T.C.
IŞIK UNIVERSTIY
SCHOOL OF GRADUATE STUDIES

MASTER THESIS
DEPARTMENT OF COMPUTER ENGINEERING
COMPUTER ENGINEERING PROGRAM

Kajal POURJALIL

EXTRACTING MEANINGFUL INFORMATION FROM STUDENT SURVEYS WITH NLP

SUPERVISOR

Assist. Prof. Emine EKİN

İSTANBUL, January 2025

T.C.
IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

MASTER THESIS
DEPARTMENT OF COMPUTER ENGINEERING
COMPUTER ENGINEERING PROGRAM

Kajal POURJALIL

(22COMP5004)

EXTRACTING MEANINGFUL INFORMATION FROM STUDENT SURVEYS WITH NLP

SUPERVISOR

Assist. Prof. Emine EKİN

İSTANBUL, January 2025

T.C.
IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

MASTER THESIS
DEPARTMENT OF COMPUTER ENGINEERING
COMPUTER ENGINEERING PROGRAM

Kajal POURJALIL

(22COMP5004)

EXTRACTING MEANINGFUL INFORMATION FROM STUDENT SURVEYS WITH NLP

Date: 29./.1./2025

Thesis Supervisor: Assist. Prof. Emine EKİN / IŞIK UNIVERSITY

Jury Members:

Assist. Prof. Tuğba ERKOÇ / IŞIK UNIVERSITY

Assist. Prof. Rahim DEHKHARGHANI / KADIR HAS UNIVERSITY

İSTANBUL, January 2025

ÖZET

NLP KULLANARAK ÖĞRENCİ ANKETLERİNDEN ANLAMLI BİLGİLER ÇIKARMAK

Bu tez, dönem sonu anketleri aracılığıyla toplanan iki dilli öğrenci geri bildirimlerini analiz etmek ve özetlemek için NLP tekniklerini uyguladı. İngilizce ve Türkçe dillerinde açık uçlu yanıtlar içeren veri kümesi, diller arası dilsel nüansları koruyabilen bir modele ihtiyaç duymuştu. Metin üretimi için özel olarak eğitilmiş Llama 2-7b-hf modeli, tutarlı ve bağlamsal olarak uygun özetler üretebilme yeteneği nedeniyle seçilmişti. Veri ön işleme aşaması, bölüm, dönem, ders adı ve şube numarası gibi üstverileri düzenlemeyi, yorumları kelime sayılarına göre ayırmayı ve gizliliği sağlamak için kişisel kimlik bilgilerini kaldırmayı içermekteydi. On kelimeden kısa yorumlar, Transformers kütüphanesinden bir ardışık düzen kullanılarak gruplandırılıp özetlenirken, daha uzun yorumlar ayrıntılı özetleme için üstveri odaklı istemlerle ince ayar yapılmıştı. Analizi daha da geliştirmek amacıyla, "cardiffnlp/twitter-roberta-base-sentiment" modeli kullanılarak duygu sınıflandırması gerçekleştirilmiş ve geri bildirimler olumsuz, tarafsız ve olumlu olmak üzere üç farklı kategoriye ayrılmıştı. Değerlendirme metrikleri arasında uzman incelemeleri, bağlamsal uygunluk ve veri kümesinin duygu dağılımıyla mantıksal tutarlılık yer almıştı. Önceki modellere kıyasla, Llama 2 modeli, yorumların genel niyetini ve tonunu koruyarak daha eksiksiz ve tutarlı özetler üretmede üstün performans sergilemişti. Sonuç olarak, bu araştırma, LLM'lerin çok dilli eğitim verilerini işlemedeki etkinliğini ve ders içeriğini geliştirmek için uygulanabilir içgörüler sağlamadaki potansiyelini net bir şekilde vurgulamıştı. Bu çalışmanın sonuçları, gelecekteki araştırmalar için de yol gösterici olacaktı.

Anahtar Kelimeler: NLP, Llama 2, Anket, Özetleme, Üretken AI

ABSTRACT

EXTRACTING MEANINGFUL INFORMATION STUDENT SURVEYS WITH NLP

This thesis applied NLP techniques to analyze and summarize bilingual student feedback collected via end-of-semester surveys. The dataset, which contained open-ended responses in both English and Turkish, required a model adept at preserving linguistic nuances across languages. The Llama 2-7b-hf model, which had been trained explicitly for text generation, was selected for its capability to produce coherent and contextually relevant summaries. Data preprocessing involved organizing metadata such as department, semester, course name, and section number, segregating comments by word count, and removing personal identifiers to ensure privacy. Shorter comments (fewer than ten words) were grouped and summarized using a pipeline from the Transformers library, while longer comments were fine-tuned with metadata-specific prompts for detailed summarization. To further enhance analysis, sentiment classification was performed using the "cardiffnlp/twitter-roberta-base-sentiment" model, categorizing feedback into negative, neutral, and positive sentiments. Evaluation metrics included expert reviews, contextual relevance, and logical consistency with the dataset's sentiment distribution. Compared to previous models, the Llama 2 model demonstrated superior performance in generating complete, coherent summaries while preserving the overall intent and tone of the comments. Ultimately, this research highlighted the effectiveness of LLMs in processing multilingual educational data and their potential to provide actionable insights for improving course content and student experiences.

Keywords: NLP, Llama 2, Survey, Summarization, Multilingual Analysis

ACKNOWLEDGEMENT

I would like to express my sincerest gratitude to my advisor Assist. Prof. Emine Ekin for introducing me to this field and also for her unwavering support throughout my journey of research and writing of this thesis.

I am deeply thankful to my parents for their unconditional love, countless sacrifices, and always believing in me even when my confidence faltered. I want to thank my friends for their inspiring ideas and for being there through all the stressful times.

Finally, to my younger self, thank you for finding the strength to hold on, even when it felt impossible.

Kajal POURJALIL

In the name of science, one must make sacrifices

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

BERT: Bidirectional Encoder Representations from Transformers

EN: English

EOS: End of Sequence

GDPR: General Data Protection Regulation

GPT: Generative Pre-Trained Transformers

Llama: Large Language Model Meta AI

LLM: Large Language Model

LSTM: Long Short-Term Memory

ML: Machine Learning

NER: Named Entity Recognition

NLP: Natural Language Processing

NLTK: Natural Language Processing Toolkit

NLU: Natural Language Understanding

RoBERTa: Robustly Optimized BERT Pretraining Approach

SA: Sentiment Analysis

SVM: Support Vector Machine

TR: Turkish

CHAPTER 1

1. INTRODUCTION

1.1 BACKGROUND

Educational surveys play a significant role in capturing students' perspectives on teaching quality, course content, and comprehension. Such surveys, conducted systematically across academic institutions, generate vast amounts of valuable feedback and insights (Bhargavi, n.d.). However, analyzing open-ended responses remains a significant challenge due to their unstructured nature and language diversity. Unorganized data from open-ended surveys can be problematic to process without appropriate tools (Jordan, 2011). In addition, the attempt to sort and categorize open-ended data is significant and requires careful consideration and time management (Alchemer, n.d.).

1.2 PROBLEM STATEMENT

Open-ended survey responses have the potential to give rich feedback but also face challenging obstacles such as their unorganized structure and multilingual diversity. Manual analysis is often illogical to use especially since traditional methods like manual coding and statistical techniques often struggle to handle qualitative data in large datasets. The unstructured attributes of open-ended responses require significant effort to sort, categorize, and process which makes manual analysis time-consuming and prone to more inconsistencies (Alchemer, n.d.). Traditional methods for qualitative analysis rely heavily on resources which leads to the development of rapid qualitative approaches to address such challenges (Nevedal et al., 2021). Moreover, the multilingual diversity of responses creates complexities, as language nuances affect the overall accuracy of manual coding and statistical analysis. These limitations

underscore the need for automated, natural language processing (NLP)-driven approaches that can effectively process unordered data, handle multilingual diversity, and extract actionable insights from open-ended survey responses.

1.3 SIGNIFICANCE OF THE STUDY

This research is motivated by the need to transform large-scale, unstructured survey data into actionable insights efficiently. The study leverages recent advancements in NLP, particularly Large Language Models (LLM), to analyze and summarize open-ended responses, thereby assisting educators in making data-driven improvements to teaching practices. By utilizing NLP methods, institutions can gain a deeper understanding of student experiences and identify areas critical for intervention and improvement. (Wang et al., 2024) LLMs not only excel at understanding the semantic context of the survey responses but also offer the scalability required to handle large datasets, empowering educators to make data-driven decisions. (Katz et al., n.d.) Automating the process of analyzing open-ended student feedback enables instructors to focus on implementing changes that will directly impact learning outcomes and foster an evidence-based approach to education. (Kastrati et al., 2021)

1.4 RESEARCH OBJECTIVES

The fundamental aim of this research is to develop and evaluate advanced methodologies for analyzing open-ended student survey responses. Specifically, this study seeks to:

- Preprocess and organize large-scale, multilingual survey data and ensure it is structured for effective analysis and reduce noise in the dataset.
- Utilize advanced NLP techniques including Sentiment Analysis (SA) and summarization in extracting meaningful insights from the unorganized textual data.

