A red and black logo

Description automatically generated

**Developing a Geospatial Sentiment Dashboard for Fashion E-Commerce to Enhance Customer Engagement and Adoption**

***Final Report***

**MSc in Computer Science (Business computing)**

George Osecalome Ugboke

[U039913m@student.staffs.ac.uk](mailto:U031510l@student.staffs.ac.uk)

**Staffordshire University**

**Supervisor: Murouj Aljamaeen**

March 2024

Contents

[Abstract 2](#_Toc1434425932)

[CHAPTER 1 INTRODUCTION 3](#_Toc1168086821)

[CHAPTER 2 LITERATURE REVIEW 7](#_Toc1906189764)

[2.1 E-Commerce 7](#_Toc76359302)

[2.1.1 Fashion E-Commerce 9](#_Toc718635025)

[2.2 Customers 10](#_Toc456980027)

[2.2.1 Customer Feedback 11](#_Toc250377204)

[2.2.2 Customer Engagement 11](#_Toc828708329)

[2.2.3 Importance of Customer Feedback / Engagement in E-commerce Success 12](#_Toc2011001819)

[2.3 Sentiment Analysis and Machine Learning in e-commerce 12](#_Toc1386651424)

[2.3.1 Geospatial Analysis in E-commerce 13](#_Toc1709845120)

[2.3.2 Sentiment Analysis in E-commerce 14](#_Toc1705226003)

[2.3.3 Geospatial Sentiment Analysis in E-commerce 16](#_Toc109718980)

[2.3.4 Advanced Technologies in E-commerce 18](#_Toc875517763)

[2.3.5 Application of Advance Technologies in Fashion E-commerce 20](#_Toc2078818478)

[2.4 Gaps and Opportunities 21](#_Toc1087940032)

[2.5 Analysis Of Problem/improvement 24](#_Toc1313450827)

[CHAPTER 3. RESEARCH METHOD 25](#_Toc1411470629)

[3.1 Research Onion Overview 25](#_Toc566266932)

[3.2 Research Philosophy 27](#_Toc1747354838)

[3.3 Research Approach 27](#_Toc2083095437)

[3.4 Research Choice 28](#_Toc1303052438)

[3.5 Research Strategy 28](#_Toc484601464)

[3.6 Time Horizon 28](#_Toc652026044)

[3.7 Data Collection and Data Analysis 29](#_Toc207455941)

[CHAPTER 4. DESIGN of DASHBOARD 29](#_Toc2033410982)

[4.1 Design of Artefact 29](#_Toc1035352018)

[4.1.1 Technological Stack 31](#_Toc621294896)

[4.1.2 Data Preparation and Modeling 32](#_Toc199492531)

[4.1.3 Geospatial Mapping Integration 33](#_Toc1684055779)

[4.1.4 User Interface and Experience 33](#_Toc928700309)

[4.1.5 Dashboard Design 34](#_Toc1476798275)

[4.2 Implementation 34](#_Toc255681479)

[4.2.1 Data Collection and Preparation 34](#_Toc1039995098)

[4.2.2 Machine Learning Model Selection 39](#_Toc1838669063)

[4.2.3 Deployment on Streamlit 41](#_Toc1302438259)

[4.3 Testing and Validation 46](#_Toc1811408972)

[Testing of models 46](#_Toc810720150)

[4.4 Critical Evaluation 52](#_Toc1893171991)

[CHAPTER 5. CONCLUSION and FUTURE WORK 60](#_Toc1047815827)

[5.1 Conclusion 60](#_Toc1482612547)

[5.2 Future work 61](#_Toc38435305)

[CHAPTER 6. REFERENCES 62](#_Toc1350023338)

[CHAPTER 7. APPENDICE 65](#_Toc212246255)

[Appendix 1 Preprocessing 66](#_Toc529232411)

[Appendix 2 Model training and evaluation 69](#_Toc524296600)

[Appendix 3 Streamlit python script 69](#_Toc26614720)

**List of Figures**

[Figure 1: B2B 6](#_Toc166103885)

[Figure 2:B2C 7](#_Toc166103886)

[Figure 3: C2C 7](#_Toc166103887)

[Figure 4:C2B 8](#_Toc166103888)

[Figure 5:B2A 8](#_Toc166103889)

[Figure 6: TYPES OF MACHINE LEARNING (SARKER, 2021) 18](#_Toc166103890)

[Figure 7 research onion (Mardiana, n.d., 2020) 25](#_Toc166103891)

[Figure 8: architectural design 30](#_Toc166103892)

[Figure 9:dataset before preprocessing 34](#_Toc166103893)

[Figure 10: missing data 35](#_Toc166103894)

[Figure 11:cleaning the dataset 36](#_Toc166103895)

[Figure 12: creating sentiment column as baseline. 36](#_Toc166103896)

[Figure 13: dataset after preprocessing. 37](#_Toc166103897)

[Figure 14:code to test and train 37](#_Toc166103898)

[Figure 15:code to vectorise 38](#_Toc166103899)

[Figure 16: sparse matrix 38](#_Toc166103900)

[Figure 17: model accuracy and classification report 39](#_Toc166103901)

[Figure 18 mode and vectoriser serialisation 40](#_Toc166103902)

[Figure 19:prediction page of the application 41](#_Toc166103903)

[Figure 20dashboard page of the app 42](#_Toc166103904)

[Figure 21: launching the app 43](#_Toc166103905)

[Figure 22search filter on dashboard 43](#_Toc166103906)

[Figure 23: sentiment information of united kindgom 44](#_Toc166103907)

[Figure 24: visualisation in dashboard 44](#_Toc166103908)

[Figure 25: balance the dataset 47](#_Toc166103909)

[Figure 26: sparse matrix of balanced dataset 48](#_Toc166103910)

[Figure 27: cross validation result on prediction page app 49](#_Toc166103911)

[Figure 28: filter function in app 51](#_Toc166103912)

[Figure 29: confusion matrix of svm\_model 52](#_Toc166103913)

[Figure 30:dataset to validate the model 54](#_Toc166103914)

[Figure 31: results of the external validation 55](#_Toc166103915)

**List of Tables**

[Table 1: table of summary of past work 24](#_Toc166099086)

[Table 2: Research onion method 28](#_Toc166099087)

[Table 3: initial model results 49](#_Toc166099088)

[Table 4: final model result 51](#_Toc166099089)

# Abstract

The rapid evolution of fashion e-commerce necessitates innovative approaches to enhance customer engagement and business strategy adaptation. This research presents the development of a real-time geospatial sentiment dashboard aimed at transforming customer feedback into actionable insights by integrating machine learning (ML) and natural language processing (NLP) with geospatial analytics. Utilising data from customer reviews and social media, this tool analyses and visualises sentiment across diverse geographic regions, enabling businesses to tailor marketing efforts and product offerings more effectively.

The application of this dashboard in the fashion e-commerce sector allows for the identification of regional sentiment trends, offering a significant competitive advantage by enabling precise market segmentation and targeted marketing strategies. The dashboard employs advanced ML algorithms to process and classify large volumes of textual data, while geospatial analysis reveals the regional distribution of consumer emotions and preferences, thus fostering a deeper understanding of the global consumer base.

Moreover, the study addresses a notable gap in current research by merging sentiment analysis with geospatial data, a synergy not extensively explored in existing e-commerce research. The implications of this research extend beyond academic interest, offering substantial benefits for industry practitioners by enhancing customer engagement through data-driven strategies.

By providing real-time insights and a user-friendly interface, the dashboard supports dynamic decision-making and strategy formulation, making it an invaluable tool for businesses aiming to thrive in the digitally competitive fashion market. This research contributes to the fields of e-commerce, data analytics, and customer relationship management by demonstrating the practical application of interdisciplinary techniques to solve industry challenges.

**Keywords**: Fashion E-Commerce, Geospatial Analysis, Sentiment Analysis, Machine Learning, Natural Language Processing, Customer Engagement.

# CHAPTER 1 INTRODUCTION

In the landscape of fashion e-commerce, understanding and engaging with consumers across the globe has become paramount for businesses aiming to stay competitive and innovative. The advent of digital commerce has not only expanded the marketplace beyond physical borders but has also introduced a new realm of opportunities and challenges for fashion retailers. One of the most significant shifts in this digital transformation is the rise of e-commerce, a sector that has seen exponential growth over the past few decades (Rita & Ramos, 2022). From the early days of Electronic Data Interchange (EDI) to the current era dominated by online marketplaces like eBay and Amazon, e-commerce has revolutionised the way businesses operate and interact with their customers (Taher, 2021). This evolution has led to the emergence of various business models, each facilitating unique transactional relationships among consumers, businesses, and administrations. Among these, fashion e-commerce stands out as a particularly dynamic sector, driven by rapidly changing trends, consumer preferences, and the need for personalised shopping experiences.

The journey towards digitalisation in fashion retailing has necessitated the adoption of advanced technologies such as Artificial Intelligence (AI) and Machine Learning (ML), enhancing customer experiences by offering personalised and engaging interactions (Van Ngoc & Dat, 2022). However, the shift towards online platforms, significantly accelerated by the COVID-19 pandemic, has highlighted not only opportunities but also challenges within the sector. Issues such as sustainability and the intricacies of digital distribution have become increasingly relevant, urging businesses to seek innovative solutions to address these challenges while enhancing customer engagement and satisfaction (Jílková & Králová, 2021).

Recognising the critical role of customer feedback and engagement in the success of e-commerce, businesses have turned to sophisticated data analysis tools to decipher consumer behaviors and preferences. Among these tools, sentiment analysis and machine learning have emerged as powerful methodologies for interpreting and leveraging customer feedback. By analysing textual data from customer reviews and social media interactions, businesses can gain insights into consumer sentiments, preferences, and behaviors, enabling them to make data-driven decisions that enhance customer satisfaction and foster a more engaging online shopping environment (Sempe & Naong, 2021).

Despite the potential of these technologies to transform the e-commerce landscape, there exists a notable gap in research, particularly in the integration of geospatial data with sentiment analysis within the fashion e-commerce context. This gap signifies a crucial opportunity for research to explore how geographical differences in consumer sentiment can profoundly affect purchasing behaviors and preferences, underscoring the need for geographically nuanced marketing strategies and product offerings.

This dissertation aims to bridge the identified gap by developing a real-time geospatial sentiment dashboard. By integrating ML and NLP to analyze geospatially tagged consumer feedback, this tool is designed to offer real-time insights into regional consumer trends. The developed dashboard allows fashion e-commerce businesses to tailor their marketing efforts and product offerings to specific geographic locations, potentially transforming how businesses engage with their regional markets.

The development of artefact is expected to significantly impact fashion e-commerce by enabling businesses to enhance their customer engagement strategies based on nuanced, location-specific consumer data. This capability will provide a competitive edge in a market where understanding and responding to regional preferences quickly can mean the difference between success and failure. The anticipated benefits include improved customer satisfaction, increased sales through better-targeted marketing campaigns, and a stronger brand loyalty cultivated through personalized consumer interactions.

**Research Motivation**

The rapid digital transformation within the fashion e-commerce sector has prompted a significant shift from traditional retail methods to a more complex, globally interconnected marketplace. This evolution underscores the necessity for robust analytical tools that can process and interpret vast amounts of consumer data, including sentiments and geographic variables. Recognizing this need, this research is motivated by the potential to harness advanced technologies such as Python, Jupyter Notebook, and Streamlit to develop a geospatial sentiment dashboard. By integrating sentiment analysis with geospatial data, the project aims to unveil intricate consumer behavior patterns across different regions, thereby enabling businesses to tailor their strategies with unprecedented precision and insight.

The importance of this artifact in the context of fashion e-commerce cannot be overstated. With the use of Python for data analysis and machine learning tasks, complemented by the dynamic visualization capabilities of Streamlit, the geospatial sentiment dashboard stands as a critical innovation. It not only facilitates a deeper understanding of regional consumer sentiments but also enhances marketing effectiveness and customer engagement. By providing fashion retailers with a tool to visualize sentiment trends across various locations, the dashboard allows for more informed decision-making. This can lead to highly optimized marketing strategies and product placements, ultimately fostering a more personalized and satisfying shopping experience for consumers globally. The potential impact on business operations and strategic planning within the fast-paced e-commerce landscape is profound, making this research both a valuable academic contribution and a practical asset for industry practitioners.

**Aim**

This project is dedicated to designing and implementing a cutting-edge geospatial sentiment dashboard tailored for the fashion e-commerce industry. By merging advanced techniques in natural language processing (NLP) and machine learning (ML) with geospatial analytics, this dashboard aims to transform raw consumer feedback into actionable geographic insights, enabling businesses to optimize their marketing strategies and product placements effectively.

**Objective**

1. **Literature Review:** To thoroughly examine existing research on sentiment analysis and geospatial data integration within the e-commerce context to identify research gaps and potential technological solutions.
2. **Technology Exploration:** To evaluate and select appropriate NLP and ML methodologies that effectively analyze and interpret sentiment data when integrated with geospatial information.
3. **System Development:** To develop and configure a cloud-based geospatial sentiment dashboard that utilizes real-time data processing to map consumer sentiments across diverse geographic locations using python and streamlit application.
4. **Performance Validation**: To rigorously test and validate the dashboard in other scenarios, ensuring its efficacy in providing precise sentiments and region-specific consumer insights.
5. **Knowledge Dissemination:** To contribute to academic and industry knowledge through comprehensive documentation of the research findings and the development process, offering insights for future enhancements and applications.

**Deliverables**

1. **Comprehensive Review Report:** A synthesis of current literature surrounding sentiment analysis and geospatial data utilization within e-commerce.
2. **Technical Framework Documentation:** Detailed documentation of the methodologies and technologies employed in the dashboard's development, including codebases and system architecture.
3. **Operational Geospatial Sentiment Dashboard:** A fully functional dashboard interface that provides real-time sentiment analysis across various geographies.
4. **Validation Report:** An assessment of the dashboard's performance through case studies and empirical data, demonstrating its impact on marketing strategies and consumer engagement.
5. **Final Project Report:** A complete report detailing the research process, findings, and recommendations for future work.

# CHAPTER 2 LITERATURE REVIEW

## 2.1 E-Commerce

E-commerce, which is also short for electronic commerce, refers to the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet. This business model encompasses a wide range of business types and consumer interactions, from retail sites and auction or music sites to business exchanges trading goods and services between corporations (Taher, 2021).

E-commerce has remarkably transformed the landscape of global trade and commerce, evolving from the early adoption of Electronic Data Interchange (EDI) in the 1960s to the proliferation of online marketplaces such as eBay and Amazon in the 1990s. This evolution reflects a significant shift toward a digital economy, revolutionising how businesses operate and interact with consumers worldwide. E-commerce encompasses a broad spectrum of models as explained by (Taher, 2021) below.

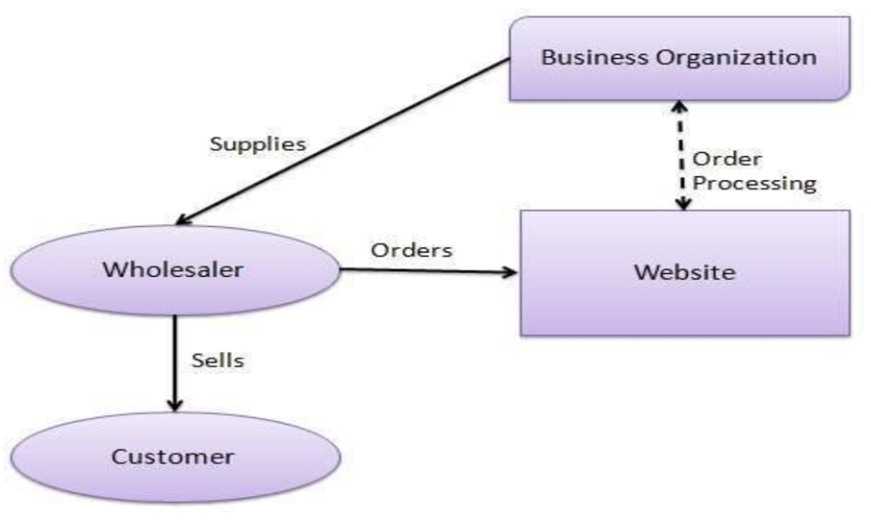


Figure 1: B2B

1. Business to business model(B2B) involves e-commerce involves transactions between two companies, such as a manufacturer selling to a wholesaler, or a wholesaler to a retailer.

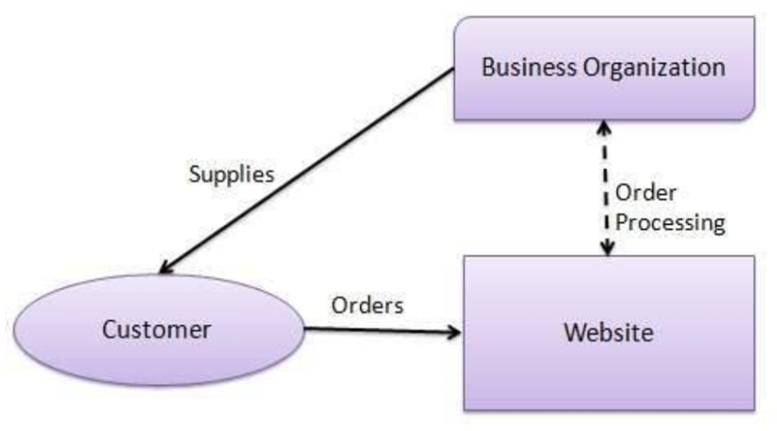


Figure 2:B2C

1. Business to customer model(B2C) refers to the direct sale of goods or services from businesses to individual consumers, typically through online storefronts.

