Survey Automation: A Machine Learning Approach for Question Classification and Open-Ended Response Analysis

By

George Kamau Maina

095729

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: **George Kamau Maina**

Sign: _____ Date: 27th May 2025

Approval

The dissertation of **George Kamau Maina** was reviewed and approved by the following:

Dr. Kennedy Senagi,
Lecturer, Insititute of Mathematical Sciences,
Strathmore Univeristy.

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University.

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University.

Abstract

Organizations increasingly rely on survey data to guide decision-making, yet manual analysis approaches—often involving spreadsheets and human coding—are time-intensive, error-prone, and struggle to scale, particularly for open-ended responses. This project developed an end-to-end automated solution to improve the accuracy and efficiency of survey analysis by integrating machine learning, natural language processing (NLP), and automated report generation. A labeled dataset of 5,785 survey questions (3,272 open-ended and 2,513 closed-ended) was used to train and evaluate four classifiers based on accuracy: Logistic Regression (97.99%), Naive Bayes (97.69%), Random Forest (97.90%), and Support Vector Machine (SVM), which achieved the best performance with a cross-validation accuracy of 98.51% and a test accuracy of 98.36%. For topic modeling of open-ended responses, 100,000 entries from the IBM Employee Reviews dataset were analyzed using five techniques. BERTopic recorded the highest topic coherence score (0.6148), outperforming NMF (0.6028), LDA (0.5449), TF-IDF + KMeans (0.5426), and GloVe-based clustering (0.4731). The solution was deployed as a web-based application featuring modules for data preprocessing, classification, topic modeling, and visualization. Users can upload raw survey files and automatically receive a fully annotated PowerPoint report—complete with narrative insights generated using GPT-3.5. This automation not only improves the consistency and depth of analysis but also eliminates manual bottlenecks, making it highly scalable and practical for organizations handling frequent or large-scale surveys and enabling faster decision cycles.

Keywords: Survey Automation, Machine Learning, NLP, Topic Modeling, BERTopic, SVM, TF-IDF, Report Generation, Open-Ended Responses

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

API Application Programming Interface

BAMS Building Automation and Management Systems

BERT Bidirectional Encoder Representations from Transformers

BLEU Bilingual Evaluation Understudy

CSV Comma-Separated Values

c-TFIDF Class-based Term Frequency-Inverse Document Frequency

CRISP-DM Cross-Industry Standard Process for Data Mining

DNN Deep Neural Network

GCRISP-DS Generalized Cross-Industry Standard Process for Data Science

GloVe Global Vectors for Word Representation

GPT Generative Pretrained Transformer

GUI Graphical User Interface

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

LLM Large Language Model

LDA Latent Dirichlet Allocation

ML Machine Learning

MLM Masked Language Modeling

NLP Natural Language Processing

NMF Non-negative Matrix Factorization

PPTX PowerPoint XML Format

RNN Recurrent Neural Network

RoBERTa Robustly Optimized BERT Pretraining Approach

SVM Support Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

UMAP Uniform Manifold Approximation and Projection

XLS Excel Spreadsheet Format

Acknowledgement

This thesis stands as a testament to the collective support I have received. I am incredibly thankful to my supervisor, Dr. Kennedy Senagi, for his invaluable guidance, support, and encouragement throughout this research. His expertise and insightful feedback were instrumental in shaping the direction and quality of this dissertation, continually pushing me to refine my thinking and improve my work at every stage.

My heartfelt appreciation extends to my beloved parents, Esther Wambui Murugu and John Maina Kamau, whose sacrifices, encouragement, and steadfast love have been a constant source of strength. Their belief in my abilities never wavered, and I dedicate this achievement to them.

Chapter 1: Introduction

1.1 Background

1.1.1 Data as a Strategic Asset

According to (Cupoli et al., 2014), an asset is a resource that holds economic value, can be owned and controlled. In today's economy and digital world, vast amounts of data are generated from day-to-day business activities. Data is now being recognised as an enterprise asset just like financial resources and is no longer treated as a by-product of business processes but as part of business operations.

Organizations are increasingly recognising data as a critical asset in their pursuit of becoming data-driven. They are leveraging data to drive strategy, understand customer behaviour, improve experiences, identify market trends, create new products, and enhance operational efficiency (Cupoli et al., 2014). A study by the MIT Center for Digital Business found that companies using data-driven decision-making were, on average, 5 percent more productive and 6 percent more profitable than competitors, even after controlling for variables such as labour and IT investment (McAfee et al., 2012). Additionally, analytics fosters competitive advantage through innovation and customer-centric strategies (Davenport et al., 2010).

1.1.2 Importance of Timely Insights

Real-time analytics is the practice of collecting, analysing, and interpreting data as it is generated. It enables organizations to gain immediate insights and respond to events as they occur (Wolniak, 2023). According to (Cupoli et al., 2014), one of the core dimensions of data quality is timeliness—i.e., whether the data is available and current when needed. Timely data supports short- and long-term decision-making and helps capitalize on emerging opportunities (Cupoli et al., 2014). Today, customers expect immediacy. Real-time analytics enhances service personalization, targeted marketing, and timely operational responses (Greasley, 2019). Technologies enabling big data have made it possible to process large volumes of information, extract actionable insights, and adapt to changing market dynamics (Henke and Jacques Bughin, 2016). Real-time capabilities are essential to stay competitive in dynamic environments (Greasley, 2019).

Faster insight extraction leads to increased value, lower costs, and improved efficiency (Yadranjiaghdam et al., 2016). However, manual data analysis introduces delays that undermine the usefulness of insights. In particular, survey data processed using traditional methods—manual cleaning, analysis, approvals, and presentation—often results in outdated findings. This can cause organizations to miss critical trends or make decisions based on stale information.

Given the time-sensitive nature of decision-making, there is a growing need for organizations to adopt end-to-end automation in how they manage survey data—from collection to analysis. Rather than continuing to rely on fragmented and manual processes, automation offers a pathway to improve consistency, scalability, and timeliness.

The automation of survey data analysis has gained significant attention as it offers solutions to the inefficiencies and complexities associated with manual data processing. Manual approaches to survey design, data collection, and analysis can be time-consuming, error-prone, and lack scalability. In contrast, automated systems integrate machine learning (ML), artificial intelligence (AI), and natural language processing (NLP) to streamline workflows and enhance the extraction of meaningful insights from both structured and unstructured responses. This project proposes the development of such an automated system tailored to classify, process, and interpret open-ended and close-ended survey feedback while generating structured outputs and visual summaries to support timely and data-informed decision-making.

The automation of survey data analysis addresses longstanding inefficiencies in the way survey data is collected and interpreted. Manual approaches to survey design, data collection, and analysis can be time-consuming, error-prone, and lack scalability. In contrast, automated systems integrate machine learning (ML), artificial intelligence (AI), and natural language processing (NLP) to streamline workflows, reduce manual effort, and enhance the accuracy and speed of insights derived from both structured and unstructured survey responses.

1.1.3 Survey as a Source of Data and its Importance

Surveys are widely recognized as scalable and effective data collection tools, particularly via online platforms. Surveys typically include close-ended and open-ended questions,

each offering distinct advantages (Reja et al., 2003). Close-ended questions provide predefined responses, facilitating quick and structured analysis (Krosnick and Presser, 2010). Open-ended questions, on the other hand, allow respondents to express themselves in their own words, offering richer and often unexpected insights. They are particularly valuable for gathering sensitive or nuanced feedback (Allen, 2017).

Despite their value, open-ended questions require substantial effort to code and analyse. They also have higher item non-response rates (Reja et al., 2003). Still, insights gleaned from surveys—especially when organizations adopt a user-centric approach—can offer competitive advantages. Understanding customer needs helps firms improve experience and loyalty, both of which are directly linked to financial performance. According to (Palmatier, 2008), strong customer relationships translate into greater loyalty, more accurate insights, and increased market share.

In today's competitive market, firms can no longer assume they understand customer needs. With numerous alternatives available, it is imperative that organizations listen to and act on customer feedback. Ignoring this is akin to navigating with a broken compass.

Voice of the Customer (VoC) programs, which involve collecting and acting on customer feedback, have become strategic priorities. They help companies decipher complex decision-making processes and act swiftly. The timeliness of response can determine whether a customer stays or leaves. For instance, Safaricom, a leading telecommunications firm in Kenya, uses Customer Relationship Management (CRM) systems to efficiently handle responses and improve service quality (Gitonga, 2016). These tools help attract new customers and retain existing ones by providing timely and effective responses, thus enhancing overall satisfaction.

1.2 Problem Statement

Survey analysis in many organizations still follows a largely manual and sequential process, beginning with survey design and distribution—typically via email or printed forms—followed by manual collation of responses and the use of spreadsheet software for basic aggregation and analysis. For open-ended questions, researchers typically rely on human coders to read and categorize responses based on emerging themes. This method can be time-consuming, subjective, and error-prone, especially when dealing with large volumes of

data.

This manual or semi-automated approach presents significant challenges in terms of timeliness, scalability, and accuracy. Large-scale surveys can take extended periods to collect, and by the time the data is ready for analysis, the insights may be outdated. Moreover, manual coding of qualitative responses is resource-intensive, highly dependent on individual judgment, and difficult to standardize. This not only delays decision-making but also risks compromising the consistency and reliability of insights.

In a dynamic business environment where swift and informed decisions are critical, these inefficiencies hinder organizations from fully leveraging their survey data. Despite the availability of advanced tools for analytics and machine learning, there remains a gap in comprehensive, end-to-end automated solutions that can handle both quantitative and qualitative survey data at scale. As a result, businesses struggle to derive timely and actionable insights, ultimately affecting their competitiveness and responsiveness.

1.3 Main Objective

The key objective of this project is to develop an end-to-end automated solution for survey data analysis that enhances the timeliness, accuracy, and scalability of deriving insights.

1.4 Specific Objective

This research aimed to address the following objectives:

(a) To review existing machine learning solutions for survey data analysis.

(b) To automate the coding of qualitative responses.

(c) To develop a data pipeline for seamless data cleaning, processing and analysis.

(d) To automate the creation of visual reports.

1.5 Research Questions

Based on the main and specific objectives of this study, the following research questions were formulated to guide the investigation:

a What existing machine learning approaches have been proposed or applied for automating survey data analysis?

b Can a machine learning model accurately classify survey questions into open-ended and closed-ended types based solely on their phrasing?

c How effective are NLP-based topic modeling techniques in extracting meaningful themes from open-ended survey responses?

d Is it possible to generate annotated, presentation-ready reports directly from raw survey data with minimal manual intervention?

1.6 Scope of the Study

1.6.1 Functional Scope

The study included the following key components:

(a) Automated Question Classification: Distinguishes between open-ended and closed-ended questions using a supervised machine learning model trained on 5,785 labeled survey questions.

(b) Topic Modeling of Open-Ended Responses: Extracts themes from qualitative feedback using multiple modeling techniques, including BERTopic, to convert unstructured data into interpAI-Powered Chart Annotation: Uses Large Language Models (LLMs) to generate narrative summaries of charts and graphs, enhancing insight delivery for non-technical users.

(c) Automated PowerPoint Report Generation: Creates a structured, insight-rich PowerPoint file containing charts and annotated narratives, ready for presentation or stakeholder reporting.

1.6.2 Target Users

This project is designed to support a diverse range of users who work with survey data in various contexts:

(a) Customer Experience Teams: Seeking to streamline analysis of customer feedback for service improvement.

(b) Academic Institutions: Conducting research surveys and requiring scalable tools for analysis and reporting.

(c) Survey Designers and Researchers: Aiming to automate parts of the survey workflow to reduce manual effort.

(d) Data Analysts and Data Scientists: Responsible for transforming raw survey responses into actionable insights.

1.7 Limitations of the Project

While this project aims at automating the survey analysis process and reducing the time that data comes to insights, there are several limitations that must be acknowledged:

(a) Model Training and Bias: Pre-trained models, may carry inherent biases from the data they were trained on, potentially affecting the fairness and objectivity of the analysis.

(b) Dependence on Data Quality: The effectiveness of the automated system relies heavily on the quality of the survey data. Poorly structured, incomplete, or inconsistent data may hinder the system's ability to generate accurate insights.

(c) Complexity of Open-Ended Responses: NLP models may struggle with highly nuanced or context-specific language, dialects, or uncommon terminologies, potentially leading to inaccuracies in the coding of qualitative responses.

1.8 Research Justification

This project is justified by the need to revolutionize the survey analysis process, making it more efficient, consistent, and responsive to the dynamic nature of business environments.The automation of repetitive tasks not only accelerates the analysis process but also frees up valuable resources for more strategic and complex aspects of data interpretation. The introduction of standardized templates ensures that reports are consistently formatted, reducing the chances of errors and promoting a culture of standardized reporting.

Moreover, the project's emphasis on enhancing data quality through automation aligns with industry best practices, addressing common challenges related to incomplete re-

sponses, data entry errors, and sampling biases. The goal of providing timely insights is crucial. This project aims to ensure that decision-makers have access to fresh and up-to-date data, empowering them to respond promptly to changing dynamics and capitalize on timely opportunities. Therefore, this project not only addresses the inherent delay in traditional survey analysis methods but also enhances the overall quality and relevance of data and reports available to decision-makers.

Chapter 2: Literature Review

Data analysis has undergone a significant transformation, shifting from manual, labor-intensive processes to highly automated, technology-driven approaches. Historically, data was processed through manual entry, cleaning, and analysis, which were time-consuming and prone to errors. The rise of automation has enabled more efficient data handling, allowing for faster processing, improved data quality, and real-time insights, which are crucial for timely decision-making across various sectors. Given this shift, this chapter explores the automation of data analysis, with a focus on survey data, discussing methods that improve efficiency and accuracy, addressing the challenges related to timeliness and quality, and evaluating the effectiveness of automated tools using key performance metrics.

2.1 Automating Survey Data Analysis

The automation of survey data analysis has gained significant attention as it offers solutions to the inefficiencies and complexities associated with manual data processing. Traditional approaches to survey design, data collection, and analysis can be time-consuming, error-prone, and lack scalability. Automation, through the use of machine learning (ML), artificial intelligence (AI), and natural language processing (NLP), addresses these challenges by streamlining the survey workflow and enhancing the extraction of meaningful insights.