- Compare the differences between traditional analysis techniques with modern LLMs in summarizing and open-ended survey responses, further highlighting the advantages and limitations of each approach.
- Assessing the effectiveness of automated methods in identifying actionable insights that can inform instructors and improve course content quality.
- Explore the potential of NLP models in handling the challenges posed by the multilingual nature of survey response to ensure inclusivity and comprehensive feedback analysis.

1.5 SCOPE OF THE STUDY

Spanning three years of data (Spring 2021 to Spring 2024), this research employs advanced NLP methodologies to analyze and summarize open-ended survey responses, focusing on English and Turkish inputs.

1.6 STRUCTURE

This thesis is organized as mentioned below:

- Chapter 2 reviews the relevant literature, including traditional survey analysis techniques and the role of NLP.
- Chapter 3 outlines the proposed methodology, including data preprocessing, NLP techniques, and tools used.
- Chapter 4 presents the previous experimental attempts, results, and their analysis.
- Chapter 5 compares the results with baseline techniques and discusses the findings.
- Chapter 6 concludes the study and provides suggestions for future work.

CHAPTER 2

2. LITERATURE REVIEW

2.1 THE ROLE OF NLP IN SURVEY ANALYSIS

This section will discuss a brief introduction to NLP and its applications, have an overview of survey analysis and reasons why it's needed in education, and talk about the challenges of extracting meaningful information from open-ended survey responses.

2.1.1 Bried Introduction to the Applications of NLP

NLP is a subfield of Artificial Intelligence (AI) focusing on the interaction between human language and computers, helping machines to comprehend, annotate, and process textual data efficiently and practically. Among the wide variety of NLP applications, machine translation, SA, text summarization, information extraction, and even conversational such as chatbots can be pointed out. In survey analysis, NLP applications such as SA can automatically determine the sentiment behind open-ended responses, categorizing them as positive, negative, or neutral (SurveyMonkey, n.d.). Recently, the development of LLMs, which are an important technique in NLP, has changed the scene in the ability to process large-scale volumes of unregulated textual data.

2.1.2 An Overview of the Use and Importance of Survey Analysis in Education

The role of Survey Analysis in education is crucial in gathering information about students' learning experiences, course content quality and monitoring teaching efficiency. Institutions tend to rely on surveys to find areas of improvement, make enhancements based on data-driven metrics, and measure overall student satisfaction. Both open-ended and close-ended responses offer

valuable insights with the former focusing on deeper and qualitative feedback portraying students' sentiments and suggestions while the latter provides statistical information. For example, an analysis of student feedback through surveys, as highlighted by (Parker et al., 2023), identified specific areas in the curriculum that required enhancement, leading to improved educational outcomes.

2.1.3 The Challenges of Extracting Meaningful Information from Open-Ended Responses

The complexity and diversity of open-ended survey responses, present significant challenges despite their value. Student responses include various formal and informal language, different sentence structures, and multilingual content, making the manual analysis task extremely time-consuming and inefficient. Hence, extracting meaningful patterns from such unorganized data requires techniques that can handle sarcasm, and refined sentiments, such as ambivalence, where a respondent expresses both positive and negative feelings in a single response, which can complicate the analysis. Contextual variations, like cultural differences in expressing feedback, can also lead to misinterpretation if not properly accounted for. Traditional methods like manual coding fail to present an equal output in large-scale surveys, bringing the need for automated approaches to the surface.

2.1.4 The Significance of Using NLP when Analyzing Student Surveys

Applying NLP to student surveys is a great approach for automated and extensible open-ended responses. Other NLP techniques such as SA and text summarization can extract valuable information from qualitative data allowing institutions to track repeating issues, elevate course quality, and increase the overall student experience. With the assistance of LLMs, NLP can find contextual nuances, handle multilingual responses, and create a deeper comprehendible understanding of the students' feedback. This research demonstrates the ways NLP-based methods will bridge the gap between

unstructured survey data and the insights that will lead to data-driven decisions for educational services.

2.2 SURVEY RESULT ANALYSIS TECHNIQUES

This section discusses the techniques used to analyze survey results, highlighting their benefits and limitations, as well as the importance of adapting more advanced methods.

2.2.1 Traditional Methods

Traditional methods depend heavily on manual coding and statistical methods to process responses. Manual coding involves researchers reading through every response and assigning categories or themes. This process is both time-consuming and prone to human error. While statistical methods like frequency analysis and regression models are efficient for handling numerical and closed-ended questions, manual coding focuses on understanding the meaning of individual responses. Statistical methods tend to fail to analyze open-ended survey responses and qualitative data, and manual coding will be problematic if practiced on large datasets overall, while these approaches provide considerable results, the limitations become visible moreover as datasets grow in both complexity and size. Manual coding of open-ended survey responses is widely recognized as time-consuming, resource-intensive, and prone to human error. For instance, (Dovetail Editorial Team, n.d.) highlights that manual coding can be "time-intensive, verging on painful," and lacks scalability, making it impractical for large datasets. Additionally, (Widmer, n.d.) notes that verbatim coding is "a tedious and time-consuming process, prone to mistakes and human biases if not handled carefully." Traditional statistical methods often fall short of capturing the nuanced sentiments expressed in open-ended responses. As noted by (Thematic, n.d.), SA uses AI to analyze large volumes of text to determine whether it expresses a positive, negative, or neutral

sentiment, highlighting the limitations of conventional statistical approaches in this context.

2.2.2 Automated Methods

To address the limitations caused by the traditional approaches, automated methods were called into action. Initially, automated approaches included rule-based methods using them to match keywords, extract the SA, and categorize responses. Then there are classical machine learning techniques like Support Vector Machines (SVM) and Naïve Bayes classifiers which automate the tasks of SA and summarization. For instance, Naïve Bayes classifiers have been effectively applied in SA tasks involving datasets like the IMDB movie reviews dataset, where it achieved accuracy rates of around 81% for binary sentiment classification tasks (Pang et al., n.d.). This method's reliance on probabilistic principles enables it to handle text classification efficiently when the independence assumption approximately holds, though it may struggle with more complex dependencies between features.

Similarly, Support Vector Machines (SVM) have shown strong performance in text classification tasks, particularly when combined with the Bag-of-Words model for feature representation. An example is the application of SVM on the Twitter SA Dataset, achieving an F1 score of 84% for sentiment classification (Go, Bhayani, & Huang, n.d.). SVM's ability to maximize the margin between classes allows for robust generalization, though it can be computationally intensive for larger datasets. They have proven to be successful when applied to classify survey responses into predefined categories, integrating the analysis process. By learning the patterns from labeled data, such models overturned previously mentioned rule-based systems. Although they may struggle with multilingual data and often require extensive preprocessing and still fail to capture slightly different meanings, leading to suboptimal performance. Machine learning techniques still need crucial preprocessing, labeled datasets, and feature engineering even though they have improved accuracy and scalability. Still, when it came to dealing with unorganized and

instructed responses that were multilingual and had nuanced sentiments, their performances were faulty.

2.2.3 The Requirement for More Advanced Technique

An impending need for more advanced methods arises with the increasing volume of survey data and the growing complexity of open-ended responses. Hence, the advancements in NLP and LLMs have made revolutionary modifications to how textual data can be both analyzed and summarized. With classical machine learning techniques and traditional methods being the groundwork for survey result analysis, this research aims to build upon such methodologies using state-of-the-art NLP techniques, mainly LLMs, to extract meaningful information from open-ended surveys efficiently, facing challenges like multilingual data, contextual nuances, and both short and long-formed responses. The inability of classical methods to effectively handle complex language structures and contextual meanings underscores the need for LLMs, which can comprehend and analyze text with greater sophistication.

2.3 THE USE OF LARGE LANGUAGE MODELS IN TEXT ANALYSIS

This section will examine LLMs, highlighting their strengths in text generation and summarization, comparing fine-tuned models with pre-trained models, and addressing the challenges associated with their practical implementations.

2.3.1 Introduction to LLMs

LLMs, as noted by (Touvron et al., 2023), represent a significant advancement in NLP due to their ability to process vast datasets and excel in tasks like text generation, translation, and question answering. They capture intricate linguistic patterns and have been highlighted as promising tools for aspect-based summarization, building on insights from prior research. These models are trained on vast quantities of text data and have achieved remarkable

performance on a range of NLP tasks, including text classification, question answering, and machine translation (Hoffmann et al., 2022; Yang et al., n.d.). The model size and the number of training tokens are two crucial parameters to choose when influencing an LLM's performance and the procedure to train it. Other important factors include learning rate, learning rate schedule, batch size, optimizer, and width-to-depth ratio (Hoffmann et al., 2022). Larger models trained on extensive datasets tend to exhibit superior language understanding capabilities.