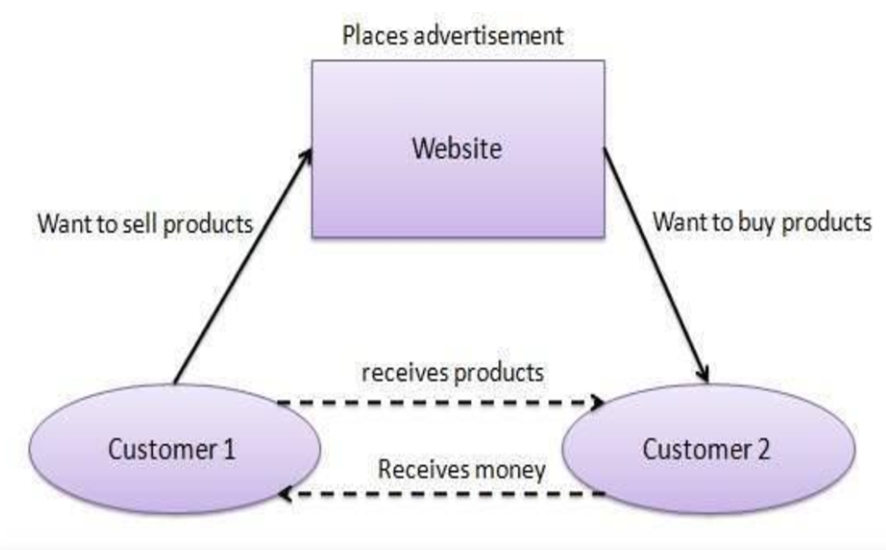


Figure 3: C2C

1. Customer to customer model (C2C) allows consumers to sell goods or services to each other, often facilitated by a third-party platform like eBay or Craigslist.

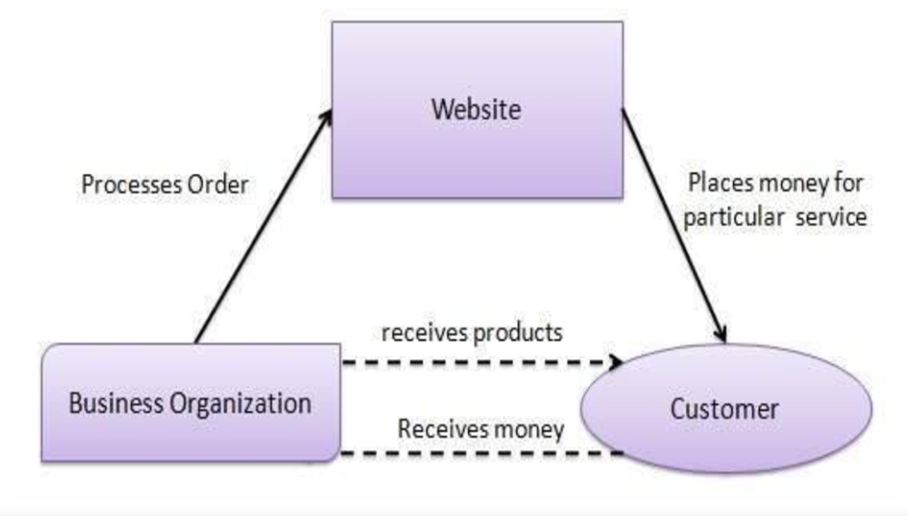


Figure 4:C2B

1. Customer to Business model (C2B) enables consumers to sell products or services to businesses, such as a freelancer offering design services to a company.

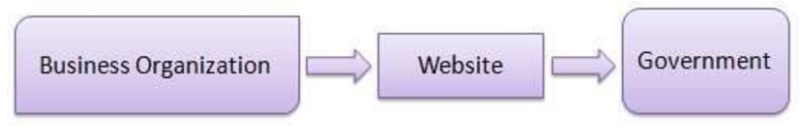


Figure 5:B2A

1. Business to Administration model (B2A) involves transactions conducted between companies and public administrations or government entities, often related to legal documents, social security, etc.

Each model facilitates unique transactional relationships among consumers, businesses, and administrations, highlighting the diverse and pervasive impact of e-commerce in today's digital commerce landscape (Tsagkias et al., n.d., 2021).

The journey of e-commerce marks a series of technological advancements and paradigm shifts that have expanded market reach and enhanced the efficiency of business transactions. The advent of the internet and subsequent technological innovations have played pivotal roles in this transformation, allowing for the seamless exchange of information and goods over digital platforms. As e-commerce continues to evolve, it not only broadens the scope of global trade but also introduces new challenges and opportunities in the digital marketplace, shaping the future of commerce in the digital age.

### 2.1.1 Fashion E-Commerce

Fashion e-commerce Specifically pertains to the segment of e-commerce dealing with the sale and purchase of fashion items, including clothing, accessories, and shoes, through online platforms. It involves a complex interplay of fashion trends, consumer preferences, and digital marketing strategies, often leveraging advanced technologies like AI and ML for enhanced customer experiences (Wang, 2023).

The journey from broad e-commerce platforms to the specialised niche of fashion e-commerce illustrates a unique evolution within the digital marketplace. This sector, focusing on the online sale and purchase of fashion items like clothing, accessories, and shoes, has seen remarkable growth driven by a confluence of fashion trends, consumer preferences, and innovative digital marketing strategies. The introduction of innovative technologies has significantly enhanced the online shopping experience, making it more personalised and engaging (Shukla, 2023).

The shift towards fashion e-commerce has mirrored broader technological advancements and changes in consumer behavior that have transformed the e-commerce landscape at large. This transformation has been significantly accelerated by the COVID-19 pandemic, which has led to an increased reliance on online platforms for shopping, thereby highlighting a critical period of adaptation and opportunity within the fashion retail sector (Nayak et al., 2023).

The fashion industry's rapid expansion online has not come without its challenges and opportunities. There is a growing demand for sustainable practices as digital consumption increases, reflecting a shift in consumer buying behaviors towards environmental consciousness (Bindoo et al., 2021). Furthermore, the industry has had to navigate the intricacies of digital distribution and mobile commerce (m-commerce), essential for the survival and growth of startup fashion brands, especially amidst global health crises (Skurpel, 2022).

Given the rapid evolution and specific challenges of fashion- e-commerce, such as sustainability and the need for effective digital marketing strategies, there's a clear indication that understanding and leveraging consumer engagement and feedback is crucial. This understanding not only helps navigate the complexities of the digital marketplace but also ensures that fashion e-commerce continues to grow and evolve in response to changing consumer habits and technological advancements.

## 2.2 Customers

In e-commerce, the customer's voice plays a pivotal role in shaping business strategies and driving success, particularly within the highly competitive fashion sector. Understanding customer behaviors, feedback, and engagement is crucial for businesses aiming to thrive in the digital marketplace. As e-commerce platforms increasingly become the primary touchpoint between fashion brands and their consumers, analysing and responding to customer interactions becomes essential. This not only aids in tailoring product offerings and enhancing user experiences but also in fostering brand loyalty and trust. Consequently, a deep dive into customer feedback emerges as a natural progression, revealing insights that empower brands to meet and exceed consumer expectations, thereby solidifying their position in the market (Sagvekar & Sharma, 2021).

### 2.2.1 Customer Feedback

Customer feedback in the e-commerce sector, particularly within the fashion industry, is an invaluable asset for businesses aiming to enhance their market presence and adapt to consumer needs effectively. As highlighted by (Shiv Ratan & Divya, n.d.), customer reviews and ratings, prominently displayed on e-commerce platforms, significantly influence purchasing decisions. The power of sentiment analysis in interpreting customer feedback enables brands to align their product offerings and marketing strategies with consumer expectations. This analysis underscores the critical role of feedback in the continuous improvement of product quality and customer service, ultimately contributing to a brand's reputation and consumer trust.

### 2.2.2 Customer Engagement

Customer Engagement in the context of e-commerce refers to the methods and strategies employed by businesses to interact with customers through various channels in a meaningful way. It aims to create a relationship beyond the transactional level, focusing on generating loyalty and long-term engagement by providing value to the customer through interactive, personalised experiences (Fedushko & Ustyianovych, 2022).

Engagement within the digital marketplace is not just about transactions but about fostering a deeper connection between consumers and brands. Research by (Sempe & Naong, 2021) emphasises the substantial impact of targeted awareness campaigns in enhancing engagement, particularly among younger consumers and females, suggesting that well-informed customers are more likely to interact positively with e-commerce environments. Furthermore, (J Tamara et al., 2023) explore the emerging technologies like chatbots in e-commerce, which have revolutionised customer service by providing timely and personalised assistance, thereby significantly enhancing customer engagement. These innovations highlight the shift towards more interactive and responsive e-commerce experiences, tailored to meet the dynamic preferences of the digital consumer.

### 2.2.3 Importance of Customer Feedback / Engagement in E-commerce Success

The amalgamation of customer feedback and engagement forms the cornerstone of e-commerce success. The insights derived from customer feedback, when analysed through advanced sentiment analysis tools, offer businesses a detailed understanding of consumer preferences and pain points. As elucidated by (Kawasaki et al., 2022), the application of machine learning and natural language processing in analysing customer interactions has paved the way for more nuanced and personalised e-commerce strategies. This meticulous approach to incorporating customer feedback and engagement metrics into business strategies not only enhances customer satisfaction but also drives brand loyalty and repeat purchases, underscoring their pivotal role in the thriving e-commerce landscape.

The synthesis of customer feedback and engagement, powered by technological advancements in sentiment analysis and machine learning, presents a transformative opportunity for e-commerce platforms, especially within the fashion sector, to refine their offerings and marketing approaches. The emphasis on leveraging customer interactions as strategic insights for business growth highlights the evolving dynamics of the digital marketplace, where consumer sentiments and behaviors significantly dictate the success and sustainability of e-commerce ventures.

2.3 Sentiment Analysis and Machine Learning in e-commerce

The advent of sentiment analysis and machine learning has revolutionised the way e-commerce platforms interpret and leverage customer feedback and engagement. These advanced analytical tools go beyond mere data collection, offering nuanced insights into customer sentiments, preferences, and behaviors. By applying machine learning algorithms to customer reviews, social media interactions, and other forms of feedback, businesses can identify trends, predict consumer behavior, and personalise the shopping experience. Sentiment analysis allows for the qualitative analysis of textual feedback, enabling e-commerce platforms to understand the emotional drivers behind customer opinions. Together, these technologies empower e-commerce entities to make data-driven decisions, enhancing customer satisfaction and fostering a more engaging online shopping environment (Rane, 2023).

## 2.3.1 Geospatial Analysis in E-commerce

Geospatial Analysis involves the gathering, display, and manipulation of imagery, GPS, satellite photography, and historical data, described explicitly in terms of geographic coordinates or implicitly in terms of a street address, postal code, or forest stand identifier as they apply to geographic models. In e-commerce, geospatial analysis can be used to understand consumer behavior patterns, market trends, and logistics optimisation across different regions (Li et al., 2020). Moreso, Geospatial data encapsulates information associated with geographic locations on Earth, enabling the analysis and visualisation of spatial phenomena and patterns. In the context of analysing social media data during emergencies, geospatial data plays a pivotal role in understanding how public behaviors and discussions are geographically distributed and evolve over time.

(Han et al., 2024) shows various methods to represent geospatial data within a comprehensive analytical framework:

1. Spatial Indexing: This technique involves converting continuous geographic coordinates into a discrete spatial index by segmenting the geographic area into a grid of longitude and latitude. Each data point is assigned to a specific grid cell, simplifying spatial analysis by categorising data into manageable units.
2. Spatial Sequences: By tracking the sequence of geographic locations or grid cells associated with user activities, spatial sequences offer insights into movement patterns and trajectories, enriching the analysis with temporal and spatial dynamics.
3. Time-Geographic-Semantic Cube (TGSC): An innovative multidimensional model that integrates spatial data with time and semantic information. The TGSC enables dynamic modeling of social media user behavior, revealing how spatial patterns intersect with temporal trends and thematic discussions.

The integration of Geospatial Analysis into e-commerce platforms represents a transformative approach to understanding and engaging with consumers across diverse geographic locations. This innovative technique allows businesses to map out and analyse customer behaviors, preferences, and trends with a regional focus, facilitating more targeted and effective marketing strategies. Geospatial data analysis offers a unique lens to view consumer behavior, revealing insights that traditional analysis methods might overlook.

One of the pivotal contributions to this field is the work by (Liu et al., 2022), which highlights how geospatial analysis can significantly impact e-commerce, from retail space valuation to customer engagement patterns. Their research underscores the utility of geospatial data in optimising logistical operations and tailoring marketing efforts to meet the specific needs of different regions, thereby enhancing customer satisfaction and operational efficiency.

Furthermore, the study by (Figueiredo et al., 2021) elaborates on the use of geospatial data to analyse the geographical distribution of consumer demand and its implications for e-commerce strategy formulation. By incorporating geospatial insights, e-commerce platforms can identify high-demand areas, optimise inventory distribution, and customise promotional activities to align with local consumer preferences, demonstrating the critical role of geospatial analysis in enhancing market penetration and revenue growth.

The convergence of geospatial analysis with advanced data analytics technologies opens new avenues for the e-commerce sector. As businesses continue to explore this integration, they unlock the potential for a more nuanced understanding of consumer sentiments across different geographical areas. This, in turn, enables the development of targeted marketing strategies and enhances customer engagement by delivering personalised experiences based on location-specific insights (Liu et al., 2022; Figueiredo et al., 2021).

The adoption of geospatial analysis in e-commerce not only aligns with the industry's ongoing digital transformation but also opens new possibilities for understanding and engaging with the consumer base more effectively. By leveraging these advanced technological approaches, fashion e-commerce businesses stand to gain a competitive edge, ensuring that their strategies are responsive to the evolving digital marketplace and the geographical nuances of consumer demand.

2.3.2 Sentiment Analysis in E-commerce

Sentiment Analysis is a field within natural language processing (NLP) and machine learning (ML) that focuses on identifying, extracting, and quantifying affective states and subjective information from text data. In e-commerce, sentiment analysis is utilised to gauge customer opinions, feedback, and emotions toward products or services, facilitating better business decision-making and strategy development (Mehta et al., 2023).

Sentiment Analysis is a crucial subset of NLP and has become an invaluable tool in the e-commerce sector for gauging consumer opinions, emotions, and attitudes towards products and services. This computational method interprets and classifies subjective information in textual data, enabling businesses to understand consumer sentiments at scale (Zhou et al., 2020). The proliferation of user-generated content on digital platforms—such as reviews, ratings, and social media posts—has underscored the importance of sentiment analysis in today’s data-driven decision-making processes.

The process of sentiment analysis typically involves several key steps (Polyakov et al., 2020):

1. Data collection: Gathering relevant data from various sources such as social media, reviews, or forums, which contain opinions or sentiments to be analysed.
2. Preprocessing: Cleaning and preparing the data for analysis by removing noise (e.g., irrelevant characters, stopwords), normalising text (e.g., lowercasing, stemming), and handling missing values or duplicates to improve the quality and efficiency of the analysis.
3. Exploratory data analysis (EDA): Performing initial investigations on the collected data to discover patterns, spot anomalies, identify important variables, and test hypotheses using statistical figures and visualisation tools.
4. Vectorisation: Converting text data into numerical format so that machine learning algorithms can process it. Common types of vectorisation include:
5. Count Vectorisation: Represents text documents as matrices of token counts but ignores the order of words.
6. TF-IDF Vectorisation (Term Frequency-Inverse Document Frequency): Weighs the words' frequency against their rarity across documents, highlighting the importance of certain terms.
7. Word Embeddings: Represents words in dense vector spaces where the semantic similarity between words is reflected in their closeness within the vector space.
8. Modeling for classification: Applying machine learning or deep learning algorithms to the vectorised data to classify sentiments into categories (e.g., positive, negative, neutral). This involves training a model on a labeled dataset and then using it to predict the sentiment of new, unseen texts.

According to these steps above, textual data is gathered from various online sources, followed by EDA to grasp the data’s characteristics, and inform preprocessing efforts. Preprocessing prepares the text for analysis, which includes tokenisation, removing stopwords, and lemmatisation, among others. Subsequently, vectorisation techniques convert the text into a numerical format suitable for machine learning models. Advanced neural networks, such as RNNs, LSTMs, and transformers like BERT and GPT, are then employed to classify the sentiment of the text accurately.

Despite the advancements in ML and NLP, the implementation of sentiment analysis in e-commerce faces inherent challenges, including the need for vast and diverse datasets to train models effectively and the complexity of accurately interpreting the nuances of human language (Tusar & Islam, 2021). Ethical considerations around data privacy and the potential for algorithmic bias further complicate the deployment of sentiment analysis technologies. Nonetheless, the integration of sentiment analysis into e-commerce platforms enables businesses to respond more effectively to consumer sentiments, tailor marketing strategies, optimise product offerings, and ultimately enhance the customer experience.

By leveraging sentiment analysis, e-commerce platforms can uncover valuable insights into consumer preferences and trends, facilitating a more personalised shopping experience. This approach not only aids in refining marketing strategies but also plays a significant role in product development and customer service improvement, signifying sentiment analysis as an essential component of the modern e-commerce ecosystem.

### 2.3.3 Geospatial Sentiment Analysis in E-commerce

The (Dejene et al., 2022) research presents a novel approach to enhancing location-based services through the integration of sentiment analysis. Utilising travel review data from the UCI machine learning repository, the study applies advanced algorithms, including Random Forest, Gaussian Naive Bayes, and LSTM, to analyse consumer feedback. Notably, the Bi-LSTM model outperforms others, showcasing the robustness of deep learning in dissecting sentiments for geospatial recommendations. The successful merger of textual with numerical data not only boosts the system's precision but also marks a significant step forward in creating more nuanced and personalised recommendation systems. While the study presents a compelling integration of sentiment analysis with geospatial data, it primarily focuses on travel reviews, which may limit its direct applicability and insights into the distinct dynamics of fashion e-commerce, where consumer preferences and trends can significantly differ. This study aligns with shows the potential benefits of leveraging deep learning techniques to interpret sentiments within specific geographic contexts, offering a foundational framework for enhancing customer engagement through tailored recommendations.

