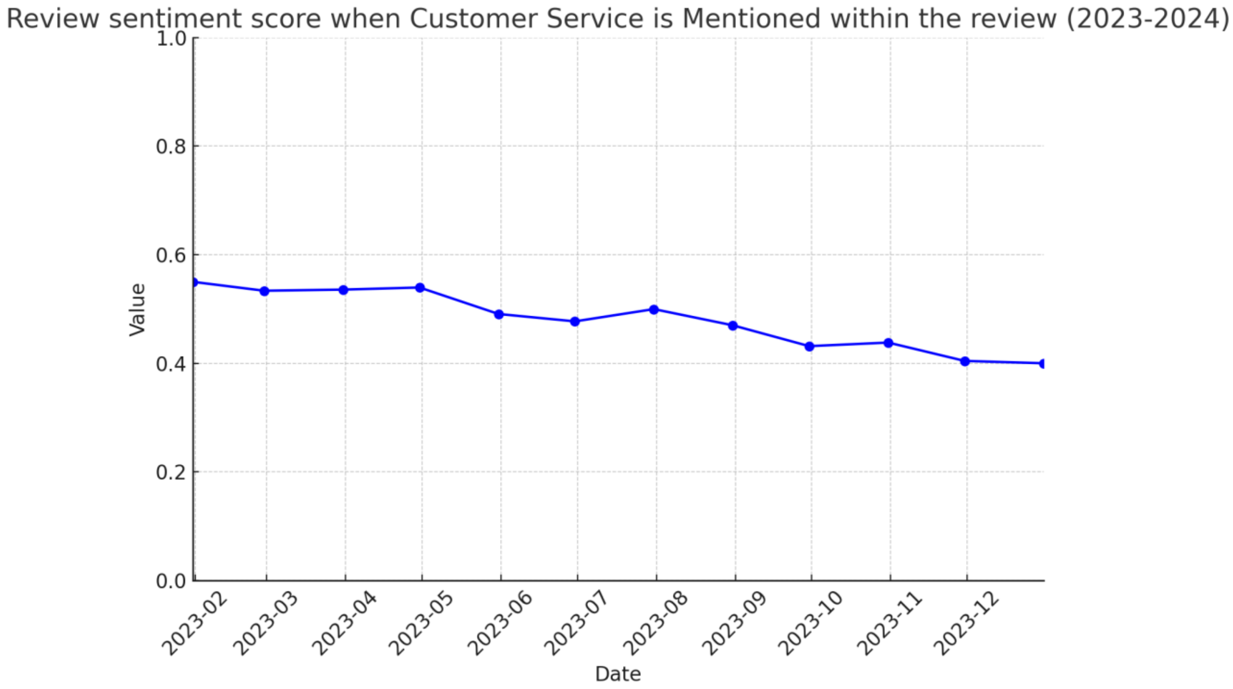


Figure 6



## GitHub Repository

ArisN123/MastersThesisRepo (github.com)