The automation of survey data analysis has gained significant attention as it offers solutions to the inefficiencies and complexities associated with manual data processing. Traditional approaches to survey design, data collection, and analysis can be time-consuming, error-prone, and lack scalability. Automation, through the use of machine learning (ML), artificial intelligence (AI), and natural language processing (NLP), addresses these challenges by streamlining the survey workflow and enhancing the extraction of meaningful insights. A key advancement in this domain is the integration of AI and large language models (LLMs) to automate and improve survey generation and analysis. (Bonorino, 2023) introduces the Smart Surveys tool, a modular Python-based platform that automates the full lifecycle of survey research. Structured around three core modules—question generation, survey distribution and data collection, and text mining with insight gen-

eration—the tool uses LLMs to automatically create questions based on user prompts, supports refinement of uploaded items, and facilitates deployment across email and social platforms. It tracks responses and applies text mining techniques such as sentiment analysis, named entity recognition, and topic extraction to derive insights, visualized through bigram networks and keyword frequency charts. Demonstrated through educational case studies, the platform targets users with minimal technical skills and is built on scalable open-source libraries like Scikit-learn, NLTK, Pandas, and Transformers. However, while it enhances the efficiency of structured survey workflows, Smart Surveys offers limited depth in processing unstructured or open-ended responses. Its emphasis is largely on administrative automation and descriptive reporting, lacking the semantic depth required for nuanced qualitative analysis. This study addresses that gap by integrating advanced NLP-based topic modeling to deepen insight generation from open-ended feedback, thus building on the foundational framework proposed by Bonorino.

In parallel, (Du et al., 2023) introduce QuestGenius, a web-based platform designed to automate the generation and analysis of surveys using AI and natural language processing. Built with a user-friendly interface using HTML, CSS, and JavaScript, the system enables users—particularly those with limited statistical expertise—to create AI-generated questionnaires from custom prompts. The platform emphasizes bias mitigation during question formulation and includes automated mechanisms for professional-level statistical analysis and visual presentation of results. In addition to simplifying question creation and survey distribution, QuestGenius allows for analysis without the need for complex statistical tools, thereby increasing accessibility for a broad audience. Experimental evaluations demonstrate the system's robustness, particularly in handling adversarial or ambiguous user inputs, while comparative analysis with platforms like SurveyMonkey and Typeform highlights QuestGenius's strength in integrating intelligent analytics with intuitive design. However, despite these advancements, the platform focuses primarily on structured data and does not deeply address challenges related to unstructured text analysis, especially in open-ended responses. This project expands on such efforts by incorporating topic modeling and advanced NLP for in-depth interpretation of qualitative survey data.

Beyond simplifying survey creation and analysis, automation is also transforming how

open-ended survey responses are processed. The complexity of analyzing textual data has been addressed by techniques that go beyond traditional keyword-based grouping. For instance, (Moreo et al., 2019) present a framework for building automated survey coders using Interactive Learning (IL), which combines active learning and incremental learning to optimize the annotation and training process. In this framework, the system iteratively selects the most informative verbatim responses for manual labeling, thereby enhancing the learning rate and minimizing annotation effort. As new manually coded data becomes available, the model is updated in real time, gradually improving classification accuracy and consistency. Notably, this IL-based approach allows for greater adaptability and supports classifier reuse across different coding tasks—making it a practical tool for longitudinal and multi-topic survey workflows. However, the framework still depends heavily on continuous human input for supervision, which constrains its scalability and automation potential. While the method significantly reduces the burden of manual labeling, it does not fully eliminate the need for human intervention. In contrast, this study proposes using pre-trained NLP models and topic modeling to automate qualitative response analysis more comprehensively, aiming to deliver scalable, end-to-end automation that minimizes manual effort even further.

Another complementary approach to handling open-ended responses is the use of context-aware clustering. (Esmaeilzadeh et al., 2022) propose a novel end-to-end framework that leverages Sentence-BERT to transform textual survey responses into semantic embeddings and applies k-means clustering to group them into thematically coherent clusters. The system improves interpretability by automatically labeling each cluster using representative keywords and providing visual summaries such as context-specific word clouds. Notably, it is designed for on-device deployment to protect user privacy and support real-time processing, making it suitable for practical applications in mobile learning environments. Evaluation on synthetic datasets demonstrates the framework's effectiveness in identifying coherent clusters and reducing the manual effort required in qualitative analysis. However, a key limitation is its dependence on a predefined number of clusters, which may not optimally reflect the natural topic distribution in diverse real-world data. Our study builds on this work by using probabilistic topic modeling techniques that automatically infer the number and composition of topics, thereby enhancing adaptability and minimizing manual configuration.

(Card and Smith, 2015) investigate the effectiveness of two machine learning models—L1-regularized logistic regression and recurrent neural networks (RNNs)—for the task of automatically coding open-ended survey responses. Using a dataset from the 2008 American National Election Studies (ANES), the authors evaluate how well these models perform in classifying verbatim responses into predefined categories, such as political issues or policy themes. The logistic regression model, optimized through Bayesian inference and L1 regularization, achieved robust performance with relatively small training sets, showing consistent accuracy across multiple labels. In contrast, the RNNs struggled to outperform this baseline, particularly in data-sparse and multi-label scenarios. This finding highlights a key challenge in applying deep learning to social science survey data—namely, the data sparsity and label imbalance that commonly occur in real-world surveys. While their study advances the case for adopting regularized linear models in automated survey coding, it stops short of addressing qualitative topic discovery or the interpretability of unstructured responses. The present study seeks to close that gap by leveraging topic modeling techniques, which not only classify responses but also extract latent themes and insights from open-ended feedback without relying on predefined categories. In doing so, it enhances interpretability and provides greater flexibility in analyzing qualitative survey data.

(Baburajan, 2021) explores the utility of topic modeling in processing open-ended survey responses, particularly in the context of public opinion on autonomous vehicles. The study applies Latent Dirichlet Allocation (LDA) to uncover underlying themes in qualitative feedback collected via surveys. Through a detailed comparative analysis, Baburajan demonstrates that open-ended questions—when analyzed with topic modeling—can reveal deeper and more varied insights than traditional close-ended formats. The study also underscores practical challenges such as text preprocessing, ambiguity in topic interpretation, and sensitivity to the number of specified topics. While the LDA approach proves effective in surfacing hidden patterns, the model's reliance on bag-of-words representation limits its capacity to capture contextual relationships within the text. Moreover, the paper does not propose mechanisms to automate report generation or integrate visual outputs. In contrast, the current study builds upon these foundations by incorporating more context-aware models such as BERTopic and aims to deliver not only topic extraction but also auto-generated visualizations and narrative summaries, thereby improving

both depth and usability of analysis.

(Pais et al., 2025) propose a hierarchical Mixture of Unigrams (hMoU) topic model specifically designed to handle the complexities of open-ended responses in surveys conducted using complex sampling designs. Their model incorporates survey weights, strata, and clustering structures into the generative process, allowing for more accurate and representative topic distributions at the population level. Applied to the 2020 American National Election Studies (ANES), the hMoU model effectively captures nuanced differences in topic prevalence across demographic subgroups while ensuring that survey design characteristics are properly reflected in the output. This is particularly valuable for public policy research, where understanding variation across populations is essential. However, while hMoU enhances the representativeness of topic estimates in complex surveys, it still assumes a static set of latent topics and does not directly address interpretability or user-facing outputs. In contrast, this study prioritizes real-time, automated generation of topic visualizations and summaries, making the insights more accessible and actionable for practitioners who require quick turnaround and interpretability in reporting.

2.1.1 Algorithm-Level Approaches

(Blei et al., 2003) LDA is one of the most widely used generative probabilistic models for topic modeling. It assumes that documents are mixtures of topics, and topics are distributions over words. In one of the foundational applications, (Blei et al., 2003) introduced LDA to model large document corpora by inferring hidden thematic structures. More recently, (Roberts et al., 2014) applied a structural topic model (an extension of LDA) to analyze open-ended responses from the American National Election Study (ANES). Their model allowed the incorporation of metadata like demographic variables to examine how topic prevalence varied across population subgroups. Despite its flexibility, LDA struggles with short text segments like survey responses due to sparsity and the assumption of word independence.

(Lee and Seung, 1999) NMF is a matrix factorization technique that decomposes a document-term matrix into two non-negative matrices capturing latent topic structures. In the context of topic modeling, (Xu et al., 2003) demonstrated its application in clustering news articles and noted that NMF often yields parts-based, interpretable topics.

However, the method operates under the assumption that word co-occurrence statistics suffice to identify thematic structure and does not account for contextual usage. This makes NMF suitable for applications where interpretability and computational efficiency are paramount, but less effective in capturing deep semantic relationships in text. In our study, we improve upon this by employing BERTopic, which benefits from contextualized embeddings.

KMeans is a widely used algorithm for partitioning text vectors into k clusters. (**?**) evaluated KMeans on several text mining tasks and found it effective for large document sets when used with TF-IDF features. In a more recent application, (Esmaeilzadeh et al., 2022) applied KMeans on Sentence-BERT embeddings to cluster open-ended educational survey responses, enabling semantic grouping of free-text answers. Nevertheless, KMeans requires specifying the number of clusters a priori and assumes that topics are equidistant in feature space. These assumptions are often violated in real survey data where the semantic distribution of responses is uneven. This project instead uses HDBSCAN within BERTopic, allowing for adaptive, density-based topic clustering without fixed cluster numbers.

Algorithm-level techniques play a critical role in automating open-ended survey analysis, particularly in tasks involving clustering and classification. (Ikotun et al., 2023) highlight k-means as a widely adopted clustering algorithm due to its simplicity and efficiency. However, they also underscore key limitations, such as sensitivity to initialization and the requirement to predefine the number of clusters—factors that reduce its flexibility in real-world data scenarios. These shortcomings support the shift toward density-based clustering methods like HDBSCAN, which are more adaptive and used in models such as BERTopic, as applied in this study.

Beyond clustering, (Sarker, 2021) categorizes key machine learning paradigms(supervised, unsupervised, and deep learning) as essential for handling structured and unstructured data. This project aligns with this view by using supervised learning for classifying survey question types and unsupervised topic modeling for analyzing open-ended responses. However, Sarker's work is primarily conceptual and lacks implementation detail specific to survey workflows.

To address practical constraints such as limited labeled data, (Adadi, 2021) advocate for

data-efficient AI strategies, including transfer learning and domain adaptation. This resonates with the current study's use of pre-trained Sentence-BERT embeddings in BERTopic, which enhance semantic understanding with minimal training overhead. Nonetheless, Adadi's recommendations are high-level; this project builds on them by applying such strategies within a complete survey automation framework, offering operational insight into their real-world utility.

2.1.2 Data Quality Assurance Techniques

Automated data quality assurance techniques are fundamental to enhancing the reliability and performance of machine learning models. (Bilal et al., 2022) argue that preprocessing is a crucial step in the data pipeline, directly impacting model accuracy and efficiency. They propose a Python-based auto-preprocessing architecture that automatically identifies and rectifies data issues, providing tailored recommendations for cleaning and preparation through visual insights. This automated approach simplifies data transformation tasks and significantly improves model performance by ensuring high-quality input data. Such automation streamlines the traditionally labor-intensive data preprocessing phase, making it more accessible and effective for varied datasets.

Further contributing to the discourse, (Whang et al., 2023) emphasize a shift towards data-centric AI, wherein data quality takes precedence alongside algorithmic development. With real-world datasets often being imperfect—containing biases, inconsistencies, or incomplete entries—the study underscores the need for automated validation, cleaning, and integration to enhance model robustness. (Geekiyanage et al., 2021) similarly focus on challenges within data analytics, such as noise, time delays, and missing data, advocating for automated techniques to preserve accuracy and validity in real-time and processed data. Across diverse applications, from AI ethics to industrial analytics, these studies collectively highlight that robust, automated data quality management is pivotal to ensuring dependable, fair, and efficient machine learning outcomes.

2.1.3 Real-Time Processing Strategies

Real-time processing strategies are increasingly crucial in data analysis, particularly as data volumes and the need for immediate insights grow. (Dubuc et al., 2020) highlight the emergence of streaming data and its challenges in processing and analysis, with

a focus on the velocity aspect of big data—requiring real-time data stream processing that traditional storage and computation cannot handle effectively. They delve into foundational concepts such as distributed computation and fault tolerance, emphasizing technological solutions that enable online analysis of data streams, thereby addressing real-world challenges in data stream mining. This underscores the evolution of data analytics, where processing pipelines must not only accommodate high data throughput but also provide timely insights without overwhelming computational resources.

(Sharma et al., 2023) further stress the importance of real-time data analysis (RDA) as a transformative discipline empowering organizations to make instant, informed decisions, thus securing a competitive edge in industries where data generation is exponential. The paper elucidates the architectural frameworks and applications of RDA, underlining its significance in today's data-driven landscape, where timely insights are invaluable. (Himeur et al., 2023) extend the concept to building automation and management systems (BAMSs), where AI and big data analytics enhance the management and operational efficiency of buildings. They address the use of real-time data in intelligent decision-making for tasks such as energy consumption monitoring, anomaly detection, and performance optimization. Collectively, these studies reveal how real-time processing strategies not only provide critical, actionable insights across varied domains but also support sustainable and efficient systems management through AI-enhanced analytics and continuous data stream processing.

2.2 Timeliness and Quality Challenges

2.2.1 Managing Data Volume and Complexity

Handling large and complex datasets presents several challenges that directly impact the scalability, processing speed, and diversity of data, thus affecting the timeliness of automated analysis. (Lwakatare et al., 2020) identify the difficulties in developing and maintaining large-scale machine learning (ML) systems, particularly when adaptability and scalability are concerned. As datasets grow, so do the requirements for flexible and scalable architectures that can handle the vast amount of information during data acquisition, training, evaluation, and deployment. The complexity of these processes often hinders timely data processing and real-time analysis, making it challenging for ML

systems to adapt to rapidly changing data conditions without performance trade-offs.