Also, (Zhao et al., 2024) discussed a sophisticated BERT-Fusion-DNN framework is introduced, designed to refine e-commerce recommendation systems by fusing BERT-extracted textual insights with numerical customer data through a Deep Neural Network. Analysing an extensive dataset of 23,486 customer reviews from the Women Clothing E-Commerce sector, the framework excels in gleaning subtle consumer preferences, significantly surpassing conventional recommendation models. This method's success in harmonising diverse data types to enhance recommendation accuracy is particularly relevant to the development of a geospatial sentiment dashboard. It underscores the importance of integrating detailed textual analysis with demographic and geographical insights to deliver highly personalised e-commerce experiences, laying a critical pathway for future research into optimising fashion e-commerce strategies.

The study (Patel, n.d.) ventures into analysing e-commerce metrics and customer sentiments using data from a significant number of orders on a Brazilian e-commerce platform. Through the employment of Logistic Regression, PCA, and other analytic techniques, the research effectively categorises customer feedback and offers product recommendations. This study highlights the indispensability of rigorous data preprocessing for the accuracy of sentiment analysis and validates Logistic Regression as a potent tool for sentiment categorisation. This research bears relevance to the proposed geospatial sentiment dashboard by illustrating the critical role of sentiment analysis combined with exploratory data analysis in understanding e-commerce dynamics. this study is fully based on Brazilian e-commerce platform sentiment analysis offers valuable insights into customer feedback and sentiment categorisation, but the research's geographical and cultural specificity to the Brazilian market do not translate seamlessly to other regions or global e-commerce contexts, limiting the universality of the proposed models and findings in a globalised fashion e-commerce environment. The methodologies and findings however can provide valuable insights into how sentiment analysis can inform targeted marketing efforts and product positioning in the fashion e-commerce sector, further supporting the project's aim to leverage machine learning for improved consumer insight and engagement.

### 2.3.4 Advanced Technologies in E-commerce

In the contemporary e-commerce landscape, the advent of Advanced Technologies such as ML and NLP heralds a paradigm shift in the realm of online retail, particularly within the fashion industry. These technologies are fundamentally reshaping how businesses engage with their consumers, personalise shopping experiences, and derive actionable insights from complex data streams.

ML, characterised by its ability to learn and make informed decisions from data without being explicitly programmed, is revolutionising the fashion e-commerce space through its various subsets (Lee & Shin, 2020). Its algorithms can be broadly categorised into supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning algorithms, for example, leverage labeled datasets to predict customer preferences and enhance product recommendations. Unsupervised learning algorithms excel in identifying hidden patterns in consumer behavior, providing e-commerce platforms with insights into emerging trends. Semi-supervised learning and reinforcement learning further refine these capabilities, offering more accurate predictions and adaptive learning mechanisms that optimise customer interactions based on real-time feedback (Sarker, 2021). These categories of machine learning can be seen in the figure below.

A diagram of a machine learning type

Description automatically generated

Figure 6: TYPES OF MACHINE LEARNING (SARKER, 2021)

The incorporation of ML techniques and geospatial analysis into sentiment analysis emerges as a novel approach in e-commerce, particularly salient for the fashion industry. By mapping consumer sentiments across different locales, businesses can gain granular insights into regional preferences and behavioral patterns, enabling a more strategic allocation of marketing resources and product offerings (Liu et al., 2022; Figueiredo et al., 2021). However, the creation of a dynamic, real-time geospatial sentiment analysis dashboard that effectively integrates and visualises this data remains an uncharted territory. Such a tool would not only empower e-commerce platforms with unparalleled strategic insights but also enhance the shopping experience for consumers through more personalised engagement and recommendations.

Despite the potential of ML and NLP to revolutionise e-commerce practices, the journey from theoretical potential to practical application is fraught with hurdles. The scalability of solutions that leverage these technologies, the management of vast and ever-expanding data repositories, and the imperative to ensure ethical usage of consumer data are critical considerations that must be addressed. Recent studies (Suresh et al., 2023; Nazeer et al., n.d.) underscore the promise of these technologies in real-time sentiment analysis and consumer engagement. Yet, the development of scalable, ethical, and effective frameworks for their application in fashion e-commerce necessitates further investigation.

In summary, the integration of Machine Learning and Natural Language Processing technologies holds immense promise for transforming the fashion e-commerce landscape by enabling a deeper, data-driven understanding of consumer behaviors and preferences. Yet, to achieve this potential fully requires a concerted effort to address the technical, ethical, and operational challenges associated with these technologies. By advancing research in this domain, particularly in the development of integrated geospatial sentiment analysis tools, this dissertation aims to bridge the gap between the theoretical capabilities of ML and NLP and their practical application in enhancing customer engagement and personalisation within the fashion e-commerce sector. The pursuit of this research not only aligns with the overarching goal of improving the e-commerce experience for consumers but also contributes to the scholarly discourse on the application of advanced technologies in retail. Ultimately, by addressing the existing challenges and exploring innovative solutions, this work seeks to provide actionable insights that can drive the future of fashion e-commerce towards more personalised, responsive, and ethically grounded practices, thereby ensuring that businesses can meet and exceed the evolving expectations of their digital consumers.

### 2.3.5 Application of Advance Technologies in Fashion E-commerce

**ASOS**

ASOS.com, a leading online fashion retailer, deployed "Promotheus," a machine learning framework, to innovate its pricing strategy, marking a departure from traditional rule-based methods. Promotheus, in conjunction with "Ithax," a tool designed for supply-side pricing, employs gradient-boosted trees to predict demand and optimise markdowns based on price elasticity. This approach allowed ASOS to execute markdown strategies with unprecedented precision, leading to an 86% increase in profitability during discount events. This example showcases the potential of machine learning to revolutionise pricing strategies in fashion e-commerce, allowing for data-driven decisions that significantly enhance operational efficiency and profitability. The integration of such advanced technologies not only bolsters profit margins but also sets new standards for strategic planning within the industry. ASOS’s experience illustrates how machine learning can be applied to tackle specific challenges in fashion e-commerce, such as optimising markdown strategies, and underscores the importance of leveraging technology to drive business success (Loh et al., 2022).

**MYNTRA**

Myntra, an Indian fashion e-commerce leader, has harnessed Artificial Intelligence (AI) to revolutionise the customer shopping experience and optimise operational efficiencies. By employing AI for functions such as personalised recommendations, visual search capabilities, virtual try-ons, and dynamic pricing, Myntra has addressed key challenges in customer satisfaction and market competitiveness. The AI-driven recommendation engine customises product suggestions based on user behavior, enhancing engagement and conversion rates, while visual search and virtual try-on features, powered by computer vision, reduce return rates by allowing customers to see how products look on them before purchase. Dynamic pricing algorithms adjust prices in real-time based on various factors, optimising revenue opportunities. These AI implementations have led to notable improvements in user engagement, product discovery efficiency, reduced returns, and increased sales, demonstrating the transformative impact of AI on fashion e-commerce. Myntra’s case underscores the significance of AI in crafting personalised customer experiences and streamlining business processes, highlighting the role of technology in advancing the fashion e-commerce sector toward more personalised, efficient, and customer-centric operations (Sungheetha et al., 2023).

## 2.4 Gaps and Opportunities

The literature review identifies a pivotal gap in applying geospatial analysis and sentiment analysis using NLP in the fashion e-commerce landscape. While sentiment analysis and ML have been recognised for their potential to transform consumer insights and business strategies across e-commerce by (Zhou et al., 2020; Tusar & Islam, 2021; Rane, 2023), there is a notable lack of studies that specifically explore the integration of these technologies with geospatial data in the fashion e-commerce context. This gap signifies a crucial research opportunity to investigate how geographical differences in consumer sentiment profoundly affect purchasing behaviors and preferences, underscoring the need for geographically nuanced marketing strategies and product offerings.

Moreover, the review indicates a marked scarcity of research focusing on the adaptation and efficacy of ML and NLP techniques for processing geospatially tagged consumer feedback within fashion e-commerce. Although the utility of ML and NLP is well-documented in broader e-commerce analyses by (Suresh et al., 2023; Nazeer et al., n.d.; Loh et al., 2022), the studies lack an in-depth exploration of these technologies' optimisation for fashion e-commerce's unique challenges. This oversight highlights an urgent need for dedicated research to tailor and refine ML and NLP methodologies for accurately capturing and analying the geographic variations in consumer sentiments within the fashion e-commerce domain. Such focused investigations are paramount for advancing the capabilities of sentiment analysis in fashion e-commerce, ensuring that these technologies are leveraged to augment the precision of geographic insights, thereby fostering more strategic and informed business decisions.

Furthermore, despite acknowledging the potential benefits of melding sentiment analysis with geospatial data as explored by (Patel, n.d.; Zhao et al., 2024; Dejene et al., 2022; Liu et al., 2022), there's a significant gap in practical explorations into creating and deploying a dynamic, real-time tool capable of efficiently processing and visualising this intricate dataset. existing literature has not addressed the development and operationalisation of a real-time geospatial sentiment analysis tool tailored for the fashion e-commerce sector. Filling this gap is crucial for empowering fashion e-commerce platforms to utilise instant insights into regional consumer sentiments, enabling the deployment of agile, precisely targeted marketing, and product development strategies.

In response to these identified research gaps, this dissertation aims to pioneer an innovative, real-time geospatial sentiment analysis dashboard, engineered specifically for the fashion e-commerce sector. By exploiting advanced NLP technologies and incorporating ML techniques, this proposed solution intends to conduct a comprehensive analysis of customer feedback, identifying regional sentiment trends to deliver unprecedented insights into the geographical nuances of consumer preferences. This research will rigorously evaluate the most suitable ML and NLP techniques for this endeavor, alongside developing an integration framework that employs cloud technologies for real-time data processing and visualisation. This initiative aspires to profoundly impact the fashion e-commerce industry, providing businesses with crucial intelligence to tailor their marketing and product strategies effectively, thereby promoting enhanced customer engagement and adoption rates. Through its novel approach and practical implementation, this research strives to bridge a significant literature gap, contributing to the progression of e-commerce strategies that resonate with the dynamic preferences of a global consumer base.

Table 1: table of summary of past work

|  |  |  |  |
| --- | --- | --- | --- |
| **Author/year** | **Major approach** | **Result achieve** | **Limitation** |
| (Dejene et al., 2022) | Machine Learning Models: The study employed various models including Random Forest and Bi-LSTM (Bidirectional Long Short-Term Memory) for sentiment analysis and recommendations based on user feedback and characteristics of locations. | Random Forest achieved 85% accuracy with a Mean Squared Error (MSE) of 0.87. Bi-LSTM provided approximately 84.7% accuracy with the lowest MSE among the models tested. | general challenges in sentiment analysis and geospatial recommendation systems could include handling diverse and unstructured data from user reviews, ensuring the accuracy of sentiment analysis, and scalability of the recommendation system. |
| (Suresh et al., 2023) | The study utilised Convolutional Neural Networks (CNNs) and Bidirectional Encoder Representations from Transformers (BERT) to perform emotion analysis from social media data. The analysis was based on the GoEmotion dataset, which includes 58,000 comments labeled for 28 emotion categories. | The BERT model showed strong performance in predicting specific emotions like gratitude, love, and admiration with F1-scores of 0.82, 0.57, and 0.48, respectively. The CNN model also demonstrated high precision, recall, and F1-scores in predicting "gratitude” | The study faced challenges with the macro-average recall and F1-score, which did not perform as well as anticipated. This suggests that while the models were precise, they might not have been as effective in covering all instances of emotions comprehensively. |
| (Tusar & Islam, 2021) | Natural Language Processing (NLP) techniques (Bag-of-Words, TF-IDF); Machine Learning algorithms (SVM, Logistic Regression, Multinomial Naive Bayes, Random Forest) | The best approaches achieved 77% accuracy using Support Vector Machine and Logistic Regression with the Bag-of-Words technique. | The study addressed the challenge of analysing sentiment on a large, imbalanced, and multi-classed dataset. |
| (Polyakov et al., 2020) | The research compared classical machine learning algorithms and deep learning approaches. They used a variety of models, including Logistic Regression, Random Forest, and deep learning models like LSTM and Bidirectional LSTM. | The best results were obtained using LSTM with deep learning, achieving a ROC-AUC of 0.9133 after hyperparameter optimisation. | The study noted that while the proposed approach improved the accuracy of sentiment analysis, the training and optimisation process was time-intensive, especially for deep learning models which required significant computational resources. |

## 2.5 Analysis Of Problem/improvement

**Analysis of Problem**

Current research as highlighted in research gap and the table 1 above shows the gap in the integration of geospatial data with sentiment analysis. This interdisciplinary approach faces challenges mainly due to the complexity of merging data that vary widely in structure and semantic content. While various studies have proposed methodologies to address these issues, they often fall short when adapted to geospatial frameworks. For instance, the GeoRSA system as shown in table 1, despite its accuracy, struggles with the diverse and unstructured data inherent in user reviews, which is vital for geospatial applications (Dejene et al., 2022). Furthermore, even high-performing models like those used by (Suresh et al., 2023) reveal deficiencies in emotion coverage, indicating potential complications in geospatial sentiment analysis where data variability is more pronounced. These challenges underscore the need for innovative approaches that can enhance data processing and sentiment interpretation in a geospatial context.

**Benefits and Improvements through the Artifact:**

The geospatial sentiment dashboard aims to address these deficiencies by enhancing the accuracy and efficiency of sentiment analysis in geospatial settings. This artifact is expected to improve the handling of linguistic complexities and mixed emotions in geographically varied datasets by:

**Advanced Data Preprocessing:** Implementing robust techniques to manage encoding issues and normalise text data, thus ensuring high-quality inputs for analysis.

**Sophisticated Feature Engineering**: Developing features that accurately capture consumer sentiments and geographic specifics, which are crucial for refining the model's predictive capabilities.

**Utilisation of Cutting-edge NLP Techniques**: Employing advanced techniques like lemmatisation to better analyse consumer feedback, which often includes slang and regional colloquialisms.

**Strategic Machine Learning Model Selection:** Choosing models adept at managing high-dimensional and complex datasets typical in geospatial analyses.

**Real-Time Processing and Visualisation:** Creating a dynamic dashboard that provides real-time insights into consumer sentiments across different regions, thereby enabling businesses to make informed decisions swiftly.

These enhancements not only address the limitations noted in previous studies but also set a new standard for integrating sentiment analysis with geospatial data. This advanced capability significantly improves the potential of fashion e-commerce platforms to engage customers more personally and effectively, driving better business outcomes.

# CHAPTER 3. RESEARCH METHOD

This chapter delineates the methodology employed to develop a geospatial sentiment dashboard tailored for the fashion e-commerce sector. Utilising advanced technological tools such as Machine Learning (ML), Natural Language Processing (NLP), and geographical analysis, the methodology is designed to thoroughly explore and address existing research gaps. It emphasises understanding regional consumer behaviors, enhancing customer engagement, and optimising targeted marketing strategies. This systematic approach not only provides insights into the dynamic interplay of consumer sentiments across various geographical regions but also fosters a deeper understanding of market trends and consumer preferences. These insights are crucial for formulating effective business strategies in the highly competitive realm of fashion e-commerce. The chapter outlines the structured research framework used, discusses the data collection and analysis processes, and details the ethical considerations adhered to throughout the study.

### 3.1 Research Onion Overview

The research onion, conceptualised by (Tengli, 2020) offers a structured framework to systematically peel back the layers of research methodology, from the outermost philosophical foundations to the more detailed aspects of data collection and analysis methods. This framework is particularly suited to methodologically sound investigations into complex research questions.

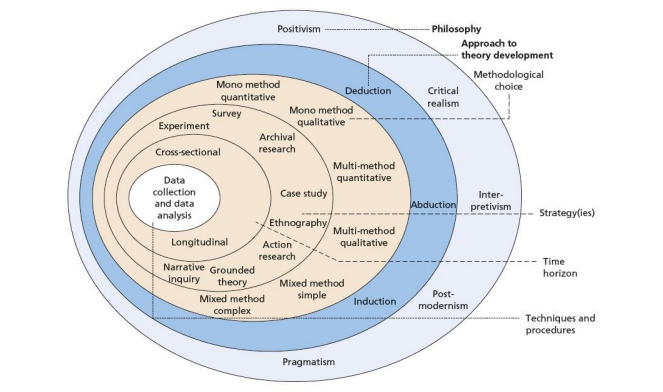


Figure 7 research onion (Mardiana, n.d., 2020)

In this dissertation, the research onion is adapted to specifically address the challenges and intricacies of geographical sentiment analysis within the fashion e-commerce industry. The adaptations ensure a thorough examination of the factors influencing consumer behavior across different regions, providing a robust methodological base for the study.

The following outlines how the research onion is tailored for this dissertation:

* Philosophy: Adapted to incorporate positivist approaches suitable for quantitative data analysis.
* Approach: Deductive methodologies are utilised to confirm hypotheses derived from existing literature.
* Choice: Emphasises a mono-method qualitative approach, enriched with quantitative data analysis to deepen the understanding of consumer sentiments.
* Strategy: Leverages archival data from extensive datasets, facilitating comprehensive sentiment analysis.
* Time Horizon: Employs a cross-sectional design to capture a snapshot of consumer sentiments at a specific time.

Each layer's specific adaptations and their implications are discussed in greater detail in subsequent sections of this chapter. This structure not only maintains the integrity of the research framework but also aligns it closely with the objectives of studying geographical sentiment analysis in an e-commerce context. Table 2 below shows the method chosen for each layer of the research onion.