2.3.2 The Strength of LLMs in Text Analysis

One of the leading strengths of LLMs is their capability to create context-based and coherent text. For example, Bidirectional Encoder Representations from Transformers (BERT)-based models perform well when working with nuanced contextual relationships in a text while Generative Pre-trained Transformer (GPT)-based models work better creating more detailed and fluent summaries out of unorganized input (Brown et al., 2020; Devlin et al., n.d.). LLMs can provide an unmatched advantage when working in the field of survey analysis since they can distinguish unorderly sentiments and generate compact summaries from open-ended responses (Conneau et al., 2019; Y. Liu et al., 2019). Such models will reduce the dependency on manual preprocessing because they can work efficiently with raw and unstructured data. This gives them the upper hand when working with summarization and SA tasks (Radford et al., n.d.; Raffel et al., 2019). In educational survey analysis, LLMs have demonstrated the ability to accurately interpret complex student feedback, outperforming traditional methods like thematic analysis and sentiment detection (Parker et al., 2023).

2.3.3 Pre-Training and Fine-Tuning the Models

Pre-trained models, while trained on large-scale corpora, may lack the domain-specific precision required for educational surveys, potentially, missing important context-dependent nuances. To address this, they undergo fine-tuning of smaller, task-specific datasets, optimizing them for applications such as text classification or question answering, for example, models like GPT-4 are designed to handle various tasks due to their broad training data, but without fine-tuning, they lack domain-specific expertise, making them less suitable for education surveys (Brown et al., 2020). However, fine-tuning allows these models to adapt to specialized fields. BERT, for instance, can be fine-tuned for SA using labeled datasets, enabling it to capture domain-relevant sentiments more effectively. (Devlin et al., n.d.).

2.3.4 The Challenges of Working with LLMs

LLMs, while being powerful, are challenging to work with. One notable obstacle is the computational cost that comes with training and inference (Brown et al., 2020). Another barrier is the fine-tuning which needs labeled data, expertly specific domain, and hyperparameter tuning to provide an optimal performance as a result (Devlin et al., n.d.). LLMs might sometimes create incoherent or inconsistent results when processing ambiguous and disorderly text which increases suspicions of their reliability on the outputs for important applications such as survey analysis (Leivada et al., 2023). There is also the matter of bias in LLMs which stems from the biases in their training data that create some ethical challenges in analyzing sensitive information like student feedback (Nozza et al., 2020; Wiedemann et al., 2019). Strategies to mitigate LLM limitations include domain adaptation, where models are fine-tuned on domain-specific data, and implementing model interpretability techniques to understand and rectify biases.

2.3.5 Reasons to Switch to LLMs

The transition to LLMs becomes vital when the limitations of traditional and classical machine learning models become more apparent. In survey analysis methodologies, such models show great progress in scalability, contextual understanding, and multilingual capabilities which exceeds the abilities of their predecessors (Minaee et al., 2024). Furthermore, the benefits of LLMs in summarizing and analyzing vast, open-ended survey data efficiently, make them indispensable in this research (*Fine-Tuning Large Language Models: Future Trends and Challenges*, n.d.). Adopting LLMs can lead to improved accuracy in text analysis, reducing the need for extensive preprocessing and enabling more efficient handling of complex language structures.

2.4 CHALLENGES IN EXTRACTING INFORMATION FROM EDUCATION SURVEYS

This section discusses the challenges of raw data, including multilingual comments, varying response lengths, and personal information. It also explores techniques for addressing these issues, such as preprocessing steps, and highlights the gaps this research aims to fill.

2.4.1 Reasons to Switch to LLMs

Raw data in open-ended surveys will include multilingual comments causing challenges for models working with monolingual datasets (Conneau et al., 2019). Surveys that have responses in multiple languages pose a challenge for analysis, as direct translations may not capture the original sentiment or cultural context accurately. Short responses may lack sufficient context while long responses on the other hand often need summarization before key insight extraction (Devlin et al., n.d.). Also, survey data might heedlessly include personal information that necessitates precise preprocessing steps to ensure

assent with data protection regulations like the General Data Protection Regulation (GDPR) (Voigt & Bussche, 2017).

2.4.2 Techniques to Handle the Complications

When addressing these challenges, preprocessing techniques like tokenization, lemmatization, and language identification are applied to normalize and clean out the data (Manning et al., 2008). More advanced methods like multilingual embeddings and translation models have been arranged to bridge the gap between comments in different languages (Conneau & Lample, 2019). Preprocessing tools such as spaCy and Natural Language Toolkit (NLTK), are commonly used for tasks such as tokenization and language identification, facilitating the preparation of textual data for analysis.

2.4.3 Gaps in this Literature

Despite the existing research exploring SA and summarization individually, there is still a lack of integrated approaches that combine such techniques to analyze open-ended survey responses efficiently (Raffel et al., 2019). This study approaches this gap by asserting fine-tuned LLMs for SA and summarization, providing a comprehensive analysis of survey data, and integrating SA and summarization techniques to provide a more comprehensive understanding of survey responses compared to previous studies that may have addressed these aspects in isolation.

2.5 THE ROLE OF SENTIMENT ANALYSIS IN SURVEY INTERPRETATION

SA has become a pivotal tool in deciphering the subjective nuances in survey data and allows organizations to transform qualitative feedback into actionable insights. This section discusses the aid of SA in this research.

2.5.1 Understanding Sentiment Analysis

Often known as "Opinion Mining", SA contains the computational study of people's emotions, opinions, and sentiments expressed in a particular text aiming to reach the goal of extracting subjective information from source materials (Tejwani, 2014). SA plays a crucial role in understanding the emotional tone of data, which can provide organizations and researchers with actionable insights into public opinion or user satisfaction.

The methodology involves various steps, such as preprocessing text data, feature extraction, and classification into sentiment categories like positive, negative, or neutral. Advanced SA incorporates techniques like deep learning, aspect-based SA, and hybrid approaches that combine machine learning and lexicon-based strategies to improve accuracy (Pang & Lee, 2008).

In addition to its computational benefits, SA has found extensive applications across industries, enabling better customer experience management, political campaign analysis, and decision-making processes in education, healthcare, and retail. It has become an indispensable tool for translating qualitative feedback into measurable and actionable insights (B. Liu, 2012).

2.5.2 Importance of its Usage in Survey Data

It enables organizations to systematically assess subjective feedback and transform qualitative opinions into quantifiable data that can inform strategic decisions. Institutions can use it to identify prominent emotions and attitudes, facilitating targeted improvements in services or products. This analytic approach enhances the understanding of customer or employee satisfaction levels, supporting the development of more effective engagement strategies (Qualtrics, n.d.).

2.5.3 Techniques and Approaches

There are many methodologies in SA such as lexicon-based approaches, Machine Learning (ML) techniques, and hybrid models which have been developed to perform SA efficiently (Upadhye, 2022). Various methodologies have been developed to perform SA efficiently, including lexicon-based approaches, machine-learning techniques, and hybrid models. Lexicon-based methods utilize predefined dictionaries of sentiment-laden words to evaluate text, while machine learning techniques involve training algorithms on labeled datasets to recognize sentiment patterns. Hybrid models combine both approaches to enhance accuracy and adaptability across different contexts (Blyakhman, n.d.).

2.5.4 Challenges in Sentiment Analysis

SA still faces multiple challenges such as context interpretation, sarcasm detection, and handling domain-specific language despite its advantages (Narayanan Venkit et al., n.d.). The complexity of human emotions, language nuances, and cultural differences further complicate the analysis, often leading to misclassification of sentiments. Additionally, the presence of ambiguous language and the need to adapt models to various industries and multilingual data add layers of difficulty to achieving precise SA (Determ, n.d.).

2.5.5 Applications in Educational Surveys

SA has been applied to student surveys to estimate the sentiments towards courses and instructors, enhance teaching strategies, and provide valuable insight for academic improvements (Jiménez et al., n.d.). This application aids in enhancing teaching strategies and provides valuable insights for academic improvements. By analyzing qualitative feedback, educational institutions can identify areas of concern, adapt curricula, and improve overall student satisfaction (Dake & Gyimah, 2022).

2.5.6 Combining Sentiment Analysis with Other NLP Techniques

Research shows that when SA integrates with aspect-based analysis, it will allow for a more nuanced understanding of specific elements within survey responses, enhancing the depth of insights gained (Yadav et al., 2020). This combination enhances the depth of insights gained, enabling organizations to pinpoint particular aspects that elicit positive or negative sentiments. Such detailed analysis supports more targeted interventions and improvements (Mello et al., 2022).

2.6 MODELS AND FRAMEWORK COMPARISON

The table below provides a comparison of the framework and models used in this research.