Additionally, cyber-physical systems (CPSs), as highlighted by (Alwan et al., 2022), exemplify large-scale data operations where data quality issues arise from dynamic conditions and diverse hardware, software, and sensor components. Scalability becomes problematic in these contexts, particularly when maintaining data accuracy, completeness, and consistency across various sensor nodes and networks. The diversity and heterogeneity of data add another layer of complexity. (Aldoseri et al., 2023) emphasize that AI-based systems, reliant on large datasets from varied sources, struggle with ensuring consistent data quality and integrity, especially when data is sourced from numerous domains with varying formats and standards. These challenges affect not just scalability but also the speed and accuracy of processing, as the systems must account for diverse data characteristics while remaining efficient. Furthermore, (Ikegwu et al., 2022) shed light on the intricacies of big data analytics (BDA), where the vast and varied data formats pose challenges in storage, visualization, and processing. Properly managing this complexity is essential for timely extraction of insights. However, the lack of comprehensive strategies to handle diverse data streams efficiently and to meet the speed requirements of real-time analysis still remains a gap, necessitating further research and technological advancements to ensure both timeliness and quality in automated data analysis.

2.3 Ensuring Data Accuracy and Reducing Bias

Maintaining data accuracy and reducing bias are critical challenges in automated survey data analysis, requiring careful consideration to preserve the integrity of responses. (Oladoyinbo et al., 2024) highlight that ethical considerations and regulatory compliance play a pivotal role in ensuring data integrity, particularly as artificial intelligence (AI) applications grow more complex. They argue that ethical awareness and adherence to regulatory frameworks significantly improve the accuracy and effectiveness of AI-driven systems by setting standards for data collection, processing, and usage. However, even with stringent ethics and regulations, (Oladoyinbo et al., 2024) note that experience alone is insufficient to address accuracy fully, indicating that more comprehensive approaches—incorporating ethical practices, proper training, and adaptable regulations—are needed to uphold data integrity effectively in AI and automated systems.

Similarly, (Chen et al., 2023) emphasize the challenges surrounding bias in AI systems, which can compromise data accuracy and fairness. The paper underscores that both human-introduced biases (e.g., during data labeling or algorithm design) and inherent machine biases (stemming from imbalanced training data) present significant obstacles to fairness and data quality. They explore various mitigation techniques, including fairness-aware training methods and bias analysis, which aim to reduce inaccuracies and align AI systems with principles of equity. The study argues that transparency in AI systems is critical for ensuring accurate, unbiased outputs and discusses how interpretability plays a central role in assessing and addressing bias.

Moreover, in the context of Big Data Analytics (BDA), (Sivarajah et al., 2017) address challenges in ensuring data quality, especially given the variety and volume of data processed. They outline that BDA faces intrinsic issues such as data veracity, complexity, and scalability, which can lead to inaccuracies if not managed effectively. The study advocates for a deep understanding of BD challenges and appropriate BDA methods to mitigate potential distortions in data interpretation and analysis. Building on this, (Balayn et al., 2021) delve into the intricacies of bias in data-driven decision-making systems, arguing that most fairness interventions are algorithmic-centered and fail to fully account for biases rooted in data management. The study suggests a shift towards a data-centric approach, where fairness constraints are imposed on datasets before training and during validation, ensuring that both the content and the structure of the data maintain integrity. Such approaches aim to eliminate inaccuracies and biases from the ground up, thereby ensuring the ethical and accurate functioning of automated survey systems.

2.4 Balancing Automation with Interpretation

The challenge of ensuring that automated data analysis retains context-sensitive interpretation lies in its inherent limitations to grasp nuanced insights embedded within survey results. (Gerdon, 2024) emphasizes the critical role of contextual norms in evaluating the legitimacy of data-driven technologies, suggesting that public opinion on fairness and privacy varies by context and timing. Automation struggles to account for these subtle societal and ethical variations, risking oversimplified interpretations that may not align with real-world complexities. (Matteucci et al., 2023) further elaborate on this by examining explainability in AI-driven automation, arguing that technical constraints and

diverse stakeholder needs make it challenging to provide contextually relevant explanations. Such limitations hinder the capacity of automated systems to respond to dynamic human perspectives and ethical demands, necessitating human intervention for deeper contextual understanding.

Moreover, while advanced models like large language models (LLMs) enhance certain aspects of survey analysis, they still fall short in interpreting the intricacies of sampling biases and context-specific insights (Jansen et al., 2023). Automation effectively addresses structural issues in data analysis but cannot wholly replace human interpretation of cultural, emotional, or situational nuances. Similarly, (Zhong et al., 2024) discuss how knowledge graph technology improves data integration and predictive capabilities in domains like intelligent auditing but lacks scalability in understanding complex contexts inherent to real-world applications. While cloud-edge collaboration may optimize computational efficiency, the depth of context-specific analysis still demands human expertise to identify subtle risks and provide accurate interpretations. These challenges collectively highlight that while automation enhances efficiency, the sophistication of contextual interpretation remains a domain where human oversight is indispensable.

2.5 Evaluation Metrics for Automated Survey Analysis

2.5.1 Accuracy and Consistency Metrics

Evaluating the accuracy and consistency of automated survey analysis is critical for ensuring that results are reliable and can be reproduced across different contexts. (Tripathi et al., 2021) emphasize the need for frameworks like the Cross-Industry Standard Process for Data Mining (CRISP-DM), and its proposed extension, the Generalized Cross-Industry Standard Process for Data Science (GCRISP-DS), to dynamically handle data and model-related challenges. Such frameworks play a pivotal role in ensuring that model development is aligned with business objectives—such as improving decision-making efficiency, reducing human effort, or enabling timely insights—while maintaining robust accuracy in data interpretation. By introducing phases that allow iterative refinement and comprehensive business understanding, these frameworks support the development of more consistent and accurate automated analysis processes. Tailored evaluation metrics integrated into such frameworks help assess how well models perform over time, enabling

better detection of inconsistencies and the enhancement of reproducibility.

Similarly, (Taylor et al., 2023) argue for a more holistic presentation of results in neuroimaging, proposing an approach where highlighting complete data sets rather than focusing solely on statistically significant findings can improve the accuracy and reliability of conclusions drawn from automated analysis. Such transparency addresses issues like selection bias and irreproducibility, providing more robust insights into variability across studies. (Scott et al., 2021) add to this by emphasizing that automation tools, such as Covidence and RevMan, improve time efficiency and accuracy in systematic reviews, although knowledge gaps remain a barrier to full adoption. Their findings suggest that continuous training on using such tools, coupled with further development to enhance accuracy and consistency, are paramount to meeting the needs of those conducting surveys and systematic analyses. Collectively, these works highlight that implementing clear, dynamic metrics and transparent practices are central to achieving accuracy and consistency in automated survey data analysis.

2.5.2 Timeliness and Efficiency Measures

Timeliness and efficiency are critical measures in automated survey analysis, directly influencing how quickly data can be processed and insights delivered for decision-making. In a landscape where vast amounts of data are collected and processed daily, ensuring swift analysis without compromising quality is vital. Jaeger and Cardello (2022) highlight that while online data collection allows for rapid data processing, there are challenges in maintaining data quality due to uncontrolled survey administration elements. To achieve efficient analysis, it is necessary to implement real-time quality checks and well-balanced survey designs that can promptly identify issues, ensuring that insights remain both timely and valid. This approach minimizes delays and maintains the reliability of data throughout the analytical pipeline.

In industries such as tourism and hospitality, where markets are highly dynamic, the rapid delivery of insights provides a significant competitive advantage. (Stylos et al., 2021) suggest that big data can drive agile decision-making, where the speed of data processing enables organizations to quickly interpret consumer behavior and respond to market trends. An integrated framework for big data analysis not only accelerates pro-

cessing but also maximizes the utility of insights, contributing to efficient value creation. Moreover, (Zhong et al., 2024) introduce the concept of "Data Analytics Competency," which emphasizes the balance between timely data processing and maintaining data quality. They argue that for open data (OD) to be leveraged effectively, organizations need to focus on the rapid delivery of insights while ensuring that data meets quality standards. This holistic approach to data timeliness and efficiency ultimately supports informed and strategic decision-making in fast-paced environments, providing both speed and accuracy in analysis.

2.5.3 Quality and Data Integrity

Data quality and integrity are fundamental to the effectiveness of automated analysis, with a focus on accuracy, completeness, and handling anomalies. (Rangineni et al., 2023) identify core dimensions of data quality such as accuracy, consistency, and reliability, emphasizing their influence on the effectiveness of data analytics. High-quality data enhances decision-making, customer satisfaction, and operational efficiency, but issues like data integration complexities and evolving sources challenge quality maintenance. To address these challenges, methodologies like data profiling, cleansing, and standardization are crucial for rectifying inconsistencies. Assessing data quality requires comprehensive metrics that measure the degree of alignment with expected values across multiple dimensions. By implementing robust data governance frameworks, organizations can proactively manage quality anomalies, thereby driving innovation and maintaining a competitive advantage.

Specific to sensor data and big data contexts, (Teh et al., 2020) and (Widad et al., 2023) underscore the importance of metrics for managing anomalies such as outliers, missing values, and faults. (Teh et al., 2020) find that common solutions for sensor data quality involve PCA, artificial neural networks, and Bayesian networks, particularly for fault detection and correction. However, the lack of uniform evaluation processes and the use of non-public datasets make direct comparison of these approaches challenging. (Widad et al., 2023) propose a holistic framework for detecting quality anomalies in big data, accounting for multiple dimensions such as completeness and consistency. They introduce the "Quality Anomaly Score" as a metric to measure the degree of anomalousness, achieving a high level of precision in anomaly detection. These contributions highlight the need for intelligent, domain-agnostic frameworks that not only detect but also effec-

tively address data quality issues, thereby preserving the integrity and completeness of datasets for accurate automated analysis.

2.6 Approaches to Addressing Timeliness and Quality Challenges in Automated Data Analysis

To address the challenges of managing data volume and complexity in automated data analysis, scalable and flexible system architectures are essential. (Lwakatare et al., 2020) advocate for adaptable frameworks that can handle the vastness and diversity of data efficiently, particularly for large-scale machine learning systems. To improve scalability, these frameworks must be able to optimize data processing pipelines at each stage(acquisition, training, evaluation, and deployment) while maintaining the ability to adapt to changes in data types and sources. (Alwan et al., 2022) also suggest the need for more robust data quality mechanisms, especially in cyber-physical systems (CPSs), to ensure that sensor data remains accurate and consistent. One approach to this is employing distributed computing solutions that can manage data in real-time, along with developing standardized protocols for data integration across hardware and software components to mitigate data inconsistencies.

Ensuring data accuracy and reducing bias requires a combination of technical, ethical, and regulatory approaches. (Oladoyinbo et al., 2024) argue that ethical guidelines and adherence to regulatory standards are critical for maintaining data integrity, particularly in AI systems that are prone to biases introduced at the design or implementation stages. Developing fairness-aware training models, as suggested by (Chen et al., 2023), is a practical approach to reduce inaccuracies by proactively addressing biases in the algorithmic process. This involves incorporating bias analysis and interpretability mechanisms that not only detect but also mitigate disparities in data, ensuring more equitable and reliable outcomes. Additionally,(Sivarajah et al., 2017) emphasize the importance of deep understanding of the challenges inherent to Big Data Analytics (BDA). By employing advanced data governance strategies and methods specifically aimed at data veracity and consistency, organizations can significantly improve accuracy. (Balayn et al., 2021) propose a data-centric approach to fairness, which means embedding fairness constraints into the dataset before training, ensuring that the data structure itself aligns with equitable principles.

Balancing automation with human interpretation remains a complex issue, necessitating approaches that combine computational efficiency with contextual awareness. While automation speeds up data processing and initial analysis, (Gerdon, 2024) suggests incorporating context-aware frameworks that account for the socio-ethical nuances in data, especially when it comes to public opinion on fairness and privacy. This can be achieved through hybrid models, where automated analysis is supplemented by human oversight to provide a more comprehensive interpretation. Similarly, (Matteucci et al., 2023) highlight the need for explainable AI that can communicate results effectively to diverse stakeholders without compromising performance. By implementing layered explainability in automated systems, where complex data is translated into human-interpretable formats, the transparency and contextual relevance of the insights are improved.

Enhancing data quality also requires domain-specific anomaly detection and correction tools that are both proactive and adaptive. Techniques like Principal Component Analysis (PCA) and Artificial Neural Networks (ANNs), as discussed by (Teh et al., 2020) , provide robust frameworks for detecting faults and anomalies in sensor data. These solutions are complemented by hybrid approaches that combine several data science methods to correct errors in real-time, thereby ensuring data consistency and completeness. Addressing outliers and missing data, as suggested by (Widad et al., 2023), is equally crucial in big data scenarios, and developing generic anomaly detection frameworks that are independent of specific domains can significantly enhance data quality across varied datasets. Employing "Quality Anomaly Scores," as they propose, provides an additional metric to quantify the degree of data quality anomalies, allowing automated systems to prioritize corrections and maintain integrity effectively.

2.7 Research Gap Covered

Traditional survey tools focus heavily on structured, close-ended questions, offering straightforward visuals and simple analytics. However, they often fall short when it comes to handling open-ended responses, which are rich with qualitative insights. These tools typically provide basic reports for quantitative data, leaving the more complex, unstructured feedback either untouched or poorly analyzed.

This project aims to fill that gap by employing advanced Natural Language Processing

(NLP) techniques. The goal is to automate not only the classification of open-ended responses but also their analysis, unlocking deeper insights from the qualitative data that is frequently overlooked. Beyond just processing the data, the project will generate customized reports with insightful annotations, eliminating the need for manual intervention and aligning the output with the organization's specific reporting requirements. This will significantly improve efficiency, enabling decision-makers to incorporate both qualitative and quantitative data seamlessly into their analysis, all while saving time and resources.

Chapter 3: Methodology

The study followed the CRISP-DM methodology, a widely adopted framework for machine learning and data mining projects. CRISP-DM comprises several essential phases organized in an iterative process.

1. **Business Understanding:** During this stage, objectives and business goals are defined to ensure alignment with desired outcomes.
2. **Data Understanding:** This phase involved collecting and examining the data utilized in the project, with the aim of obtaining a thorough understanding of the dataset.
3. **Data Preparation:** Activities like feature engineering and selection were undertaken to prepare the data for modeling.
4. **Modeling:** Diverse models are employed on the prepared data to construct predictive or descriptive models.
5. **Evaluation:** The fitted models undergo evaluation to gauge their accuracy and performance in alignment with the project's objectives.
6. **Deployment:** After obtaining a satisfactory model, it is deployed and integrated into the appropriate business processes to extract practical insights and facilitate decision-making.