Table 2: Research onion method

|  |  |
| --- | --- |
| **Research Onion** | **Approach chosen** |
| Philosophy | Positivism |
| Approach | Deductive |
| Choice | Mono-Method Qualitative |
| Strategy | Archival |
| Time Horizon | Cross-sectional |
| Data Collection and Analysis | Data Collection and Analysis |

### 3.2 Research Philosophy

The research philosophy is based on positivism, which means it relies on collecting and analysing data in an objective way to create dependable and measurable results. Positivism is useful in this situation because it enables the examination of data collected from customer reviews and geospatial information, making it easier to draw accurate conclusions about different consumer opinions in various geographic locations. This viewpoint is in accordance with the scientific process, which involves evaluating hypotheses using observable, empirical evidence gathered from structured data sources.

### 3.3 Research Approach

This dissertation employs a deductive research approach, which begins by formulating specific hypotheses derived from established ideas and literature about sentiment analysis and geospatial consumer behaviour. This deductive logic is employed to establish a clear direction for empirical inquiry, with the goal of validating or disproving these prior hypotheses. This approach enables a targeted evaluation of the connections between geographical places and consumer sentiments by utilising structured data analysis. As a result, it facilitates the testing of theoretical concepts using real-world data.

### 3.4 Research Choice

This dissertation focuses on using a mono-method qualitative approach to conduct a thorough and in-depth investigation of textual data obtained from customer interactions online. This decision is crucial since it enhances the research by providing in-depth understanding of consumer emotions and subjective experiences, which are sometimes not adequately captured using quantitative methodologies. The study reveals complex patterns of consumer sentiment that are closely linked to specific regional and cultural contexts through the analysis of text from online reviews, social media comments, and forum conversations.

### 3.5 Research Strategy

The research strategy involves utilising an archival approach, focusing on the analysis of existing datasets that encompass detailed consumer reviews coupled with geospatial tags. This dataset, sourced from online fashion platforms available to the public, provide a robust foundation for conducting in-depth sentiment analysis across different geographical areas. By leveraging this pre-collected data, the research circumvents the labor-intensive process of primary data collection, facilitating a more efficient exploration of vast data volumes. This approach not only accelerates the research process but also enhances its breadth by providing access to a diverse range of consumer interactions spread across various countries. The datasets enable a systematic examination of sentiment trends and patterns, allowing for a nuanced understanding of how geographic contexts influence consumer perceptions and behaviors. This method is particularly effective in mapping the regional distribution of sentiments and identifying specific local factors that can affect consumer responses, offering valuable insights that are critical for tailoring region-specific marketing strategies in the fashion e-commerce sector. This will be represented with an interactive dashboard built with streamlits and python for text prediction and visualisation.

### 3.6 Time Horizon

The study used a cross-sectional time horizon, which effectively captures a momentary overview of consumer sentiments across several areas. This methodological decision is beneficial for discovering present patterns and local idiosyncrasies in consumer behavior without the complexities linked to long term investigations. It offers a strong basis for examining how geographical elements impact customer attitudes during a specific and defined period, which is crucial for the timely and pertinent formulation of marketing strategies.

### 3.7 Data Collection and Data Analysis

Data for this research was sourced from TeePublic through Kaggle, known for its rich datasets that support academic research. Compliance with ethical standards ensured the protection of privacy and confidentiality through anonymising reviewer identities.

In the analysis phase, advanced Natural Language Processing (NLP) techniques like tokenisation, sentiment analysis, and entity recognition were employed. Machine learning models analysed the text and geospatial markers to predict sentiments across different regions, revealing distinct consumer behavior patterns. These insights help understand the impact of geographical contexts on consumer emotions, aiding in crafting targeted marketing strategies for the fashion e-commerce industry.

The findings underpin the development of a geographic sentiment dashboard using Streamlit, which visualises regional consumer sentiments dynamically. This approach not only upholds academic rigor but also enhances practical marketing strategies by providing actionable insights.

# CHAPTER 4. DESIGN of DASHBOARD

The chapter covers the design and development process of the geospatial sentiment dashboard crafted for the nuanced needs of the fashion e-commerce industry. It unfolds the journey from conceptualisation to the concrete realisation of a tool that amalgamates machine learning algorithms with user-centric visualisation techniques, delivering actionable insights into consumer sentiment and mapping them regionally.

## 4.1 Design of Artefact

The conceptual design of the geospatial sentiment dashboard is rooted in a systematic process that encompasses everything from data acquisition to the final interactive application. The project initiates with the goal of crafting a dashboard that vividly represents customer sentiments across geographical landscapes, specifically for the fashion e-commerce arena.

The journey begins with data collection from a prominent online fashion platform, gathering a rich dataset that serves as the cornerstone for sentiment analysis and geospatial mapping. Key features crucial for sentiment analysis, such as customer feedback and geospatial coordinates, are carefully selected and extracted from the dataset.

To prepare the data for machine learning models, a comprehensive preprocessing phase is undertaken. This involves cleansing the data, ensuring consistency, and transforming textual information into a numerical format through vectorisation. The preprocessed dataset is then split, dedicating a portion to training the model to discern sentiment patterns and another to testing the model's accuracy in predicting sentiments.

A variety of machine learning models are evaluated to determine the most effective means of classifying sentiments expressed in customer reviews. Once the optimal model is identified, it undergoes training to learn from the dataset and gain the ability to accurately interpret consumer sentiments.

The trained models are scored against the test data to evaluate their performance, focusing on their precision and effectiveness in sentiment prediction. This evaluation leads to the selection of the best-performing model, which is then poised for integration into the dashboard.

Leveraging the Streamlit framework, an interactive dashboard is designed, merging the elements of sentiment analysis with geospatial visualisation capabilities. The deployment of the model within the Streamlit application renders the sentiment analysis accessible via a web-based interface, enhancing user experience and engagement.

The dashboard itself is comprised of a predictive page, which allows for real-time sentiment predictions, and a visualisation page that employs interactive maps to display sentiment data. Users are granted the power to filter through the sentiment data by date and location, enabling them to perform tailored analysis and gain insights into consumer feedback.

A separate validation dataset ensures the reliability of the model, corroborating the consistency and accuracy of the predictions made by the dashboard. Further, an independent dataset from a disparate e-commerce platform is employed to externally validate the model, testing its robustness and ability to adapt to different data sources within the fashion retail sphere.

The final results are presented on the dashboard, effectively showcasing the tool's capabilities in providing a detailed and insightful representation of consumer sentiment trends across the globe. This design not only highlights the technical proficiency of the dashboard but also its user-centric approach, providing a seamless and informative experience for stakeholders.

A diagram of a flowchart

Description automatically generated

Figure 8: architectural design

### 4.1.1 Technological Stack

The technological stack deployed in the creation of the geospatial sentiment dashboard is a testament to the power of modern programming tools and libraries. Each component was chosen for its strength and synergy within the stack:

1. **Python:** The core programming language, renowned for its extensive support in data analysis and machine learning applications.
2. **Streamlit:** A powerful tool that enabled the rapid transformation of complex Python scripts into interactive and accessible web applications.
3. **Pandas:** Utilised for its exceptional data manipulation capabilities, allowing for sophisticated data structuring and operations.
4. **NumPy:** Incorporated for its comprehensive support for numerical computing, essential for handling large data arrays and matrices.
5. **joblib:** Selected for its efficiency in serialising Python objects, crucial for saving and reloading the machine learning models and other jobs like vectorisations.
6. **re (Regular Expressions)**: This module provided robust pattern matching and text manipulation functions, fundamental for the preprocessing stages.
7. **NLTK (Natural Language Toolkit):** A suite of libraries and programs that facilitated complex linguistic processing, such as tokenisation and lemmatisation, crucial for accurate sentiment analysis.
8. **Plotly:** A versatile graphing library that was indispensable for creating interactive, publication-quality visualisations, particularly in mapping sentiments geospatially.
9. **PyCountry:** Provided a comprehensive database of country identifiers, vital for translating country codes into names during the geospatial mapping process.

This stack not only ensured that the backend of the dashboard was capable of managing complex computations but also that the frontend remained responsive and user-friendly. The blend of these technologies fostered an environment where robust analytical capabilities met interactive and engaging visualisation.

### 4.1.2 Data Preparation and Modeling

Data is the lifeblood of any sentiment analysis engine. The TeePublic dataset from Kaggle with link <https://www.kaggle.com/datasets/nelgiriyewithana/shoppersentiments> comprising over a quarter-million customer reviews, presented a rich yet challenging canvas for the project. The data preparation phase saw the cleansing, transformation, and structuring of this data, ensuring that it could effectively train the chosen machine learning models.

1. Model selection was a meticulous process. Several models were tested.   
   Logistic Regression: A fundamental statistical approach that models the probability of a binary outcome. It's valued for its simplicity, interpretability, and robustness in various conditions.
2. Support Vector Machine (SVM): Renowned for its efficacy in high-dimensional spaces, which is characteristic of text data. SVMs are particularly adept at finding the optimal hyperplane that maximises the margin between classes.
3. LinearSVC: A variant of SVM that specialises in handling linear separations, this model is optimised for scalability and performance on large datasets.
4. Extra Trees Classifier: This ensemble model employs many randomised decision trees, aimed at reducing the variance and overfitting common in singular decision trees.
5. Random Forest: Another ensemble method that constructs multiple decision trees at training time and outputs the mode of the classes for classification. It provides a good balance between accuracy and overfitting.

After evaluating several candidates, the Support Vector Machine (SVM) model was selected for its adeptness at classifying high-dimensional textual data. SVM's ability to discern subtle sentiment cues within the dataset made it the linchpin of the sentiment analysis component.

### 4.1.3 Geospatial Mapping Integration

A significant feature of the dashboard is its geospatial mapping capability, which is crucial for depicting sentiment analysis results on a global scale. This component involved the integration of geographical coordinates from the dataset with sentiment data to plot visualisations on an interactive map. Users can filter and drill down into specific regions, exploring the sentiment trends through a geospatial lens.

*Plotly's* mapping functions were used to create choropleth maps, which highlighted sentiment distribution across countries, and scatter geo plots, which placed sentiment markers on the globe. Users were thus given tools to see and interact with global sentiment data.

### 4.1.4 User Interface and Experience

*Streamlit* was a strategic enabler that transformed complex data operations into seamless user experiences. The user interface was crafted with the ethos of simplicity and functionality. Interactive elements such as filters for sentiment and date ranges, as well as visualisation options, were integrated, allowing users to engage with the data in a dynamic and insightful manner.

### 4.1.5 Dashboard Design

The dashboard's design traversed through rigorous coding, testing, and validation cycles. Each feature was implemented with a sharp focus on performance and user engagement, ensuring that the final product was not just functional but also intuitive and insightful. The deployment was staged through a local server, setting the stage for a global rollout.

This chapter articulates the thoughtful and strategic process of designing and developing the geospatial sentiment dashboard. It narrates the harmonisation of machine learning precision with interactive data visualisation, a synthesis that stands to offer the fashion e-commerce industry a transformative tool for understanding and leveraging consumer sentiment across geographical boundaries.

## 4.2 Implementation

### 4.2.1 Data Collection and Preparation

The integrity of any analytical model is intrinsically linked to the quality of the data it processes. The initial phase of crafting the geospatial sentiment dashboard - an artefact designed to distill consumer sentiment across geographical landscapes for the fashion e-commerce industry - began with meticulous data collection and methodical preparation. This stage was pivotal, serving as the keystone for the artefact's eventual performance and reliability.

#### Data Acquisition

Our dataset, a dataset of over 271,000 reviews, was sourced from TeePublic via Kaggle as described in **section 4.1**- a nodal point for datasets facilitating academic and research pursuits. Each record within this dataset encapsulated a unique customer perspective, quantified through a rating scale and qualitatively enriched by review texts. The extraction process was undergirded by stringent adherence to ethical standards of data use, maintaining the anonymity of reviewers in compliance with data protection regulations.

Attributes such as ‘reviewer\_id’, ‘store\_location’, and geospatial coordinates (‘latitude’ and ‘longitude’), alongside temporal markers (‘date’, ‘month’, ‘year’), were procured to enable a multi-faceted analysis. Textual elements, the ‘title’ and ‘review’ fields, serve as the primary substrates for sentiment analysis. The overview of the dataset is shown in figure 3 below.

A screenshot of a computer

Description automatically generated

Figure 9:dataset before preprocessing

#### Data Preprocessing

**Challenges in Data Collection and Preparation:** During the data preparation phase, one of the challenges faced was related to encoding. The dataset initially exhibited encoding issues that necessitated the use of 'latin1' encoding to read the files correctly. This challenge often arises when dealing with datasets that contain a mix of characters from various languages, which may not be properly captured by the default UTF-8 encoding. 'Latin1' encoding was utilised as a workaround to accommodate special characters found in the customer reviews, ensuring no data was lost or misinterpreted during the import process.

Another significant challenge was the presence of missing values and non-ASCII characters within critical text fields. Such discrepancies can introduce biases or errors during analysis and needed to be resolved to maintain the integrity of the model's predictions. The approach was two-fold:

* **Handling Missing Data:** The dataset contained records with missing entries in essential columns such as 'title' and ‘review’ as shown in figure 4 below. To ensure a robust dataset for analysis, rows with missing values in these key fields were dropped. This measure prevented potential distortions in sentiment analysis caused by incomplete data.

A screenshot of a computer

Description automatically generated

Figure 10: missing data

* **Text Data Normalisation:** The text data within the 'title' and 'review' columns contained various non-ASCII characters, which can be problematic for natural language processing tasks. A regular expression pattern was implemented to clean the text data, removing unwanted characters such as non-printable symbols, numerics, and underscores while preserving words, spaces, and apostrophes. This normalisation process ensured the dataset only contained clean, ASCII-encoded text, which is crucial for accurate tokenisation and sentiment analysis downstream.

These preparatory steps were instrumental in overcoming the dataset's initial quality issues, thereby setting a solid foundation for the subsequent machine learning processes. The cleaning steps which involve removing missing data and text normalisation is as shown in figure 5.

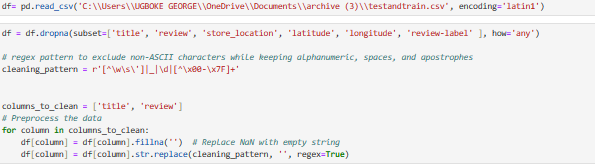


Figure 11:cleaning the dataset

**Feature Engineering**: Feature engineering played a critical role in preparing the dataset for sentiment analysis. The engineered feature is to introduced a new column called the 'sentiment' column which will act as a base line for the artefact model. It is distinguished by two categories which are negative sentiments (ratings 1 to 3) with a label of 0 and positive sentiments (ratings 4 to 5) with a label of 1. This binary labeling facilitated a focused approach for the machine learning model, enabling it to discern consumer sentiment with higher accuracy. This sentiment classification was essential for establishing a clear baseline for the model's training and validation. Ratings were carefully transformed into sentiment labels based on established customer feedback patterns, ensuring consistency across the predictive model.

The 'sentiment' column arose from cleaned and normalised data to prevent model skew. It became crucial for gauging the model’s accuracy, playing a pivotal role in both training and evaluating the machine learning pipeline. The strategic creation of this feature was vital in streamlining the sentiment analysis and improving the model's efficiency and effectiveness. The code used to create this new column is in figure 6.



Figure 12: creating sentiment column as baseline.

**Lemmatisation:** Lemmatisation was a critical preprocessing step aimed at reducing words to their base or dictionary form, which is essential for standardising varying inflections of words in the sentiment analysis. The process utilised the *WordNetLemmatiser* and *NLTK's* part-of-speech tagging to ensure that each word was accurately returned to its lemma according to its context in speech. This approach enhances the model's ability to interpret and analyse the sentiment accurately by normalising the linguistic variations in the dataset. The lemmatisation was applied to both 'review' and 'title' columns of the dataframe, transforming each entry into a cleaner, more analytically valuable form. The lemmatisation code is meticulously scripted in the project's Jupyter Notebook to guarantee reproducibility and efficiency in text processing. The code used during this stage is available in appendix.

The figure below shows the cleaned and lemmatised dataset.



Figure 13: dataset after preprocessing.

#### Test and Train Split

Following lemmatisation, the dataset was partitioned into features (x) representing the reviews and the target variable (y) denoting sentiments. Utilising *train\_test\_split* library from sklearn with a test size of 20%, the data was stratified into training and test subsets, ensuring diverse sentiment representation for robust model training and validation. This process is shown in figure 8 below.

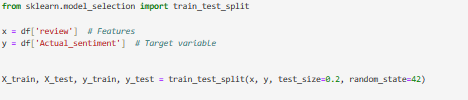


Figure 14:code to test and train

#### Vectorisation

Following lemmatisation, the text data was transformed into numerical vectors using TF-IDF Vectorisation, which evaluates the importance of a word in relation to the corpus. A custom list of stop words was curated, thoughtfully retaining negative connotations crucial for sentiment classification, ensuring the vectoriser did not overlook words pivotal to negative sentiment. This nuanced approach to stop word inclusion played a significant role in enriching the model's understanding of subtle negative expressions. The vectorisation was performed on both the training and testing datasets, resulting in TF-IDF weighted matrices that serve as refined inputs for the machine learning models. The vectorisation process as done in jupyter notebook is shown in figure 9 below.



Figure 15:code to vectorise

Following vectorisation, the sparse matrix of the vectorised test and train set is shown in figure 10 below.

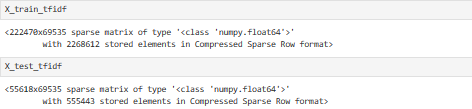


Figure 16: sparse matrix

### 4.2.2 Machine Learning Model Selection

The effectiveness of the geospatial sentiment dashboard hinges significantly on the machine learning model utilised for sentiment classification. The primary criterion guiding the selection process was the model's proficiency in handling high-dimensional text data, which is crucial for accurate sentiment analysis in diverse geographic contexts.