Table 2.1 Comparing the Framework and Models

Feature	Keras	BERT	GPT-2	Llama 2-7b-hf
Type	Framework for building deep learning models	Transformer-based model for Natural Language Understanding (NLU)	Transformer-based generative language model	Transformer-based generative model
Primary Purpose	Facilitate easy development and experimentation of neural networks	- SA - Question answering - Classification	- Text generation - Text completion - Creative writing	- Text generation - Conversational AI
Architecture	Supports various architectures: - CNN - RNN - Transformers	Bidirectional Transformer	Autoregressive Transformer	Autoregressive Transformer
Training Data	Depends on user input	Pretrained on a vast corpus of text including books and Wikipedia	Pretrained on a large-scale, diverse internet text corpus	Pretrained on an extensive dataset of publicly available internet text
Use cases	- Vision - Speech - Text-related	- SA - Named Entity Recognition (NER) - Translation	- Chatbots - Creative writing - Interactive storytelling	- Chatbots - Summarization - Document generation - Multilingual understanding
Advantages	- Simple API - Supports multiple backend frameworks	- Bidirectional context understanding - Strong pretraining	- Powerful text generation - Large-scale pretraining	- Multilingual support - State-of-the-art performance - Open access for fine-tuning
Limitations	Requires user expertise for optimal performance	Limited for generative tasks	Less efficient for bidirectional context understanding	- Requires significant computational resources

2.7 SUMMARY AND RESEARCH GAP

This section will summarize the key insights from previous sections, highlight the gap this research aims to fill, and explain how the methodology addresses this gap.

2.7.1 Summary of Key Insights

As mentioned before, existing approaches to survey analysis still face challenges with raw data, including multilingual comments, different response lengths, and potential privacy concerns requiring extensive preprocessing (Conneau et al., 2019; Voigt & Bussche, 2017). Recent techniques often approach SA and summarization separately, leaving a gap in unified methods able to handle the complexity of vast and large-scale, open-ended survey data (Raffel et al., 2019).

2.7.2 Stating the Research Gap

By applying fine-tuned LLMs specifically designed for multilingual survey summarization, a critical gap has been addressed in an area that has received limited attention in the literature (Devlin et al., n.d.).

2.7.3 Proposing the Methodology

The method aims to provide accurate, context-aware insights from compound survey data by integrating preprocessing techniques with fine-tuned LLMs, setting the stage for detailed discussions in Chapter 3. The methodology bridges the gap by leveraging state-of-the-art LLMs, fine-tuned on multilingual and domain-specific datasets to perform SA and summarization synchronously (Conneau & Lample, 2019; Raffel et al., 2019).

CHAPTER 3

3. PROPOSED METHODOLOGY

3.1 RESEARCH DESIGN

This research aims to analyze open-ended survey responses to extract meaningful information and sentiment. Its exploratory nature allowed for the refinement of methods, while the experimental design validated the effectiveness of the chosen approaches.

3.2 MODEL SELECTION AND FINE-TUNING

The model used in this study is the Large Language Model Meta AI (Llama) 2-7b-hf model which is trained explicitly for text generation (*Meta-Llama/Llama-2-7b-Hf · Hugging Face*, n.d.).

The Llama 2-7b-hf is a state-of-the-art transformer-based language model developed by Meta. It has 7 billion parameters, making it capable of handling complex natural language understanding and generation tasks. Designed for both research and practical applications, the model generates coherent, contextually relevant responses and summarizations across various domains. Its architecture is optimized for computational efficiency, allowing it to be fine-tuned effectively for specialized tasks such as summarization and SA.

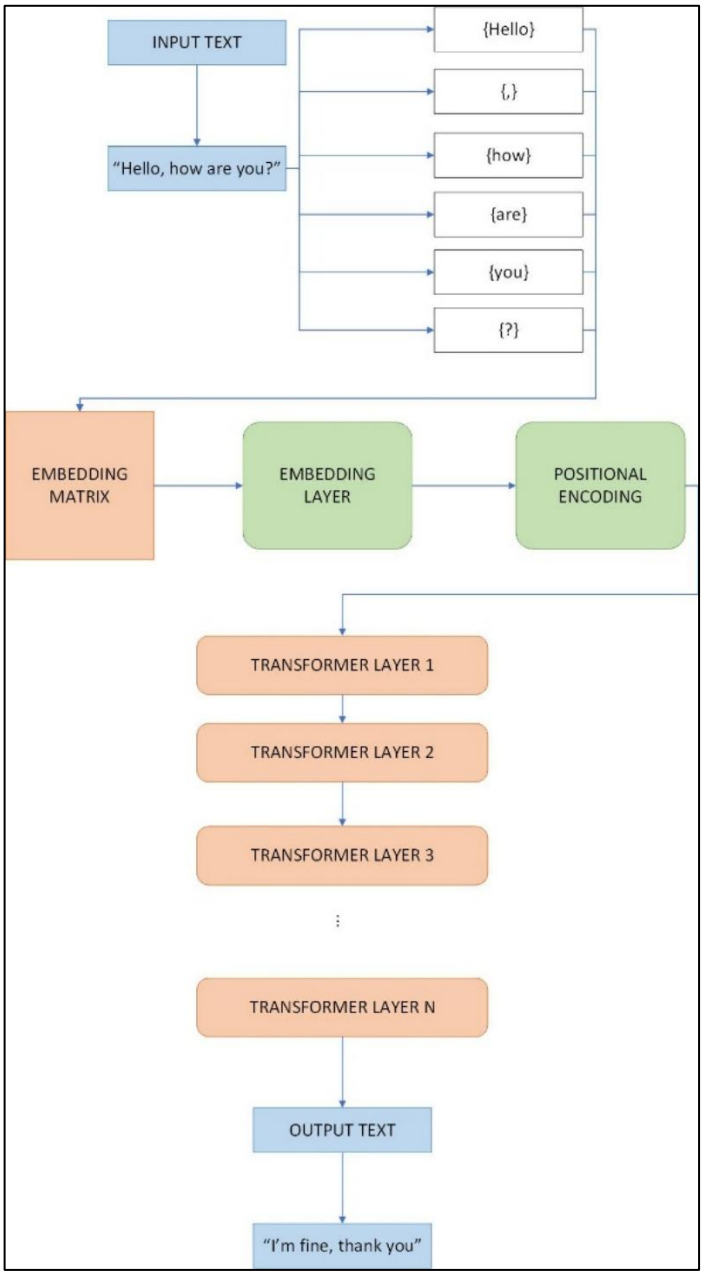

Figure 3.1 The Llama 2-7b-hf Model Workflow

Figure 3.5 shows the model's workflow in simple terms. The model starts with receiving an input text broken down into smaller tokens. Next, the tokens are passed to the embedding layer, and mapped to dense vectors using a pre-embedding matrix. Since the model processes all tokens simultaneously, it adds information about the position of each token in sequence to the embeddings. This

ensures the model understands word order. The embeddings are passed through several transformer layers where they will be processed using attention-masks and feed-forward networks. Finally, the model predicts the most likely next tokens step-by-step to generate a coherent response.

3.3 DATA COLLECTION AND PREPARATION

The data used in this research is collected via the university's learning management system. At the end of every semester, the survey designed by the university will become accessible to the students, allowing them to share their opinions anonymously. Students are encouraged to exclude any personal information. The survey includes 27 close-ended questions. The first 25 questions are answered based on the Likert scale. Question 26 asks whether the students attended at least 70% of the sessions and completed the course material. Question 27 asks which grade the students think they will receive at the end. But the main questions that this research focused on are the last two, Questions 28 and 29 which are open-ended, and the students are free to answer them however they please. Respectively, the questions are as follows:

- Q28: "What do you suggest to improve course content?"
- Q29: "To improve student comprehension in the course, what do you suggest?"

A few extra columns were added to help handle the data better such as:

- Department
- Semester
- Section
- Lang_Q28
- Lang_Q29
- Q28_no
- Q29_no

In the Department column, the 4 first letters from the course names, representing the department names were added. The Semester column was filled with the name and number of the semesters and in the Section column, the last number after the dot from the course name which represents the section number of a course was added.

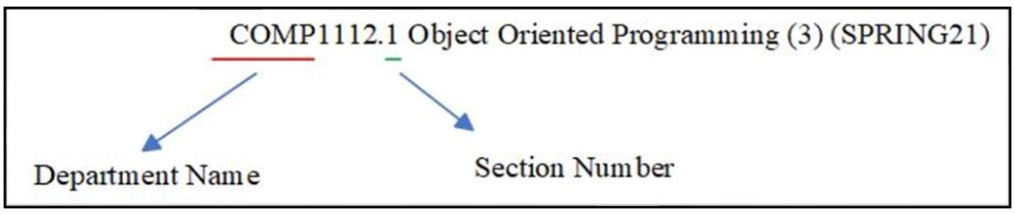

Figure 3.2 Extracting the Department Name and Section Number

In the columns where the students did not respond to the Q28 and Q29 questions, the word “null” was written. Emojis, tags, and links were deleted along with all personal information in the comments including the instructors’ names for privacy. The instructors’ names were replaced with unique IDs in the form of ID_1, ID_2, etc. On some occasions, students responded in both languages. They would write in one language and add the exact translation after. In such cases, only one of the two was chosen. The Llama 2-7b-hf model struggles with sentences shorter than 10 words, often leading to overfitting. To address this, the dataset was split based on comment length. The comments now have either equal or more than 10 words or less than 10 words in them and two separate approaches were used for the two types of comments. Comments were categorized as either fewer than 10 words or 10 words or more, with separate approaches applied to each group. Shorter comments were grouped based on metadata, including department name, course name, semester, and section number. These grouped comments were then concatenated and summarized using a summarization pipeline from Transformers.