3.1 Business Understanding

Manual approaches to survey analysis frequently encounter inefficiencies such as time-intensive data cleaning, inconsistent response categorization, and delays in insight generation. These limitations impede timely and informed decision-making. This project proposes to leverage Natural Language Processing (NLP) to automate the analysis process and ensure deeper insights from both structured and unstructured survey data. The overall objective as discussed in Chapter 1, Section 1.3 is to automate survey analysis and enhance efficiency, reduce manual intervention, and deliver timely, interpretable, and high-quality reports that inform strategic decisions. The system integrates NLP for semantic understanding, Large Language Models (LLMs) for intelligent annotation,

and Python for scripting, data processing, and report generation using the `python-pptx` library.

Overall, the business understanding stage provided a comprehensive understanding of the problem domain, the needs of stakeholders, and the objectives and requirements for the automation task.

3.2 Data Understanding

To build a robust and generalizable system, data was sourced from a diverse range of publicly available survey instruments. These included; The World Bank Open Survey Dataset , The General Social Survey (GSS), Pew Research Center datasets, World Values Survey (WVS), Kaggle repositories (e.g., IBM Employee Reviews). From these sources, a total of 5,785 survey questions were compiled for model training 3,272 open-ended and 2,513 closed-ended. The questions were selected across various domains such as health, education, customer service, and employee engagement to ensure diversity.

For classification, the survey questions were analyzed based on the text of the question itself, not the responses. This was important to enable pre-response classification and downstream routing into appropriate analysis workflows.

Additionally, a separate dataset containing 100,000 survey responses from the IBM Employee Reviews dataset was used specifically for topic modeling experiments. The variables in the Employee Review Survey are highlighted in Table 3.1.

Table 3.1: Description of Columns from Employee Reviews Dataset

Column Name	**Data Type**	**Description**
Job Title	String	The employee's role or job position within the company
Location	String	The geographical location where the review originated
Employment Status	String	Indicates whether the reviewer is a current or former employee
Review Title	String	A short title summarizing the employee's review
Pros	String (Open Text)	Positive aspects of working at IBM, as described by the employee
Cons	String (Open Text)	Negative aspects or challenges encountered while working at IBM
Advice to Management	String (Open Text)	Suggestions or recommendations provided by the employee to the leadership
Rating	Float	Overall rating score (typically on a scale from 1 to 5)
Work-Life Balance	Integer	Numeric rating specifically assessing work-life balance
Job Security	Integer	Numeric rating reflecting the employee's perception of job stability

3.3 Data Preparation

3.3.1 Data Cleaning

Initial preprocessing included removal of missing values (via deletion or imputation), elimination of duplicate entries, and standardization of date formats and text encodings. This ensured that the input to the modeling pipelines was uniform and error-free.

3.3.2 Text Cleaning and Normalization

Text preprocessing was conducted using the Natural Language Toolkit (NLTK) and consisted of the following steps:

a Lowercasing: Converted all text to lowercase to eliminate case-based distinctions (e.g., "Excellent" and "excellent" become equivalent).

b Stopword Removal: Removed common words (e.g., "the", "and", "is") that do not contribute significant meaning, thus focusing the analysis on informative terms.

c Stemming (PorterStemmer): Reduced words to their root forms (e.g., "running"

to "run") by stripping common suffixes. This helped reduce lexical diversity and dimensionality.

d Lemmatization (WordNetLemmatizer): Transformed words to their canonical form based on part-of-speech context (e.g., "better" to "good"). This improved linguistic accuracy over stemming alone.

e Removal of Special Characters and Punctuation: Stripped out non-alphabetic characters and punctuation to clean the textual input.

These preprocessing steps were crucial for eliminating noise, standardizing inputs, and improving the quality of text representations used for downstream modeling.

3.3.3 Tokenization and Embedding

Tokenization segmented the preprocessed text into individual units (typically words or subwords). This allowed models to interpret the text in manageable fragments for further analysis. Embedding techniques were then applied to convert these tokens into numerical vectors that preserve semantic meaning.

Three types of embeddings were explored:

a) TF-IDF (Term Frequency-Inverse Document Frequency): This traditional technique assigns a weight to each word based on how frequently it appears in a document relative to the entire corpus. While resulting in high-dimensional and sparse vectors, TF-IDF is interpretable and computationally efficient, making it suitable for baseline classification models.

b) Contextual Embeddings (BERT, Sentence-BERT): Unlike TF-IDF, contextual embeddings capture the semantic meaning of a word based on its surrounding words. BERT-based models generate dense, low-dimensional vectors where the representation of each token dynamically changes depending on its usage in the sentence. This made them especially effective for topic modeling in open-ended responses, where linguistic nuance and word order are important.

c) GloVe (Global Vectors for Word Representation): GloVe is a pre-trained embedding method that constructs word vectors based on global word co-occurrence statistics from a large corpus. Each word is represented as a fixed-dimensional dense vec-

tor, capturing semantic relationships (e.g., "king" - "man" + "woman" "queen"). In this study, GloVe embeddings were averaged at the document level to form input vectors, which were later clustered using KMeans for topic modeling. While GloVe embeddings are not contextual like BERT, they offer strong generalization capabilities and are computationally efficient.

These embedding strategies served as foundational inputs for both classification and topic modeling tasks, enabling the models to capture syntactic structure, semantic meaning, and contextual relationships necessary for accurate and interpretable analysis of survey responses.

3.4 Machine Learning Modeling: Survey Question Classification

To automate the differentiation of open-ended versus closed-ended survey questions, a supervised learning pipeline was designed. This pipeline involved transforming raw text into structured numerical vectors using Term Frequency–Inverse Document Frequency (TF-IDF) and applying various machine learning classifiers. The classifiers evaluated included Logistic Regression, Support Vector Machine (SVM), Naive Bayes (Multinomial), and Random Forest. These models are widely used in Natural Language Processing (NLP) and have proven effective for short-text classification tasks such as this.

3.4.1 Logistic Regression

Logistic Regression is a linear model used for binary classification tasks. It estimates the probability that a given input belongs to a particular class using the sigmoid activation function:

$$P(y = 1|x) = \frac{1}{1 + e^{-w^T x}}, \tag{1}$$

where:

- x is the TF-IDF feature vector of a survey question,
- w is the learned weight vector,
- $P(y = 1|x)$ gives the probability that the question is open-ended.

The model is trained by minimizing the binary cross-entropy loss and is particularly interpretable since each weight can be directly linked to a specific term. It performs well when the input features and class label exhibit a linear relationship in the log-odds space.

3.4.2 Support Vector Machine (SVM)

SVM is a margin-based classifier that seeks to find the hyperplane that best separates data points of different classes by maximizing the margin between the classes. The soft-margin optimization problem is given by:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{2}$$

where:

- x_i are the TF-IDF features,
- $y_i \in \{-1, 1\}$ are the class labels,
- ξ_i are slack variables for misclassified points,
- C is a regularization parameter controlling the trade-off between maximizing margin and minimizing misclassification.

SVM is especially effective for high-dimensional, sparse feature spaces such as those resulting from TF-IDF vectorization of short text, where it tends to outperform many alternatives due to its robustness to overfitting.

3.4.3 Naive Bayes (Multinomial)

The Multinomial Naive Bayes classifier is a probabilistic model particularly suited for word-frequency features such as TF-IDF. It applies Bayes' Theorem with a strong (naive) assumption of conditional independence among features:

$$P(y|x_1, x_2, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y), \tag{3}$$

where:

- $P(y)$ is the prior probability of the class,

- $P(x_i|y)$ is the likelihood of feature x_i given class y,
- The prediction is made by selecting the class with the highest posterior probability.

Despite the unrealistic independence assumption, Naive Bayes is computationally efficient and has demonstrated strong performance in text classification tasks, especially when the training data is limited.

3.4.4 Random Forest

Random Forest is an ensemble model consisting of a collection of decision trees, where each tree is trained on a random subset of the data using bagging (bootstrap aggregation). The final prediction is made via majority voting. Each tree split is based on impurity measures such as Gini Index:

$$Gini = 1 - \sum_{k=1}^{K} p_k^2, \tag{4}$$

where p_k is the proportion of samples of class k in a given node.

Random Forest is robust to overfitting, handles nonlinear feature interactions well, and is less sensitive to noise in the input. Although not optimized for sparse textual features, it was included in this study to evaluate its effectiveness compared to linear and probabilistic models.

3.4.5 Feature Representation and Model Input-Output Mechanism

Each survey question was treated as an individual text sample. After preprocessing, the question texts were transformed into feature vectors using TF-IDF vectorization. In this representation:

a Each feature corresponds to a unique term (word) from the vocabulary.

b The feature value indicates the importance of the term in the question relative to the corpus.

The resulting TF-IDF vector served as the input to the classification models. This input was typically a sparse vector of dimension, where is the total number of unique terms in

the training corpus. The output of the model was a binary label indicating the predicted type of the survey question:

a `0` for closed-ended questions (e.g., multiple choice, Likert scale).

b `1` for open-ended questions (e.g., those requiring free-text responses).

For probabilistic models such as Logistic Regression and Naive Bayes, the output also included a probability score associated with each class, enabling threshold-based decision-making. In contrast, SVM and Random Forest output the class label directly, though probabilities could be estimated using calibration techniques when required. This architecture enabled real-time classification of any new survey instrument by parsing each question, applying text transformations, and passing it through the trained SVM classifier to determine its type. Based on the predicted type, questions were routed into distinct processing pipelines for either statistical analysis (closed-ended) or NLP-based topic modeling (open-ended).

3.5 NLP: Topic Modeling of Open-Ended Responses

3.5.1 Objective

The goal was to identify latent themes in qualitative responses. A sample of 100,000 responses from the IBM Employee Reviews dataset (fields like Pros, Cons, Advice to Management) was used.

3.5.2 Topic Modelling Techniques

To effectively analyze open-ended survey responses, various models and techniques were used for text embedding and topic modeling. Each model has unique strengths in terms of capturing semantics, computational efficiency, and scalability. This section explores the most relevant models and provides a comparative overview to guide the selection of the optimal approach for embedding and topic modeling.

3.5.3 BERT and Variants

BERT (Bidirectional Encoder Representations from Transformers) has been widely adopted for its ability to produce deep contextual embeddings, capturing the subtle relationships

between words in a sentence. Developed by (Devlin, 2018), BERT uses a transformer-based architecture to model bidirectional contexts, which enables it to generate word representations that change based on surrounding words. This capability makes BERT particularly effective for understanding complex open-ended survey responses. The core of BERT's transformer model is the self-attention mechanism, which is mathematically expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

Where:

- Q (Query), K (Key), and V (Value) are matrices derived from the input embeddings through learned linear transformations.
- QK^T computes the similarity between the query and key vectors.
- $\sqrt{d_k}$ is a scaling factor where d_k is the dimensionality of the key vectors, used to prevent extremely large dot products that would push softmax into regions with small gradients.
- softmax$(\cdot)$ normalizes the similarity scores across all keys.
- The result is a weighted sum of the values (V), where higher similarity scores contribute more to the output.

However, its computational cost can be high, especially for large datasets. Variants of BERT, such as RoBERTa (Liu, 2019), enhance this contextual understanding by using additional training steps and dynamic masking, making it more effective in certain applications. The RoBERTa model utilizes the same masked language modeling (MLM) objective function as BERT, given by:

$$L_{\text{MLM}} = -\sum_{i=1}^{N} \log p(w_i|\text{context}(w_i)) \tag{6}$$

where:

- N is the total number of masked tokens in the input sequence.
- $p(w_i|\text{context}(w_i))$ is the probability that the model correctly predicts the masked word w_i based on its surrounding words (context).

- The negative log-likelihood penalizes incorrect predictions more heavily, encouraging the model to improve its contextual understanding.

This objective function measures the performance of the model in predicting the original masked word w_i given its surrounding context. By optimizing this loss function, RoBERTa learns to better model the relationship between words in context, which is vital for accurately interpreting natural language in open-ended survey responses.

Alternatively, DistilBERT (Sanh, 2019) offers a lighter version of BERT, sacrificing some depth of contextualization for faster processing times and reduced computational demands. The distillation process for DistilBERT can be described by the combined loss function:

$$L_{\text{distillation}} = \lambda L_{\text{teacher-student}} + (1 - \lambda) L_{\text{supervised}} \tag{7}$$

In this equation:

- $L_{\text{teacher-student}}$ is the loss that measures how well the student model (DistilBERT) mimics the soft predictions of the larger teacher model (BERT).
- $L_{\text{supervised}}$ is the traditional supervised loss based on ground truth labels.
- λ is a weighting coefficient that balances the two components.

The distillation loss $L_{\text{teacher-student}}$ is further defined as:

$$L_{\text{teacher-student}} = -\sum_{i=1}^{N} \text{softmax}\left(\frac{z^{(T)}}{T}\right) \log\left(\text{softmax}\left(\frac{z^{(S)}}{T}\right)\right) \tag{8}$$

where:

- $z^{(T)}$ and $z^{(S)}$ represent the logits (pre-softmax outputs) of the teacher and student models, respectively.
- T is a temperature parameter that smooths the probability distributions—higher T values result in softer probabilities.
- The softmax operation ensures that these logits are converted into comparable probability distributions.
- The cross-entropy loss compares the teacher's and student's output distributions to

guide the student toward emulating the teacher.

BERT and its Variants (RoBERTa and DistilBERT) begin with pre-processing open-ended responses through tokenization and normalization, followed by converting the text into embeddings that capture contextual relationships. These embeddings can then be used for various tasks, such as sentiment analysis, classification, or clustering. The output from this model may consist of predicted classes or clusters that reflect the underlying themes in the responses.

For topic modeling and grouping open-ended survey responses, BERTopic (Grootendorst, 2020) is a state-of-the-art method that combines BERT embedding with clustering techniques to create coherent topic clusters. It employs UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction, which can be represented as:

$$Y = \text{UMAP}(X) \tag{9}$$

where X is the high-dimensional input data and Y is the low-dimensional representation. For clustering, it utilizes HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). In **BERTopic** ,the input data is processed by first generating BERT embeddings and then utilizing UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction. Following this, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is applied to cluster the reduced embeddings into distinct topics. The output consists of clusters that represent distinct themes in the survey responses, with each cluster encompassing semantically similar responses.