#### Model Evaluation and Selection

Various models were considered, with a focus on evaluating their capabilities to process and classify complex sentiment data efficiently. The Support Vector Machine (SVC) model was particularly noted for its effectiveness in high-dimensional spaces, which makes it ideal for text classification tasks where feature spaces are extensive and intricate. This model was rigorously tested against others in its ability to classify sentiment data accurately, considering factors such as precision, recall, and the F1 score as shown in figure 11.

The SVC model demonstrated superior performance during empirical tests, excelling in both the accuracy of classification and its computational efficiency in processing large datasets. These attributes made it particularly suitable for the real-time processing needs of the geospatial sentiment dashboard.

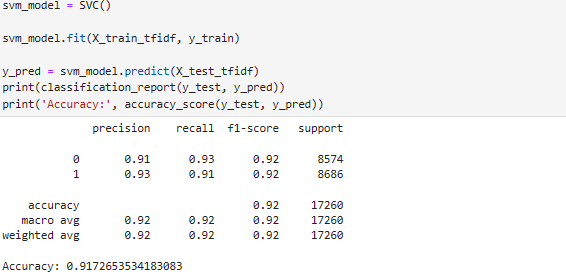


Figure 17: model accuracy and classification report

#### Model Finalisation

Following the selection, the SVC model along with the associated vectoriser was serialised using the *joblib* library. This step was crucial for preserving the model's configuration and ensuring its performance remains consistent when deployed. Serialisation also facilitates the seamless integration of the model into the Streamlit application, supporting the dashboard's interactive features and real-time data processing capabilities. The Serialisation steps for both the svm\_model and vectorisation is shown in figure 12 below

The machine learning model selection phase was marked by a blend of empirical testing and theoretical analysis, ensuring the chosen model met the detailed requirements of the project. The SVC model stood out as the optimal choice, offering a robust foundation for the geospatial sentiment dashboard. This selection not only aligns with the project's objectives but also enhances the dashboard's functionality, making it a powerful tool for analysing and visualising sentiment across different geographical landscapes.

A screenshot of a computer program

Description automatically generated

Figure 18 mode and vectoriser serialisation

### 4.2.3 Deployment on Streamlit

After meticulously preparing our data and selecting a suitable machine learning model, the next step was deploying our solution. Streamlit, an open-source app framework tailored for machine learning and data science teams, for its user-friendly interface and seamless transition from scripts to interactive web applications.

**Installation and Setup:**

We began by installing Streamlit directly within our Jupyter Notebook environment using the command *pip install streamlit*, integrating it seamlessly into our workflow. Streamlit's capability to quickly convert data scripts into shareable web apps expedited our transition from concept to product.

**Preparation for Deployment:**

The deployment preparation involved consolidating all preprocessing techniques, filters, visualisations, and predictive functionalities into a single Python script. This script served as the foundation for our web application, encompassing modules for data cleaning, text preprocessing, sentiment analysis, and geospatial mapping. It also facilitated real-time sentiment prediction by loading the serialised SVM model and vectoriser.

**Ensuring Consistency:**

Maintaining consistency in model output was paramount, necessitating that incoming data undergo the same preprocessing pipeline as the training data. Thus, our script mirrored the preprocessing steps applied during training, including tokenisation, lemmatisation, and vectorisation, before passing the data to the machine learning model for sentiment prediction.

**Streamlit Dashboard Development**

The development of the Streamlit dashboard was approached with the user experience as the focal point. The design philosophy was anchored in simplicity and functionality. The dashboard was structured to have two main interactive components:

1. **Prediction Page:** This page facilitated the prediction of sentiments using either direct text input or CSV file uploads. It displayed real-time predictions and allowed users to engage with the sentiment analysis model interactively.

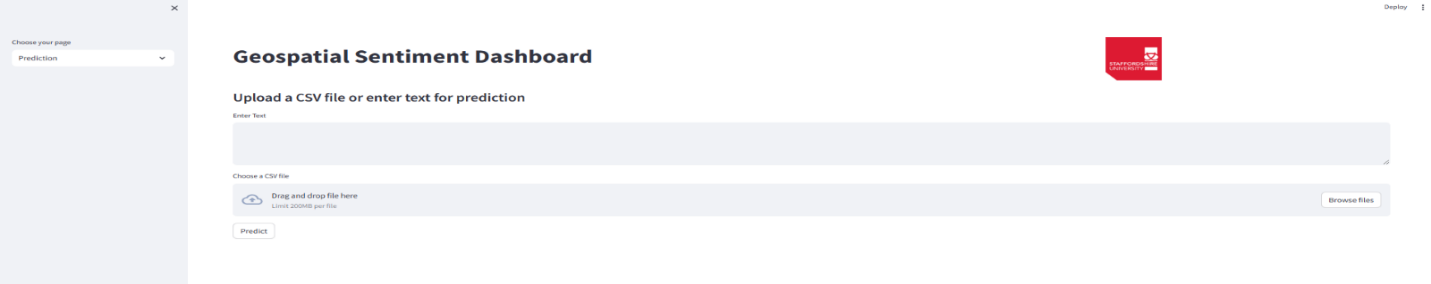


Figure 19:prediction page of the application

1. **Data Visualisation Dashboard:** The second component was a comprehensive dashboard that visually represented sentiment analysis results. Utilising *Plotly* for interactive graphs, users could filter results based on sentiment labels, timeframes, and geographical locations, displaying the processed data in an easily digestible format.

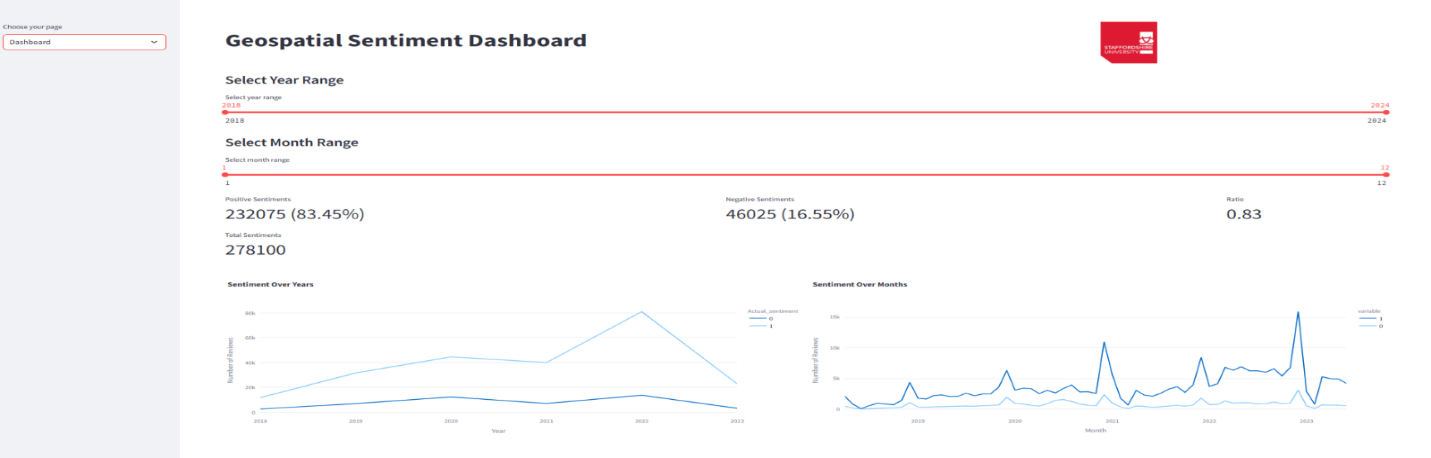


Figure 20dashboard page of the app

**Streamlit Web App Deployment**

The culmination of our efforts was the deployment of the Streamlit web application. This stage involved extensive local testing to confirm the application's functionality and user interface design. Ensuring data security, system stability, and responsiveness were paramount, leading to the implementation of rigorous testing protocols.

Once the local testing on jupyter notebook application confirmed the robustness of the application, we proceeded to launch the web application using the command *streamlit run app.py*. This command initiated the Streamlit local server and rendered our interactive dashboard in a web browser, marking the transition from a series of data science tasks to an integrated analytical tool.

The deployment on Streamlit signified a key milestone, as it brought the geospatial sentiment analysis model out of the development environment and into a practical, real-world application.



Figure 21: launching the app

This dashboard now stands as a testament to the potential of modern data science tools to transform raw data into actionable insights, showing the use of complex machine learning models for a broader audience.

The python script used to develop the geospatial sentiment dashboard is in appendix 3.

**Geospatial Mapping**

Geospatial mapping was integrated into the dashboard to provide users with intuitive visual representations of data across different geographic regions. This feature allows users to interact with the data through zooming, panning, and clicking on specific areas to get more detailed insights. Efforts were made to ensure that the dashboard is user-friendly, with responsive design features that adapt to different device screens and user inputs. Specific region can be searched to see its sentiment information and other features of that region as shown in 16 and 17 below.

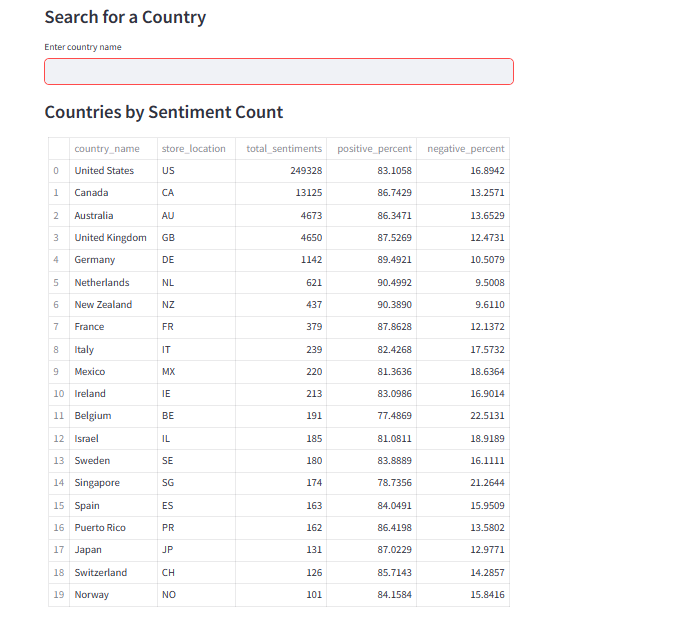


Figure 22search filter on dashboard

A screenshot of a computer

Description automatically generated

Figure 23: sentiment information of united kindgom

A screenshot of a computer

Description automatically generated

Figure 24: visualisation in dashboard

The geospatial mapping feature of the dashboard employs advanced geographic visualisation tools to overlay sentiment data onto interactive maps. Utilising libraries like Plotly for rendering and geopandas for geospatial data processing, the dashboard maps sentiment indicators such as positive and negative reviews across various global locations. This feature is technically implemented through color-coded markers that signify different sentiment levels, providing immediate visual cues that help users discern patterns and trends.

Key technical aspects include:

1. Data Integration and Processing: Geospatial coordinates are extracted from the dataset and integrated with sentiment data to create a comprehensive mapping dataset.
2. Geographic regions(countries) filtering which dynamically updates the map to reflect these sentiments in those regions including concerning reviews.
3. Visualisation Enhancements: Techniques such choropleth maps are utilised to represent density and distribution of sentiments accross different countries, enhancing the analytical depth of the visualisations. Figure 18 shows the element of the filtered datset in term of total number of reviews made by countries.
4. User Interaction: The implementation includes zooming and panning capabilities, allowing users to explore detailed regional sentiment analytics intuitively.

## 4.3 Testing and Validation

Ensuring the reliability and accuracy of the geospatial sentiment dashboard required a robust testing and validation framework. This section outlines the methodologies and results of the internal and external validation processes, which were crucial for the artefact’s quality assurance.

### Testing of models

This project began with a comprehensive testing phase using Python's Scikit-learn library within a Jupyter Notebook environment. The initial tests focused on evaluating several machine learning algorithms known for their text classification capabilities. These included Logistic Regression, Support Vector Machine (SVM) and its linear variant LinearSVC, as well as ensemble methods like Random Forest and Extra Trees Classifier. The primary metrics for assessment were precision, recall, F1 Score, and accuracy. The dataset as stated earlier was split into three which are the test and train set. and the cross validation set. The table 3 below shows the result of training and testing this model with the test and train set.

Table 3: initial model results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** |  | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| Random Forest | 0 | 0.89 | 0.60 | 0.71 | 0.92 |
|  | 1 | 0.93 | 0.99 | 0.95 |  |
| SVC | 0 | 0.85 | 0.75 | 0.80 | 0.94 |
|  | 1 | 0.95 | 0.97 | 0.96 |  |
| linearSVC | 0 | 0.83 | 0.74 | 0.78 | 0.93 |
|  | 1 | 0.95 | 0.97 | 0.96 |  |
| Extra Trees Model | 0 | 0.89 | 0.60 | 0.72 | 0.92 |
|  | 1 | 0.93 | 0.99 | 0.95 |  |
| Logistic regression | 0 | 0.84 | 0.73 | 0.78 | 0.93 |
|  | 1 | 0.95 | 0.97 | 0.96 |  |

**Challenges Encountered**

During the initial testing phase, we encountered several challenges:

1. Class Imbalance: The dataset exhibited a significant skew towards positive reviews, leading to high bias where models predominantly predicted the majority class. This imbalance risked distorting performance metrics, falsely suggesting high accuracy.
2. Performance Metrics Disparity: The models initially showed a discrepancy in precision and recall, especially for less represented negative sentiments, resulting in unacceptably low F1 scores for this crucial sentiment class.
3. Overfitting Risk: There was a tendency for models to overfit to the majority positive class, which could result in poor performance on unseen data.

**Adjustments Made**

To address these issues, we implemented several strategies:

1. Random Undersampling: We balanced the class distribution by undersampling the majority class to match the size of the minority class, ensuring each sentiment class was equally represented.
2. Model Metrics Re-evaluation: After class balancing, the models were re-evaluated. This adjustment provided a more accurate picture of the models’ capabilities, showing improved precision and recall and enhanced F1 scores for negative sentiment.
3. Cross-validation: We employed cross-validation methods to mitigate overfitting, enhancing the models' generalisability and fine-tuning hyperparameters for optimal performance.

The summary of this adjustment are shown in figure 20 below.

A screenshot of a computer program

Description automatically generated

Figure 25: balance the dataset

After resmapling the dataset to make it balance, vectorisation was again performed and now the corpus sparse matrix for both test and train dataset is shown in figure 21 below..

A screenshot of a computer

Description automatically generated

Figure 26: sparse matrix of balanced dataset

**Final Model Selection Rationale**

Post-adjustment, the models demonstrated enhanced performance, proving the effectiveness of the corrective measures. The SVC model, in particular, showed commendable precision and recall across sentiment classes, reaffirming its selection for the project’s requirements.

Table 4: final model result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** |  | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| Random Forest | 0 | 0.87 | 0.93 | 0.90 | 0.89 |
|  | 1 | 0.92 | 0.86 | 0.89 |  |
| SVC | 0 | 0.91 | 0.93 | 0.92 | 0.92 |
|  | 1 | 0.93 | 0.91 | 0.92 |  |
| linearSVC | 0 | 0.90 | 0.92 | 0.91 | 0.91 |
|  | 1 | 0.92 | 0.90 | 0.91 |  |
| Extra Trees Model | 0 | 0.87 | 0.94 | 0.90 | 0.90 |
|  | 1 | 0.94 | 0.86 | 0.90 |  |
| Logistic regression | 0 | 0.90 | 0.92 | 0.91 | 0.91 |
|  | 1 | 0.92 | 0.90 | 0.91 |  |

#### Cross-validation with streamlit application

The internal validation was carried out on the streamlit application with the validation dataset kept from the main dataset before the training. This multifaceted approach was designed to thoroughly assess the model's performance across various dimensions of accuracy.

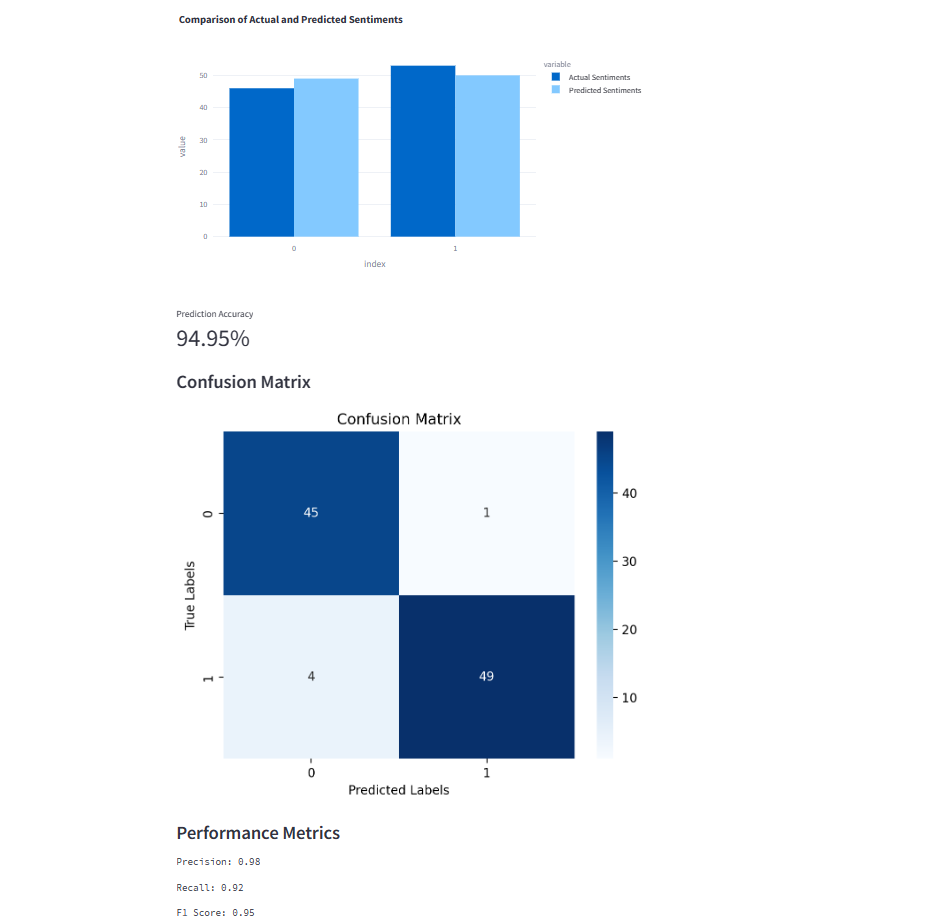


Figure 27: cross validation result on prediction page app

**Model Performance Metrics**

The validation process was meticulously structured to measure the precision, recall, and F1 scores, offering a comprehensive assessment of the model's predictive capabilities. Notably, a classification report generated from the Streamlit application demonstrated a high precision of 0.98 for positive sentiments and an overall recall rate of 0.92, indicating the model's ability to accurately identify and classify sentiments. The F1 score, standing at 0.95, confirmed the model's balanced performance.