For longer comments, the same metadata-based grouping was used, but instead of direct concatenation, the model was prompted with: 'Summarize the feedback in a few sentences and generate insights based on them.' Since Llama 2 models are prompt-based, this step was necessary. The model was fine-tuned accordingly, and summaries and insights were generated. Another example of prompts used:

- "Rewrite the feedback into an actionable summary for instructors. Focus on key themes and insights without repeating the exact words."
- "Based on the above feedback, generate actionable insights and key takeaways in Turkish. To improve the course, highlighting recurring themes and the overall tone."

Response	Departmer	Course	Semester	Lang_Q28	Section	Q1	...	Q27	Q28	Q28_no
Student1	SOFT	SOFT3101.1 Software Engineering (4) (FALL21)	Fall21	TR	1	5	...			62
Student2	SOFT	SOFT3101.1 Software Engineering (4) (FALL21)	Fall21	EN	1	3	...			19
Student3	SOFT	SOFT3101.1 Software Engineering (4) (FALL21)	Fall21	TR	1	4	...			40

Figure 3.3 Sample of the Dataset with Q28 and Long Comments

Response	Departmer	Course	Semester	Lang_Q29	Section	Q1	...	Q27	Q29	Q29_no
Student1	COMP	COMP1101.1 Introduction to Programming (4) (SPRING21)	Spring21	EN	1	1	...			0
Student2	COMP	COMP1101.1 Introduction to Programming (4) (SPRING21)	Spring21	EN	1	1	...			1
Student3	COMP	COMP1101.1 Introduction to Programming (4) (SPRING21)	Spring21	TR	1	4	...			5

Figure 3.4 Sample of the Dataset with Q29 and Short Comments

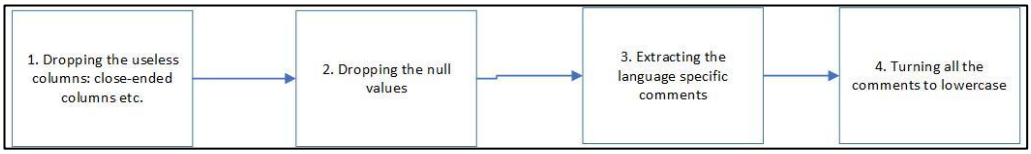

Figure 3.5 Data Preparation Workflow

To clean the data, the first step was to load it using the Pandas library. Measures were taken to check for any mismatched rows. After ensuring none remained, all null and irrelevant columns such as those containing close-ended questions, response numbers, and submission dates were dropped. The data was split based on the two columns of Q28 and Q29. Since the comments were either in English or Turkish, the English and Turkish responses were processed separately for each question. To facilitate this, two new columns, Lang_Q28 and Lang_Q29, were added, with values "EN" for English and "TR" for Turkish. All comments were converted to lowercase to standardize the text for easier processing. Finally, the pre-processed data was saved in pickle format for future use.

Table 3.1 Distribution of EN, TR, NULL, and Total Entries in the Dataset

CATEGORY	COUNT	PERCENTAGE
Total number of EN comments	2019	8.89%
Total number of TR comments	5471	24.09%
Total number of NULL comments	15,219	67.01%
Total number of entries	22,709	----------

Table 3.2 shows the dataset files and their respective sizes, total number of rows, and approximate insight generation time.

- Q28_ge_10: Question 28 with comments of more than 10 words
- Q28_lt_10: Question 28 with comments less than 10 words
- Q29_ge_10: Question 29 with comments of more than 10 words
- Q29_lt_10: Question 29 with comments less than 10 words

Table 3.2 Dataset Size and Generation Time

Dataset	SIZE	Total number of rows	≈ Generation Time
Q28_ge_10	1112 KB	2292	2m 4s
Q28_lt_10	4230 KB	20418	1m 5s
Q29_ge_10	810 KB	1751	2m 34s
Q29_lt_10	4334 KB	20,959	12s

3.4 APPLYING SENTIMENT ANALYSIS

In a previous attempt, the "DistilBERT for Sentiment Analysis" model was used to achieve the sentiment of the comments. It is a fine-tuned version of the "distilbert-base-uncased" model on a social media dataset. However, due to the model only providing binary sentiment, another model was used. The model used for SA on both approaches is "cardiffnlp/twitter-roberta-base-sentiment" which is trained on Twitter comments and provides fine-grained SA. The labels are as follows:

- LABEL_0: Negative
- LABEL_1: Neutral
- LABEL_2: Positive

The labels will be assigned to the sentences (Hugging Face, n.d.).

3.5 TOOLS AND MODELS

Tools used in this research:

- Jupyter Notebook for computational tasks and experimentation.
- The Transformers library for summarization pipelines and model fine-tuning,
- The Pandas library for data preprocessing and cleaning.
- The Pickle library for saving and loading the preprocessed data.

Models used in the research:

- Llama-2-7b-hf
- cardiffnlp/twitter-roberta-base-sentiment

3.6 EVALUATION METRICS

Metrics such as the percentage of missing data handled, and duplicates removed were recorded to ensure the data was clean and ready for analysis. Expert opinion was used to assess the quality of the insights and summaries extracted from the LLM and SA model. The outputs were evaluated based on their contextual relevance and alignment with domain knowledge. The distribution of sentiment labels was analyzed to ensure logical consistency with the overall sentiment observed in the dataset.

CHAPTER 4

4. EXPERIMENTAL EVALUATIONS

At the beginning of this research, the goal was to build a generative AI model. Therefore, before working with the Llama 2-7b-hf model, other tools were experimented with in hopes of leading to the generated summaries. The first step was to find a tool that would aid in preprocessing the data. spaCy was used for this task. It completed the tasks of tokenization, removal of stop words and punctuation marks, part-of-speech tagging, and lemmatization. Next, appropriate spaCy models for both the English and Turkish languages were found to process the comments. The model used for the English comments is the "en_core_web_md" and the model used for the Turkish comments is the "tr_core_news_md". An obstacle occurring at this point was that, unlike the English model, the Turkish model wasn't compatible with the latest spaCy version. Since no other compatible Turkish models were found, an older version (version 3.4.2) had to be installed. This further complicated the process of installing the models. Due to the mismatch in spaCy versions, while working on either the English comments or the Turkish comments, the spaCy models had to uninstall the previously installed model and reinstall the other one. To prevent this, all the Turkish data and notebooks were uploaded to a second Drive account. The initial aim was to apply SA on the comments and then combine them with the processed output data from spaCy and build the model based on them. However, the problem was that SA works only on full sentences and not tokens. This would create a huge data ambiguity which the model couldn't process since the number of the sentiments derived from the comments were less than the number of all the tokens achieved from the preprocessing step. Instead, the SA had to be applied at the end of the generated summaries.

4.1 BUILDING A MODEL WITH KERAS

The first attempt was to build the model with Keras, an open-source deep-learning framework that provides a user-friendly interface for building and training neural network models (Chollet & others, 2015). The approach began by installing the necessary libraries from Transformers and TensorFlow. The preprocessed tokens from spaCy were converted to Transformer-compatible tokens using "bert-base-uncased" as the tokenizer, generating the input IDs and attention masks. Next, sentence paddings were created and the sentences were truncated to optimize the model's efficiency. To build the model, the pre-trained BERT model, "bert-base-uncased" was used as the encoder, with some BERT layers frozen to improve processing speed. Encoder inputs were defined, and a Lambda layer was used to wrap the BERT model, process the embeddings, and specify the output shape explicitly. An LSTM decoder was then implemented to predict the next words in a sentence, and a dense layer was added to output the probabilities of the next words. In the last step, the complete model was defined and compiled using the encoder, decoder, and categorical cross entropy for sequence prediction.

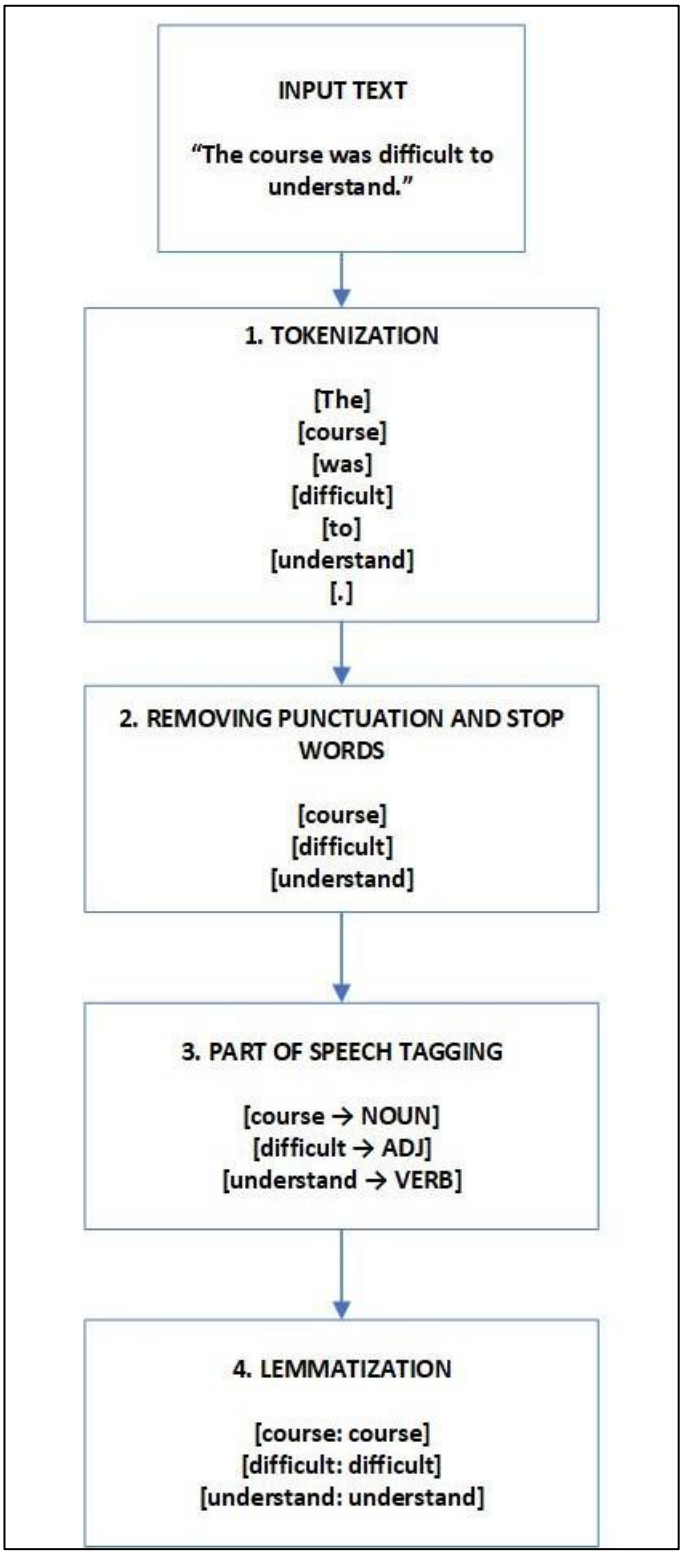

Figure 4.1 Example of Preprocessing an EN Comment with spaCy

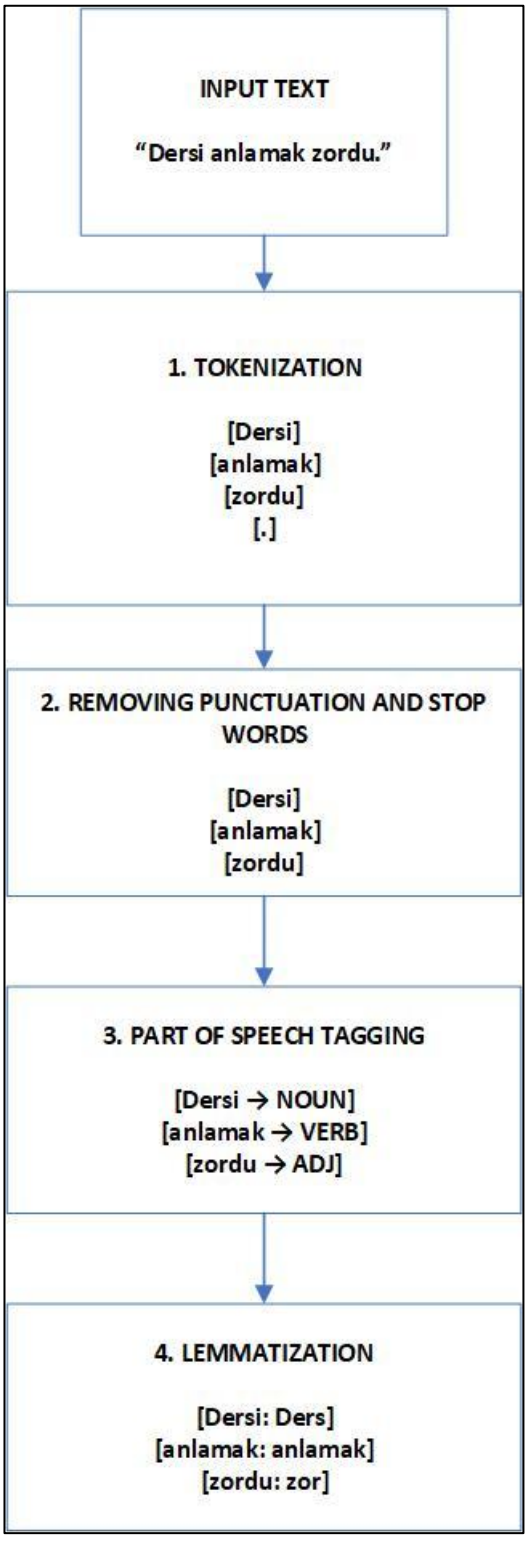

Figure 4.2 Example of Preprocessing a TR Comment with spaCy

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	(None, 128)	0	-
attention_mask (InputLayer)	(None, 128)	0	-
decoder_input (InputLayer)	(None, None)	0	-
lambda (Lambda)	(None, 128, 768)	0	input_ids[0][0], attention_mask[0][0]
embedding (Embedding)	(None, None, 128)	1,280,000	decoder_input[0][0]
get_item (GetItem)	(None, 768)	0	lambda[0][0]
get_item_1 (GetItem)	(None, 768)	0	lambda[0][0]
lstm (LSTM)	(None, 768)	2,755,584	embedding[0][0], get_item[0][0], get_item_1[0][0]
dense (Dense)	(None, 30522)	23,471,418	lstm[0][0]

Total params: 27,507,002 (104.93 MB)
Trainable params: 27,507,002 (104.93 MB)
Non-trainable params: 0 (0.00 B)

Figure 4.3 Model Summary built with Keras

As shown in Figure 4.3, the layer column portrays the name and type of each layer in the model, defining the architecture and data flow. The output shape column indicates the shape of the data output by each layer. The first dimension is usually "None" because it shows the batch size which varies during training and inference. The param column shows the number of trainable parameters in each layer that the model uses during the training phase. They are calculated as follows:

- Dense layer: (input units + 1 for bias) x output units
- Embedding layer: vocabulary size x embedding size
- LSTM layer: A combination of weights and biases for the gates (input, forget, cell, and output)

The last column lists the connections between layers, showing how the data flows through the network. After compiling the model, the input-output pairs were created and shifted to predict the next words. The padded sequences were set to maximum length as the input data and the output data was converted into a NumPy array. Since it is a single token, there is no need for padding. The

output data is converted into a one-hot encoded format where each token is transformed into a binary vector of size equal to the vocabulary. This helps the model predict probabilities for each word in the vocabulary at each step. Next, the input data is changed into a NumPy array format for training compatibility. Finally, the training phase started on 3 epochs with batches of size 32 and a 0.2 validation split.

```
Epoch 1/3
299/299 ━━━━━━━━━━━━━━━━━━━━ 6400s 21s/step - accuracy: 0.2302 - loss: 3.4304 - val_accuracy: 0.2801 - val_loss: 2.4476
Epoch 2/3
299/299 ━━━━━━━━━━━━━━━━━━━━ 6253s 21s/step - accuracy: 0.2872 - loss: 2.2742 - val_accuracy: 0.2855 - val_loss: 2.0937
Epoch 3/3
299/299 ━━━━━━━━━━━━━━━━━━━━ 5908s 20s/step - accuracy: 0.3116 - loss: 2.0288 - val_accuracy: 0.3300 - val_loss: 1.9980
<keras.src.callbacks.history.History at 0x780be979d780>
```

Figure 4.4 Training Details from the Model built with Keras

As presented in Figure 4.4, both the training and validation accuracy improve over the three epochs, indicating the model is learning and generalizing better. The loss values in both of them decreased steadily showing that the model is optimizing its parameters effectively. Validation metrics follow a similar trend to the training metrics, with accuracy improving and loss decreasing. Each epoch takes slightly less time than the previous one. This could be due to improved hardware efficiency such as slightly smaller batch processing time as the model becomes more stable. After three epochs, the training accuracy reached ~31.16% and validation accuracy reached ~33.00%.

Despite multiple fine-tuning attempts, the model was unable to generate complete or coherent sentences and instead produced concatenated tokens and random characters.

4.2 BUILDING A TRANSFORMER-BASED MODEL WITH BERT

The second approach involved building the model using BERT, a transformer-based model developed by Google. Designed for NLU tasks, such as question answering and SA, BERT captures contextual information from both directions (left-to-right and right-to-left) in text (Devlin et al., 2018). The

approach began with loading the preprocessed data and converting it into Transformer-compatible tokens using the "BertTokenizer". The list of preprocessed tuples was then transformed into a list of strings, and their lengths were checked before unpacking to prevent potential errors. The tokenizer was applied to pad the sentences and extract the input IDs and attention masks. Next, the "bert-base-uncased" model was set as the encoder, with certain layers frozen to optimize the training process. The input layers were defined and a custom BertEncoderLayer was used in place of a Lambda layer. After defining the decoder input and layers, the model compilation process began.

Layer (type)	Output Shape	Param #	Connected to
decoder_input (InputLayer)	(None, None)	0	-
input_ids (InputLayer)	(None, 5)	0	-
attention_mask (InputLayer)	(None, 5)	0	-
embedding_3 (Embedding)	(None, None, 128)	3,906,816	decoder_input[0][0]
bert_encoder_layer_2 (BertEncoderLayer)	(None, 5, 768)	0	input_ids[0][0], attention_mask[0][0]
dropout (Dropout)	(None, None, 128)	0	embedding_3[0][0]
get_item_6 (GetItem)	(None, 768)	0	bert_encoder_layer_2[…
get_item_7 (GetItem)	(None, 768)	0	bert_encoder_layer_2[…
lstm_3 (LSTM)	(None, 768)	2,755,584	dropout[0][0], get_item_6[0][0], get_item_7[0][0]
dropout_1 (Dropout)	(None, 768)	0	lstm_3[0][0]
dense_3 (Dense)	(None, 30522)	23,471,418	dropout_1[0][0]

Total params: 30,133,818 (114.95 MB)
Trainable params: 30,133,818 (114.95 MB)
Non-trainable params: 0 (0.00 B)

Figure 4.5 Model Summary built with BERT

An early-stopping callback function was defined to reduce the completion time for each epoch to create the epoch loops. The input IDs and attention masks were then sliced to match the input and output data lengths. After ensuring that the output data was formatted as an integer array without one-hot encoding, the

shapes of the input data, input IDs, attention masks, and output data were verified before initiating the epoch loops.

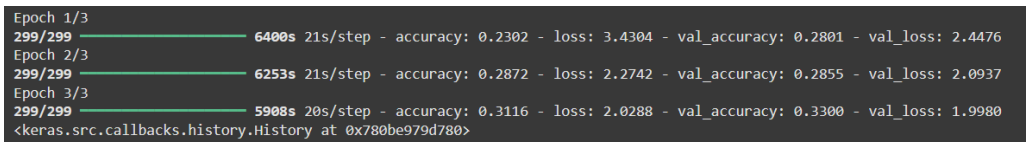

Figure 4.6 Training Details from the Model built with BERT

Unfortunately, since the BERT model was not designed for text generation, it struggled to produce cohesive and complete sentences. Instead, it generated iterations of single, random words and unstructured tuples containing jargon characters, which led to the decision to switch to the GPT-2 model.

4.3 BUILDING A MODEL WITH GPT-2

The third approach was to build the model with GPT-2. It is a large transformer-based language model developed by OpenAI. It is designed for text generation tasks, trained to predict the next word in a sentence, enabling it to generate coherent and human-like text (*OpenAI GPT2 — Transformers 3.0.2 Documentation*, n.d.).

After loading the pre-processed data saved from the pickle files, the data was split into three groups:

- Training set (70%)
- Test set (15%)
- Validation set (15%)

Next, the GPT-2 tokenizer model was used to process the data. The padding token was set to the End of Sequence (EOS) token for consistency. A preprocessing function was used to tokenize the data and generate input-output pairs for the GPT-2 model. The preprocessed tuples were converted into a list of strings, which were then tokenized and either truncated or padded to the maximum sequence length. The input sequences were then padded and

converted into a NumPy array suitable for training. To accelerate the model training, an optimizer from the Transformers library was used to fine-tune GPT-2 on a lower learning rate.

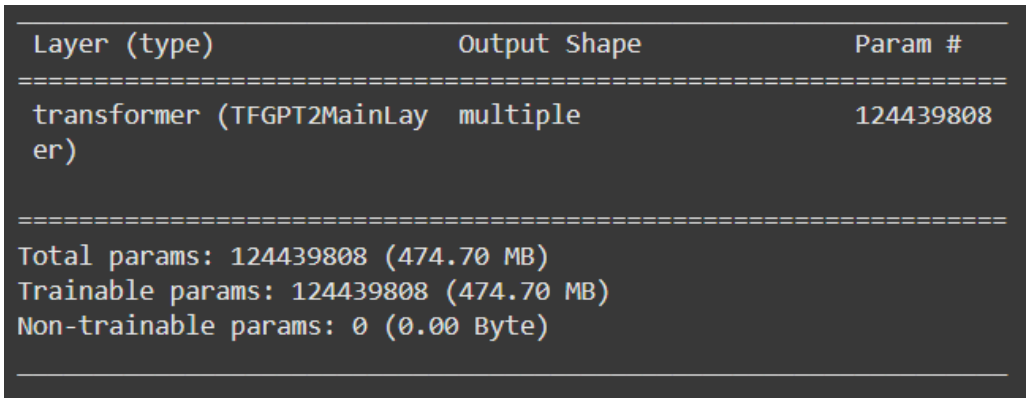

Figure 4.7 The GPT-2 Model build Summary

To reduce memory usage, techniques such as using a mixed precision policy, reducing batch size, decreasing the sequence length, and applying gradient accumulation to simulate larger batch sizes were used. After implementing the training loops for each epoch, the average training loss and validation loss were calculated to summarize the model performance.

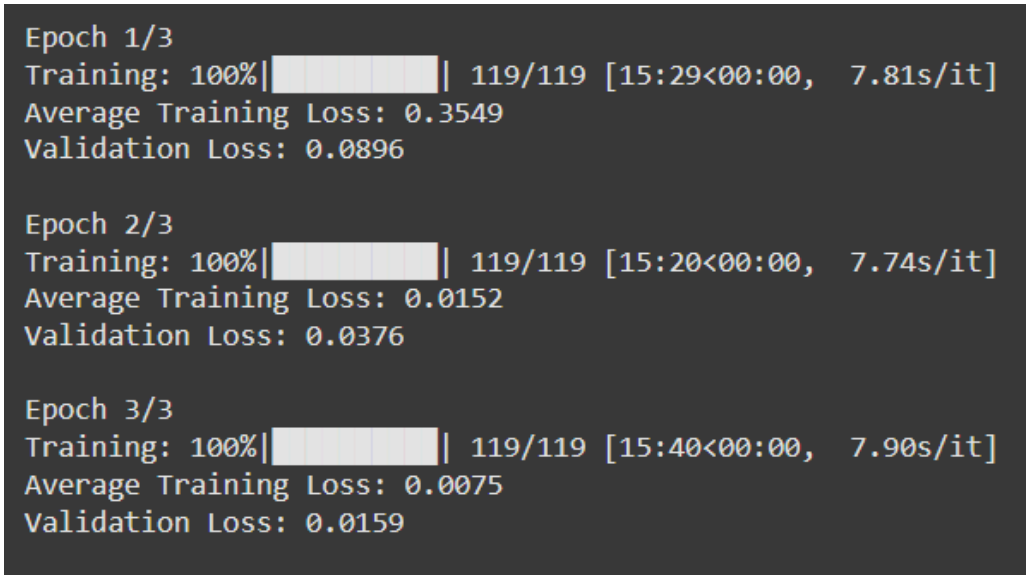

Figure 4.8 Training Detail from the Model built with GPT-2

Although this model was able to produce complete and coherent sentences, it struggled to utilize the provided dataset due to its relatively small size for this research due to the size compared to the overall model. As a result, the model generated sentences based on its pre-trained data rather than the dataset intended for this research.

CHAPTER 5

5. DISCUSSION

The analysis of student comments revealed several recurring themes such as the need for more practical examples and improved course organization. These findings align with the prior research on student feedback mechanisms, underscoring their importance in curriculum development. A significant observation was the difference in feedback quality between comments of fewer than 10 words and those exceeding this threshold. Longer comments provided more actionable insights, justifying the decision to handle them separately in the analysis pipeline. Grouping the responses by metadata such as department and semester was a crucial step in the analysis process since it enabled a more structured approach to identify trends in the data. Although the overall trends in sentiment could not be conclusively determined, the grouping methodology highlights the potential for future investigations into course-specific or section-specific feedback trends. The combination of Llama 2 and RoBERTa models proved effective in summarizing and analyzing the data, particularly when applied to open-ended responses. This approach demonstrated the potential of using advanced NLP techniques for improving feedback analysis in educational settings. The preprocessing techniques, including language-based separation and metadata grouping, enhanced the clarity and utility of the dataset, ensuring more accurate and relevant insights. One of the primary limitations encountered in this study was the insufficient size of the dataset, which impacted the performance of the models. LLMs and deep learning architectures such as BERT and GPT-2, typically require substantial amounts of data to fine-tune effectively. The limited data availability constrained the model's ability to generalize and capture nuanced patterns in the feedback. The lack of diverse and representative data also made it challenging to achieve robust SA and opinion categorization, potentially leading to underperformance in detecting subtle trends or variations.

This limitation underscored the importance of obtaining larger datasets in future work to enhance model accuracy and reliability of insights. Compared to traditional manual methods of analyzing student feedback, the automated approach significantly reduced processing time while maintaining high relevance and accuracy in identifying key themes. Unlike prior studies that focused solely on quantitative feedback, this research emphasizes the value of qualitative analysis, offering a more comprehensive view of student experiences.

CONCLUSION AND SUGGESTIONS

This research explored the use of advanced NLP techniques to analyze open-ended student survey responses and demonstrated the effectiveness of combining summarization and SA to extract meaningful information. The analysis highlights that working with LLMs is significantly more user-friendly compared to Keras, BERT, and GPT-2 models in this domain, particularly in terms of implementation and adaptability to the task. The study underscores the potential of leveraging LLMs for handling complex NLP tasks with greater ease and efficiency than traditional architectures. Insights generated from this research can inform instructors and administrators about specific areas of improvement, enabling targeted interventions to enhance course delivery and student satisfaction. Future work should focus on integrating more objective evaluation metrics to validate insights generated by the models. Exploring other NLP models or hybrid approaches such as combining traditional statistical methods with deep learning, could enhance the analysis. Extending this approach to include longitudinal data analysis would provide insights into how feedback trends evolve as time passes. Incorporating visual or audio-based feedback in addition to textual responses could broaden the scope of this research, providing a more comprehensive understanding of student opinions. In conclusion, this research has demonstrated the potential of NLP techniques in educational feedback analysis, paving the way for more effective and efficient approaches to understanding student needs and improving learning experiences.

REFERENCES

Alchemer. (n.d.). *Challenges of Analyzing Open-Ended Survey Responses*. Retrieved December 25, 2024, from https://www.alchemer.com/resources/blog/the-challenge-of-analyzing-open-ended-survey-questions

Bhargavi. (n.d.). *Spotlight on Student Course Evaluation Surveys: Why They Matter - piHappiness*. Retrieved December 25, 2024, from https://www.pihappiness.com/spotlight-on-student-course-evaluation-surveys-why-they-matter

Blyakhman, A. (n.d.). *Five Approaches to Sentiment Analysis | IMA*. Retrieved January 2, 2025, from https://www.sfmagazine.com/articles/2023/june/five-approaches-to-sentiment-analysis

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *2020-December*. https://arxiv.org/abs/2005.14165v4

Chollet, F., & others. (2015). *Keras*. https://keras.io/

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, A., & Lample, G. (2019). Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems*, *32*. https://arxiv.org/abs/1901.07291v1

Dake, D. K., & Gyimah, E. (2022). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies*, *28*(4), 4629. https://doi.org/10.1007/S10639-022-11349-1

Determ. (n.d.). *Sentiment Analysis Challenges: Solutions and Approaches - Determ*. Retrieved January 2, 2025, from https://determ.com/blog/sentiment-analysis-challenges-solutions-and-approaches

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*. https://arxiv.org/abs/1810.04805v2

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

Dovetail Editorial Team. (n.d.). *How to Code & Analyze Open-Ended Questions*. Retrieved January 2, 2025, from https://dovetail.com/surveys/how-to-code-open-ended-survey-questions

Fine-Tuning Large Language Models: Future Trends and Challenges. (n.d.). Retrieved December 21, 2024, from https://gpttutorpro.com/fine-tuning-large-language-models-future-trends-and-challenges

Go, A., Bhayani, R., & Huang, L. (n.d.). *Twitter Sentiment Classification using Distant Supervision*. Retrieved February 16, 2025, from https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems*, *35*. https://arxiv.org/abs/2203.15556v1

Hugging Face. (n.d.). *cardiffnlp/twitter-roberta-base-sentiment-latest · Hugging Face*. Retrieved December 5, 2024, from https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

Jordan, D. W. (2011). *Re-thinking Student Written Comments in Course Evaluations: Text Mining Unstructured Data for Program and Institutional Assessment Certification of Approval* Unpublished Doctor of Education Thesis, California State University. https://scholarworks.calstate.edu/downloads/3b591952b

Kastrati, Z., Dalipi, F., Imran, A. S., Nuci, K. P., & Wani, M. A. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences*, *11*(9), 3986. https://doi.org/10.3390/APP11093986

Katz, A., Norris, M., Alsharif, A. M., Klopfer, M. D., & Grohs, J. R. (n.d.). *Using Natural Language Processing to Facilitate Student Feedback Analysis*. https://peer.asee.org/using-natural-language-processing-to-facilitate-student-feedback-analysis.pdf

Leivada, E., Marcus, G., Günther, F., & Murphy, E. (2023). *A Sentence is Worth a Thousand Pictures: Can Large Language Models Understand Hum4n L4ngu4ge and the W0rld behind W0rds?* https://arxiv.org/abs/2308.00109v2

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. https://doi.org/10.1007/978-3-031-02145-9

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692v1

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. https://nlp.stanford.edu/IR-book/

Mello, C., Gullal, ·, Cheema, S., Gaurish Thakkar, ·, Cheema, G. S., & Thakkar, G. (2022). Combining sentiment analysis classifiers to explore multilingual news articles covering London 2012 and Rio 2016 Olympics. *International Journal of Digital Humanities*, *5*(2). https://doi.org/10.1007/S42803-022-00052-9

meta-llama/Llama-2-7b-hf · Hugging Face. (n.d.). Retrieved December 10, 2024, from https://huggingface.co/meta-llama/Llama-2-7b-hf

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large Language Models: A Survey*. http://arxiv.org/abs/2402.06196

Narayanan Venkit, P., Srinath, M., Gautam, S., Venkatraman, S., Gupta, V., Passonneau, R. J., & Wilson, S. (n.d.). *The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis*. https://arxiv.org/abs/2310.12318

Nevedal, A. L., Reardon, C. M., Opra Widerquist, M. A., Jackson, G. L., Cutrona, S. L., White, B. S., & Damschroder, L. J. (2021). Rapid versus traditional qualitative analysis using the Consolidated Framework for Implementation Research (CFIR). *Implementation Science*, *16*(1). https://doi.org/10.1186/S13012-021-01111-5/TABLES/5

Nozza, D., Bianchi, F., & Hovy, D. (2020). *What the [MASK]? Making Sense of Language-Specific BERT Models*. https://arxiv.org/abs/2003.02912v1

OpenAI GPT2 — transformers 3.0.2 documentation. (n.d.). Retrieved February 1, 2025, from https://huggingface.co/transformers/v3.0.2/model_doc/gpt2.html

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(2). https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

Pang, B., Lee, L., & Vaithyanathan, S. (n.d.). *Thumbs up? Sentiment Classification using Machine Learning Techniques*. Retrieved February 1, 2025, from http://reviews.imdb.com/Reviews/

Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2023). A Large Language Model Approach to Educational Survey Feedback Analysis. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00414-0

Qualtrics. (n.d.). *Sentiment Analysis and How to Leverage It - Qualtrics*. Retrieved January 2, 2025, from https://www.qualtrics.com/experience-management/research/sentiment-analysis

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language Models are Unsupervised Multitask Learners*. Retrieved December 18, 2024, from https://github.com/codelucas/newspaper

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*. https://arxiv.org/abs/1910.10683v4

SurveyMonkey. (n.d.). *3 natural language processing use cases for surveys | SurveyMonkey*. Retrieved December 26, 2024, from https://www.surveymonkey.com/mp/natural-language-processing

Tejwani, R. (2014). Sentiment Analysis: A Survey. *International Journal for Research in Applied Science and Engineering Technology*, *V*(VIII), 1957–1963. https://doi.org/10.22214/ijraset.2017.8276

Thematic. (n.d.). *A complete guide to Sentiment Analysis approaches with AI | Thematic*. Retrieved January 2, 2025, from https://getthematic.com/sentiment-analysis

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://arxiv.org/abs/2307.09288v2

Upadhye, A. (2022). A Comprehensive Survey of Sentiment Analysis Methods. *International Journal of Science and Research (IJSR), 11*(2), 1318–1322. https://doi.org/10.21275/SR24401234546

Voigt, P., & Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. https://link.springer.com/book/10.1007/978-3-319-57959-7

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). *Large Language Models for Education: A Survey and Outlook*. http://arxiv.org/abs/2403.18105

Widmer, B. (n.d.). *Blix*. Retrieved January 2, 2025, from https://blix.ai/blog/verbatim-coding

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*. https://arxiv.org/abs/1909.10430v2

Yadav, K., Kumar, N., Maddikunta, P. K. R., & Gadekallu, T. R. (2020). A Comprehensive Survey on Aspect Based Sentiment Analysis. *International Journal of Engineering Systems Modelling and Simulation*, *12*(4). https://doi.org/10.1504/IJESMS.2021.119892

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (n.d.). *Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization*. https://arxiv.org/abs/2302.08081

CURRICULUM VITAE