3.5.4 GloVe with RNN or LSTM Models

Another approach to text embedding is using GloVe (Global Vectors for Word Representation) embedding in combination with recurrent neural networks (RNNs) or long short-term memory (LSTM) networks (Pennington et al., 2014). GloVe can be mathematically represented as :

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{10}$$

where:

- J is the cost function to minimize,
- V is the vocabulary size,
- X_{ij} is the number of times word J appears in the context of word i,
- $\mathbf{w}_i$ and $\mathbf{w}_j$ are the word vectors,
- $\mathbf{b}_i$ and $\mathbf{b}_j$ are the biases associated with the words.

When combined with RNNs or LSTMs, these embeddings effectively capture sequential dependencies in text, striking a balance between contextual understanding and computational efficiency. The hidden state h_t of an RNN at time step t is defined as:

$$f(W_h h_{t-1} + W_x x_t + b) \tag{11}$$

where:

- h_t: Hidden state at time t
- h_{t-1}: Hidden state at time $t-1$
- x_t: Input at time t
- W_h: Weight matrix for the hidden state
- W_x: Weight matrix for the input
- b: Bias vector
- f: Activation function (e.g., tanh or ReLU)

For LSTMs, the equations governing the cell state c_t and hidden state h_t are more complex, including gates that control the flow of information:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$$

$$h_t = o_t \cdot \tanh(c_t)$$

Where:

- i_t, f_t, and o_t are the input, forget, and output gates, respectively,
- $\tilde{c}_t$ is the candidate cell state,
- σ is the sigmoid activation function,
- W_i, W_f, W_o are weight matrices,
- b_i, b_f, b_o are bias vectors.

This approach is particularly useful for tasks that require capturing word order while keeping computational costs manageable, making it a viable alternative for analyzing open-ended survey data without the need for transformer-based models.

Another widely used method for encoding text in survey analysis is the combination of Global Vectors for Word Representation (GloVe) with sequential models like Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs). GloVe embeddings, trained on aggregated global word-word co-occurrence statistics from a large corpus, provide a dense vector representation that captures semantic similarity between words. These embeddings are then passed into RNN or LSTM layers that process the input sequentially, retaining temporal dependencies and word order context. This pipeline enables the model to learn from both global semantic features and sequential relationships, making it well-suited for downstream tasks such as topic detection or sentiment classification based on open-ended survey responses. However, unlike transformers, these models may struggle with capturing long-range dependencies and require more careful tuning to avoid issues like vanishing gradients.

3.5.5 Latent Dirichlet Allocation(LDA)

It was developed by (Blei et al., 2003), as a probabilistic model that assumes each document d is a mixture of a limited number of topics, defined as:

$$p(w|d) = \sum_{k=1}^{K} p(w|k) \cdot p(k|d) \tag{12}$$

where:

- w is the word,
- d is the document,
- K is the number of topics,
- $p(w|k)$ is the probability of word w given topic k,
- $p(k|d)$ is the probability of topic k given document d.

Its analysis involves creating a document-term matrix from the pre-processed text data. LDA analyzes this matrix and generates topic distributions for each document, along with word distributions for each topic, enabling clear insights into the main themes present in the responses.

3.5.6 Non-Negative Matrix Factorization (NMF)

NMF provides an alternative by decomposing the text data matrix V into non-negative factors W and H:

$$V \approx W \cdot H \tag{13}$$

where:

- V is the original data matrix,
- W is the matrix of topic-word distributions,
- H is the matrix of document-topic distributions.

Non-Negative Matrix Factorization (NMF) also starts with a document-term matrix, which it decomposes into two non-negative matrices. The output includes interpretable

topic-word distributions and document-topic distributions, offering insights into the underlying themes in the data.

3.6 Model Evaluation: (Machine Learning Model Survey Question Classification)

To assess the effectiveness of the model in classifying survey questions as either open-ended or close-ended, several evaluation metrics were be employed. These metrics helped determine how well the model generalized to new data and whether it met the expected performance standards. The primary metrics used were **accuracy**, **precision**, **recall**, **F1-score**, and **confusion matrix**.

3.6.1 Accuracy

Accuracy refers to the ratio of correctly predicted samples to the total number of predictions made by the model. It offers a simple and widely used measure of overall model performance. However, its reliability can be compromised in cases where class distributions are highly imbalanced, as the metric may be skewed by the dominant class. In contrast, when applied to balanced datasets, accuracy remains a valid and effective indicator of classification performance (**?**).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Where:

- TP: True Positives (correctly classified open-ended questions)
- TN: True Negatives (correctly classified close-ended questions)
- FP: False Positives (close-ended questions incorrectly classified as open-ended)
- FN: False Negatives (open-ended questions incorrectly classified as close-ended)

A high accuracy score indicates that the model is correctly classifying a majority of the questions. However, accuracy alone may not provide a complete picture if the dataset is imbalanced (e.g., if there are significantly more close-ended than open-ended questions).

3.6.2 Precision

Precision measures the proportion of true positive predictions (correctly classified open-ended questions) out of all positive predictions made by the model (Obi, 2023).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

where TP represents True Positives and FN represents False Negatives.

High precision indicates that when the model predicts a question as open-ended, it is usually correct. Precision is particularly useful in cases where false positives (misclassifying a close-ended question as open-ended) are costly.

3.6.3 Recall

Recall (also known as sensitivity or true positive rate) measures the proportion of actual positive instances (open-ended questions) that were correctly classified by the model. High recall is desirable when the cost of missing positive instances (false negatives) is high (Obi, 2023).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

High recall means the model is able to correctly identify most of the open-ended questions, minimizing the number of false negatives.

3.6.4 F1-score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It serves as a comprehensive metric, especially valuable in datasets where class imbalances exist, as it accounts for both Type I and Type II errors (Obi, 2023).

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

The F1-score is particularly useful when there is an uneven class distribution (i.e., when

one class is more prevalent than the other). A higher F1-score indicates a better balance between precision and recall.

3.6.5 Confusion matrix

The confusion matrix provides a detailed breakdown of the model's classification performance by showing the true positives, true negatives, false positives, and false negatives.

By examining the confusion matrix, we can assess whether the model is more prone to false positives or false negatives, which will guide us in tuning the model to achieve better results.

A sample confusion matrix for this classification problem would look like this:

	Predicted Open-Ended	**Predicted Close-Ended**
Actual Open-Ended	True Positives (TP)	False Negatives (FN)
Actual Close-Ended	False Positives (FP)	True Negatives (TN)

Table 3.2: Confusion Matrix for Classification Model

3.6.6 Model Optimization

By utilizing these metrics, we were able to gauge the overall performance of the classification models and fine-tuned the hyperparameters to improve their efficacy. All models were evaluated using 5-fold cross-validation. SVM emerged as the best performer, achieving an accuracy of **0.9851** as discussed in Chapter 6. The key hyperparameter tuned was the regularization strength, with values tested in 0.1, 1, 10. The optimal value was c = 1, which balanced underfitting and overfitting. Regularization was particularly critical given the high-dimensional TF-IDF feature space, as it helped constrain model complexity and enhance generalizability. Based on its margin-maximizing nature, robust performance in sparse domains, and superior cross-validation accuracy, SVM was selected as the final classifier for integration into the automated pipeline.

3.7 Model Evaluation : NLP (Topic Modeling of Open-Ended Responses)

The selection of text embedding and topic modeling techniques was based on evaluating their performance in terms of coherence, contextual accuracy, computational efficiency, and scalability. By considering different approaches, the goal was to find a model that best captures the thematic structure of open-ended survey responses while balancing the

trade-offs between processing time, accuracy, and computational resource requirements. This comparative evaluation guided the decision-making process of ensuring the most suitable and efficient models were employed for high-quality survey analysis. The key metrics that were used to measure the performance of the model include **perplexity**, **BLEU score**, **cosine similarity**, and **topic coherence**.

3.7.1 Perplexity

Perplexity is a measure of how well a language model predicts a sample. It calculates the likelihood of the model predicting the next word given the previous words. It can be represented by the formula below:

$$\text{Perplexity} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i|w_1,w_2,...,w_{i-1})} \tag{18}$$

Where $P(w_i)$ is the probability of the word at position i, and N is the total number of words.

A lower perplexity score indicates that the model is better at predicting word sequences, meaning that it has a stronger understanding of the language. While perplexity is not commonly used for BERT (as BERT is not a generative model), it can be helpful in evaluating certain aspects of its language understanding capabilities.

3.7.2 Bilingual Evaluation Understudy (BLEU)

The BLEU score is a metric that compares the output of the model with a reference, typically used for evaluating machine translation but also applicable for other text generation tasks.

It can be represented by the formula below:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{19}$$

Where BP is the brevity penalty, and p_n represents the precision of n-grams.

A higher BLEU score indicates that the model's output is closer to human-like responses, especially when analyzing text for coherence and relevancy.

3.7.3 Cosine Similarity

Cosine similarity measures the similarity between two vectors, typically used to assess how similar BERT embeddings of two pieces of text are. It calculates the cosine of the angle between the two vectors in a high-dimensional space.

It can be represented by the formula below:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{20}$$

where $\|A\|$ and $\|B\|$ represent the Euclidean norms (magnitudes) of vectors A and B, respectively, and $A \cdot B$ denotes the dot product between the two vectors.

A cosine similarity score close to 1 means the two texts have a high degree of similarity, while a score close to 0 indicates little to no similarity. This is particularly useful when comparing semantic embeddings generated by BERT for various survey responses.

3.7.4 Topic Coherence

Topic coherence measures the quality of topics generated in unsupervised topic modeling, such as those generated by BERTopic using BERT embeddings. It evaluates how well the words within a topic are semantically related.

$$\text{Coherence} = \frac{1}{\binom{|W|}{2}} \sum_{1 \le i < j \le |W|} \text{Similarity}(w_i, w_j) \tag{21}$$

where:

- W is the set of top words in a given topic,
- N is the total number of unique word pairs (i, j) in W,
- $\text{Similarity}(w_i, w_j)$ denotes a semantic similarity function (e.g., based on word embeddings or co-occurrence statistics) between words w_i and w_j.

A higher topic coherence score indicates that the topics generated are more interpretable and meaningful to human evaluators. This metric is crucial for determining how well BERT captures the contextual meaning of text in the grouping and classification of

open-ended responses.

By applying these metrics, we were able to evaluate the effectiveness of the models in generating meaningful representations of text data as discussed in Chapter 5. These evaluations guided further fine-tuning of the model to optimize its performance in tasks such as comment grouping, topic modeling, and generating accurate insights from open-ended responses.

3.8 Libraries and Tools Used

Several specialized libraries and tools were employed to support different components of the automated survey analysis system. The scikit-learn library was used for implementing the machine learning classification pipeline, including TF-IDF vectorization, classifier models such as SVM and Random Forest, cross-validation, and hyperparameter tuning through grid search.

For preprocessing and linguistic normalization of text data, the NLTK (Natural Language Toolkit) library was utilized. It provided modules for tokenization, stopword removal, stemming, and lemmatization, which were essential for preparing the survey responses for downstream modeling.

In the domain of topic modeling, several advanced libraries were integrated.BERTopic was central to the project, combining sentence embeddings from SentenceTransformers, dimensionality reduction using UMAP, and clustering using HDBSCAN. These tools collectively enabled high-quality, context-aware topic modeling. The python-pptx library facilitated automated report generation, enabling the creation of PowerPoint slides that summarized both open- and closed-ended question analysis.

To enrich visualizations with descriptive narrative insights, the OpenAI API was used to query the GPT-3.5 model. This allowed the system to generate coherent summaries, bullet points, and strategic recommendations based on the chart data. Finally,Streamlit served as the front-end framework, offering an intuitive web-based interface where users could upload survey data, run the full analysis pipeline, and download the automatically generated report.

Based on these evaluations and tool integrations, the system provided a robust, end-to-

end solution for survey analysis, leveraging both classical machine learning and cutting-edge NLP tools.

3.9 Deployment

The survey analysis application was deployed as a web-based tool, enabling users to seamlessly upload their survey data in formats such as CSV or Excel. Once the data was uploaded, users could specify the type of survey, data collection dates, and target populations—parameters essential for tailoring the analysis. By utilizing a web-based interface, we ensured a responsive and engaging user experience, allowing stakeholders to extract meaningful insights efficiently and enhance decision-making processes based on the survey data.

3.10 Expected Outcomes

The expected outcomes of this project were multifaceted, aimed at addressing the limitations of traditional survey tools while delivering a comprehensive, automated solution for survey data analysis.

1. **Improved Insights from Qualitative Data:** By leveraging NLP models, the project enabled more effective analysis of open-ended responses, offering deeper insights into the sentiments and themes captured in qualitative feedback.
2. **Customized and Annotated Reports:** The application generated reports tailored to organizational needs and reporting standards, complete with annotated graphs and visualizations that reflected the key patterns and insights from the survey data. This enhanced decision-making by ensuring that both quantitative and qualitative data were interpreted meaningfully.
3. **Automated Survey Analysis:** The application streamlined the analysis process by automatically classifying and processing both closed-ended and open-ended survey questions, significantly reducing the need for manual effort.

Chapter 4: System Design and Architecture

4.1 System Overview

The automated survey analysis system developed in this project is a modular, end-to-end framework designed to transform raw survey data into structured, interpretable insights. Its core functionality encompasses preprocessing, classification of survey questions, natural language processing (NLP) for open-ended responses, insight generation using GPT models, and automated report generation. The architecture was built for scalability, interpretability, and usability by both technical and non-technical users. The system was developed using Python and deployed through a Streamlit web application for easy accessibility.