**Prediction Accuracy**

An accuracy of 94.95% was observed, reflecting the model's efficiency in sentiment classification. The validation also included a confusion matrix visualisation, affirming substantial agreement between actual and predicted sentiments and underscoring the model's robustness.

**Geospatial Sentiment Analysis testing**

The geospatial functionality of the dashboard was rigorously tested to validate the accuracy of sentiment mapping for different countries. This testing phase focused on ensuring that the dashboard correctly interprets and displays sentiment data linked to geographic locations, particularly through the search functionality which allows users to explore sentiment analysis by country.

**Visualisation and Interaction Verification**

The dashboard was subjected to visual inspection to verify the accurate representation of sentiment data across various countries. The search functionality was specifically tested, allowing users to select a country and view detailed sentiment distributions, as well as related word clouds highlighting common themes in feedback. For example, when searching for the United Kingdom, the dashboard successfully displayed the distribution of positive and negative sentiments along with a word cloud that emphasises prevalent terms in negative reviews such as "poor quality" and "bad.", it also shows the context of the most used negative words as shown in figure 23.

A screenshot of a computer

Description automatically generated

Figure 28: filter function in app

The testing confirmed the effective functionality of geospatial features within the dashboard, demonstrating its utility in providing clear and actionable sentiment analysis for specific regions. This capability enhances the dashboard's applicability for targeted marketing and customer feedback analysis in the fashion e-commerce sector. The results from these tests affirm the dashboard’s reliability in delivering accurate geospatial sentiment insights, making it a valuable tool for businesses looking to enhance their global customer engagement strategies.

## 4.4 Critical Evaluation

This section critically evaluates the methodologies, results, a case study and implications of the study on developing a geospatial sentiment dashboard for the fashion e-commerce sector. The aim is to assess the efficiency of the chosen models, the accuracy of the results, and the potential improvements that could enhance future research and application.

**Evaluation of Model Selection**

The choice to utilise the Support Vector Machine (SVC) model was primarily driven by its robust performance in high-dimensional text classification tasks, which are essential for handling the intricate sentiment data inherent in this project. This capability is critical for the operational demands of the geospatial sentiment dashboard, which requires real-time processing of complex datasets. The confusion matrix provided below illustrates the model's performance:

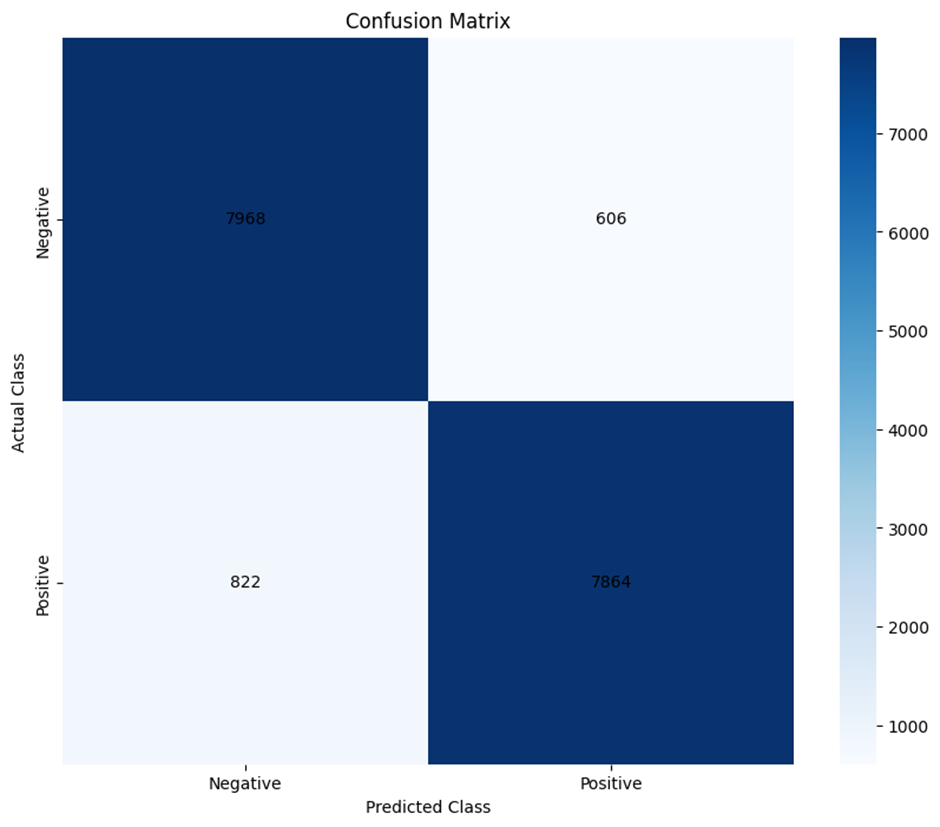


Figure 29: confusion matrix of svm\_model

**Performance Analysis of Model**

The SVC model underwent rigorous testing and adjustments, resulting in significant improvements in accuracy and reliability. The adjustments from section 4.3 addressed initial challenges such as class imbalance and model overfitting, thus enhancing the model's applicability to diverse datasets. The following metrics were used to evaluate the model's performance:

True Positives (TP): 7864 - Instances where positive sentiment was correctly identified.

True Negatives (TN): 7968 - Instances where negative sentiment was correctly identified.

False Positives (FP): 606 - Instances where negative sentiment was incorrectly labeled as positive.

False Negatives (FN): 822 - Instances where positive sentiment was incorrectly labeled as negative.

**Precision:** Precision measures the accuracy of positive or negative predictions.

Precision (positive sentiment): TP / (TP +FP) = 7864 / 8470 = 0.93

Precision (negative sentiment): TN / (TN + FN) = 7968 / 8790 = 0.91

**Recall:** Recall (or sensitivity) measures the ability of a model to find all the relevant cases (positive or negative).

Recall (positive sentiment): TP / (TP+FN) = 7864 / 8686 = 0.91

Recall (Negative sentiment): TN / (TN+FP) = 7968 / 8574 = 0.93

F1 score: The F1 Score is the harmonic mean of Precision and Recall.

F1 positive = 2 \* (precision positive \* recall positive)/ (precision positive + recall positive)

= 2\* 0.8463/ 1.84 = 0.92

F1 negative = 2 \* (precision negative \* recall negative)/ (precision negative + recall negative)

= 2\* 0.8463/1.84 = 0.92

**Model Accuracy:** Accuracy measures the overall effectiveness of a classifier.

Accuracy = (TP +TN) / (TP +TN +FP +FN) = 15832 /17260 = 0.92 = 92%

#### Case Study

To comprehensively assess the generalisability and adaptability of our machine learning model, external validation was conducted using a distinct dataset from another fashion e-commerce platform. This dataset sample is shown in the figure below, with the url from Kaggle- <https://www.kaggle.com/datasets/mansithummar67/flipkart-product-review-dataset>. This dataset varied in customer expressions and sentiment distributions, posing a fresh challenge to the model's ability to interpret and classify sentiments accurately under different linguistic contexts.

A screenshot of a computer

Description automatically generated

Figure 30:dataset to validate the model

Results Analysis

The evaluation was carried out in the web app and the external validation yielded the following key metrics and results, which are pivotal in understanding the model's performance in a new environment:

Prediction Accuracy: The model achieved an overall accuracy of 87.46%. This high accuracy rate is indicative of the model's robustness, even when applied to data with different characteristics from the training set.

**Confusion Matrix:**

True Positives (TP): 132,041 - the model correctly identified many positive sentiments.

True Negatives (TN): 29,981 - the model accurately recognised many negative sentiments.

False Positives (FP): 10,419 - instances where negative sentiments were incorrectly classified as positive.

False Negatives (FN): 12,802 - positive sentiments that were incorrectly labeled as negative.

This matrix reveals a relatively balanced detection capability for both positive and negative sentiments, albeit with some room for improvement in minimising false classifications.

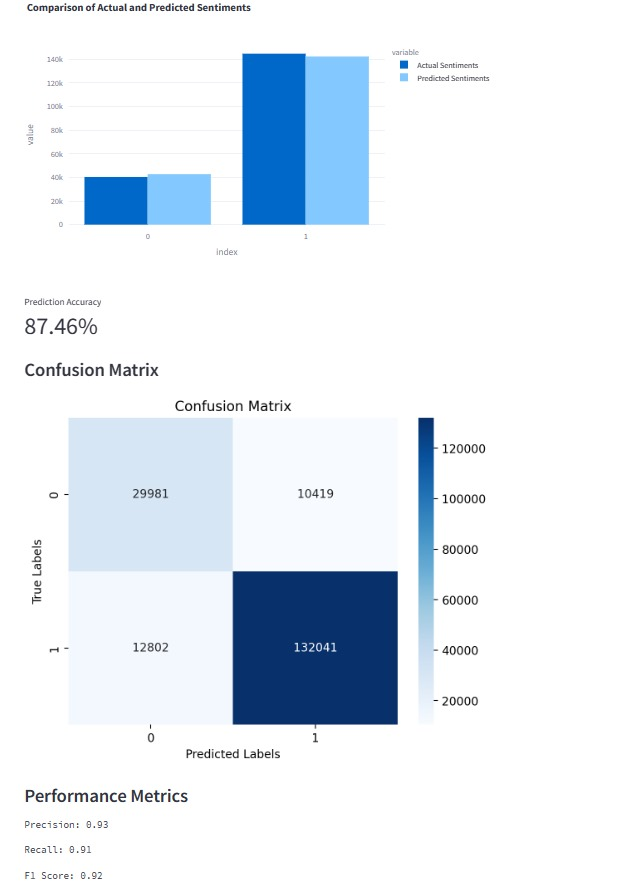


Figure 31: results of the external validation

Performance Metrics:

**Precision for Class 0 (Negative Sentiment):** 0.71

Recall for Class 0: 0.74

F1 Score for Class 0: 0.72

**Precision for Class 1 (Positive Sentiment):** 0.93

Recall for Class 1: 0.91

F1 Score for Class 1: 0.92

Macro Average Precision: 0.82

Macro Average Recall: 0.83

Macro Average F1 Score: 0.82

Weighted Average Precision: 0.88

Weighted Average Recall: 0.87

Weighted Average F1 Score: 0.88

These metrics illustrate a high degree of precision, particularly for positive sentiments, where the model demonstrates exceptional specificity in identifying true positives. The recall rates indicate the model's effectiveness in capturing a substantial proportion of relevant instances across both sentiment classes.

The case study underscores the model's capability to generalise across different datasets effectively, demonstrating strong predictive accuracy and robust classification metrics. However, the presence of false positives and false negatives highlights challenges in achieving perfect model fidelity, particularly in translating its learning to varied linguistic expressions and sentiment constructs inherent in external data.

The evaluation of this case study confirms that while the model performs well in unfamiliar settings, continuous enhancements and adjustments, such as refining the preprocessing techniques or experimenting with different or more complex algorithms, are essential for improving accuracy and reducing classification errors.

The case study of the geospatial sentiment dashboard using a completely independent dataset highlights its superior performance compared to previous studies, significantly advancing the state-of-the-art in sentiment analysis, particularly within geospatial contexts. The overall prediction accuracy of 87.46% stands out, particularly when compared to the best-performing models in earlier studies, such as the 77% accuracy achieved by SVM and Logistic Regression in handling imbalanced and multi-class datasets like the US airline Twitter data (Tusar & Islam, 2021).

The precision of 93% for positive sentiments and 71% for negative sentiments, along with recall rates of 91% for positive and 74% for negative sentiments, exceed the benchmarks set by (Suresh et al., 2023), who reported lower macro-average recall and F1-scores. This demonstrates a significant improvement in the model’s ability to correctly identify and classify sentiments with fewer errors.

Moreover, the F1 score, which balances the precision and recall, was particularly strong at 92% for positive sentiments, compared to general metrics seen in prior studies, such as the high ROC-AUC of 0.9133 achieved by (Polyakov et al., 2020) using both classical and deep learning techniques. The macro-average and weighted-average scores across precision, recall, and F1 (all approximately 82% to 88%) also indicate a more consistent performance across different sentiment classes than what has typically been documented, such as in studies that struggled with high-dimensional and complex datasets.

This case study, thus not only showcases the dashboard's robustness and reliability across various datasets but also confirms its capability to generalise better than previous models. This generalisability, coupled with high precision and recall, supports its deployment across different e-commerce platforms, making it an invaluable tool for businesses seeking to enhance their customer engagement strategies through nuanced, geographically informed sentiment analysis.

#### Limitations of the artefact

1. **AWS Deployment Challenges**

Initially, the project proposed the use of AWS SageMaker for deploying the machine learning model to leverage its scalable cloud computing capabilities. However, due to a knowledge gap, particularly in troubleshooting errors associated with AWS SageMaker, the deployment could not be realised. This led to opting for Python and Streamlit for visualisation and interaction, which, while effective, lacked the robust cloud-based infrastructure and scalability that AWS could potentially offer.

2. **Data Imbalance and Bias**

During the model training phase, the dataset exhibited a significant imbalance with a predominance of positive sentiments. This imbalance led to issues of model overfitting, where the model excessively learned to predict the majority class at the expense of the minority class accuracy. Efforts to mitigate this bias by scaling down the positive samples helped reduce overfitting but possibly at the cost of achieving higher overall accuracy. The trade-off between bias reduction and potential loss of informative data was a significant challenge.

3. **Geographic Data Handling**

The dataset initially contained only country codes, which limited the interpretability and direct application of geographic insights. To address this, the PyCountry library was employed to translate country codes into full country names, adding an extra step in the data preprocessing phase. This requirement introduced additional complexity to the data handling process, potentially impacting the efficiency of data processing.

4. **Local Deployment Limitations**

The project's final deployment was localised, primarily using Streamlit run through a Python command prompt. This setup requires users to have Python installed and to execute specific commands to launch the dashboard. This approach, while functional, limits the accessibility of the dashboard to users who are comfortable with such technical setups and restricts the potential for broader, non-technical user engagement.

5. **Performance and Responsiveness of the Dashboard**

The interactive nature of the dashboard, although a strength in terms of user experience, also presents a limitation. Each action on the dashboard necessitates rerunning a set of underlying codes, which can lead to delays in loading and responding to user interactions. This issue affects the dashboard's overall user experience, particularly in scenarios requiring rapid data updates or when used on lower-specification hardware.

These limitations underscore some of the practical challenges faced during the project's execution, from technical hurdles in cloud deployment to data handling and user interface performance issues. While they represent areas for potential improvement, they also reflect the real-world complexities of developing and deploying machine learning applications. Future iterations of this project could focus on overcoming these limitations, possibly through enhanced cloud integration, advanced data balancing techniques, and optimisation of the dashboard's performance to enhance usability and accessibility.

# CHAPTER 5. CONCLUSION and FUTURE WORK

### 5.1 Conclusion

The development of the geospatial sentiment dashboard has marked a significant advancement in the use of technology for enhancing customer engagement in the fashion e-commerce sector. This project has successfully integrated machine learning (ML), natural language processing (NLP), and geospatial analysis to create a robust tool that provides real-time insights into regional consumer sentiments. The implementation of this dashboard is poised to transform the way fashion retailers understand and cater to their diverse customer base.

The research has shown that the incorporation of geospatial data with sentiment analysis can provide a more nuanced understanding of consumer behaviors and preferences across different geographical locations. This is particularly valuable in an industry driven by fast-changing trends and regional tastes. By analysing geospatially tagged consumer feedback, businesses can now tailor their marketing strategies and product offerings more effectively, ensuring that they meet the specific demands of different regions.

The dashboard’s ability to process and visualise data in real-time allows businesses to react quickly to emerging trends and consumer sentiments, providing them with a competitive edge in the fast-paced e-commerce environment. The application of advanced data processing technologies like Python, Jupyter Notebook, and Streamlit has facilitated the handling of large datasets, enabling the extraction of meaningful insights from complex data patterns.

This project has also contributed to academic knowledge, offering a detailed exploration of integrating diverse technological tools to address specific business needs. It highlights the potential of interdisciplinary approaches in solving complex problems and advancing the state of technology in business applications.

In conclusion, the geospatial sentiment dashboard represents a milestone in the application of advanced technologies to enhance consumer understanding and engagement in the fashion e-commerce sector. It stands as a testament to the benefits of integrating sentiment analysis with geospatial data, providing a comprehensive tool that can drive innovation and strategic decision-making in the industry.

### 5.2 Future work

Looking ahead, there are several areas where future research could expand upon the current findings and technologies used in this project. One key area is the exploration of additional data sources and types to enrich the analysis capabilities of the geospatial sentiment dashboard. Incorporating multimedia feedback, such as images and videos, could provide deeper insights into consumer preferences and behaviors, particularly in a visually driven industry like fashion.

Further development could also involve enhancing the dashboard’s predictive capabilities by incorporating more sophisticated machine learning models, such as deep learning algorithms. These models could improve the accuracy of sentiment analysis, especially in interpreting complex and nuanced consumer feedback.

Another avenue for future work is the expansion of the dashboard’s functionality to include predictive analytics for market trends. By analysing historical data alongside real-time feedback, the dashboard could forecast upcoming trends, enabling businesses to proactively adjust their strategies.

Improving the dashboard’s user interface and interaction design could also enhance its usability and accessibility, making it more intuitive for users with varying levels of technical expertise. This could broaden the dashboard’s user base, increasing its impact and effectiveness in the industry.

Additionally, exploring the scalability of the dashboard to other sectors beyond fashion e-commerce could demonstrate the versatility of this approach. Industries such as hospitality, travel, and retail could also benefit from geospatial sentiment analysis, suggesting a broad potential impact of this research.

Finally, addressing ethical considerations in data collection and analysis, particularly concerning consumer privacy and data security, would be crucial as the dashboard evolves. Establishing robust protocols and guidelines would ensure that the technology continues to serve the interests of both businesses and consumers responsibly.

These directions not only promise to enhance the functionality and impact of the geospatial sentiment dashboard but also contribute to the broader field of e-commerce and business intelligence, pushing the boundaries of how data-driven technologies can be leveraged for strategic advantage.

# CHAPTER 6. REFERENCES

Bindoo, M., Vipin, J. & Satyendra, A. (2021). An Overview of Electronic Commerce (e-Commerce). *Journal of Contemporary Issues in Business and Government*. 27 (3).

​Dejene, F.M., Sahoo, A.K., Barik, R.K. & Mishra, B.S.P. (2022). GeoRSA: Geospatial Recommendation System using Sentimental Analysis. In: *PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing*. 2022, Institute of Electrical and Electronics Engineers Inc., pp. 496–500.

​Fedushko, S. & Ustyianovych, T. (2022). E-Commerce Customers Behavior Research Using Cohort Analysis: A Case Study of COVID-19. *Journal of Open Innovation: Technology, Market, and Complexity*. 8 (1).

​Figueiredo, L.J.A.S., Dos Santos, G.B., Souza, R.P.P.M., Silva, F.A. & Silva, T.R.M.B. (2021). MoreData: A Geospatial Data Enrichment Framework. In: *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. 2 November 2021, Association for Computing Machinery, pp. 419–422.

​Han, X., Wang, J., Zhang, X., Wang, L. & Xu, D. (2024). Mining public behavior patterns from social media data during emergencies: A multidimensional analytical framework considering spatial–temporal–semantic features. *Transactions in GIS*. 28 (1). p.pp. 58–82.

​J Tamara, C.A., F A Tumbuan, W.J., Gunawan, E.M., J Tamara, C.A., F Alfa Tumbuan, W.J. & Margareth Gunawan, E. (2023). CHATBOTS IN E-COMMERCE: A STUDY OF GEN Z CUSTOMER EXPERIENCE AND ENGAGEMENT-FRIEND OR FOE? *Jurnal EMBA*. 11 (3). p.pp. 161–175.

​Jílková, P. & Králová, P. (2021). Digital Consumer Behaviour and eCommerce Trends during the COVID-19 Crisis. *International Advances in Economic Research*. 27 (1) p.pp. 83–85.

​Kawasaki, T., Wakashima, H. & Shibasaki, R. (2022). The use of e-commerce and the COVID-19 outbreak: A panel data analysis in Japan. *Transport Policy*. 115. p.pp. 88–100.

​Landim, A.R.D.B., Pereira, A.M., Vieira, T., de, E., Moura, J.A.B., Wanick, V. & Bazaki, E. (2022). Chatbot design approaches for fashion E-commerce: an interdisciplinary review. *International Journal of Fashion Design, Technology and Education*. 15 (2). p.pp. 200–210.

​Lee, I. & Shin, Y.J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*. 63 (2). p.pp. 157–170.

​Li, Z., Tang, W., Huang, Q., Shook, E. & Guan, Q. (2020). Introduction to big data computing for geospatial applications. *ISPRS International Journal of Geo-Information*. 9 (8).

​Liu, X., Tong, D., Huang, J., Zheng, W., Kong, M. & Zhou, G. (2022). What matters in the e-commerce era? Modelling and mapping shop rents in Guangzhou, China. *Land Use Policy*. 123.

​Loh, E., Khandelwal, J., Regan, B. & Little, D.A. (2022). Promotheus: An End-to-End Machine Learning Framework for Optimizing Markdown in Online Fashion E-commerce. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 14 August 2022, Association for Computing Machinery, pp. 3447–3457.

​Mardiani, E., Rahmansyah, N., Desmana, S. & Rifqi, A. (2023). *Analysis of Buyer’s Trust in E-Commerce Shop And Travel Web*. 20 (2). p.pp. 850–857.

Mardiana, S. (n.d.). *Modifying Research Onion for Information Systems Research*. https://www.researchgate.net/publication/359542575

​Mehta, N., Mane, M., Satpute, P., Mahale, A. & Bhutekar, S. (2023). Sentiment Analysis for E-Commerce. *International Journal for Research in Applied Science and Engineering Technology*. 11 (4). p.pp. 2891–2895.

​Nazeer, H., Iqbal, W., Bokhari, F., Bukhari, F., Ur, S. & Baig, R. (n.d.). *Real-time Text Analytics Pipeline Using Open-source Big Data Tools*.

​Van Ngoc, D. & Dat, N.T. (2022). A design in system architecture based on mobile cloud computing for a virtual try-on solution. *International Journal of Advanced and Applied Sciences*. 9 (6). p.pp. 36–42.

​Pakpahan, D., Siallagan, V. & Siregar, S. (2023). Classification of E-Commerce Product Descriptions with The Tf-Idf and Svm Methods. *sinkron*. 8 (4). p.pp. 2130–2137.

​Patel, A. (n.d.). *Exploratory Data Analysis and Sentiment Analysis on Brazilian E-Commerce Website Item Type Thesis*. [Online]. Available from: http://hdl.handle.net/20.500.12648/1652.

​Polyakov, E. V., Voskov, L.S., Abramov, P.S. & Polyakov, S. V. (2020). Generalized approach to sentiment analysis of short text messages in natural language processing. *Informatsionno-Upravliaiushchie Sistemy*. (1). p.pp. 2–14.

​Rane, N. (2023). Enhancing Customer Loyalty through Artificial Intelligence (AI), Internet of Things (IoT), and Big Data Technologies: Improving Customer Satisfaction, Engagement, Relationship, and Experience. *SSRN Electronic Journal*.

​Rita, P. & Ramos, R.F. (2022). Global Research Trends in Consumer Behavior and Sustainability in E-Commerce: A Bibliometric Analysis of the Knowledge Structure. *Sustainability (Switzerland)*. 14 (15).

​Sagvekar, V. & Sharma, P. (2021). *Study on Product Opinion Analysis for Customer Satisfaction on E-Commerce Websites*.

​Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. 2 (3).

​Sempe, S.L. & Naong, M.N. (2021). Exploring the impact of customer awareness on customer engagement in E-commerce: Moderating effect of customer age and gender. *Journal of Contemporary Management*. 18 (2). p.pp. 212–230.

​Shiv Ratan, A. & Divya, M. (n.d.). *Optimizing customer engagement content strategy in retail and E-tail: Available on online product review videos-ScienceDirect Journal of Retailing and Consumer Services*. [Online]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0969698922000595?via%3Dihub.

Shukla, B.N. (2023). E-Commerce Adoption and Consumer Behavior. *Stallion Journal for Multidisciplinary Associated Research Studies*. [Online]. 2 (1). p.pp. 16–20. Available from: https://sjmars.com/index.php/sjmars/article/view/50.

​Skurpel, D. (2022). Digital Distribution as an E-commerce Sales Channel. *Comparative Economic Research*. 25 (1). p.pp. 95–107.

​Sungheetha, A., Bharathi, B., Ganesan, D., Karthikeyan, T., Bindu Madhavi, N. & Chairma Lakshmi, K.R. (2023). E-COMMERCE BUSINESS MODEL ANALYSIS AND SUCCESS IN URBAN AREAS USING AI-DISTRIBUTED MACHINE LEARNING. In: *2023 IEEE International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering, RMKMATE 2023*. 2023, Institute of Electrical and Electronics Engineers Inc.

​Suresh, M.M., Chooramun, N. & Sharif, M.S. (2023). Real-Time Customer Emotion Analysis in E-Commerce based on Social Media Data: Insights and Opportunities. In: *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2023*. 2023, Institute of Electrical and Electronics Engineers Inc., pp. 276–281.

​Taher, G. (2021). E-Commerce: Advantages and Limitations. *International Journal of Academic Research in Accounting, Finance and Management Sciences*. 11 (1).

Tengli, M. B. (2020). *RESEARCH ONION: A SYSTEMATIC APPROACH TO DESIGNING RESEARCH METHODOLOGY*. https://www.researchgate.net/publication/357284560

​Tusar, M.T.H.K. & Islam, M.T. (2021). A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. In: *Proceedings of International Conference on Electronics, Communications and Information Technology, ICECIT 2021*. 2021, Institute of Electrical and Electronics Engineers Inc.

​Wang, Y. (2023). Fast but Slow: Analysis of Fast Fashion Development Path from the Consumer Perspective. *Advances in Economics, Management and Political Sciences*. 36 (1). p.pp. 78–84.

​Zhao, Z., Zhang, N., Xiong, J., Feng, M., Jiang, C. & Wang, X. (2024). Enhancing E-commerce Recommendations: Unveiling Insights from Customer Reviews with BERTFusionDNN. *Journal of Theory and Practice of Engineering Science*. [Online]. 4 (02). p.pp. 38–44. Available from: https://centuryscipub.com/index.php/jtpes/article/view/493.

​Zhou, M., Duan, N., Liu, S. & Shum, H.Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*. 6 (3) p.pp. 275–290.

# CHAPTER 7. APPENDICE

## Appendix 1 Preprocessing

**1a Data cleaning**

df = df.dropna(subset=['title', 'review', 'store\_location', 'latitude', 'longitude', 'review-label' ], how='any')

*# regex pattern to exclude non-ASCII characters while keeping alphanumeric, spaces, and apostrophes*

cleaning\_pattern = r'[^\w\s\']|\_|\d|[^\x00-\x7F]+'

columns\_to\_clean = ['title', 'review']

*# Preprocess the data*

for column in columns\_to\_clean:

df[column] = df[column].fillna('') # Replace NaN with empty string

df[column] = df[column].str.replace(cleaning\_pattern, '', regex=True)

**1b Balancing dataset**

df\_negative = df[df['sentiment'] == 0]

df\_positive = df[df['sentiment'] == 1]

*# Undersample the positive class to match the number of negative instances*

df\_positive\_undersampled = df\_positive.sample(n=len(df\_negative), random\_state=42)

*# Random state for reproducibility*

*# Concatenate the undersampled positive class with the negative class*

df\_balanced = pd.concat([df\_negative, df\_positive\_undersampled])

*# Shuffle the dataset to mix positive and negative entries*

df\_balanced = df\_balanced.sample(frac=1, random\_state=42).reset\_index(drop=True)

*# Verify the new balance*

balanced\_counts = df\_balanced['sentiment'].value\_counts()

print(balanced\_counts)

**1c Lemmatisation**

*# Function to map NLTK's POS tags to WordNet's POS names*

def get\_wordnet\_pos(treebank\_tag):

if treebank\_tag.startswith('J'):

return wordnet.ADJ

elif treebank\_tag.startswith('V'):

return wordnet.VERB

elif treebank\_tag.startswith('N'):

return wordnet.NOUN

elif treebank\_tag.startswith('R'):

return wordnet.ADV

else:

return None *# If the POS tag is not recognized return None*

*# Function to lemmatize a text with POS tags*

def lemmatize\_text(text):

lemmatizer = WordNetLemmatizer()

word\_pos\_tags = pos\_tag(word\_tokenize(text)) *# Tokenize the text and get POS tags*

lemmatized\_words = [

lemmatizer.lemmatize(word, get\_wordnet\_pos(pos\_tag) or wordnet.NOUN)

for word, pos\_tag in word\_pos\_tags

]

return ' '.join(lemmatized\_words) *# Join the lemmatized words back into a single string*

df['review'] = df['review'].astype(str)

df['review'] = df['review'].apply(lemmatize\_text)

df['title'] = df['title'].astype(str)

df['title'] = df['title'].apply(lemmatize\_text)

**Spliting test and train set**

x = df['review'] *# Features*

y = df['sentiment'] *# Target variable*

*# Split the data into training and test sets*

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=42)

**1d vectorisation**

*# list of stop words*

custom\_stop\_words = list(ENGLISH\_STOP\_WORDS.difference(["but", "few", "not", "down", "under", "no", "none", "nothing", "neither", "never", "nobody", "nowhere", "cannot","not", "no", "never", "none", "nothing", "neither", "nowhere", "hardly", "scarcely", "barely", "don’t", "isn’t", "wasn’t", "shouldn’t", "wouldn’t", "couldn’t", "won’t", "can’t", "doesn’t"]))

*# Initialize TfidfVectorizer with the custom stop words list*

vectorizer = TfidfVectorizer(stop\_words=custom\_stop\_words)

*# Fit on the training data and transform both training and test data*

X\_train\_tfidf = vectorizer.fit\_transform(X\_train)

X\_test\_tfidf = vectorizer.transform(X\_test)

## Appendix 2 Model training and evaluation

from sklearn.svm import SVC

from sklearn.metrics import classification\_report, accuracy\_score

svm\_model = SVC()

svm\_model.fit(X\_train\_tfidf, y\_train)

y\_pred = svm\_model.predict(X\_test\_tfidf)

print(classification\_report(y\_test, y\_pred))

print('Accuracy:', accuracy\_score(y\_test, y\_pred))

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

plt.figure(figsize=(10, 7))

sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap="Blues",

xticklabels=['Negative', 'Positive'],

yticklabels=['Negative', 'Positive'])

plt.title('Confusion Matrix')

plt.xlabel('Predicted Class')

plt.ylabel('Actual Class')

plt.show()

## Appendix 3 Streamlit python script

import streamlit as st

import pandas as pd

import numpy as np

import joblib

import re

from nltk import pos\_tag, word\_tokenize

from nltk.corpus import wordnet

from nltk.stem import WordNetLemmatizer

import nltk

import plotly.express as px

import pycountry

from datetime import datetime

import plotly.graph\_objs as go

import plotly.express as px

from wordcloud import WordCloud

import matplotlib.pyplot as plt

from sklearn.metrics import accuracy\_score

import seaborn as sns

from sklearn.metrics import confusion\_matrix, precision\_score, recall\_score, f1\_score, classification\_report

nltk.download('averaged\_perceptron\_tagger')

nltk.download('punkt')

nltk.download('wordnet')

nltk.download('omw-1.4')

nltk.download('vader\_lexicon')

# Load your data

@st.cache\_data

def load\_data():

return pd.read\_csv(r"C:\Users\UGBOKE GEORGE\OneDrive\Documents\archive (3)\Teepublic\_review.csv", encoding= "latin1")

df = load\_data()

def get\_country\_name(alpha\_2\_code):

try:

return pycountry.countries.get(alpha\_2=alpha\_2\_code).name

except AttributeError:

return "Unknown"

def create\_wordcloud(text):

wordcloud = WordCloud(width=800, height=400, background\_color ='white').generate(text)

return wordcloud

def plot\_wordcloud(wordcloud):

plt.figure(figsize=(10, 5))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis('off')

plt.tight\_layout(pad=0)

plt.show()

def show\_wordcloud\_for\_negative\_reviews(df):

negative\_reviews = df[df['Actual\_sentiment'] == 0]['title']

# Join all reviews into a single string

combined\_reviews = ' '.join(negative\_reviews)

# Create word cloud

wordcloud = create\_wordcloud(combined\_reviews)

# Display using Streamlit

fig, ax = plt.subplots()

ax.imshow(wordcloud, interpolation='bilinear')

ax.axis('off')

st.pyplot(fig)

#preprocesing functions

def preprocess(text\_data):

if pd.isnull(text\_data):

return ""

cleaning\_pattern = r'[^\w\s\']|\_|\d|[^\x00-\x7F]+'

cleaned\_text = re.sub(cleaning\_pattern, '', text\_data)

return cleaned\_text

def get\_wordnet\_pos(treebank\_tag):

if treebank\_tag.startswith('J'):

return wordnet.ADJ

elif treebank\_tag.startswith('V'):

return wordnet.VERB

elif treebank\_tag.startswith('N'):

return wordnet.NOUN

elif treebank\_tag.startswith('R'):

return wordnet.ADV

else:

return wordnet.NOUN

def lemmatize\_text(text):

lemmatizer = WordNetLemmatizer()

word\_pos\_tags = pos\_tag(word\_tokenize(text))

lemmatized\_words = [lemmatizer.lemmatize(word, get\_wordnet\_pos(tag)) for word, tag in word\_pos\_tags]

return ' '.join(lemmatized\_words)

def load(vectorizer\_path, model\_path):

vectorizer = joblib.load(vectorizer\_path)

model = joblib.load(model\_path)

return vectorizer, model

def main():

logo = r'C:\Users\UGBOKE GEORGE\Music\Picture1.png'

col1, col2 = st.columns([3,1])

with col2:

st.image(logo, width=100)

with col1:

st.title('Geospatial Sentiment Dashboard')

df = load\_data()

df['title'] = df['title'].apply(preprocess)

df['review'] = df['review'].apply(preprocess)

# Navigation / page layout options

page = st.sidebar.selectbox("Choose your page", ["Prediction", "Dashboard"])

if page == "Prediction":