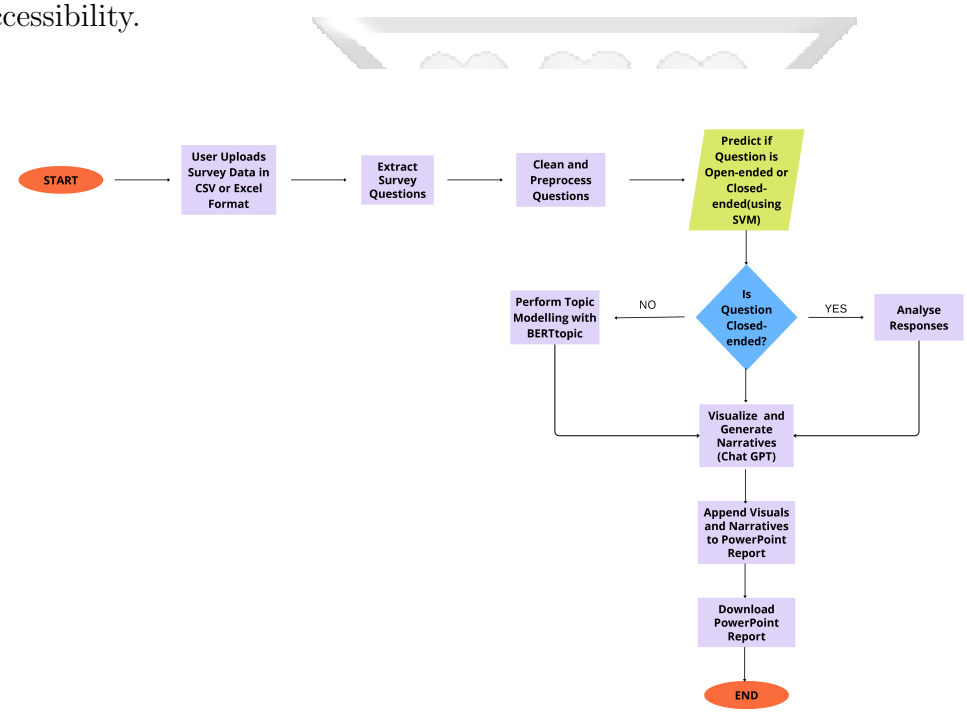


Figure 4.1: System Architecture

4.2 System Modeling Framework

4.2.1 Functional Modeling

The system can be decomposed into several functional modules:

1. **Ingestion Module:** Handles uploading of survey data in various formats (e.g., Excel, CSV).

2. **Classification Module:** Automatically classifies each question as open-ended or closed-ended using a trained SVM model.
3. **Preprocessing Module:** Cleans and normalizes textual data for downstream processing.
4. **Topic Modeling Module:** Applies NLP models such as BERTopic to cluster and label open-ended responses.
5. **Insight Generation Module:** Uses GPT-3.5 to summarize results into bullet points and narratives.
6. **Reporting Module:** Generates PowerPoint slides that compile visualizations and insights.

4.2.2 Data Flow Modeling

The data flows through the following stages:

- **Input:** User uploads raw survey data via the Streamlit interface.
- **Preprocessing:** Text is cleaned, tokenized, and embedded using models like BERT or TF-IDF.
- **Classification:** The system uses the SVM model to label questions as open- or closed-ended.
- **Branching:** Closed-ended questions are routed to descriptive analytics; open-ended responses go to NLP pipelines.
- **Topic Modeling:** Embeddings are reduced (UMAP) and clustered (HDBSCAN), then labeled with c-TFIDF.
- **Narration:** GPT-3.5 summarizes both open- and closed-ended results.
- **Output:** Charts and narratives are compiled into a downloadable PowerPoint report.

4.3 System Components

4.3.1 Preprocessing

All uploaded data undergoes preprocessing. For structured data, missing values are handled, and data types are validated. For unstructured text:

- Lowercasing ensures uniformity.
- Stopword removal eliminates common, semantically empty words.
- Punctuation and special characters are removed to reduce noise.
- Stemming and lemmatization consolidate variations of the same root word.

4.3.2 Analysis

Closed-ended responses are analyzed using value counts and percentage distributions. Open-ended responses are embedded with Sentence-BERT, clustered using UMAP and HDBSCAN, and labeled using class-based TF-IDF. The result is a list of coherent, semantically meaningful topics.

4.3.3 Insight Generation

For each open-ended question, the most frequent topics are passed to GPT-3.5 to generate bullet-point summaries. For closed-ended questions, the system summarizes the most selected response options and provides strategic recommendations based on distribution patterns.

An example prompt used in the annotation process:

```
You are a data analyst writing insights for a survey report. Below is the survey ques
Question: Why do you feel disengaged at work?
Topic Distribution: ‘lack of recognition‘: 42, ‘poor communication‘: 35,
‘limited growth‘: 30, ‘workload‘: 28
Write 3-5 bullet points summarizing the implications of this distribution.
```

4.3.4 Visualization and Reporting

Charts are generated in Python using `python-pptx`, and insights are added as textboxes. The final output is a PowerPoint file with:

- Slide per question
- Visual summary (bar chart or pie chart)
- AI-generated summary of key findings

4.4 User Interface

The front-end was implemented in Streamlit. Users interact with the system through:

- **Data Upload Panel:** Allows selection of survey files.
- **Metadata Entry:** Users input survey title, type, and data collection date.

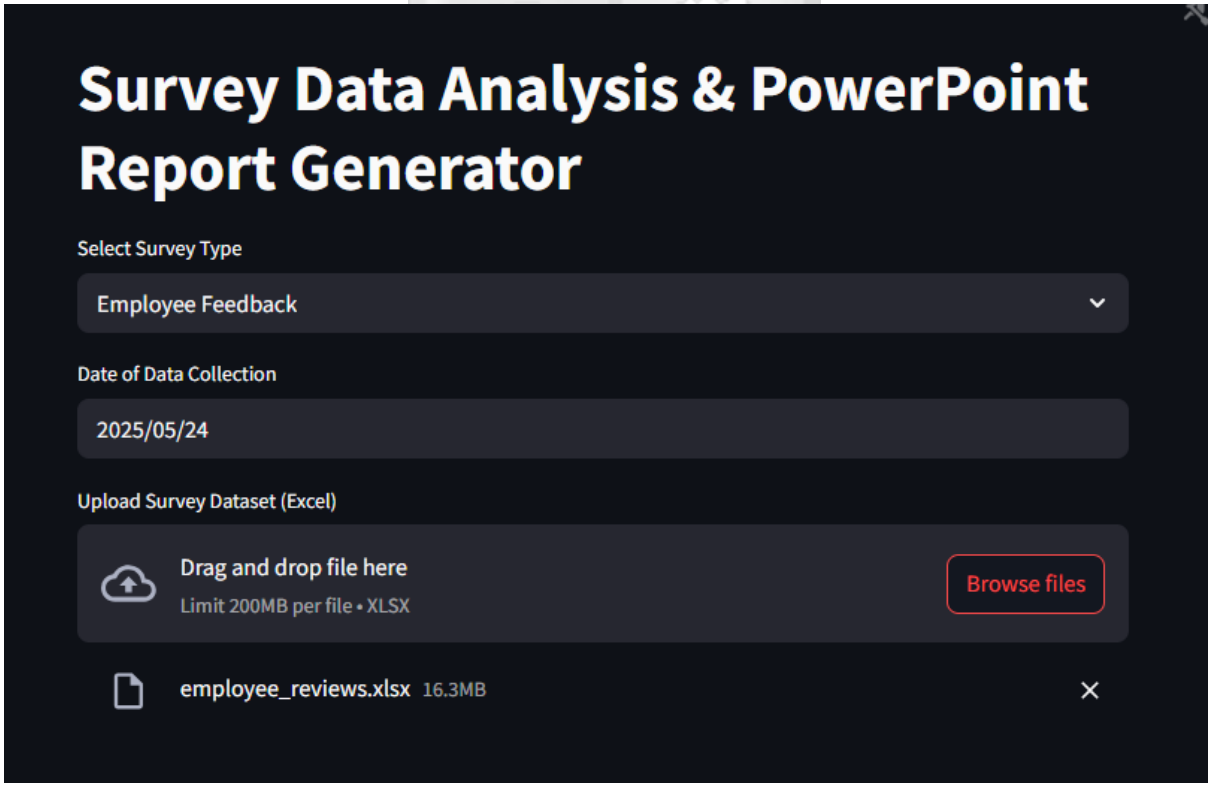


Figure 4.2: Data Upload and Metadata Entry

- **Classification Review:** Users can override the system's classification and correct mislabels.

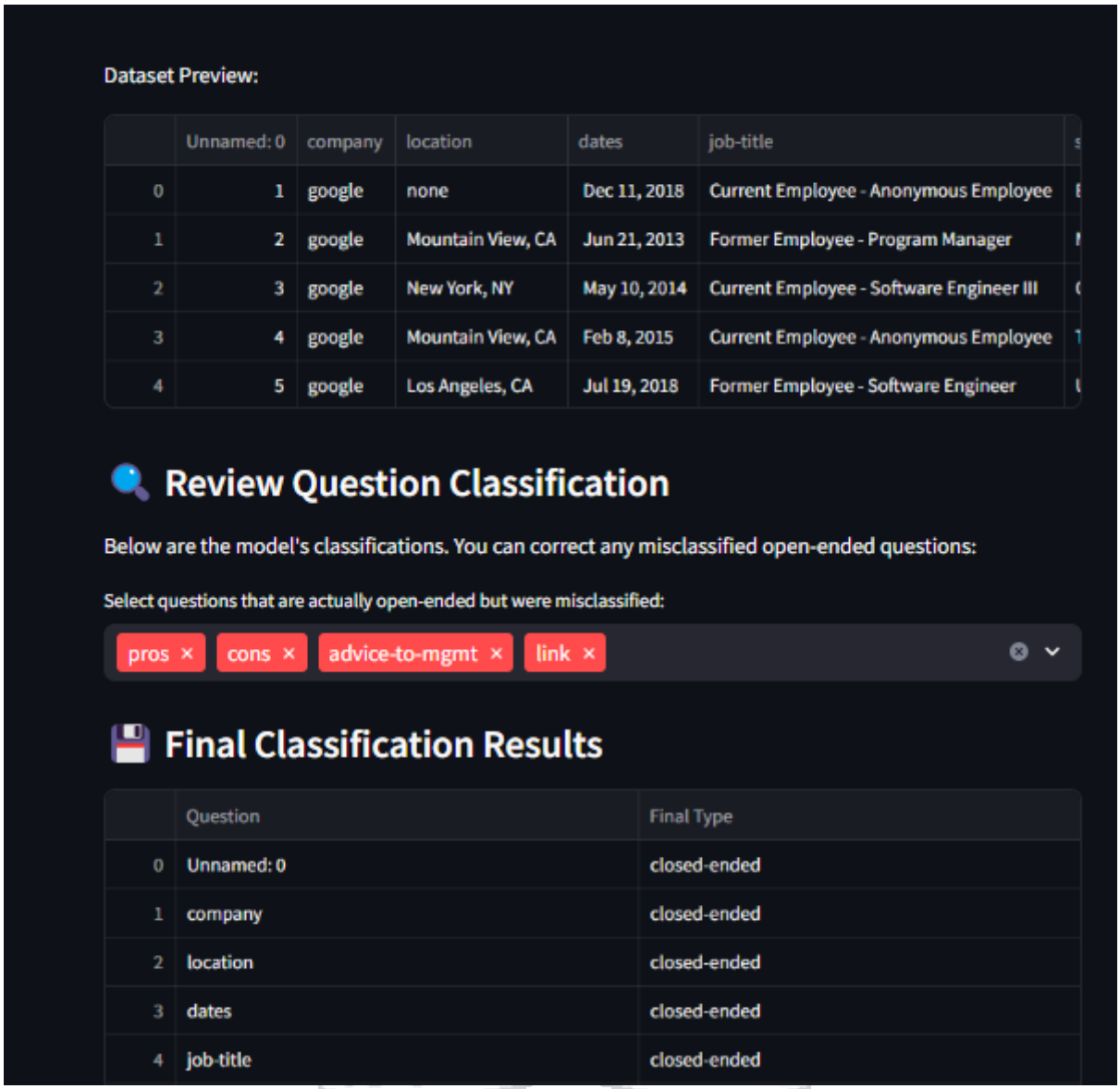


Figure 4.3: Data Preview and Classification Review

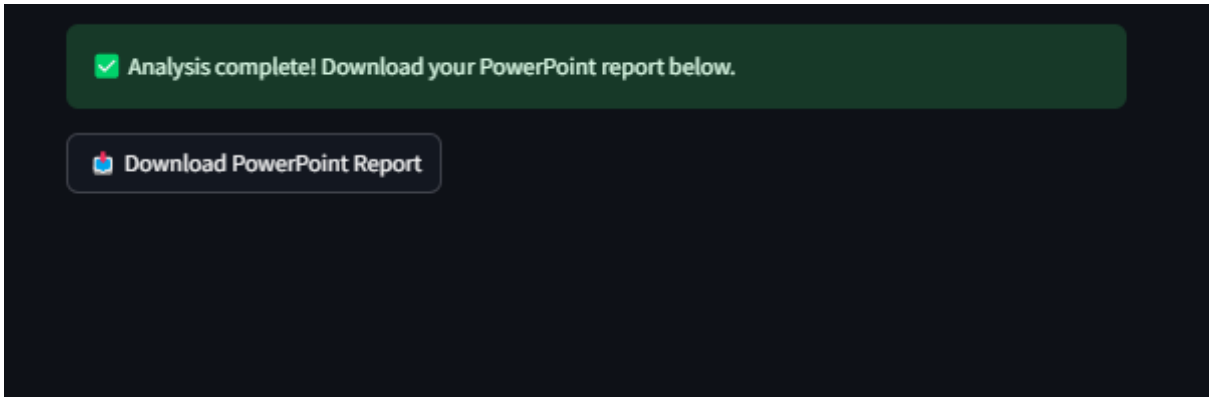


Figure 4.4: Download Option

- **Download Section:** Provides access to the generated PowerPoint report.

The interface also logs the number of open and closed-ended questions, shows topic model summaries, and displays a preview of generated slides.

4.5 Backend and Deployment

The backend was implemented entirely in Python. Models were trained and serialized using `joblib`. Key frameworks used include:

- **Scikit-learn:** For classification, pipelines, and hyperparameter tuning.
- **BERTopic:** For unsupervised topic modeling.
- **python-pptx:** For slide automation.
- **OpenAI API:** For GPT-3.5 text generation.

The application is served via Streamlit Cloud, enabling real-time interaction without requiring users to install local dependencies. This setup facilitates fast deployment and scalability.