# Include your existing text input and prediction features

st.subheader('Upload a CSV file or enter text for prediction')

text\_input = st.text\_area("Enter Text")

uploaded\_file = st.file\_uploader("Choose a CSV file")

vectorizer\_path = r"C:\Users\UGBOKE GEORGE\Music\vectorizer.joblib"

model\_path = r"C:\Users\UGBOKE GEORGE\Music\svm\_model.joblib"

if st.button('Predict'):

if uploaded\_file is not None:

data = pd.read\_csv(uploaded\_file)

data['review'] = data['review'].apply(preprocess) # Apply preprocessing to the review column

predictions = predict\_data(data, vectorizer\_path, model\_path)

data['predictions'] = predictions

# Map numeric predictions to textual labels

data['Sentiment\_label'] = data['predictions'].map({1: 'Positive Sentiment', 0: 'Negative Sentiment'})

st.write(data[['review','sentiment', 'predictions', 'Sentiment\_label']])

# Calculate the value counts for both actual and predicted sentiments

actual\_sentiments = data['sentiment'].value\_counts().sort\_index()

predicted\_sentiments = data['predictions'].value\_counts().sort\_index()

# Create a dataframe for plotting

comparison\_df = pd.DataFrame({

'Actual Sentiments': actual\_sentiments,

'Predicted Sentiments': predicted\_sentiments

})

# Create a bar chart

fig = px.bar(comparison\_df, barmode='group',

title='Comparison of Actual and Predicted Sentiments')

st.plotly\_chart(fig, use\_container\_width=True)

accuracy = accuracy\_score(data['sentiment'], data['predictions'])

# Display accuracy in a card

st.metric(label="Prediction Accuracy", value=f"{accuracy:.2%}")

cm = confusion\_matrix(data['sentiment'], data['predictions'])

precision = precision\_score(data['sentiment'], data['predictions'], zero\_division=0)

recall = recall\_score(data['sentiment'], data['predictions'], zero\_division=0)

f1 = f1\_score(data['sentiment'], data['predictions'], zero\_division=0)

st.subheader('Confusion Matrix')

fig\_cm, ax = plt.subplots()

sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', ax=ax)

ax.set\_xlabel('Predicted Labels')

ax.set\_ylabel('True Labels')

ax.set\_title('Confusion Matrix')

st.pyplot(fig\_cm)

st.subheader('Performance Metrics')

st.text('Precision: {:.2f}'.format(precision))

st.text('Recall: {:.2f}'.format(recall))

st.text('F1 Score: {:.2f}'.format(f1))

# Display classification report

st.subheader('Classification Report')

report = classification\_report(data['sentiment'], data['predictions'], output\_dict=True)

st.json(report)

# Display accuracy in a card

accuracy = accuracy\_score(data['sentiment'], data['predictions'])

st.metric(label="Prediction Accuracy", value=f"{accuracy:.2%}")

elif text\_input:

processed\_text = preprocess(text\_input) # Preprocess the text input

prediction = predict\_data(processed\_text, vectorizer\_path, model\_path)

sentiment = 'Positive Sentiment' if prediction[0] == 1 else 'Negative Sentiment'

st.write("Prediction:", sentiment)

elif page == "Dashboard":

df['country\_name'] = df['store\_location'].apply(get\_country\_name)

# Date range selection for year only

st.subheader('Select Year Range')

start\_year, end\_year = st.slider(

'Select year range', min\_value=2018, max\_value=2024, value=(2018, 2024)

)

# Filter the data based on the selected year range

filtered\_year\_data = df[(df['date'] >= start\_year) & (df['date'] <= end\_year)]

st.subheader('Select Month Range')

# You can use a range slider for the month selection

start\_month, end\_month = st.select\_slider(

'Select month range',

options=list(range(1, 13)),

value=(1, 12)

)

filtered\_month\_data = filtered\_year\_data[

(filtered\_year\_data['month'] >= start\_month) &

(filtered\_year\_data['month'] <= end\_month)

]

#Calculations for the cards based on filtered\_df

filtered\_df = filtered\_month\_data

# st.write(filtered\_df.head())

# Calculate the metrics

total\_sentiment\_count = len(filtered\_df)

positive\_count = filtered\_df[filtered\_df['Actual\_sentiment'] == 1].shape[0]

negative\_count = filtered\_df[filtered\_df['Actual\_sentiment'] == 0].shape[0]

positive\_percentage = (positive\_count / total\_sentiment\_count) \* 100

negative\_percentage = (negative\_count / total\_sentiment\_count) \* 100

average\_sentiment = filtered\_df['Actual\_sentiment'].mean()

# Create five columns

col2, col3, col4 = st.columns([3,3,1])

with col2:

st.metric("Positive Sentiments", f"{positive\_count} ({positive\_percentage:.2f}%)")

with col3:

st.metric("Negative Sentiments", f"{negative\_count} ({negative\_percentage:.2f}%)")

with col4:

st.metric("Ratio", f"{average\_sentiment:.2f}")

st.metric(label="Total Sentiments", value=total\_sentiment\_count)

sentiment\_over\_years = filtered\_df.groupby('date')['Actual\_sentiment'].value\_counts().unstack().fillna(0)

# Create a line chart with the sentiment over years

fig\_years = px.line(

sentiment\_over\_years,

x=sentiment\_over\_years.index,

y=sentiment\_over\_years.columns,

labels={'value': 'Number of Reviews', 'date': 'Year'},

title='Sentiment Over Years'

)

# Customize x-axis to show only integer year values

fig\_years.update\_xaxes(

dtick=1, # Set one tick mark per year interval

tick0=min(sentiment\_over\_years.index), # Start tick marks at the minimum year

tickvals=sentiment\_over\_years.index # Set the tick values to the years in the index

)

filtered\_df['month'] = filtered\_df['month'].astype(str).str.zfill(2) # Adds leading zeros

# Combine year and month into a single column for more detailed grouping

filtered\_df['year\_month'] = filtered\_df['date'].astype(str) + '-' + filtered\_df['month']

# Group by 'year\_month' and 'Actual\_sentiment'

sentiment\_over\_months = filtered\_df.groupby(['year\_month', 'Actual\_sentiment']).size().unstack().fillna(0)

# Create a line chart with the sentiment over months

fig\_months = px.line(sentiment\_over\_months, x=sentiment\_over\_months.index, y=[1, 0], labels={'value': 'Number of Reviews', 'year\_month': 'Month'}, title='Sentiment Over Months')

col1, col2 = st.columns(2)

with col1:

st.plotly\_chart(fig\_years, use\_container\_width=True)

with col2:

st.plotly\_chart(fig\_months, use\_container\_width=True)

top\_reviews\_by\_country = filtered\_df.groupby('country\_name')['Actual\_sentiment'].count().nlargest(5)

# Calculate the bottom 5 reviewed countries

bottom\_reviews\_by\_country = filtered\_df.groupby('country\_name')['Actual\_sentiment'].count().nsmallest(5)

# Create a bar chart for top 5 reviewed countries

fig\_top\_reviews = px.bar(top\_reviews\_by\_country, orientation='v',

title="Top 5 Reviewed Countries",

labels={'value':'Number of Reviews', 'index':'Country'})

fig\_top\_reviews.update\_layout(xaxis\_title="Country", yaxis\_title="Number of Reviews")

fig\_top\_reviews.update\_traces(marker\_color='blue')

# Create a bar chart for bottom 5 reviewed countries

fig\_bottom\_reviews = px.bar(bottom\_reviews\_by\_country, orientation='v',

title="Bottom 5 Reviewed Countries",

labels={'value':'Number of Reviews', 'index':'Country'})

fig\_bottom\_reviews.update\_layout(xaxis\_title="Country", yaxis\_title="Number of Reviews")

fig\_bottom\_reviews.update\_traces(marker\_color='red')

# Create two columns

col1, col2 = st.columns(2)

# Display the charts in Streamlit columns

with col1:

st.plotly\_chart(fig\_top\_reviews, use\_container\_width=True)

with col2:

st.plotly\_chart(fig\_bottom\_reviews, use\_container\_width=True)

if st.button('Show Word Cloud for worst concerning words '):

show\_wordcloud\_for\_negative\_reviews(filtered\_df)

# Define the function to determine the color based on count ranges

def determine\_color(count):

if count < 100:

return 'red'

elif count < 1000:

return 'black'

elif count < 10000:

return 'blue'

else:

return 'orange'

# Group by country and sentiment

country\_sentiment\_counts = filtered\_df.groupby('country\_name')['Actual\_sentiment'].value\_counts().unstack(fill\_value=0)

# Make sure the columns are named after unstacking as expected

country\_sentiment\_counts.columns = ['negative', 'positive'] if 0 in country\_sentiment\_counts.columns else ['positive']

country\_sentiment\_counts.reset\_index(inplace=True)

country\_sentiment\_counts['positive\_color'] = country\_sentiment\_counts['positive'].apply(determine\_color)

country\_sentiment\_counts['negative\_color'] = country\_sentiment\_counts['negative'].apply(determine\_color) if 'negative' in country\_sentiment\_counts.columns else 'black'

# Define the color legend as a Markdown string

color\_legend = """

#### Color Legend:

- \*\*Red\*\*: Less than 100 counts

- \*\*Black\*\*: Less than 1,000 counts

- \*\*Blue\*\*: Less than 10,000 counts

- \*\*Orange\*\*: Above 10,000 counts

"""

# Create separate figures for positive and negative sentiments

fig\_positive = px.scatter\_geo(

country\_sentiment\_counts,

locations="country\_name",

locationmode='country names',

text="country\_name", # Display country names as text

hover\_name="country\_name",

hover\_data={

'positive': True, # Show this data as is, without formatting

'negative': True # Show this data as is, without formatting

},

projection="natural earth",

title="Positive Sentiment Reviews by Country",

size\_max=15,

color='positive\_color',

color\_discrete\_map={'red': 'red', 'black': 'black', 'blue': 'blue', 'orange': 'orange'}

)

fig\_negative = px.scatter\_geo(

country\_sentiment\_counts,

locations="country\_name",

locationmode='country names',

text="country\_name", # Display country names as text

hover\_name="country\_name",

hover\_data={

'positive': True, # Show this data as is, without formatting

'negative': True # Show this data as is, without formatting

},

projection="natural earth",

title="Negative Sentiment Reviews by Country",

size\_max=15,

color='negative\_color',

color\_discrete\_map={'red': 'red', 'black': 'black', 'blue': 'blue', 'orange': 'orange'}

)

# Update both figures to have uniform size markers

fig\_positive.update\_traces(marker=dict(size=10))

fig\_negative.update\_traces(marker=dict(size=10))

# Create a two-column layout to display the maps and the color legend

col1, col2 = st.columns([3, 1])

with col1:

st.subheader('Positive Sentiment Map')

st.plotly\_chart(fig\_positive, use\_container\_width=True)

st.subheader('Negative Sentiment Map')

st.plotly\_chart(fig\_negative, use\_container\_width=True)

with col2:

st.markdown(color\_legend, unsafe\_allow\_html=True)

# Calculate the total sentiments and sum of positive sentiments for each country

country\_sentiments = filtered\_df.groupby('store\_location')['Actual\_sentiment'].agg(

total\_sentiments='count',

positive\_sentiments='sum'

)

country\_sentiments['positive\_percent'] = (country\_sentiments['positive\_sentiments'] / country\_sentiments['total\_sentiments']) \* 100

country\_sentiments['negative\_percent'] = 100 - country\_sentiments['positive\_percent']

country\_sentiments['negative\_sentiments'] = country\_sentiments['total\_sentiments'] - country\_sentiments['positive\_sentiments']

# Add a column with country names using the get\_country\_name function

country\_sentiments['country\_name'] = country\_sentiments.index.map(get\_country\_name)

sorted\_countries = country\_sentiments.sort\_values(by='total\_sentiments', ascending=False).reset\_index()

# Search functionality

st.subheader('Search for a Country')

search\_query = st.text\_input('Enter country name').lower()

if search\_query:

search\_results = sorted\_countries[sorted\_countries['country\_name'].str.lower().str.contains(search\_query)]

if not search\_results.empty:

# Selecting the first matching country if there are multiple matches

country\_data = search\_results.iloc[0]

# Generating a pie chart for the selected country

store\_location = country\_data['store\_location']

country\_reviews = filtered\_df[filtered\_df['store\_location'] == store\_location]

total\_positive\_sentiment = country\_reviews[country\_reviews['Actual\_sentiment'] == 1].shape[0]

total\_negative\_sentiment = country\_reviews[country\_reviews['Actual\_sentiment'] == 0].shape[0]

total\_sentiments = country\_reviews.shape[0]

col1, col2, col3 = st.columns(3)

with col1:

st.metric("Total Positive Sentiment", total\_positive\_sentiment)

with col2:

st.metric("Total Negative Sentiment", total\_negative\_sentiment)

with col3:

st.metric("Total Sentiment", total\_sentiments)

fig = px.pie(

values=[country\_data['positive\_percent'], country\_data['negative\_percent']],

names=['Positive Percent', 'Negative Percent'],

title=f"Sentiment Distribution for {country\_data['country\_name']}"

)

st.plotly\_chart(fig)

negative\_reviews = filtered\_df[(filtered\_df['store\_location'] == country\_data['store\_location']) & (filtered\_df['Actual\_sentiment'] == 0)]

if not negative\_reviews.empty:

button\_key = f"show\_wordcloud\_{country\_data['store\_location']}"

if st.button('Show concerning words for the above country', key=button\_key):

if negative\_reviews['title'].isna().all():

st.error("No titles available to generate a word cloud.")

else:

# Filter out possible NaN values from the title column

titles = negative\_reviews['title'].dropna()

# Generate a word cloud for titles in negative reviews

wordcloud = WordCloud(width=800, height=400, background\_color='white').generate(' '.join(titles))

# Display the word cloud using matplotlib

plt.figure(figsize=(10, 5))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis("off")

plt.title("Word Cloud for Negative Sentiments")

st.pyplot(plt)

top\_words = list(wordcloud.words\_.keys())[:5]

st.subheader("Titles containing top words from the word cloud:")

for word in top\_words:

# Get titles containing the top word

relevant\_titles = titles[titles.str.contains(word, case=False, na=False)]

st.markdown(f"#### Titles containing the word: \*\*{word}\*\*")

for title in relevant\_titles.head(5):

st.write(title)

else:

st.warning("No negative reviews found. Please adjust your search or selection.")

else:

search\_results = sorted\_countries.head(20) # Default view is top 10 countries if no search performed

st.subheader('Countries by Sentiment Count')

st.table(search\_results[['country\_name', 'store\_location', 'total\_sentiments', 'positive\_percent', 'negative\_percent']])

top\_positive = sorted\_countries.nlargest(20, 'positive\_percent')

# For top 10 negative sentiment percentages

top\_negative = sorted\_countries.nlargest(20, 'negative\_percent')

# Create the bar chart for positive sentiments

fig\_positive = px.bar(top\_positive, x='positive\_percent', y='country\_name',

orientation='h', title="Top 20 Countries by Positive Sentiment Percent",

text='positive\_percent')

fig\_positive.update\_layout(yaxis={'categoryorder':'total ascending'}, xaxis\_title="Positive Sentiment Percent", yaxis\_title="Country")

fig\_positive.update\_traces(texttemplate='%{text:.2s}%', textposition='outside')

# Create the bar chart for negative sentiments

fig\_negative = px.bar(top\_negative, x='negative\_percent', y='country\_name',

orientation='h', title="Top 20 Countries by Negative Sentiment Percent",

text='negative\_percent')

fig\_negative.update\_layout(yaxis={'categoryorder':'total ascending'}, xaxis\_title="Negative Sentiment Percent", yaxis\_title="Country")

fig\_negative.update\_traces(texttemplate='%{text:.2s}%', textposition='outside')

# Display the charts in Streamlit

st.plotly\_chart(fig\_positive, use\_container\_width=True)

st.plotly\_chart(fig\_negative, use\_container\_width=True)

top\_10\_countries = sorted\_countries.head(10)

# Create a grouped bar chart

fig = px.bar(

top\_10\_countries,

x="country\_name",

y=["positive\_percent", "negative\_percent"],

title="Positive and Negative Sentiment Percentages for Top 10 Countries",

labels={"value": "Percentage", "variable": "Sentiment Type", "country\_name": "Country"},

barmode='group'

)

# Customize the layout

fig.update\_layout(

xaxis\_title="Country",

yaxis\_title="Sentiment Percentage",

legend\_title="Sentiment Type",

legend=dict(orientation="h", yanchor="bottom", y=1.02, xanchor="right", x=1)

)

# Display the chart in Streamlit

st.plotly\_chart(fig, use\_container\_width=True)

bottom\_10\_countries = sorted\_countries.tail(10)

# Create a grouped bar chart for the bottom 10 countries

fig = px.bar(

bottom\_10\_countries,

x="country\_name",

y=["positive\_percent", "negative\_percent"],

title="Positive and Negative Sentiment Percentages for Bottom 10 Countries",

labels={"value": "Percentage", "variable": "Sentiment Type", "country\_name": "Country"},

barmode='group'

)

# Customize the layout

fig.update\_layout(

xaxis\_title="Country",

yaxis\_title="Sentiment Percentage",

legend\_title="Sentiment Type",

legend=dict(orientation="h", yanchor="bottom", y=1.02, xanchor="right", x=1)

)

# Display the chart in Streamlit

st.plotly\_chart(fig, use\_container\_width=True)

def predict\_data(input\_data, vectorizer\_path, model\_path):

vectorizer, model = load(vectorizer\_path, model\_path)

if isinstance(input\_data, pd.DataFrame):

text\_data = input\_data['review'].apply(lambda x: lemmatize\_text(preprocess(x)))

else:

text\_data = pd.Series([lemmatize\_text(preprocess(input\_data))])

text\_features = vectorizer.transform(text\_data)

predictions = model.predict(text\_features)

return predictions

#predictions = ['Positive Sentiment' if pred == 1 else 'Negative Sentiment' for pred in predictions]

#return predictions

if \_\_name\_\_ == "\_\_main\_\_":

main()