4.6 Conclusion

The system's architecture reflects a balance between modular design, performance, and usability. Each component supports automation at a specific phase of the pipeline—from ingestion to insight delivery. By integrating machine learning, NLP, and visual storytelling into a cohesive platform, the system delivers a scalable and intelligent framework for modern survey analytics.

Chapter 5: Discussion of Results

This chapter presents the empirical findings resulting from the implementation of the automated survey analysis system developed in alignment with the CRISP-DM methodology. It provides a comprehensive evaluation of the system's performance across key components, including classification accuracy, topic modeling effectiveness, and automated insight generation. The discussion highlights comparative results from multiple machine learning models, assesses the quality of discovered topics in open-ended responses, and examines the interpretability and operational viability of the proposed solution. Emphasis is placed on interpreting diagnostic metrics, analyzing emergent themes in qualitative data, and reflecting on the system's effectiveness in meeting the research objectives.

5.1 Model Evaluation and Selection Rationale

5.1.1 Classification of Survey Questions

The classification task focused on automatically distinguishing between open-ended and closed-ended questions based solely on their phrasing. Four models were trained using TF-IDF vectorized text features: Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM). The models were evaluated using 5-fold cross-validation, with metrics including accuracy, precision, recall, and F1-score to ensure a balanced performance assessment. Table 5.1 presents the average cross-validation scores for each model.

Table 5.1: Performance Metrics of Classification Models (5-Fold Cross-Validation)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.9799	0.9781	0.9806	0.9793
Naive Bayes	0.9769	0.9743	0.9755	0.9749
Random Forest	0.9790	0.9776	0.9789	0.9782
SVM	**0.9851**	**0.9840**	**0.9850**	**0.9845**

As shown in Table 5.1, all models performed well with accuracy exceeding 97%. However, the Support Vector Machine (SVM) classifier outperformed the others across all metrics. Its margin-maximizing nature, robustness to high-dimensional sparse data, and consistent performance across folds made it the most suitable model for deployment in the automated classification pipeline.

The classifier was integrated into the Streamlit-based application. When a user uploads a survey, each question is automatically classified. This routing dictates whether NLP or descriptive statistics should be applied. This automated gating is critical in mixed-format survey analysis, ensuring appropriate downstream processing.

5.1.2 Topic Modeling of Open-Ended Responses

To extract latent themes from open-ended survey responses, five topic modeling techniques were applied to a corpus of 100,000 qualitative entries. These included classical frequency-based models (TF-IDF + KMeans, LDA, NMF), semantic embedding-based methods (GloVe + KMeans), and transformer-driven techniques (BERTopic). Each model was evaluated using four performance metrics: topic coherence, average cosine similarity between topic keywords, BLEU score for summary alignment, and perplexity (for LDA only). The results are shown in Table 5.2.

Table 5.2: Evaluation Metrics for Topic Modeling Techniques

Model	**Coherence**	**Cosine Similarity**	**BLEU Score**	**Perplexity**
TF-IDF + KMeans	0.5426	0.5491	0.0042	—
LDA	0.5449	0.5522	0.0084	-8.1279
NMF	0.6028	0.5779	0.0042	—
BERTopic	**0.6148**	0.5262	0.0053	—
GloVe + KMeans	0.4731	**0.6215**	0.0042	—

Topic Coherence evaluates how semantically related the top words in a topic are. BERTopic achieved the highest coherence score (0.6148), indicating that its generated topics were the most interpretable to human analysts. NMF followed closely (0.6028), suggesting that matrix factorization methods can produce well-formed topics when properly tuned. In contrast, GloVe-based clustering recorded the lowest coherence, likely due to averaging embeddings at the document level, which may dilute thematic signals.

Cosine Similarity measured intra-topic similarity based on GloVe embeddings. Interestingly, GloVe + KMeans achieved the highest score (0.6215), suggesting that while the topics it created were not as interpretable (low coherence), the keywords within each topic were semantically compact. This highlights a trade-off between tight clustering and human-readability.

BLEU Score served as a proxy for evaluating how closely model-generated topic labels

aligned with human-curated reference summaries. LDA had the highest BLEU score (0.0084), suggesting its probabilistic nature helps align with real-world thematic labels, although all models yielded relatively low BLEU values due to the abstract nature of the task.

Perplexity, specific to LDA, assesses the model's likelihood of predicting unseen data. LDA's perplexity score of -8.1279 indicates moderate generalizability, but lower (more negative) values in log-space often reflect better performance. However, perplexity alone is insufficient for qualitative tasks like topic coherence.

Rationale for Choosing BERTopic. Despite GloVe + KMeans showing high cosine similarity, its poor coherence rendered its topics less interpretable for non-technical audiences. Similarly, while NMF and LDA offered interpretable topics, they depend on bag-of-words assumptions, which ignore word order and context—limitations that impact nuanced understanding in open-ended responses.

BERTopic's transformer-based Sentence-BERT embeddings allowed it to retain deep contextual signals from survey responses, outperforming other models in both coherence and BLEU score. Additionally, its use of UMAP for dimensionality reduction and HDBSCAN for density-based clustering offered adaptive topic separation without requiring a fixed number of clusters—a major advantage over KMeans.

Therefore, BERTopic was selected not due to architectural novelty, but because of its superior empirical performance, higher coherence, and ability to produce thematically distinct and contextually grounded topics from large-scale open-ended survey data. Its design aligns well with the practical goals of this study: delivering fast, interpretable, and accurate topic summaries that can be visualized and presented without manual intervention.

5.1.3 Insights Enabled by Automation

The integrated pipeline automated:

- Identification of dominant qualitative themes
- Conversion of unstructured responses into structured insight
- Export of annotated PowerPoint slides summarizing open-ended content

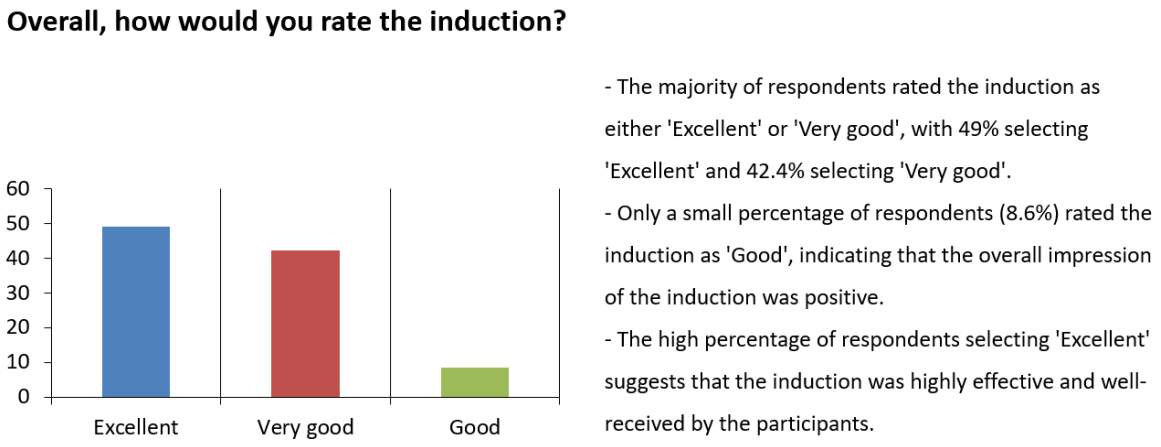


Figure 5.1: Example of closed ended question

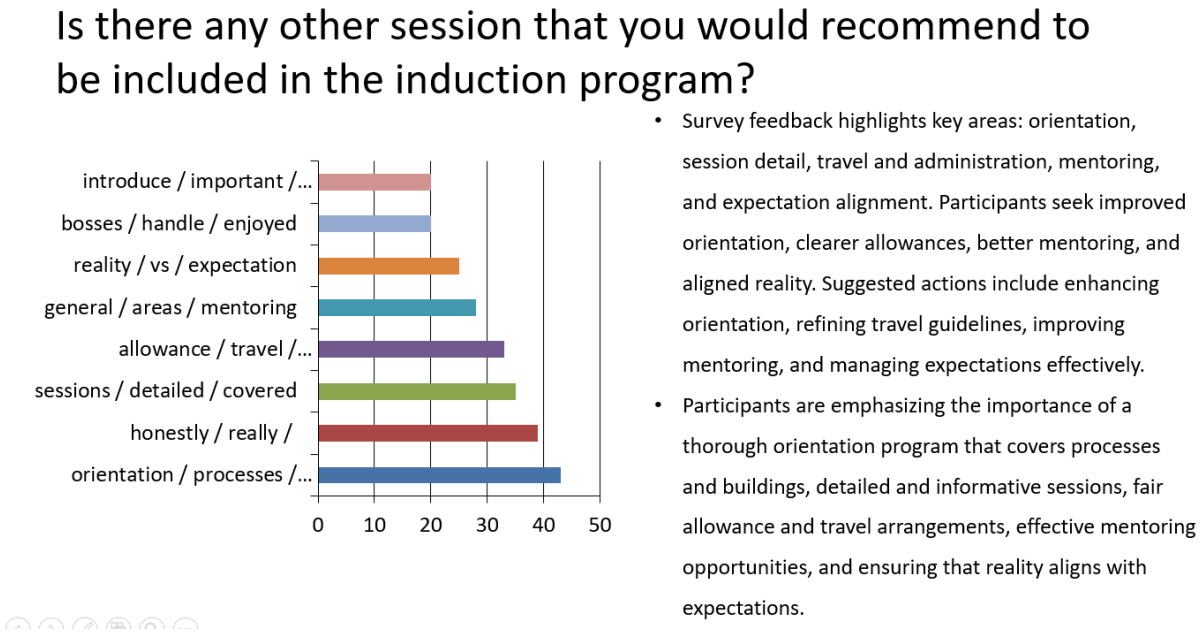


Figure 5.2: Example of open ended question

This automation proved especially valuable for high-volume surveys where manual review is impractical. The ability to surface emerging themes and employee concerns in seconds provides a significant advantage in responsiveness.

5.2 Model Strengths and Comparative Justification

SVM was preferred for classification due to its superior cross-validation results. Its ability to generalize in sparse, high-dimensional TF-IDF spaces made it ideal for this binary text classification problem. Although Logistic Regression and Random Forest were considered, their slightly lower accuracy and sensitivity to feature correlations were limiting.

BERTopic emerged as the top-performing topic model based on coherence and interpretability metrics. Unlike traditional models that rely solely on word frequency, BERTopic leverages contextual embeddings to model semantic relationships. Its architecture also avoids the rigid topic allocation of LDA by clustering based on semantic density in embedding space.

5.3 Conclusion

The results demonstrate that the combined use of SVM for question classification and BERTopic for topic modeling offers a scalable, interpretable, and high-performing solution for automated survey analysis. The deployment-ready architecture ensures not only analytical rigor but also usability for decision-makers.

Chapter 6: Conclusions, Recommendations and Future Work

6.1 Conclusion

This research set out to automate the analysis of survey data by addressing key challenges outlined in the project objectives: timely insight generation, improved data quality, and reduced manual effort. The automated system developed in this study successfully met these goals by implementing machine learning and natural language processing (NLP) techniques for classifying survey questions and analyzing open-ended responses.

In line with the first objective, a Support Vector Machine (SVM) model combined with TF-IDF features was used to classify survey questions as open-ended or closed-ended. The model achieved a high cross-validation accuracy of 98.51%, enabling precise differentiation of questions and facilitating specialized processing for each type.

Addressing the second objective, this study deployed and compared several topic modeling methods on open-ended responses. BERTopic emerged as the most suitable model, achieving the highest topic coherence and producing semantically rich clusters. This confirmed the third objective, which was to enhance the interpretability and depth of insights derived from qualitative data.

Finally, the integration of the models into a Streamlit application, alongside automated PowerPoint report generation, addressed the fourth objective: enabling end-to-end survey analysis with minimal human intervention. The resulting system not only streamlined the analysis workflow but also empowered decision-makers with accurate and timely insights.

6.2 Recommendations

Based on the successful implementation and evaluation of the system, the following recommendations are proposed for organizations and researchers seeking to improve their survey analysis processes:

a Implement automated question classification: Employ machine learning models to distinguish open-ended from closed-ended questions early in the pipeline, optimizing downstream processing.

b Use BERTopic for topic discovery: Given its contextual depth and high coherence

scores, BERTopic is recommended for organizations processing large volumes of open-ended text.

c Deploy integrated user interfaces: Tools like Streamlit provide interactive, visual feedback to users and allow real-time report generation for faster decision-making.

d Standardize preprocessing pipelines: Consistent use of techniques such as stop-word removal, lemmatization, and stemming contributes significantly to model accuracy and should be institutionalized.

6.3 Future Work

To enhance the platform's robustness and extend its applicability, several future directions are identified:

a Multilingual capabilities: Incorporate automatic language detection and translation modules to support diverse linguistic inputs.

b Domain-specific fine-tuning: Fine-tune transformer-based models (e.g., BERT or RoBERTa) on domain-specific corpora, such as healthcare or education survey responses, to improve relevance and model performance.

c Interactive topic refinement: Introduce features allowing users to merge, split, or rename topics post-modeling, making the insights more actionable.

d Survey platform integration: Develop APIs for seamless integration with tools like Google Forms, Typeform, or SurveyMonkey, automating data ingestion.

e Bias and fairness auditing: Incorporate fairness-aware modeling practices and evaluate the models for bias to ensure inclusive and ethical survey analysis.

f Distributed processing for scalability: Employ frameworks such as Apache Spark or Dask to handle larger datasets in real time without compromising responsiveness.

By incorporating these enhancements, the system could evolve into a full-fledged survey intelligence platform suitable for real-time, multilingual, and domain-specific data analytics.

Bibliography

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1):24.

Aldoseri, A., Al-Khalifa, K. N., and Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *Applied Sciences*, 13(12):7082.

Allen, M. (2017). *The SAGE encyclopedia of communication research methods*. SAGE publications.

Alwan, A. A., Ciupala, M. A., Brimicombe, A. J., Ghorashi, S. A., Baravalle, A., and Falcarin, P. (2022). Data quality challenges in large-scale cyber-physical systems: A systematic review. *Information Systems*, 105:101951.

Baburajan, V. (2021). *Automated Text Analysis on Open-ended Response Surveys: Measuring Attitudes Regarding Autonomous Vehicles*. PhD thesis, INSTITUTO SUPERIOR TÉCNICO.

Balayn, A., Lofi, C., and Houben, G.-J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768.

Bilal, M., Ali, G., Iqbal, M. W., Anwar, M., Malik, M. S. A., and Kadir, R. A. (2022). Auto-prep: efficient and automated data preprocessing pipeline. *IEEE Access*, 10:107764–107784.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bonorino, A. G. (2023). Smart surveys: An automatic survey generation and analysis tool. *SciTePress*.

Card, D. and Smith, N. A. (2015). Automated coding of open-ended survey responses. *draft paper*.

Chen, P., Wu, L., and Wang, L. (2023). Ai fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences*, 13(18):10258.

Cupoli, P., Earley, S., and Henderson, D. (2014). Dama-dmbok2 framework. *Dama International*.

Davenport, T. H., Harris, J. G., and Morison, R. (2010). *Analytics at work: Smarter decisions, better results*. Harvard Business Press.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, T. et al. (2023). Questgenius: An intelligent questionnaire generation and result analyzing system for investigation and research using artificial intelligence and nature language. In *CS & IT Conference Proceedings*, volume 13. CS & IT Conference Proceedings.

Dubuc, T., Stahl, F., and Roesch, E. B. (2020). Mapping the big data landscape: technologies, platforms and paradigms for real-time analytics of data streams. *IEEE Access*, 9:15351–15374.

Esmaeilzadeh, S., Williams, B., Shamsi, D., and Vikingstad, O. (2022). Providing insights for open-response surveys via end-to-end context-aware clustering. In *International Conference on Artificial Intelligence in Education*, pages 526–532. Springer.

Geekiyanage, S. C., Tunkiel, A., and Sui, D. (2021). Drilling data quality improvement and information extraction with case studies. *journal of petroleum Exploration and Production*, 11:819–837.

Gerdon, F. (2024). Challenges of data-driven technologies for social inequality and privacy: empirical research on context and public perceptions. *University of Mannheim*.

Gitonga, P. K. (2016). *Customer relationship management practices and performance of safaricom Limited in Kenya*. PhD thesis, University Of Nairobi.

Greasley, A. (2019). *Simulating business processes for descriptive, predictive, and prescriptive analytics*. Walter de Gruyter GmbH & Co KG.

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. *Zenodo, Version v0*, 9(10.5281).

Henke, N. and Jacques Bughin, L. (2016). The age of analytics: Competing in a data-driven world. *McKinsey Global Institute Research*.

Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., and Amira, A. (2023). Ai-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. *Artificial Intelligence Review*, 56(6):4929–5021.

Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., and Okonkwo, O. R. (2022). Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing*, 25(5):3343–3387.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210.

Jaeger, S. R. and Cardello, A. V. (2022). Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research. *Food Quality and Preference*, 102:104676.

Jansen, B. J., Jung, S.-g., and Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.

Krosnick, J. and Presser, S. (2010). Question and questionnaire design, vol. 2. *Bingley: Emerald Group Publishing Limited*.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lwakatare, L. E., Raj, A., Crnkovic, I., Bosch, J., and Olsson, H. H. (2020). Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. *Information and software technology*, 127:106368.

Matteucci, M., Mentasti, S., Schiaffonati, V., and Fossa, F. (2023). Contextual challenges to explainable driving automation: The case of machine perception. In *Connected and Automated Vehicles: Integrating Engineering and Ethics*, pages 37–61. Springer.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10):5–6.

Moreo, A., Esuli, A., and Sebastiani, F. (2019). Building automated survey coders via interactive machine learning. *International Journal of Market Research*, 61(4):408–429.

Obi, J. C. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, 8(1):308–314.

Oladoyinbo, T. O., Olabanji, S. O., Olaniyi, O. O., Adebiyi, O. O., Okunleye, O. J., and Ismaila Alao, A. (2024). Exploring the challenges of artificial intelligence in data integrity and its influence on social dynamics. *Asian Journal of Advanced Research and Reports*, 18(2):1–23.

Pais, N. V., Holan, S. H., and Parker, P. A. (2025). Topic modeling for free-response text data from a complex survey. *arXiv preprint arXiv:2501.13777*.

Palmatier, R. W. (2008). Interfirm relational drivers of customer value. *Journal of marketing*, 72(4):76–89.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rangineni, S., Bhanushali, A., Suryadevara, M., Venkata, S., and Peddireddy, K. (2023). A review on enhancing data quality for optimal data analytics performance. *International Journal of Computer Sciences and Engineering*, 11(10):51–58.

Reja, U., Manfreda, K. L., Hlebec, V., and Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in applied statistics*, 19(1):159–177.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Sanh, V. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

Scott, A. M., Forbes, C., Clark, J., Carter, M., Glasziou, P., and Munn, Z. (2021). Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. *Journal of clinical epidemiology*, 138:80–94.

Sharma, K., Madhavi, B. G., Goyal, A., Parashar, D., Shrotriya, L., and Gupta, A. (2023). Real-time data analysis, significance, architectures and applications for informed decision-making. In *2023 International Conference on Power Energy, Environment & Intelligent Control (PEEIC)*, pages 298–302. IEEE.

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of business research*, 70:263–286.

Stylos, N., Zwiegelaar, J., and Buhalis, D. (2021). Big data empowered agility for dynamic, volatile, and time-sensitive service industries: the case of tourism sector. *International Journal of Contemporary Hospitality Management*, 33(3):1015–1036.

Taylor, P. A., Reynolds, R. C., Calhoun, V., Gonzalez-Castillo, J., Handwerker, D. A., Bandettini, P. A., Mejia, A. F., and Chen, G. (2023). Highlight results, don't hide them: enhance interpretation, reduce biases and improve reproducibility. *Neuroimage*, 274:120138.

Teh, H. Y., Kempa-Liehr, A. W., and Wang, K. I.-K. (2020). Sensor data quality: A systematic review. *Journal of Big Data*, 7(1):11.

Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., and Emmert-Streib, F. (2021). Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in artificial intelligence*, 4:576892.

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4):791–813.

Widad, E., Saida, E., and Gahi, Y. (2023). Quality anomaly detection using predictive techniques: An extensive big data quality framework for reliable data analysis. *IEEE Access*.

Wolniak, R. (2023). Functioning of real-time analytics in business. *Zeszyty Naukowe. Organizacja i Zarzadzanie/Politechnika Ślaska*.

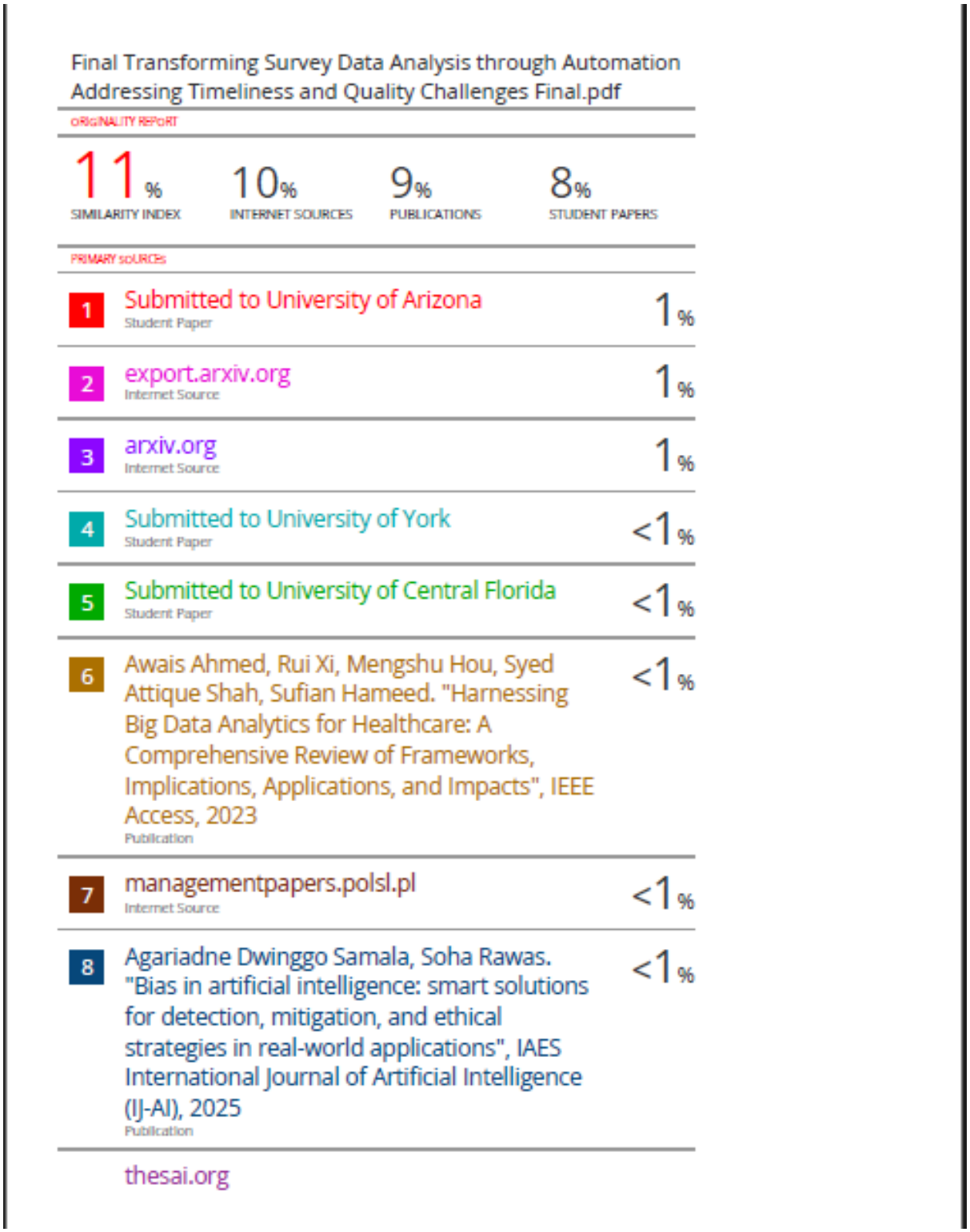
Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273.

Yadranjiaghdam, B., Pool, N., and Tabrizi, N. (2016). A survey on real-time big data analytics: applications and tools. In *2016 international conference on computational science and computational intelligence (CSCI)*, pages 404–409. IEEE.

Zhong, H., Yang, D., Shi, S., Wei, L., and Wang, Y. (2024). From data to insights: the application and challenges of knowledge graphs in intelligent audit. *Journal of Cloud Computing*, 13(1):114.

Appendices

Appendix A: Similarity Report



Final Transforming Survey Data Analysis through Automation Addressing Timeliness and Quality Challenges Final.pdf

ORIGINALITY REPORT

11%	10%	9%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Arizona Student Paper	1%
2	export.arxiv.org Internet Source	1%
3	arxiv.org Internet Source	1%
4	Submitted to University of York Student Paper	<1%
5	Submitted to University of Central Florida Student Paper	<1%
6	Awais Ahmed, Rui Xi, Mengshu Hou, Syed Attique Shah, Sufian Hameed. "Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts", IEEE Access, 2023 Publication	<1%
7	managementpapers.polsl.pl Internet Source	<1%
8	Agariadne Dwinggo Samala, Soha Rawas. "Bias in artificial intelligence: smart solutions for detection, mitigation, and ethical strategies in real-world applications", IAES International Journal of Artificial Intelligence (IJ-AI), 2025 Publication	<1%
	thesai.org	

9	Internet Source	<1%
10	Submitted to Victorian Institute of Technology Student Paper	<1%
11	www.studysmarter.co.uk Internet Source	<1%
12	wsj.westscience-press.com Internet Source	<1%
13	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
14	hvlopen.brage.unit.no Internet Source	<1%
15	Salma Maataoui, Ghita Bencheikh, Ghizlane Bencheikh. "Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models", 2023 Fifth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), 2023 Publication	<1%
16	cajitmf.centralasianstudies.org Internet Source	<1%
17	docs.upb.ro Internet Source	<1%
18	Cristiano E. Caon, Jie Li, Yong Chen. "Effective Management of Time Series Data", 2023 IEEE 16th International Conference on Cloud Computing (CLOUD), 2023 Publication	<1%
19	Submitted to Letterkenny Institute of Technology Student Paper	<1%

20	Yazan Abu Huson, Laura Sierra García, María Antonia García Benau, Nader Mohammad Aljawarneh. "Cloud-based artificial intelligence and audit report: the mediating role of the auditor", VINE Journal of Information and Knowledge Management Systems, 2025 Publication	<1%
21	essay.utwente.nl Internet Source	<1%
22	yong-wang.org Internet Source	<1%
23	sintef.brage.unit.no Internet Source	<1%
24	Submitted to Consorcio CIXUG Student Paper	<1%
25	Submitted to Monash University Student Paper	<1%
26	repositori.upf.edu Internet Source	<1%
27	core.ac.uk Internet Source	<1%
28	csitcp.net Internet Source	<1%
29	Submitted to Johns Hopkins Unversity Student Paper	<1%
30	www.zhouyujia.cn Internet Source	<1%
31	Weisi Chen, Zoran Milosevic, Fethi A. Rabhi, Andrew Berry. "Real-Time Analytics: Concepts, Architectures and ML/AI Considerations", IEEE Access, 2023 Publication	<1%

32	ro.uow.edu.au Internet Source	<1%
33	"Artificial Intelligence in Education", Springer Science and Business Media LLC, 2022 Publication	<1%
34	Submitted to University of Wales, Lampeter Student Paper	<1%
35	ijrpr.com Internet Source	<1%
36	Submitted to University of Edinburgh Student Paper	<1%
37	deepai.org Internet Source	<1%
38	ijisae.org Internet Source	<1%
39	web.realinfo.tv Internet Source	<1%
40	Submitted to Nottingham Trent University Student Paper	<1%
41	sydneyacademics.com Internet Source	<1%
42	www.coursehero.com Internet Source	<1%
43	www.pacificdataintegrators.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	< 25 words
Exclude bibliography	Off		

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

10th April 2025

Mr Maina George,
george.maina716@strathmore.edu

Dear Mr Maina,

RE: Transforming Survey Data Analysis Through Automation: Addressing Timeliness and Quality Challenges

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2420/24.** The approval period is from **10th April 2025 to 9th April 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu