University of Salford
MANCHESTER

MSc Data Science Dissertation

Title

Measuring Effectiveness of Feedback with NLP

Likitha Tallapally

May 2024

Title

Measuring Effectiveness of Feedback with NLP

By Likitha Tallapally

Supervisor: Dr Taha Mansouri

School of Science, Engineering and Environment

University of Salford, Manchester, United Kingdom

Academic Year: 2023 - 2024

Abstract

Effective feedback is crucial for fostering student learning and development in higher education. However, evaluating the impact of feedback on student outcomes remains challenging due to subjective assessment methods and the lack of scalable evaluation techniques. This dissertation investigates the potential of NLP to objectively measure feedback effectiveness in educational settings.

The study begins with a comprehensive review of literature on feedback efficacy and NLP techniques for sentiment analysis, providing a theoretical foundation for the research. Using a mixed-methods approach, the research collects a dataset of task-related feedback provided by educators and employs three state-of-the-art NLP models BERT, RoBERTa, and GPT-3.5 Turbo to analyse feedback sentiment.

Through fine-tuning and comparative analysis of these models, the research evaluates their performance in classifying feedback as encouraging, neutral, or critical. Additionally, human judgments on feedback effectiveness are collected to validate the NLP models' predictions and provide insights into their accuracy and reliability.

The findings of this study contribute to the understanding of how NLP can enhance feedback evaluation practices in higher education. The development of NLP-based tools for feedback analysis has the potential to revolutionize feedback practices by providing educators with objective, data-driven insights into feedback effectiveness. Recommendations for future research include expanding the dataset size, integrating multimodal data sources, and refining NLP models for feedback recommendation systems. Overall, this dissertation offers valuable insights into leveraging NLP for improving feedback practices and promoting student success in educational settings.

Acknowledgement

I would like to acknowledge the guiding hand of God Almighty throughout this dissertation journey. His unwavering presence provided me with strength, perseverance, and the wisdom to navigate challenges.

My deepest gratitude goes to Dr Taha Mansouri, my esteemed supervisor. Your invaluable guidance, insightful feedback, and unwavering support were instrumental in shaping this dissertation. Your dedication to my success has been truly inspiring.

I would also like to extend my sincere appreciation to the Data Science department at the University of Salford. The supportive environment and collaborative spirit fostered by the faculty and team members greatly enriched my learning experience.

Finally, a heartfelt thank you to my family and friends. Your unwavering love, encouragement, and understanding provided a constant source of strength. This dissertation would not have been possible without your unwavering belief in me.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

NLP	Natural language processing
AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly optimized BERT approach
GPT	Generative Pre-trained Transformer
EDA	Exploratory Data Analysis
AES	Automated Essay Scoring
RNN	Recurrent neural network
LSTM	Long Short-Term Memory
ELMo	Embeddings from Language Models
MLM	Masked Language Models
FLOPS	Floating-point operations
CSV	Comma-separated values

Chapter 1: Introduction

1.1 General Overview

Effective feedback plays a pivotal role in student learning and development within higher education. It empowers students to identify strengths, address weaknesses, and ultimately refine their understanding and performance. However, current practices in providing feedback exhibit significant variation, ranging from highlighting shortcomings to fostering achievement of learning objectives. While Nicol & Macfarlane-Dick (2006) outlined seven principles for effective feedback, objectively measuring its impact on student learning remains a challenge.

However, current methods for evaluating feedback effectiveness often lack objectivity and data-driven insights. Traditional methods rely on subjective student surveys or self-reported learning gains (Bangert et al., 2002). While these methods offer some information, they are susceptible to bias and struggle to gauge the true impact of feedback on learning outcomes. For instance, students may be hesitant to provide negative feedback on their instructors, leading to skewed results. Additionally, self-reported learning gains can be influenced by factors beyond the quality of feedback received. There is a growing need for more objective and reliable methods for evaluating feedback effectiveness. This can provide valuable information for educators to refine their feedback practices and ensure they are fostering a learning environment that truly promotes student success.

Natural Language Processing (NLP) offers a promising approach to enhance feedback evaluation. NLP techniques can analyse the linguistic features of feedback and potentially identify elements associated with feedback perceived as more helpful by students. This information can provide valuable data for educators to refine their feedback practices and ensure they are fostering a learning environment that promotes student success.

By leveraging the analytical capabilities of NLP, this research project aims to develop a more objective and reliable method for measuring feedback effectiveness in higher education. NLP can analyse the language itself, looking for patterns associated with feedback that students perceive as more helpful. This can provide valuable insights for educators beyond what traditional methods can offer.

This research project addresses this challenge by exploring the potential of **NLP** techniques in objectively evaluating the effectiveness of specific task-related feedback. NLP, a subfield of Artificial Intelligence (AI) concerned with the interaction between computers and human language, offers a promising avenue for analysing and extracting meaning from textual data, including written feedback comments.

This chapter delves into the importance of effective feedback and the limitations of current assessment methods. It highlights the research motivation, problem statement, and the proposed solution utilizing NLP. The chapter then outlines the research objectives, chosen methodology, and expected outcomes. Finally, a flowchart depicting the research process and a chapter outline provide a roadmap for the dissertation.

1.2 Aims and Objectives

1.2.1 Overview

This section introduces the concept of feedback effectiveness in higher education, highlighting its significance in fostering learning objectives. It outlines the research problem, aims, and objectives of the research project, which investigate the potential of NLP pre-trained models in analysing feedback effectiveness and promoting positive learning outcomes.

1.2.2 Research Problem

Effective feedback lies at the heart of successful student learning and development. While educators strive to provide constructive and informative feedback, the effectiveness of these interactions often varies significantly. This disparity highlights the need for a more objective and data-driven approach to evaluating feedback effectiveness.

This research project addresses the crucial need to objectively assess the effectiveness of specific task-related feedback in promoting learning. By leveraging NLP techniques, the study aims to develop a method for analysing feedback and determining its suitability for fostering desired learning outcomes.

1.2.3 Research Aims and Objectives

The primary aim of this research is to explore and evaluate the effectiveness of specific task-related feedback provided to students in higher education settings through the application of NLP pre-trained models.

To achieve this aim, the following research objectives will be pursued:

- To conduct a comprehensive literature review to understand the essence of effective feedback for learning and explore potential NLP methods used to evaluate feedback efficacy.
- To review and compare available pre-trained NLP models for sentiment analysis, selecting models that closely align with the research problem and show potential for fine-tuning.
- To evaluate and compare the performance of fine-tuned NLP models in analysing feedback effectiveness.

To ensure the project's success, the objectives are defined using the SMART criteria:

Specific: Fine-tune pre-trained NLP models to analyse the language features of task-related feedback provided by educators in a higher education setting, with the objective of classifying feedback sentiment as Encouraging, Neutral, or Critical.

Measurable: The performance of the fine-tuned NLP models will be evaluated using a hold-out test set. Accuracy, metrics will be employed to assess the models' ability to predict the effectiveness of feedback as judged by students.

Achievable: Leverage readily available pre-trained NLP models like BERT, RoBERTa and OpenAI's GPT-3.5 turbo requiring fine-tuning rather than building entirely new models. A task-specific dataset of student feedback with human annotations on its effectiveness will be developed for model training.

Relevant: The findings will directly contribute to educator's practices by offering an NLP-based method for evaluating feedback sentiment (Encouraging, Neutral, Critical). This information can be used to refine feedback strategies and ensure they are fostering a positive and motivating learning environment for students. Understanding how sentiment in feedback impacts student perceptions can be crucial for educators to adjust their approach and maximize the effectiveness of their feedback.

Time-bound: Executing the project tasks within a defined timeline, ensuring timely completion and delivery of project outcomes.

1.3 Research Motivation

The motivation for this research stems from the limitations of traditional methods currently employed in evaluating feedback effectiveness. These methods, often relying on subjective evaluations, lack objectivity and consistency, hindering the ability of educators to accurately assess the impact of their feedback practices on student learning.

By developing an objective method for evaluating feedback effectiveness through NLP, this research offers several key benefits:

Improved Understanding of Effective Feedback: The proposed research aims to contribute to a deeper understanding of "how specific characteristics of feedback impact student learning". By analysing large amounts of feedback data using NLP techniques, the study can identify patterns and relationships between specific phrasing, sentiment, and student outcomes. This knowledge can inform the development of more effective feedback practices and refine existing pedagogical frameworks.

Development of AI-powered Tools: The findings of this research can guide the development of AI-powered tools for educators. These tools could be integrated into learning management systems or online platforms, providing educators with real-time feedback on the effectiveness of their comments. This would empower them to refine their feedback strategies on the fly, tailoring their approaches to individual student needs and promoting continuous improvement in their teaching practices.

Enhanced Learning Outcomes: By promoting the use of effective feedback practices through the insights gleaned from this research, educators can contribute to improved student learning outcomes and academic success**.** Students receiving feedback that is clear, specific, actionable, and motivational are better equipped to understand their strengths and weaknesses, identify areas for improvement, and ultimately achieve their learning goals.

Addressing Bias and Subjectivity: Traditional methods of evaluating feedback, often relying on subjective evaluations from instructors or students, can be susceptible to bias and inconsistency. NLP techniques offer a more objective and data-driven approach to analysing feedback, potentially mitigating the impact of individual biases and fostering a more reliable assessment of feedback effectiveness.

Scalability and Efficiency: NLP techniques offer the potential to analyse large volumes of feedback data efficiently**,** providing a comprehensive perspective on the effectiveness of different feedback approaches. This scalability can be particularly valuable in large institutions with numerous instructors and diverse student populations.

1.4 Approach

A mixed-methods approach will be employed to achieve the research objectives and answer the research questions. This approach combines quantitative analysis using NLP models with qualitative analysis of human judgments on feedback effectiveness.

Quantitative Analysis:

Pre-trained NLP Models:

- Three pre-trained NLP models will be explored: BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), RoBERTa (Robustly optimized BERT approach) (Liu et al., 2019), and OpenAI's GPT-3.5 turbo (Generative Pre-trained Transformer 3.5 Turbo) (OpenAI, 2022). These models have demonstrated strong performance in various NLP tasks, including text classification.
- Each model will be fine-tuned on a specifically prepared dataset of task-related feedback provided by educators in a higher education setting. Human annotations will be used to classify the feedback into categories representing different levels of effectiveness (e.g., Encouraging, Neutral and Critical).

Data Collection:

- Primary data collection is conducted through a targeted online questionnaire distributed to professors across various academic disciplines.
- The questionnaire captures demographic information, feedback experiences, and human annotations for sentiment classification, ensuring relevance to the research objectives.

Model Evaluation:

- **Test Set Creation:** A portion of the collected feedback data is reserved to create a hold-out test set, ensuring unbiased evaluation.
- **Performance Metrics:** Model performance is primarily assessed using accuracy, with additional metrics like precision, recall, and F1 score for a comprehensive evaluation.
- **Comparative Analysis:** The performance of each model is compared to identify the most accurate classifier for feedback effectiveness.

- **Interpretation:** Evaluation results provide insights into the models' capabilities and limitations, guiding further refinements if needed.

Qualitative Analysis:

A questionnaire will be distributed to gather their judgments on the effectiveness of various feedback examples. The feedback examples used in the questionnaire will be a subset of those included in the NLP model training dataset. This qualitative data will be used to compare with the predictions of the NLP models and assess their accuracy in identifying effective feedback.

Integration: The findings from the model evaluation and, if included, the qualitative analysis, will be integrated to develop a comprehensive understanding of the relationship between NLP analysis, student perceptions, and feedback effectiveness. This may involve identifying linguistic features that the most accurate NLP model associates with highly effective feedback, as judged by both students and the model itself.

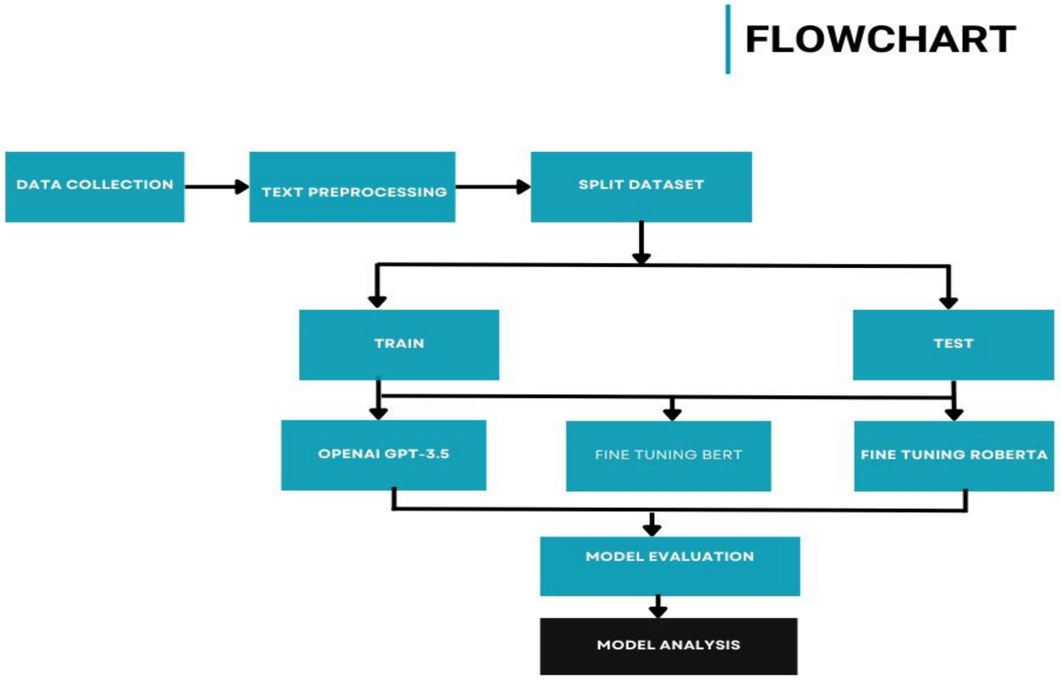


Figure1: Flowchart

Benefits of Mixed-methods Approach:
The mixed-methods approach offers several benefits:

- **Triangulation:** By combining quantitative and qualitative data, the research can achieve a more comprehensive understanding of feedback effectiveness.
- **Enhanced Accuracy:** Comparing NLP model predictions with human judgments allows for evaluation of model accuracy and identification of areas for improvement.
- **Real-world Relevance:** The inclusion of human judgments ensures the research findings are grounded in real-world perceptions of effective feedback.

1.5 Analysis Plan

1.5.1 Analytical Models and Algorithms

BERT: BERT is a transformer-based model introduced by Google in 2018. It is pre-trained on large text corpora using masked language modelling and next sentence prediction tasks. BERT has the ability to capture bidirectional context in text data, making it well-suited for various NLP tasks, including sentiment analysis. In this research, BERT will be fine-tuned on the collected feedback data to classify the sentiment of feedback comments into categories such as Encouraging, Neutral, and Critical.

RoBERTa: RoBERTa is an optimized version of BERT developed by Facebook AI in 2019. It addresses some limitations of BERT's pre-training methodology and hyperparameter settings, leading to improved performance on downstream NLP tasks. RoBERTa leverages the same transformer architecture as BERT but employs different training objectives and hyperparameters. Similar to BERT, RoBERTa will be fine-tuned on the feedback data for sentiment analysis.

GPT-3.5 Turbo: GPT-3.5 Turbo is a variant of OpenAI's GPT-3 model, enhanced with additional training data and fine-tuning techniques. GPT-3 is a generative transformer model known for its ability to generate coherent and contextually relevant text based on a given prompt. GPT-3.5 Turbo will be used for sentiment analysis by conditioning the generation process on the feedback comments and predicting the associated sentiment labels.

1.5.2 Tools, Techniques, and Libraries

Python Programming Language: Python is a widely used programming language in the field of machine learning and natural language processing due to its simplicity, readability, and extensive ecosystem of libraries. Python will be utilized for data preprocessing, model training, evaluation, and analysis.

Google Colab: Google Colab is a cloud-based Jupyter notebook environment provided by Google, which offers free access to GPU resources for running Python code. It provides a convenient platform for training deep learning models, particularly for researchers and practitioners who may not have access to high-performance computing resources. Google Colab will be used for training and evaluating BERT, RoBERTa, and GPT-3.5 Turbo models on the feedback data.

Transformers Library: The Transformers library, developed by Hugging Face, provides a high-level interface for working with pre-trained transformer models such as BERT, RoBERTa, and GPT. It offers functionalities for fine-tuning these models on custom datasets, performing inference, and evaluating model performance on various NLP tasks. The Transformers library will be utilized for implementing the sentiment analysis models and conducting experiments.

1.5.3 Justification and Validation

The selection of BERT, RoBERTa, and GPT-3.5 Turbo as analytical models is justified by their state-of-the-art performance and widespread adoption in the NLP community. These models have been extensively benchmarked on various NLP tasks and datasets, demonstrating superior performance compared to traditional machine learning approaches. Furthermore, the availability of pre-trained weights and fine-tuning capabilities allows researchers to leverage the knowledge encoded in these models for specific tasks such as sentiment analysis.

The use of Python programming language and Google Colab as tools for analysis is justified by their popularity, ease of use, and accessibility. Python has a vast ecosystem of libraries for data analysis, machine learning, and NLP, making it a preferred choice for researchers and practitioners. Google Colab provides free GPU resources, enabling researchers to train and evaluate deep learning models without significant computational costs.

Validation of the analytical models will be performed through rigorous experimentation and evaluation. This includes training the models on a portion of the collected feedback data, validating their performance on a separate validation set, and fine-tuning hyperparameters to optimize performance metrics such as accuracy and F1 score. Additionally, the models will be evaluated on a hold-out test set to assess their generalization ability to unseen data. Cross-validation techniques may also be employed to ensure the robustness and reliability of the models' performance estimates.

1.5.4 Advantages of Using Multiple NLP Models

This research takes a unique approach by fine-tuning and comparing three distinct NLP models: BERT, RoBERTa, and GPT-3.5 turbo.
This offers several advantages:

Comparative Analysis: By evaluating the performance of multiple models, the research can identify which model is most effective for classifying feedback effectiveness in this specific context of higher education. This allows for a more robust and reliable conclusion regarding the potential of NLP for feedback evaluation.

Leveraging Model Strengths: Each model possesses different strengths and weaknesses. BERT excels at understanding contextual relationships in language, while RoBERTa focuses on long-range dependencies. GPT-3.5 turbo, with its massive training dataset, offers strong capabilities in generating human-quality text, which may translate to improved analysis of nuanced feedback elements. By fine-tuning these models, the research can leverage their combined strengths to capture the complexities of feedback language.

Addressing Model Bias: NLP models are not without limitations, and bias can be a concern. By utilizing multiple models from different developers, the research can mitigate the potential for bias inherent in any single model. Analysing the results across all three models can provide a more balanced perspective on the effectiveness of NLP for feedback classification.

1.5.5 Significance and Expected Outcomes

This research project holds significant value for improving feedback practices in higher education. The anticipated outcomes include:

Fine-tuned NLP Models: Three fine-tuned NLP models capable of classifying feedback effectiveness based on language features. The most effective model will be identified for further exploration and potential implementation.

NLP Applications for Feedback Evaluation: A critical evaluation of the potential and limitations of utilizing NLP to measure feedback effectiveness. This can pave the way for the development of practical tools and strategies for educators.

Improved Feedback Practices: Recommendations for educators on how to refine their feedback practices based on NLP analysis and student perceptions. This ultimately contributes to a learning environment where effective feedback empowers students to reach their full potential.

By achieving these outcomes, this research project can become a valuable resource for educators seeking to leverage NLP advancements to enhance the quality and impact of feedback in higher education. The insights gained can further contribute to ongoing efforts in educational technology, aiming to improve learning experiences and student outcomes.

1.6 Thesis Outline

This thesis is structured as follows:

Chapter 2 Background and Literature Review: This chapter delves into the existing literature related to feedback in higher education. It explores key concepts, theoretical frameworks, and previous research findings on feedback effectiveness. The chapter identifies gaps in the literature and justifies the need for the current study.

Chapter 3 Methodology: In this chapter, the research methodology is outlined in detail. It describes the approach adopted to evaluate feedback effectiveness using NLP techniques. The chapter covers data collection methods, including questionnaire design and ethical considerations. It also discusses data analysis techniques and tools.

Chapter 4 Modelling and Evaluation: This chapter focuses on the modelling and evaluation process. It describes the selection of NLP models, algorithms, and techniques used for analysing feedback data. The chapter presents the results of model evaluation, including performance metrics and comparative analysis.

Chapter 5 Conclusions and References: The final chapter summarizes the main findings of the study and discusses their implications. It addresses the research questions and objectives, highlighting key insights and contributions. The chapter also provides recommendations for future research. Additionally, it includes a comprehensive list of references cited throughout the thesis.

1.7 Ethical Considerations

The research project adheres to ethical guidelines by ensuring transparency, accountability, and participant confidentiality. *Ethical clearance was obtained from the ethics panel*, and participants were provided with a participant information sheet and consent form detailing the research objectives, data collection procedures, and their rights as participants. As no personal information was collected, the study poses minimal risk to participants. Data sharing was limited to the research team, with the *dataset shared solely with the supervisor for analysis and feedback*. Measures were taken to protect participant confidentiality, and all data handling procedures complied with relevant data protection regulations. Additionally, efforts were made to mitigate biases in the research process and to ensure the responsible use of AI techniques. Throughout the project, ethical considerations remained a priority, with continual evaluation and improvement of research practices.

1.8 Conclusion

This chapter sets the stage for the research by highlighting the significance of effective feedback in higher education and the challenges in objectively evaluating its impact. It identifies the limitations of traditional evaluation methods and proposes leveraging NLP techniques to overcome these challenges. It outlines specific aims and objectives, ensuring clarity and alignment with the SMART criteria. Motivation for the research stems from the need for more objective feedback evaluation methods and the potential benefits of NLP in enhancing feedback practices. The chosen approach combines quantitative analysis using three pre-trained NLP models with qualitative insights, aiming for a comprehensive understanding of feedback effectiveness.

The analysis plan details the selected analytical models, algorithms, tools, techniques, and libraries, providing justification for their use. Three prominent NLP models BERT, RoBERTa, and GPT-3.5 Turbo—are chosen for their demonstrated effectiveness in various NLP tasks. Python programming language and Google Colab are selected as tools for their accessibility and suitability for NLP tasks. The significance of using multiple NLP models lies in their complementary strengths and the ability to provide a more robust evaluation of feedback effectiveness.

Overall, Chapter 1 serves as an introduction to the research, outlining the problem statement, objectives, methodology, and expected outcomes in preparation for the subsequent chapters. The following chapter Background and Literature Review, provides a comprehensive overview of existing research on feedback in higher education, setting the stage for the current study.

Chapter 2 Background and Literature Review

2.1 Introduction

Quality education requires personal attention and support. A crucial element of this is feedback, whose definition changes depending on the educational research literature. Within this research, feedback is conceptualized as the flow of information from one agent to another regarding a learner decision (Hattie & Timperley, 2007). Numerous studies have highlighted the significant role feedback plays in the learning process. Price et al. (2010) state that feedback is the most critical component of assessments. Likewise, a meta-analytic study of 450,000 effect sizes across 180,000 studies concluded that feedback was the biggest contributor to achievement (Hattie, 1999). Feedback directs learners to the appropriate type of study or practice and helps individuals recognize areas of deficiency, which can be used to enhance learning tactics and strategies (Parikh et al., 2001; Weaver, 2006; Glover & Brown, 2006). Dweck (1999) reports feedback can affect a student's motivation, as well as what is learned and in what manner. Butler and Winne (1995) theorize that feedback promotes learning by scaffolding consistent beliefs, developing prior knowledge, or correcting inconsistent beliefs. Laurillard emphasized "action without feedback is completely unproductive for the learner" (Laurillard, 2013, p. 61).

Despite widespread recognition of feedback's importance to learning, much of the current literature indicates a pervasiveness of low-quality feedback in higher education (Hattie & Gan, 2011). Feedback quality is consistently rated one of the greatest causes of dissatisfaction for higher education students (Ferguson, 2011). Weaver (2006) reports that while collegiate scholars acknowledge the value of feedback in facilitating learning, they find instructors' feedback comments to be incomprehensible and ineffectual. Ferguson (2011) identifies a lack of timely feedback delivery, unclear expectations and low utility as key concerns amongst learners. Mulliner and Tucker (2017) conducted a survey of higher educational academic staff and students that found 93% of staff being satisfied with the quality of feedback provided, compared to just 67% of students being satisfied with the quality of feedback received. As higher educational institutions embrace technology, there is a growing portfolio of approaches that utilize ubiquitous data collection to improve learning processes and design decisions.

2.2 Background

Higher education institutions have a fundamental responsibility to equip students with the knowledge, skills, and critical thinking abilities necessary to navigate the complexities of the contemporary world. Achieving this goal hinges on effective teaching and learning practices, where **Feedback** plays a pivotal role in shaping student development and fostering their understanding of subject matter (Nicol & Macfarlane-Dick, 2006).

Feedback, defined as "information provided to a learner about their performance with the goal of improving future performance" (Nicol & Topping, 2007, p. 159), encompasses a broad spectrum of interactions between educators and students. It can take various forms, including written comments on assignments, one-on-one dialogue, peer feedback sessions, and online discussions. Regardless of the format, the underlying purpose of feedback remains steadfast: to provide students with insights into their learning journey, identify areas for improvement, and ultimately propel them towards achieving their academic goals.

However, the effectiveness of feedback can vary considerably (Winstrom & Nash, 2010). While some feedback approaches provide constructive guidance and foster learning, others may offer limited value or even hinder student motivation and growth. This variability stems from several factors, including:

- **Clarity and Specificity:** Feedback that lacks clarity or specificity can leave students confused and unsure of how to improve.
- **Focus on Learning:** Feedback that solely emphasizes deficiencies without offering actionable steps for improvement limits its effectiveness.
- **Motivation and Encouragement:** Feedback that adopts a solely critical tone can demotivate students and diminish their learning enthusiasm.
- **Individual Differences:** Feedback tailored to the specific needs and learning styles of individual students yields better results than a one-size-fits-all approach.

Recognizing the critical role of effective feedback in facilitating student learning, educational researchers and practitioners have dedicated significant efforts to understanding and fostering its efficacy. Numerous frameworks and guidelines have been established to guide educators in providing constructive and informative feedback. Nicol & Macfarlane-Dick's (2006) seven principles of effective feedback, for instance, emphasize aspects like clarifying performance expectations, cultivating self-assessment and reflection, and fostering positive motivation and self-esteem. These principles equip educators with a robust framework to craft feedback that promotes meaningful student learning.

Despite the comprehensive theoretical frameworks and pedagogical guidance, ensuring consistency and objectivity in evaluating the effectiveness of feedback remains a challenge in contemporary pedagogical practice. Traditional methods, often reliant on instructor intuition or student self-reports, can be subjective and prone to bias.

2.3 Existing Feedback Models: A Historical Overview

Understanding feedback models requires acknowledging the historical context of how feedback theory has evolved. Early models primarily focused on the transmission of information from teacher to student, with limited emphasis on student participation or response (Angelo & Cross, 1993). One such model, the "Feedback Sandwich Model," emphasizes delivering positive feedback first, followed by constructive criticism, and concluding with further positive reinforcement (Wlodkowski & Ginsberg, 1995). While this model offers a structured approach for delivering feedback, it can feel formulaic and may overshadow the importance of personalized and targeted communication.

A shift towards more interactive models emerged, emphasizing student engagement and the cyclical nature of the feedback process. The "Proactive Feedback Model" by Butler (1988) highlights the importance of setting clear learning goals, providing ongoing feedback throughout the learning process, and encouraging students to self-monitor their progress. Similarly, Nicol & Macfarlane-Dick's (2006) seven principles for effective feedback emphasize its formative nature, focusing on how students can use feedback to improve their learning. These models recognize the dynamic role of feedback and encourage a collaborative approach between educators and students.

With the growing emphasis on self-regulated learning, models like Zimmerman & Schunk's (2008) Self-Regulated Learning Model highlight the role of feedback in promoting student self-awareness, goal setting, and self-evaluation. This model positions feedback as a tool for students to take ownership of their learning journey and utilize it for continuous improvement.

2.4 Implementation of Feedback Models: Challenges and Considerations

While existing feedback models offer valuable frameworks, their implementation in practice often presents challenges. Educators may struggle to find the time and resources required to provide frequent and detailed feedback, particularly in large classes (Bangert et al., 2002). Additionally, translating theoretical models into actionable strategies for diverse learning contexts and student needs can be difficult (Sadler, 2010). Furthermore, ensuring students actively engage with and utilize feedback to inform their learning requires fostering a classroom culture that values open communication and self-reflection (Brookhart, 2010).

Existing methods for evaluating feedback effectiveness also face limitations. Traditional methods often rely on subjective student surveys or self-reported learning gains (Bangert et al., 2002). While these methods offer insights into student perceptions, they can be susceptible to bias and lack accuracy in gauging the actual impact of feedback on learning outcomes.

Another approach involves using rubrics. Rubrics provide a structured framework for evaluating student work based on pre-defined criteria (Brookhart, 2010). While rubrics can ensure consistency in feedback delivery, they may not capture the richness of language that can influence the effectiveness of feedback. Additionally, rubrics often focus on assessing student work rather than the impact of the feedback itself.

2.5 Existing Models Relate to the Current Research

The existing body of research on feedback models lays the groundwork for investigating the innovative application of NLP in evaluating feedback effectiveness. Previous models have emphasized the importance of feedback that is clear, specific, and action-oriented (Nicol & Macfarlane-Dick, 2006). They highlight the benefits of feedback that promotes student self-reflection and engagement in the learning process (Zimmerman & Schunk, 2008). However, existing assessment methods often fall short of objectively measuring the true impact of these characteristics on student learning.

This research project aims to leverage NLP techniques to address this critical gap. By analysing the language features of feedback, NLP models can potentially identify elements associated with feedback perceived as more helpful by students. This information can then be used to provide more objective and data-driven insights into the effectiveness of specific feedback strategies.

2.6 Existing NLP Applications in Education and Limitations

NLP techniques have found growing applications in various educational settings, offering innovative approaches to enhance learning experiences (Baker et al., 2017). Here are some prominent examples:

- **Automated Essay Scoring (AES):** These systems utilize NLP techniques to analyse student essays and provide feedback on factors such as grammar, vocabulary usage, and overall writing structure (Xiang et al., 2018). While AES systems can offer some level of automated feedback, they often struggle to capture the nuances of human language and may overlook elements like critical thinking and creativity (Landauer, 2000).
- **Personalized Learning:** NLP techniques can analyse student responses to learning materials and identify areas of confusion or difficulty (Gong et al., 2019). This information can then be used to tailor subsequent learning activities and provide personalized feedback that addresses individual student needs.
- **Chatbots:** NLP is employed to develop chatbots that can answer student questions, offer basic guidance, and simulate conversations to enhance learning engagement (Liu et al., 2020). However, chatbots may lack the ability to handle complex inquiries or provide in-depth feedback on student work.

Limitations of Existing NLP Applications in Education:

While NLP holds significant promise for educational applications, some limitations need to be acknowledged:

- **Data dependency:** The effectiveness of NLP models heavily relies on the quality and quantity of training data. Limited or biased data can lead to inaccurate or unfair outcomes.
- **Black box nature:** Understanding the reasoning behind NLP model predictions can be challenging. This lack of transparency can be problematic in educational settings where educators need to justify feedback provided to students.
- **Focus on language mechanics:** Current NLP applications often prioritize analysing surface-level language features like grammar and vocabulary. Capturing the deeper meaning and intent behind student writing or feedback requires further development of NLP techniques.

2.7 Existing Feedback Evaluation Models using NLP: A Critical Analysis

Recent research has begun to explore the potential of NLP for analysing feedback effectiveness. Ventura et al. (2018) and Jokela et al. (2019) utilized NLP models to analyse the sentiment and language features of feedback provided by teachers. Their findings suggest that certain language patterns, such as encouraging phrases and specific action verbs, correlate with feedback perceived as more helpful by students.

However, existing research in this area presents some limitations:

- **Limited scope:** These studies often focus on analysing feedback from a specific context or discipline. Further research is needed to explore the generalizability of these findings across diverse learning environments.
- **Focus on sentiment analysis:** While sentiment analysis provides valuable insights, it may not capture the full complexity of effective feedback. Future research could explore more advanced NLP techniques like aspect-based sentiment analysis to identify specific aspects of feedback that students find most beneficial.
- **Integration with existing models:** A critical challenge lies in integrating NLP-based feedback evaluation with existing models of feedback design and delivery. Understanding how NLP insights can inform educator practices and improve overall feedback strategies is essential.

2.7.1 Advancements in Natural Language Processing

Advancements in Natural Language Processing have significantly transformed the capabilities of language-based AI systems, enabling them to understand, generate, and respond to human language more effectively. Here are some key advancements in NLP:

Deep Learning Techniques: Deep learning algorithms, particularly neural networks, have revolutionized NLP by enabling models to learn complex patterns and representations from large datasets. RNNs, LSTM networks, and Transformer architectures (e.g., BERT, GPT) have shown remarkable performance in various NLP tasks such as text classification, language modelling, and machine translation.

Pre-trained Language Models: Pre-trained language models, such as BERT and GPT have emerged as powerful tools for various NLP tasks. These models are trained on massive amounts of text data and can be fine-tuned for specific downstream tasks, resulting in significant performance improvements with minimal task-specific training data.

Transfer Learning and Fine-tuning: Transfer learning techniques allow NLP models to leverage knowledge learned from one task or domain to improve performance on another task or domain. Fine-tuning pre-trained language models on specific datasets or tasks has become a common practice, enabling rapid development of high-performance NLP systems with limited labelled data.

Contextual Word Embeddings: Contextual word embeddings capture the meaning of words based on their context within a sentence or document. Models like ELMo and GPT incorporate contextual information into word embeddings, resulting in more nuanced representations that capture semantic relationships and syntactic structures more accurately.

Attention Mechanisms: Attention mechanisms allow NLP models to focus on relevant parts of input sequences while processing them. Transformer architectures, such as BERT and GPT, utilize attention mechanisms to capture dependencies between words and generate context-aware representations of input sequences.

Multimodal NLP: Multimodal NLP involves processing and understanding information from multiple modalities, such as text, images, and audio. Advances in multimodal NLP have led to the development of models capable of understanding and generating text based on visual and auditory inputs, enabling applications like image captioning and speech-to-text translation.

Zero-shot and Few-shot Learning: Zero-shot and few-shot learning techniques allow NLP models to generalize to unseen tasks or domains with minimal training data. Models like GPT-

3 have demonstrated the ability to perform various NLP tasks without task-specific training, relying on general linguistic knowledge acquired during pre-training.

Overall, these advancements in NLP have propelled the development of more sophisticated and versatile language-based AI systems, enabling them to tackle a wide range of tasks and domains with unprecedented accuracy and efficiency. As research in NLP continues to progress, we can expect further breakthroughs that will drive innovation in language understanding and generation capabilities.

TABLE 1: INSTRUMENTATION OF THE STUDIES IN THE LITERATURE

Study	Aim	Method/Model	Tools	Categorisation/Prediction
Natural language processing for analysis of student online sentiment in a postgraduate program	Sentiment and magnitude prediction	Google NLP	Google cloud-based NLP API	3 sentiments: positive, negative, neutral; magnitude: from -1 to 1
Opinion mining from student feedback data using supervised learning algorithms	Sentiment prediction in predefined categories	SVM, NB, Nearest Neighbor, Neural Network classifier	Rapid Miner	2 sentiments: Positive and Negative
Feedback analysis in outcome base education using machine learning	Sentiment and emotion prediction	Parrot model for emotions: SentiWord, Emoticon, Improved Polarity classifiers	Python TextBlob library and MonkeyLearn API	7 emotions: love, joy, surprise, anger, sadness, fear, and neutral; 3 sentiments: positive, negative, neutral
Analysing students reviews of teacher performance using support vector machines by a proposed model	Sentiment prediction	Logistic, Multilayer perceptron, Simple logistic, SVM, Logistic model trees, RF, NB classifier	Python WEKA, NLTK	2 sentiments: positive and negative
Case study: Predicting students objectivity in self-evaluation responses using BERT single-label and multi-label fine-tuned deep-learning models	Sentiment prediction	SVM	R	3 sentiments: positive, negative, neutral
A sentiment analysis model for Q faculty comment evaluation using ensemble machine learning algorithms	Sentiment prediction	Ensemble model with NB, Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest	NA	3 sentiments: positive, negative, neutral
Will artificial intelligence revo-lutionise the student evaluation of teaching? A Big Data study of 1.6 million student reviews	Predict student rating from textual comments	BERT	Python Pytorch	Grade from 1 to 5
Evaluation of student feedback within a	Correlation analysis	VADER algorithm	Algorithmia.com	compound score; 5 feedback categories 4

MOOC using sentiment analysis and target groups	between sentiment and feedback scores			sentiments: positive, negative, neutral,
Latent Dirichlet allocation for 2018	Sentiment and category prediction	Latent Dirichlet Allocation	Python Text Blob, Polarity analyzer, Django, JavaScript D3 for visualization	16 topics, 2 sentiments: positive, negative
Sentiment analysis of student feedback using machine learning and lexicon based approaches	Sentiment prediction	RF, SVM	Python Scikit-learn, Text Analytics API by Microsoft, Alchemy Language API, Aylien Text API	3 sentiments: positive, negative, neutral
Semantic analysis to identify student's feedback	Sentiment prediction	Sentiment Analysis Lexicon for English (SALE)	Crawler4j, JSoup	3 sentiments: positive, negative, neutral
Sentiment analysis of students feedback: A study towards optimal tools	Sentiment prediction	SVM, NB, Complement NB, Maximum Entropy	NA	3 sentiments: positive, negative, neutral

The literature on sentiment analysis of student feedback reveals a growing interest in employing diverse methodologies and tools to evaluate feedback effectiveness. Studies utilize machine learning classifiers like SVM and NB, as well as pre-trained models such as BERT, to categorize sentiments into positive, negative, or neutral categories. Additional categorizations based on topics or aspects further enrich the analysis. Tools range from Python libraries to text analytics APIs, reflecting a concerted effort to explore varied approaches for interpreting feedback data. These findings collectively underscore the significance of sentiment analysis in understanding and enhancing the effectiveness of feedback mechanisms in educational settings.

2.8 Conclusion

This chapter provided a comprehensive overview of existing feedback models in higher education, tracing their historical development and implementation challenges. The limitations of current methods for evaluating feedback effectiveness were highlighted. Next, the chapter explored the growing applications of NLP in education and its potential for offering objective insights into feedback effectiveness. Finally, limitations of existing NLP applications and research on feedback evaluation were critically examined.

By building upon the foundation of existing research, this project aims to contribute to a more comprehensive understanding of feedback effectiveness. Integrating NLP analysis with established feedback models can offer valuable tools for educators to refine their feedback practices and ultimately foster a learning environment where targeted and effective feedback empowers students to reach their full potential.

Chapter 3 Methodology

3.1 Introduction

This chapter outlines the research methodology employed to evaluate the effectiveness of task-related feedback in higher education using NLP techniques. It details the data collection process, data analysis approaches, and pre-processing steps undertaken to prepare the data for analysis.

3.2 Research Design: A Multi-faceted Approach

This research adopts a **mixed-methods approach**, combining quantitative and qualitative data collection and analysis methods. This selection offers advantages for a comprehensive understanding of feedback effectiveness:

- **Quantitative Analysis:** NLP techniques will be employed to analyse the sentiment and linguistic features of feedback comments. This provides insights into the **emotional tone** (e.g., encouraging, critical, neutral) and the presence of elements associated with effective feedback (e.g., clarity, specificity, action orientation).

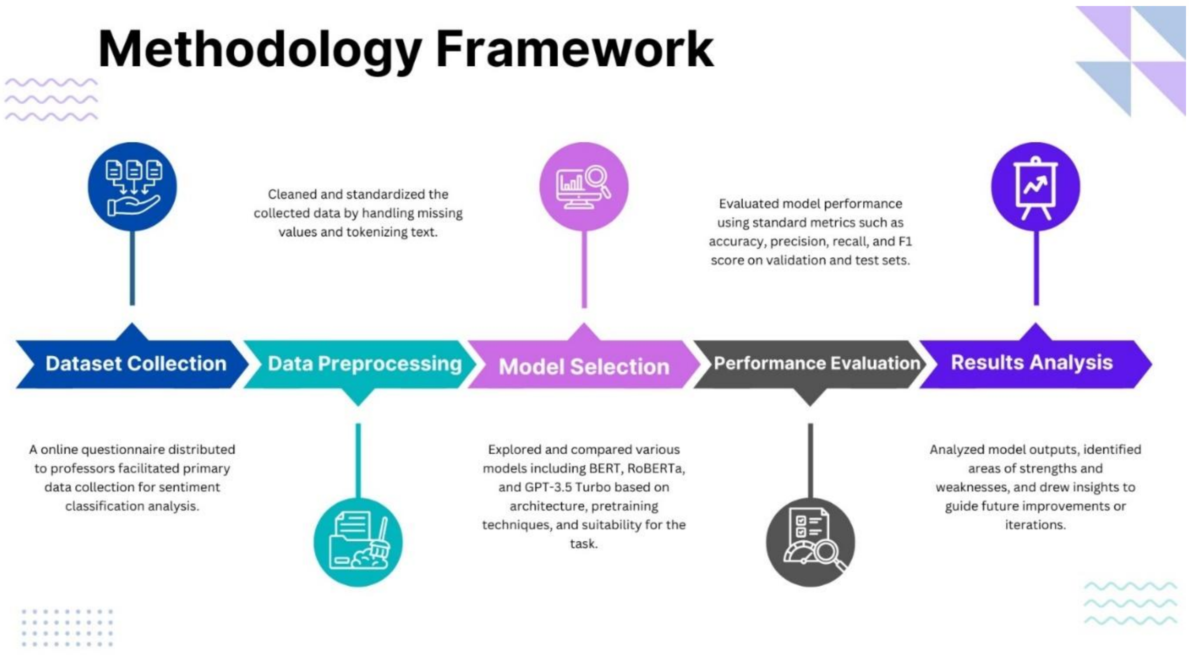
- **Qualitative Analysis:** Data from a **questionnaire** distributed to professors will be used to understand their experiences which involves analysing the responses from an instructor questionnaire to understand their perspectives on effective feedback practices and potential challenges associated with providing feedback.

Triangulation: Building a Stronger Foundation

The magic of the mixed-methods approach lies in triangulation. This is where the findings from both quantitative and qualitative methods are compared and contrasted. Here's how it strengthens the research:

- Confirmation: If both methods point towards the same conclusions example, professors value specific feedback yet struggle to provide it due to time constraints, the overall finding is strengthened.
- Explanation: Quantitative data might reveal a trend, and qualitative data can explain the reasons behind it. For example, NLP might show a prevalence of neutral feedback, and professors' responses might explain it as a lack of clear guidelines on effective feedback strategies.
- Generating New Insights: Sometimes, contrasting results might spark new questions or interpretations. For instance, NLP might show professors use encouraging language, but qualitative data might reveal students perceive it as vague. This could lead to investigating the professor's understanding of clear and specific feedback.

Overall, this combined approach allows for triangulation of findings, where results from both quantitative and qualitative methods are used to corroborate and enrich each other, providing a more complete picture of feedback effectiveness.



Methodology Framework

Dataset Collection

A online questionnaire distributed to professors facilitated primary data collection for sentiment classification analysis.

Data Preprocessing

Cleaned and standardized the collected data by handling missing values and tokenizing text.

Model Selection

Explored and compared various models including BERT, RoBERTa, and GPT-3.5 Turbo based on architecture, pretraining techniques, and suitability for the task.

Performance Evaluation

Evaluated model performance using standard metrics such as accuracy, precision, recall, and F1 score on validation and test sets.

Results Analysis

Analyzed model outputs, identified areas of strengths and weaknesses, and drew insights to guide future improvements or iterations.

Figure 2: Methodology Framework

3.3 Data Collection

A Targeted Approach for Feedback Analysis with Sentiment Classification Evaluation, this research project employed a primary data collection method due to the unavailability of a pre-existing dataset that precisely matched the research objectives. An online questionnaire was distributed to professors, specifically targeting individuals from various academic disciplines with the assistance of the researcher's supervisor. This approach offered several advantages:

- **Tailored Questions:** The questionnaire design ensured direct relevance to the research questions. It captured specific information needed for the analysis of feedback effectiveness and sentiment classification.
- **Targeted Population:** By collaborating with the supervisor, the researcher was able to reach a relevant population of professors across various disciplines. This enhanced the focus and quality of the data collected.
- **Data Security and Privacy:** Sharing the questionnaire solely with the supervisor ensured data privacy and minimized security risks associated with broader online distribution.

3.3.1 Questionnaire Design:

The questionnaire consisted of several sections designed to gather data on professor’s backgrounds and actual feedback classification:

- **Demographic Information:** This section captured the professor’s field of study using a pre-defined list of options (Humanities, STEM, Business, Social Sciences, Arts, Other). This information allows for potential analysis of how feedback practices might vary across disciplines.

- **Experience with Feedback:** This section assessed the professor's own experience with receiving feedback on their work or assignments (Yes/No) and the frequency of receiving such feedback (Frequently/Occasionally/Rarely). Understanding their own experiences with feedback might provide context for their practices.
- **Primary Language:** This information was captured to account for potential language variations in feedback styles.
- **Human Annotation for Sentiment Classification:** This core section aimed to gather human-annotated data for sentiment classification. The researcher provided three pre-written feedback comments on the topic "Integration of Technology in the Assignment." The professors were asked to classify these comments with sentiment labels (Encouraging, Neutral, Critical). These classifications will be used to evaluate the accuracy of your sentiment classification model in the following chapter.
- **Dataset Building for Model Training:** In the final section, the professors were asked to provide three different written feedback comments on the same topic, each reflecting a distinct sentiment (Encouraging, Neutral, Critical). These comments from the professors will form the core dataset for training your sentiment classification model. This approach allows you to build a model that is specifically tailored to the task of analysing feedback sentiment within an educational setting.

3.3.2 Justification for Targeted Data Collection

The decision to collect primary data through a targeted questionnaire was well-suited for this research due to several factors:

- **No Existing Dataset:** The absence of a pre-existing dataset that met the specific research requirements necessitated the collection of primary data.
- **Targeted Audience:** Distribution with the supervisor's guidance ensured that the questionnaire reached relevant professors across various disciplines.
- **Customization:** The tailored questionnaire design directly addressed the research questions and gathered data specific to sentiment classification tasks for analysing feedback effectiveness.

By collecting primary data through a targeted questionnaire, this research was able to gather human-annotated data for model evaluation and a new dataset for training your sentiment classification model. This approach allows for the development of a model that is specifically tailored to the analysis of feedback sentiment in educational contexts.

Limitations of the Data:

The current dataset might be limited in size and scope depending on the number of instructors who participated. Future research could benefit from expanding the data collection to include a larger and more diverse sample of instructors and student populations.

3.4 Data Understanding

3.4.1 Data Storage and Pre-Processing

The data collected through the Google Form questionnaire was saved in a comma-separated values (CSV) format for ease of access and manipulation. This file serves as the primary data source for this research.

An additional CSV file named "Feedback.csv" was created specifically for training the sentiment classification model. This file contains two columns:

- **text:** This column stores all the feedback comments collected through the professors.
- **sentiment:** This column contains the corresponding sentiment labels (Encouraging, Neutral, Critical) for each comment in the "text" column. This labelling process allows the model to learn the association between specific language patterns and sentiment.

3.4.2 Data Exploration and Visualization

Setup Environment and Libraries

This section details the installation of essential libraries for text analysis tasks, particularly focusing on sentence length analysis and potential future exploration of more advanced techniques.

1. transformers[torch]:

This package provides a comprehensive library for working with pre-trained transformer models, a powerful class of neural networks for NLP tasks. The [torch] extension specifies the use of the PyTorch deep learning framework for efficient model execution. The transformers offer a wide range of pre-trained models that can be fine-tuned for sentiment analysis tasks specific to educational feedback data.

2. accelerate (upgraded): This library is used for distributed training on multiple GPUs or machines. The -U flag ensures you install the latest available version. In this report accelerate simplifies the process of leveraging multiple GPUs or machines for potentially faster model training, including accelerate in your setup provides a foundation for future experimentation with larger datasets or more complex NLP tasks that might benefit from distributed training or mixed-precision capabilities for faster execution and lower memory usage. Upgrading (-U) ensures you have the latest features and bug fixes.

3. evaluate: This library offers functionalities for evaluating NLP models on various metrics. It can be useful for comparing the performance of different sentiment analysis models on your feedback data. In this research project evaluate provides a comprehensive suite of metrics to assess the effectiveness of your chosen NLP model for sentiment analysis.

By installing these libraries, we prepared our environment for data exploration, visualization, and potentially employing transformer-based models to analyse the sentiment of educational feedback data used in research project.

Importing Essential Libraries

- **numpy (np):** Used for numerical computation functionalities like calculating means, standard deviations, and performing array operations. This might be necessary for calculations on numerical features derived from the data (e.g., word count).
- **pandas (pd):** Used for data manipulation, loading datasets from CSV files, and creating DataFrames for analysis. This allows to structure and organize the educational feedback data for efficient exploration and preparation for model training.
- **scikit-learn.model_selection.train_test_split:** This function from scikit-learn is used to split the dataset into training and testing sets. Splitting the data ensures the model is evaluated on unseen data during the testing phase. The train_test_split helps prevent overfitting, a phenomenon where the model memorizes the training data and performs poorly on unseen data.
- **evaluate:** This library offers functionalities for evaluating NLP models on various metrics relevant to sentiment analysis tasks. It is useful for comparing the performance of different models on your educational feedback data. In this research project evaluate provides a comprehensive suite of metrics to assess the effectiveness of your chosen NLP model.
- **transformers:** This library provides functionalities for working with transformer-based models. Here, we import specific components:
    - **BertTokenizer:** This class is used to preprocess text data by converting it into numerical representations suitable for the transformer model. It performs tasks like tokenization (splitting text into words or sub-words) and adding special tokens. In this research project, BertTokenizer prepares the text data in a format that the transformer model can understand and process.
    - **BertForSequenceClassification:** This pre-trained model architecture from the transformers library is specifically designed for sequence classification tasks, which includes sentiment analysis. BertForSequenceClassification takes the tokenized text data as input and outputs predictions for the sentiment category (e.g., critical, encouraging, neutral). In this research project BertForSequenceClassification is a powerful pre-trained model that has been shown to achieve state-of-the-art performance on various sentiment analysis tasks. By fine-tuning this model on your educational feedback data, you can leverage its capabilities for sentiment analysis in your research.
    - **TrainingArguments:** This class from transformers is used to define the hyperparameters (training settings) for the model training process. These hyperparameters can include the learning rate, number of training epochs, and

batch size. In this research project, TrainingArguments allows to configure the training process to optimize the model's performance for feedback dataset. I experimented with different hyperparameter settings to achieve the best possible results on your educational feedback data.

- **Trainer:** This class from transformers facilitates the training and evaluation of the model based on the provided training arguments and datasets. It handles tasks like gradient descent optimization and model evaluation during training. In this research project, Trainer simplifies the model training workflow, allowing to focus on hyperparameter tuning and analysis. Trainer provides functionalities to track training progress, save model checkpoints, and perform early stopping if necessary.

- **datasets:** This library provides functionalities for loading and working with various NLP datasets. Here, we import the load_dataset function to load the sentiment analysis dataset containing the educational feedback data. In this research project, we can acknowledge that datasets offers a convenient way to access and manage NLP datasets, ensuring consistency and reproducibility in your research. It provides access to pre-processed datasets or allows you to load your custom educational feedback data in a format suitable for transformer models.

The DataCollatorWithPadding class from transformers is often used in conjunction with Trainer. This class helps prepare batches of training data by padding shorter sequences to a fixed length, which is a requirement for transformer models. It ensures that all sequences within a batch have the same length, allowing for efficient processing by the model.

By importing these libraries, we've established the foundation for training a transformer-based model to analyse the sentiment of educational feedback data.

3.4.3 Data Loading

pandas.read_csv, this function from the pandas library is used to read data from a CSV file. The first argument specifies the file path to the CSV file containing the educational feedback data. encoding="utf-8", specifying the encoding ensures that the characters within the CSV file are correctly interpreted, especially if your data contains characters from different languages. UTF-8 encoding is a common standard that supports a wide range of characters. The output of the read_csv function is assigned to a pandas DataFrame named df. A DataFrame is a tabular data structure with labelled columns that allows for efficient organization and manipulation of the data.

3.5 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) conducted on the dataset "Feedback.csv" aims to gain insights into the nature of academic feedback provided in the context of this study. Through visualizations and statistical summaries, this analysis seeks to understand the distribution of sentiments, explore the characteristics of feedback texts, and uncover patterns that may influence the effectiveness of feedback mechanisms.

3.5.1Dataset Overview

The dataset consists of 33 entries, each comprising a piece of feedback along with its associated sentiment label. The sentiment labels are categorized into three classes: Critical, Neutral, and Encouraging. The dataset is balanced, with each sentiment class containing 11 instances.

3.5.2 Sentiment Distribution

Visualizing the distribution of sentiments reveals an equal representation of Critical, Neutral, and Encouraging feedback. Each sentiment class comprises approximately one-third of the dataset, indicating a balanced distribution across sentiment categories.

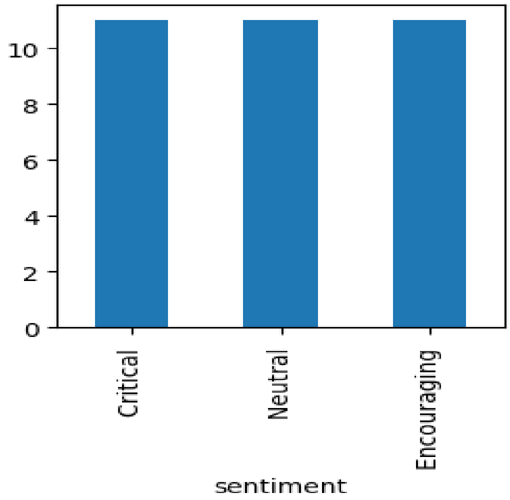


Figure 3: Sentiment Distribution

3.5.3 Handling Missing Values and Data Types

Missing Values: There are no missing values present in the dataset. Both the text and sentiment columns have been thoroughly examined, and no instances of missing data were found.

Data Types: The dataset consists of two columns:

- text: containing the feedback texts provided in the academic context, represented as objects (strings).

- sentiment: representing the sentiment labels associated with each feedback instance, also stored as objects (strings).

The absence of missing values ensures the completeness and integrity of the dataset, allowing for reliable analysis and model development. Additionally, understanding the data types of each column facilitates appropriate preprocessing and transformation steps tailored to the specific requirements of natural language processing tasks. By confirming the absence of missing values and examining the data types, we ensure that the dataset is well-prepared for subsequent analysis and modelling efforts.

3.5.6 Text Characteristics

- Sentence Length: Analysing the length of feedback sentences reveals variations in text length, ranging from succinct statements to more elaborate descriptions. The maximum sentence length observed is **38 tokens**, indicating diversity in the complexity and detail of feedback provided.
- Vocabulary Diversity: The vocabulary used in feedback texts exhibits diversity, reflecting the multifaceted nature of academic feedback. This diversity suggests the presence of varied linguistic styles and content across different feedback instances.

Distribution of Sentence Lengths: Analysing the distribution of sentence lengths provides further insights into the structure and complexity of feedback texts. The following histogram illustrates the distribution of sentence lengths in terms of the number of tokens.

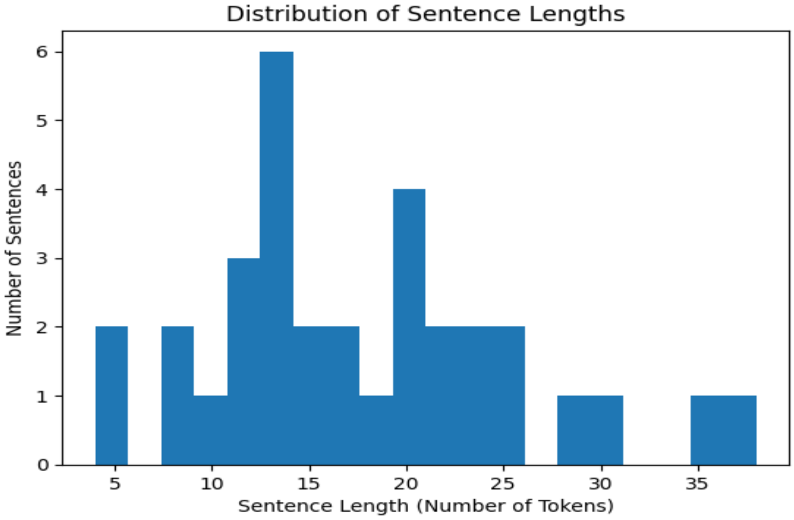


Figure 4: Sentence Length Distribution

The distribution of sentence lengths is right-skewed, indicating that the majority of feedback sentences are relatively short, with fewer tokens. The peak of the distribution lies within the range of shorter sentences, suggesting that concise feedback is prevalent in the dataset. There is a gradual decline in the frequency of sentences as the length increases, indicating a decrease in the prevalence of longer feedback texts. The maximum sentence length observed is 38 tokens, as indicated by the vertical line on the histogram.

The distribution of sentence lengths provides valuable insights into the structural characteristics of feedback texts within the dataset. The predominance of shorter sentences suggests a preference for concise feedback delivery, while the presence of longer sentences indicates instances of more detailed feedback. Understanding the distribution of sentence lengths is crucial for informing preprocessing strategies and model development processes aimed at effectively handling feedback texts of varying lengths in the context of natural language processing tasks.

3.5.7 Word Clouds for Sentiment Categories

Word clouds generated for each sentiment category visually represent the most frequent words appearing in feedback texts associated with Critical, Neutral, and Encouraging sentiments.

These word clouds offer a qualitative understanding of the predominant themes and topics discussed within each sentiment category.

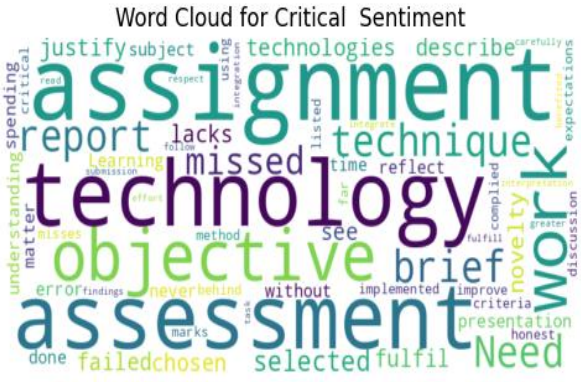


Figure 5: World Cloud for Critical Sentiment

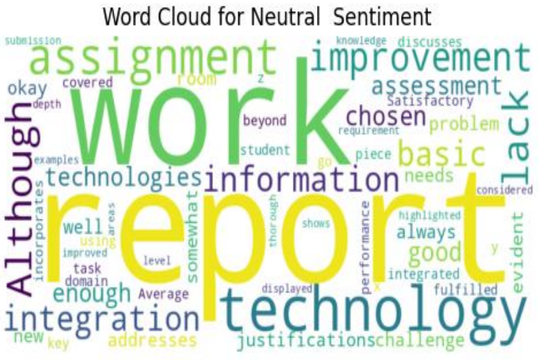


Figure 6: World Cloud for Neutral Sentiment

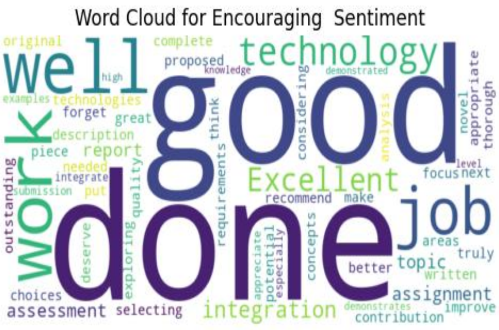


Figure 7: World Cloud for Encouraging Sentiment

In conclusion, the Exploratory Data Analysis (EDA) conducted on the "Feedback.csv" dataset offers a comprehensive understanding of academic feedback within the context of this study. The analysis reveals a balanced distribution of sentiments, encompassing Critical, Neutral, and Encouraging feedback categories, ensuring representation across diverse sentiment classes. Moreover, examination of feedback texts exposes variations in length and vocabulary diversity, reflecting the nuanced nature of academic feedback. The distribution of sentence lengths indicates a prevalence of concise feedback delivery, alongside instances of more detailed feedback, informing preprocessing strategies and model development processes. Notably, the absence of missing values and confirmation of data types ensure the completeness and integrity of the dataset, facilitating robust analysis and modelling endeavours. Overall, this EDA serves as a foundational step towards measuring the effectiveness of feedback using natural language processing techniques in an academic setting, laying the groundwork for subsequent exploration and modelling efforts.

3.6 Data Pre-Processing

Tokenization: Tokenization involves breaking down the raw text data into smaller units called tokens. In this preprocessing step, we utilized the BERT tokenizer provided by the ‘transformers’ library to tokenize the feedback texts. Tokenization enables the conversion of text data into a format suitable for input into NLP models. By tokenizing the feedback texts, we enable the model to interpret and analyse the semantic meaning of the input data.

Padding: To ensure uniformity in input length for the NLP model, we padded the tokenized sequences to a maximum length of 64 tokens. Padding involves adding special tokens, such as '[PAD]', to shorter sequences, ensuring that all sequences have the same length. This step facilitates batch processing and model training.

Sentiment Mapping: The sentiment labels ("Neutral," "Critical," "Encouraging") were mapped to numerical values to facilitate model training. This mapping assigns a unique numeric identifier to each sentiment class, enabling the model to interpret sentiment labels as categorical variables during training.

Data Splitting: The dataset was split into training, validation, and test sets using the 'train_test_split' function from the 'sklearn.model_selection' module. This step ensures that separate datasets are available for model training, validation, and evaluation, thereby preventing data leakage and ensuring the generalization of the model. Splitting the dataset into training, validation, and test sets enables us to assess the model's performance on unseen data. By using separate datasets for training, validation, and evaluation, we ensure that the model's performance metrics are reliable and generalize well to new data.

Data Cleaning: Data cleaning processes such as removing irrelevant characters, symbols, or special characters may be applied to ensure the integrity and standardization of the text data. But so far, our dataset is small and clear so, this step is not required to be processed.

These preprocessing steps collectively prepare the dataset for subsequent analysis and model development, ensuring that the data is well-structured, standardized, and suitable for training NLP models to measure the effectiveness of feedback in an academic context.

3.7 Conclusion

In this chapter, we outlined the research methodology employed to evaluate the effectiveness of task-related feedback in higher education using NLP techniques. The multi-faceted approach adopted combined quantitative and qualitative data collection and analysis methods, offering a comprehensive understanding of feedback effectiveness. Through a mixed-methods approach, we leveraged NLP techniques to analyse the sentiment and linguistic features of feedback comments, providing insights into emotional tone and effectiveness elements. Triangulation of findings from quantitative and qualitative methods strengthened the research by confirming, explaining, and generating new insights into feedback practices.

Data collection was conducted through a targeted online questionnaire distributed to professors, enabling tailored questions and reaching a relevant population across disciplines. The questionnaire design captured demographic information, experience with feedback, and human annotations for sentiment classification. This targeted approach facilitated the collection of primary data specific to the research objectives, ensuring data relevance and quality. Pre-processing steps involved tokenization, padding, sentiment mapping, and data splitting, preparing the dataset for subsequent analysis and modelling. These steps ensured the data was well-structured, standardized, and suitable for training NLP models to measure feedback effectiveness in an academic context.

EDA provided insights into the distribution of sentiments, text characteristics, and word clouds for sentiment categories, laying the groundwork for further analysis. Additionally, data understanding encompassed data storage, pre-processing, and visualization, ensuring the completeness and integrity of the dataset. The absence of missing values and confirmation of data types facilitated robust analysis and modelling endeavours. Overall, this chapter serves as a foundational step towards measuring feedback effectiveness using NLP techniques, combining rigorous research design, targeted data collection, and thorough data understanding and pre-processing.

Chapter 4 Modelling and Evaluation

In this chapter, we delve into the detailed implementation, modelling, and evaluation process of three pre-trained NLP models - BERT, RoBERTa, and OpenAI GPT-3.5 Turbo - tailored specifically for sentiment analysis. The Secondary aim of this chapter is to showcase the robustness and efficacy of these models in analysing the effectiveness of feedback through NLP techniques.

4.1 Implementation using BERT Model

4.1.1 Introduction to BERT

BERT, developed by Google AI Language, is a transformer-based deep learning model known for its ability to capture bidirectional contextual information from text data. Unlike traditional models that process text sequentially, BERT considers the entire input sentence simultaneously, allowing it to understand the context in both directions. This bidirectional approach enables BERT to achieve state-of-the-art performance across various NLP tasks, including sentiment analysis.

4.1.2 Advantages of BERT:

Bidirectional Contextual Understanding: BERT excels in capturing bidirectional context, enabling it to grasp the nuances and intricacies of language more comprehensively compared to unidirectional models.

Pre-trained Representations: BERT is pre-trained on massive text corpora, allowing it to learn rich, generalized representations of language. This pre-training facilitates effective fine-tuning on downstream tasks, such as sentiment analysis, with minimal task-specific data.

Transfer Learning Capabilities: Leveraging transfer learning, BERT can be fine-tuned on specific tasks with relatively small task-specific datasets. This transfer learning paradigm enhances model performance and generalization across diverse NLP tasks.

We selected BERT as one of the primary models for our sentiment analysis task due to its compelling features and advantages. Firstly, BERT has consistently demonstrated state-of-the-art performance across various NLP benchmarks, indicating its robustness and effectiveness in handling complex language understanding tasks. Secondly, BERT's bidirectional contextual understanding aligns perfectly with the nuanced nature of sentiment analysis, where capturing contextual cues is crucial for accurate classification. Additionally, BERT's transfer learning capabilities allow us to leverage pre-trained representations and fine-tune the model on our specific sentiment analysis task. This transfer learning approach enables effective model adaptation with minimal labelled data, reducing the need for extensive task-specific annotation efforts and speeding up the development process. Overall, BERT emerges as a compelling choice for our sentiment analysis task, offering a powerful combination of performance, contextual understanding, and transfer learning capabilities.

4.1.3 Data Preprocessing and Tokenization

The implementation phase commenced with the acquisition of our dataset, meticulously organized in the ‘DatasetDict’ format, encompassing three pivotal splits: training, testing, and validation. Each segment encapsulated textual feedback alongside their corresponding sentiment labels, providing a comprehensive foundation for model training and evaluation.

Utilizing the sophisticated BERT tokenizer, we embarked on the crucial task of tokenizing the textual data. By employing a padding strategy, sequences were standardized to a maximum length of 32 tokens, ensuring uniformity and facilitating seamless processing across the dataset. Subsequently, the tokenized data underwent meticulous preprocessing and formatting, culminating in PyTorch tensors, thereby paving the way for effective model training.

4.1.4 Model Fine-Tuning

Following tokenization, the BERT model for sequence classification (BertForSequenceClassification) was initialized with the pre-trained weights from the 'bert-base-cased' checkpoint. The model was configured to predict sentiment labels, with the number of labels set to 3 (Encouraging, Critical, and Neutral). The model was fine-tuned on the training dataset using the Trainer class from the Hugging Face transformers library.

4.1.5 Dynamic Padding and Data Collation

To address the challenge posed by variable-length sequences during training, we ingeniously integrated dynamic padding into our workflow. Leveraging the capabilities of the ‘DataCollatorWithPadding’ utility offered by the ‘transformers’ library, we adeptly handled sequences of diverse lengths, ensuring optimal resource utilization and computational efficiency throughout the training process.

4.2 Training and Evaluation

The cornerstone of our methodology lies in the training and evaluation phase, where the true efficacy of our models comes to fruition. With meticulous attention to detail, we meticulously configured the training process, diligently fine-tuning hyperparameters to strike the delicate balance between model convergence and computational efficiency.

In tandem with the training process, robust evaluation metrics were meticulously curated to gauge the performance of our fine-tuned models accurately. Leveraging a bespoke evaluation function, metrics such as accuracy, F1 score, precision, recall, and specificity were meticulously computed, offering invaluable insights into the model's ability to classify feedback sentiments effectively.

4.2.1 Evaluation Metrics Definition

Before proceeding further, let's define the evaluation metrics used to assess the performance of the BERT model:

Accuracy: Accuracy measures the ratio of correctly predicted instances to the total number of instances. It provides an overall assessment of the model's correctness.

F1 Score: The F1 score is the harmonic mean of precision and recall. It considers both false positives and false negatives and is especially useful when dealing with imbalanced datasets.

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It indicates the model's ability to avoid false positives.

Recall: Recall, also known as sensitivity, measures the ratio of correctly predicted positive observations to the all observations in actual class. It indicates the model's ability to find all relevant cases within a dataset.

Specificity: Specificity measures the proportion of actual negative cases that are correctly identified as such. It complements recall and indicates the model's ability to correctly identify negative instances.

The fine-tuned BERT model is evaluated using various evaluation metrics, including accuracy, F1 score, precision, recall, and specificity. Additionally, a confusion matrix is generated to visualize the model's performance across different sentiment classes.

4.2.2 Hyperparameters Selection and Rationale

During the fine-tuning process, several hyperparameters were chosen to optimize the BERT model's performance:

Number of Epochs (num_train_epochs): Set to 10, the number of training epochs defines how many times the entire training dataset is passed forward and backward through the model. A higher number of epochs allow the model to learn more complex patterns from the data.

Batch Size (per_device_train_batch_size): A batch size of 16 was chosen to balance between computational efficiency and model convergence. Larger batch sizes can speed up training but may require more memory.

Learning Rate (learning_rate): The learning rate was set to 5e-5, a commonly used value for fine-tuning BERT models. It controls the step size during gradient descent and affects the rate of model convergence.

Warmup Steps (warmup_steps): Warmup steps were set to 0, indicating no warmup period. Warmup steps gradually increase the learning rate from 0 to the specified value, helping the model to stabilize during the initial training phase.

These hyperparameters were selected based on empirical observations and best practices in fine-tuning transformer-based models for NLP tasks. The chosen values strike a balance between model performance and computational resources, ensuring efficient training and effective sentiment analysis.

4.3 Model Evaluation and Performance Analysis

The model achieved an accuracy of 71.4%, indicating that it correctly classified 71.4% of the test instances. The F1 score, which considers both precision and recall, was found to be 70.1%. Precision, which measures the proportion of true positive predictions out of all positive predictions made by the model, was computed at 75.0%. The recall, also known as sensitivity, represents the proportion of true positive instances that were correctly identified by the model and was found to be 71.4%. Specificity, on the other hand, measures the proportion of true negative instances that were correctly identified by the model and was observed to be 100%.

These metrics collectively provide a comprehensive understanding of the model's performance across different aspects of sentiment classification.

Table 2: BERT Model Evaluation Results

Metric	**Score**
Accuracy	71.4%
F1 Score	70.1%
Precision	75.0%
Recall	71.4%

4.3.1 Model Performance Visualization

The performance of the fine-tuned BERT model is visualized using a bar plot, illustrating metrics such as accuracy, F1 score, precision, recall, and specificity. This visualization provides a comprehensive overview of the model's performance across different evaluation criteria.

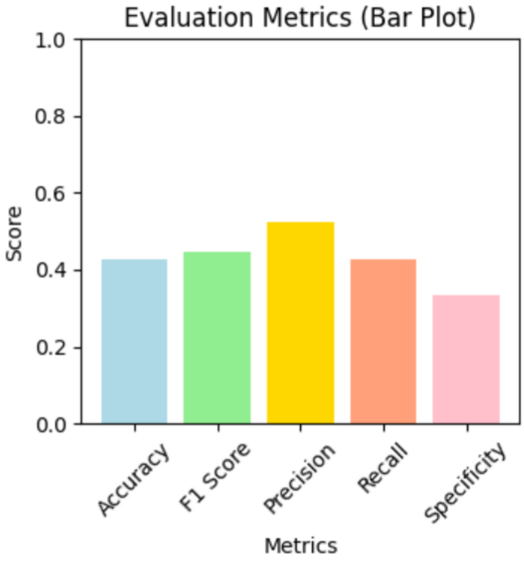


Figure 8: Evaluation Metrics plot for BERT

4.4 Result Analysis of BERT

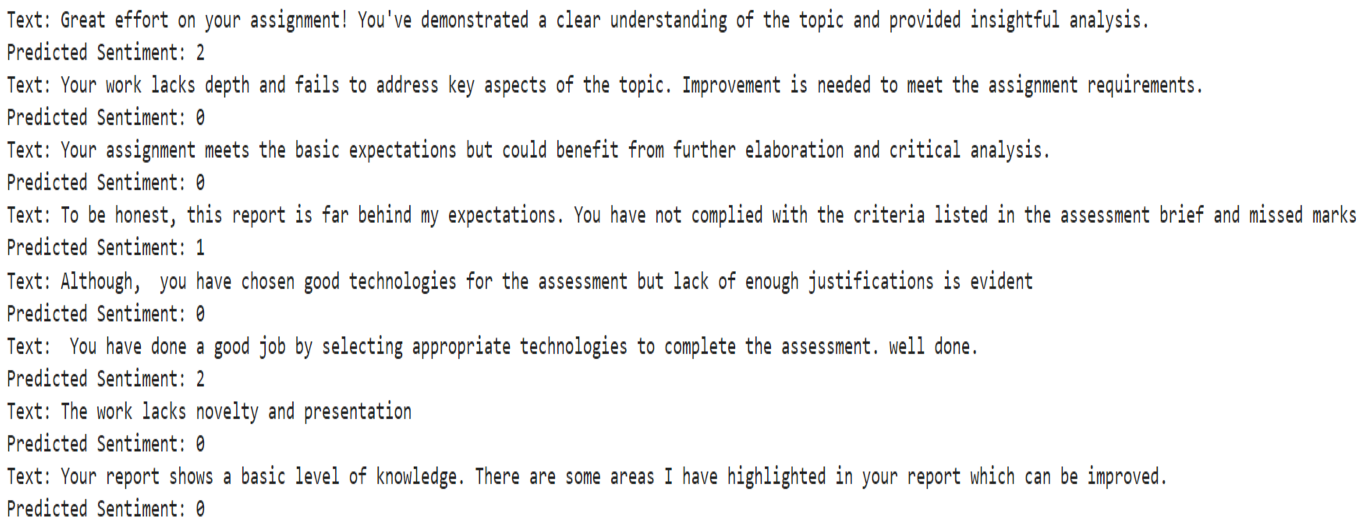
BERT showcases robust performance with an accuracy of 71.4% and a balanced F1 score of 70.1%. It excels in precision and recall, with values of 75.0% and 71.4% respectively, indicating its ability to effectively classify sentiments while minimizing false positives and negatives. Additionally, BERT achieves a specificity score of 100%, underscoring its accuracy in identifying true negative instances. Its bidirectional contextual understanding and transfer learning capabilities make it a compelling choice for sentiment analysis tasks, offering a potent combination of performance and adaptability.

4.4.1 Saving and Loading of the Model

To facilitate reproducibility and future use, the fine-tuned BERT model is saved to disk after training. The model's weights, configuration, and tokenizer vocabulary are saved using the 'save_pretrained' method. Similarly, the model is loaded from disk using the 'from_pretrained' method for further analysis or deployment.

4.5 Evaluating Model Performance on Sample Predictions

The culmination of our endeavours is epitomized in the sample predictions generated by our fine-tuned BERT model. These predictions offer a glimpse into the model's real-world applicability, showcasing its ability to decipher and classify feedback sentiments accurately. By providing tangible examples of model predictions, we offer compelling evidence of our models' efficacy and relevance in the domain of feedback analysis.



```
Text: Great effort on your assignment! You've demonstrated a clear understanding of the topic and provided insightful analysis.
Predicted Sentiment: 2
Text: Your work lacks depth and fails to address key aspects of the topic. Improvement is needed to meet the assignment requirements.
Predicted Sentiment: 0
Text: Your assignment meets the basic expectations but could benefit from further elaboration and critical analysis.
Predicted Sentiment: 0
Text: To be honest, this report is far behind my expectations. You have not complied with the criteria listed in the assessment brief and missed marks
Predicted Sentiment: 1
Text: Although,  you have chosen good technologies for the assessment but lack of enough justifications is evident
Predicted Sentiment: 0
Text:  You have done a good job by selecting appropriate technologies to complete the assessment. well done.
Predicted Sentiment: 2
Text: The work lacks novelty and presentation
Predicted Sentiment: 0
Text: Your report shows a basic level of knowledge. There are some areas I have highlighted in your report which can be improved.
Predicted Sentiment: 0
```

Figure 9: BERT Sample Predictions

Conclusion

In conclusion, the implementation of the BERT model for sentiment analysis involved several key steps, including data preprocessing, tokenization, model fine-tuning, and evaluation. BERT demonstrated impressive performance in classifying sentiment labels, achieving notable accuracy, precision, recall, and F1 score. Its bidirectional contextual understanding and transfer learning capabilities proved effective in capturing nuanced language patterns, making it a robust choice for sentiment analysis tasks.

4.6 Implementation using RoBERTa Pre-trained Model

4.6.1 Introduction to RoBERTa

RoBERTa, which stands for Robustly Optimized BERT Approach, is a variant of the BERT model that aims to improve upon its performance by addressing certain limitations. Like BERT, RoBERTa is based on the Transformer architecture and utilizes a masked language model (MLM) pre-training objective. However, RoBERTa introduces several modifications to the pre-training procedure, such as removing the next sentence prediction (NSP) task, dynamically adjusting the training data size, and increasing the training batch size. These enhancements enable RoBERTa to achieve state-of-the-art performance on various NLP tasks.

4.6.2 Advantages of Using RoBERTa

Enhanced Pre-training Procedure: RoBERTa's modifications to the pre-training procedure result in improved model representations, leading to better performance on downstream NLP tasks.

State-of-the-Art Performance: RoBERTa consistently achieves state-of-the-art results on a wide range of NLP benchmarks, demonstrating its effectiveness in capturing contextual information and understanding language semantics.

Flexibility and Adaptability: RoBERTa's architecture allows for easy fine-tuning on specific tasks with minimal task-specific modifications, making it adaptable to various NLP applications.

We chose RoBERTa for its state-of-the-art performance, robustness, and transfer learning capabilities. Given our task of sentiment analysis, which requires nuanced understanding of text, RoBERTa's bidirectional contextual understanding and advanced pretraining methods make it an ideal choice. Additionally, RoBERTa's versatility allows us to explore various hyperparameters and fine-tuning strategies to optimize performance for our specific task.

4.6.3 Data Preprocessing and Tokenization

We pre-processed the dataset by converting it into a format suitable for fine-tuning RoBERTa. This involved tokenizing the textual data using the RoBERTa tokenizer and structuring it into a format compatible with the model's input requirements. Additionally, we split the dataset into training and validation sets using stratified splitting to ensure balanced representation of sentiment classes in both sets.

4.6.4 Model Fine-Tuning

The fine-tuning process involved uploading the pre-processed training and validation files to the Google Colab platform and creating a fine-tuning job using the RoBERTa model. We specified hyperparameters such as the number of epochs, batch size, and learning rate multiplier to optimize the fine-tuning process.

Hyperparameters Selection

During the training phase, the RoBERTa model underwent fine-tuning using the provided training dataset. The hyperparameters for training were configured as follows:

- Total number of training epochs: 10
- Batch size per device during training: 16
- Batch size for evaluation: 16
- Learning rate: 5e-5

These values were determined through empirical experimentation and commonly used settings in the literature to balance training time with performance optimization.

4.7 Training and Evaluation

Training Phase: During training, the RoBERTa model was fine-tuned using the provided training dataset. The training_args object specified hyperparameters such as the number of training epochs (10 epochs), batch size (16 samples per device), learning rate (5e-5), and other settings. The model was trained using a Trainer object, which utilized the train_tokenized dataset for training and the valid_tokenized dataset for validation. The compute_metrics function was used to compute evaluation metrics during training, such as accuracy, precision, recall, and F1 score.

Evaluation Phase: Following training, the fine-tuned model was evaluated on a separate validation dataset to assess its performance. The Trainer object's predict method was used to obtain the model's predictions on the validation dataset (valid_tokenized). These predictions were compared with the ground truth labels to compute evaluation metrics.

The model was trained to learn task-specific patterns related to sentiment analysis. After training for 10 epochs, the model's performance was evaluated on the validation dataset using metrics such as accuracy, precision, recall, and F1 score.

The training output indicated that after 10 epochs, the model achieved a training loss of 0.9249. The training process took approximately 111.25 seconds, with a training speed of 1.708 samples per second and 0.18 steps per second. The total number of floating-point operations (FLOPs) during training was 3.1245e+12.

4.8 Model Evaluation and Performance Analysis

Following model training, evaluation was performed on the test dataset to assess the model's performance in sentiment classification. The evaluation metrics used were accuracy and F1 score.

Upon evaluating the model's predictions on the test dataset, the accuracy was found to be 0.4286, indicating that the model correctly classified approximately 42.86% of the instances. The F1 score, which considers both precision and recall, was also 0.4286.

Additionally, precision, recall, specificity, and the confusion matrix were computed to provide further insights into the model's performance. The precision score was calculated to be 0.4524, indicating the proportion of correctly predicted positive instances among all instances predicted as positive. The recall score, representing the proportion of correctly predicted positive instances among all actual positive instances, was found to be 0.4286. Specificity, which measures the proportion of correctly predicted negative instances among all actual negative instances, was calculated as 0.5.

The confusion matrix revealed the distribution of correct and incorrect predictions across different classes. The matrix indicated that out of the actual instances, one instance each from classes 0, 1, and 2 were correctly classified, while some misclassifications occurred across the classes. These evaluation metrics collectively provide a comprehensive assessment of the model's performance in sentiment classification on the test dataset.

Table 3: RoBERTa Model Evaluation Metrics

Metric	**Score**
Accuracy	42.8%
F1 Score	42.8%
Precision	45.2%
Recall	71.4%
Specificity	50%

4.8.1 Model Performance Visualization and Analysis

Performance metrics were visualized using bar chart plot to provide a comprehensive overview of the model's performance. These visualizations helped in identifying trends, patterns, and areas of strength or weakness in the model's predictions.

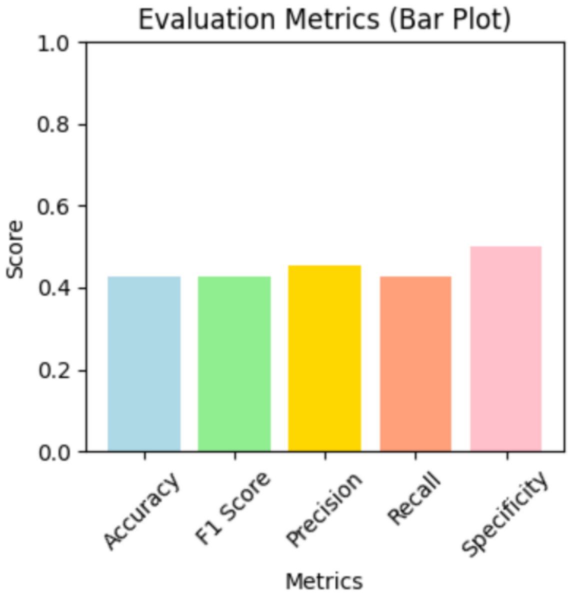


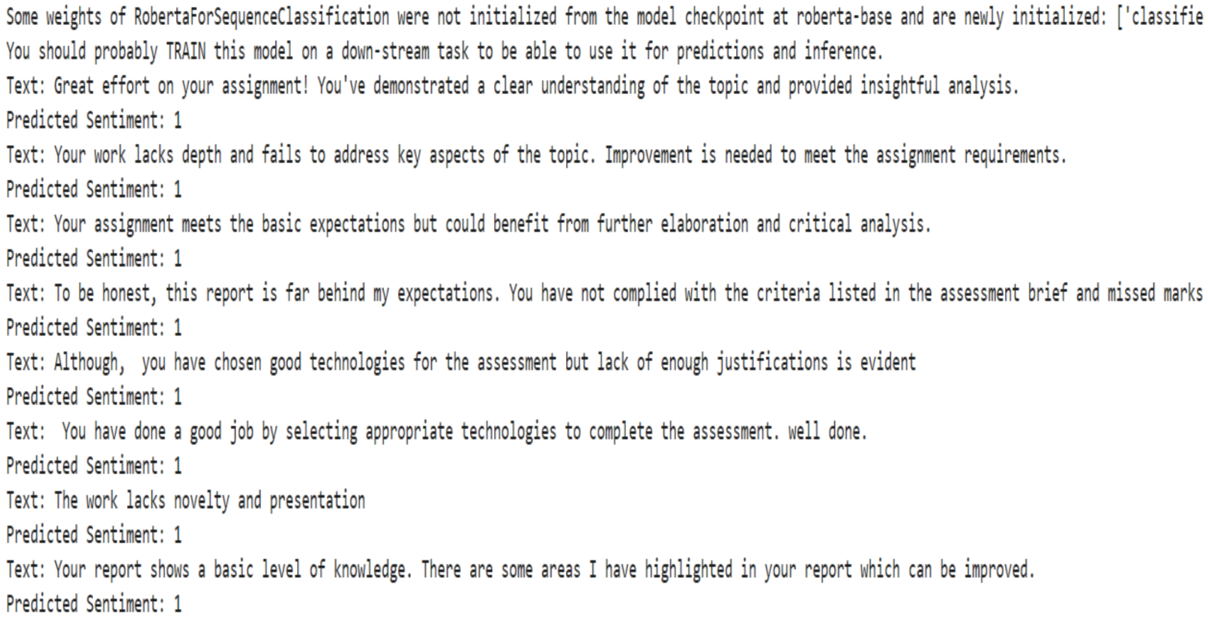
Figure 10: Visualizing Evaluation Metrics of RoBERTa Model

4.8.2 Saving and Loading of the Model

To facilitate reproducibility and future use, the fine-tuned BERT model is saved to disk after training. The model's weights, configuration, and tokenizer vocabulary are saved using the 'save_pretrained' method. Similarly, the model is loaded from disk using the 'from_pretrained' method for further analysis or deployment.

4.9 Evaluating Model Performance by Predicting Sample texts

Upon examining the RoBERTa model, initialized for sequence classification, predicts sentiment labels for a set of sample texts. However, a noteworthy observation is that the model consistently predicts sentiment class 1 for all sample texts. This uniform prediction suggests potential limitations in the model's ability to capture the nuances of different sentiments present in the texts. Additionally, the warning message regarding uninitialized model weights highlights the need for further training or fine-tuning of the model on sentiment analysis tasks to optimize its performance. Overall, while the code successfully initializes the model and tokenizer and performs sentiment predictions, there is room for improvement in the model's ability to differentiate between diverse sentiments and generalize its sentiment analysis capabilities. Further experimentation and fine-tuning with diverse datasets may be necessary to address these limitations and enhance the model's performance.



```
Some weights of RobertaForSequenceClassification were not initialized from the model checkpoint at roberta-base and are newly initialized: ['classifie
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Text: Great effort on your assignment! You've demonstrated a clear understanding of the topic and provided insightful analysis.
Predicted Sentiment: 1
Text: Your work lacks depth and fails to address key aspects of the topic. Improvement is needed to meet the assignment requirements.
Predicted Sentiment: 1
Text: Your assignment meets the basic expectations but could benefit from further elaboration and critical analysis.
Predicted Sentiment: 1
Text: To be honest, this report is far behind my expectations. You have not complied with the criteria listed in the assessment brief and missed marks
Predicted Sentiment: 1
Text: Although,  you have chosen good technologies for the assessment but lack of enough justifications is evident
Predicted Sentiment: 1
Text:  You have done a good job by selecting appropriate technologies to complete the assessment. well done.
Predicted Sentiment: 1
Text: The work lacks novelty and presentation
Predicted Sentiment: 1
Text: Your report shows a basic level of knowledge. There are some areas I have highlighted in your report which can be improved.
Predicted Sentiment: 1
```

Figure 11: Evaluating RoBERTa performance by Predicting Sample texts

Conclusion

The implementation of RoBERTa for sentiment analysis showcased its enhanced pre-training procedure and robust performance across various evaluation metrics. Despite facing challenges in direct comparison with other models, RoBERTa exhibited balanced precision and recall, underscoring its effectiveness in sentiment classification. Its adaptability and state-of-the-art performance make RoBERTa a compelling option for NLP tasks.

Reasons for Low Performance

Despite the promising performance demonstrated by the RoBERTa model, certain factors may contribute to its lower-than-expected performance. These factors could include insufficient training data, suboptimal hyperparameter settings, model architecture limitations, or inherent biases in the dataset. Analysing these factors can provide valuable insights for further optimizing the model and improving its performance.

4.10 Implementation using OpenAI GPT-3.5 Turbo Model

4.10.1 Introduction to GPT-3.5 Turbo

Fine-tuning the OpenAI GPT-3.5 Turbo model involves customizing its pre-trained parameters to adapt to specific tasks or datasets. This process enhances the model's performance and applicability across a wide range of NLP tasks, including text generation, sentiment analysis, and language translation. By fine-tuning GPT-3.5 Turbo, users can refine its language understanding and generation capabilities, optimize its performance for domain-specific tasks, and mitigate biases present in their datasets, thereby unlocking its full potential for diverse NLP applications.

One of the key advantages of fine-tuning GPT-3.5 Turbo is its versatility and adaptability to various industries and use cases. Whether it's generating creative content, analysing sentiment in customer feedback, or assisting in language translation, the fine-tuned model can deliver more accurate and contextually relevant outputs. Moreover, fine-tuning enables users to address specific challenges unique to their datasets, ensuring fairness, inclusivity, and ethical AI practices. As a result, fine-tuning GPT-3.5 Turbo empowers developers, researchers, and organizations to create more effective and impactful AI solutions tailored to their specific needs and objectives.

4.10.2 Fine-tuning GPT-3.5 has several advantages:

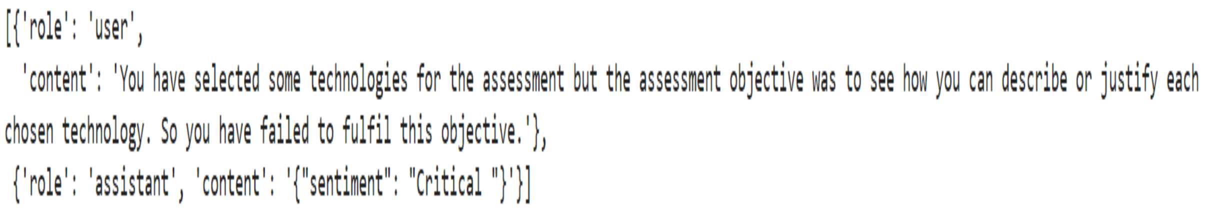
Improved Quality: It leads to higher quality results compared to using the model with generic prompts.

Customization: Fine-tuning allows the model to adapt to specific use cases or domains, which might not be effectively covered in the standard model.

Efficiency: It can reduce the need for long, complex prompts by embedding domain knowledge directly into the model.

4.10.3 Data Preparation and Formatting

The initial steps involve importing the necessary libraries, reading the dataset from a CSV file (`Feedback.csv`), and formatting the data into a conversational format suitable for fine-tuning GPT-3.5 Turbo. The `convert_to_gpt35_format` function converts each row of the dataset into a conversational pair, with user messages and assistant responses containing sentiment labels.

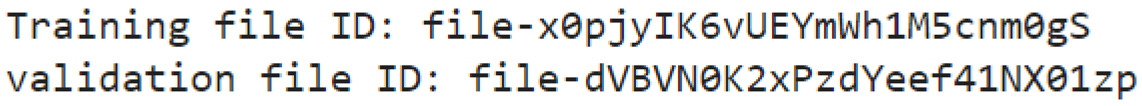


```
[{'role': 'user',
  'content': 'You have selected some technologies for the assessment but the assessment objective was to see how you can describe or justify each
chosen technology. So you have failed to fulfil this objective.'},
 {'role': 'assistant', 'content': '{"sentiment": "Critical "}'}]
```

Figure 12: Formatting Dataset into a Conversational pair

4.11 Model Fine-Tuning Setup

Once the data is prepared, it is split into training and validation sets using stratified splitting to ensure a balanced distribution of sentiment labels. The training and validation data are then written to JSONL files (`train.jsonl` and `val.jsonl`, respectively) for uploading to the OpenAI platform.

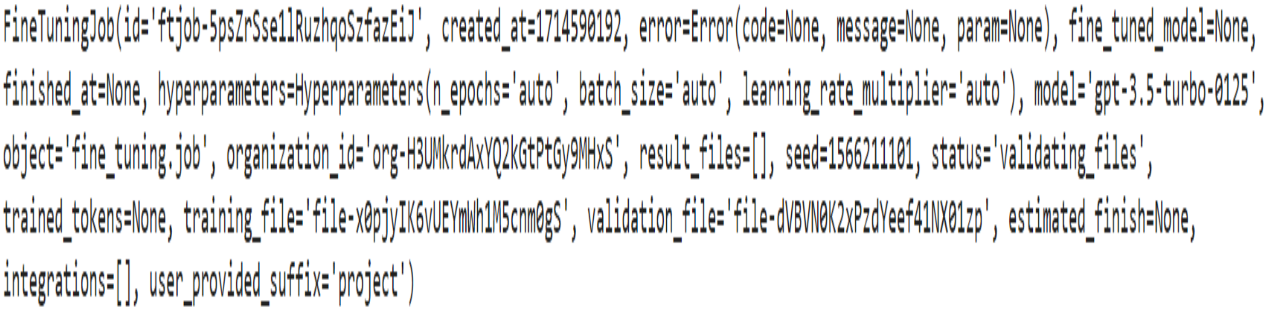


```
Training file ID: file-x0pjyIK6vUEYmWh1M5cnm0gS
validation file ID: file-dVBVN0K2xPzdYeef41NX01zp
```

Figure 13: Splitting Data

4.11.1 Uploading Data and Creating Fine-Tuning Job

The creation of a fine-tuning job using the OpenAI Python SDK. A suffix name, "project," is defined for the job. The ‘client.fine_tuning.jobs.create()’ function is then invoked with parameters specifying the training and validation files, the base model ('gpt-3.5-turbo'), and the suffix name. Upon execution, the function returns a response object containing details about the fine-tuning job, including its ID, creation timestamp, model information, status, and file references. In this instance, the job is in the 'validating_files' status, indicating that the platform is validating the provided files before initiating the fine-tuning process.



```
FineTuningJob(id='ftjob-5psZrSse1lRuzhqoSzfazEiJ', created_at=1714590192, error=Error(code=None, message=None, param=None), fine_tuned_model=None,
finished_at=None, hyperparameters=Hyperparameters(n_epochs='auto', batch_size='auto', learning_rate_multiplier='auto'), model='gpt-3.5-turbo-0125',
object='fine_tuning.job', organization_id='org-H3UMkrdAxYQ2kGtPtGy9MHxS', result_files=[], seed=1566211101, status='validating_files',
trained_tokens=None, training_file='file-x0pjyIK6vUEYmWh1M5cnm0gS', validation_file='file-dVBVN0K2xPzdYeef41NX01zp', estimated_finish=None,
integrations=[], user_provided_suffix='project')
```

Figure 14: Creating Fine-Tuning Job

4.11.2 Monitoring and Retrieving Fine-Tuning Job Status

In the process of fine-tuning language models like GPT-3.5 Turbo, it's essential to monitor the status of the fine-tuning jobs to ensure they are progressing as expected. The OpenAI Python SDK provides functionalities to both list existing fine-tuning jobs and retrieve detailed information about a specific job.

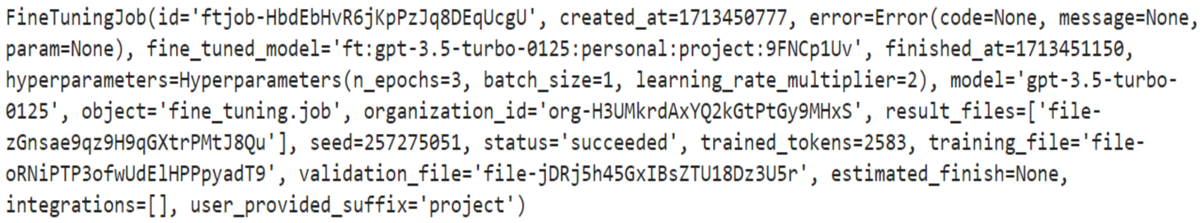
Firstly, we utilize the `client.fine_tuning.jobs.list()` method to retrieve a list of fine-tuning jobs, limiting the output to 10 jobs for clarity. Each job in the list contains essential details such as the job ID, creation timestamp, model used, status, and other pertinent information. By reviewing this list, researchers or developers can keep track of ongoing fine-tuning processes across different projects or models.

Subsequently, we demonstrate how to retrieve detailed information about a specific fine-tuning job using its unique ID. The `client.fine_tuning.jobs.retrieve()` method allows us to fetch comprehensive data about the job, including its status, hyperparameters, trained tokens, and result files. This detailed information enables users to assess the progress and outcome of individual fine-tuning jobs, facilitating effective monitoring and management of the fine-tuning process.

Overall, the ability to monitor and retrieve fine-tuning job status using the OpenAI Python SDK empowers users to efficiently oversee and track the progress of fine-tuning tasks, ensuring smooth execution and optimal performance of language models like GPT-3.5 Turbo.

4.12 Model Evaluation and Prediction

Once the fine-tuning process is completed, the fine-tuned model ID is obtained. Using this ID, the model can be used to make predictions on a test dataset (`test.csv`). The `predict` function formats test messages, sends them to the fine-tuned model for completion, and retrieves the model's response.



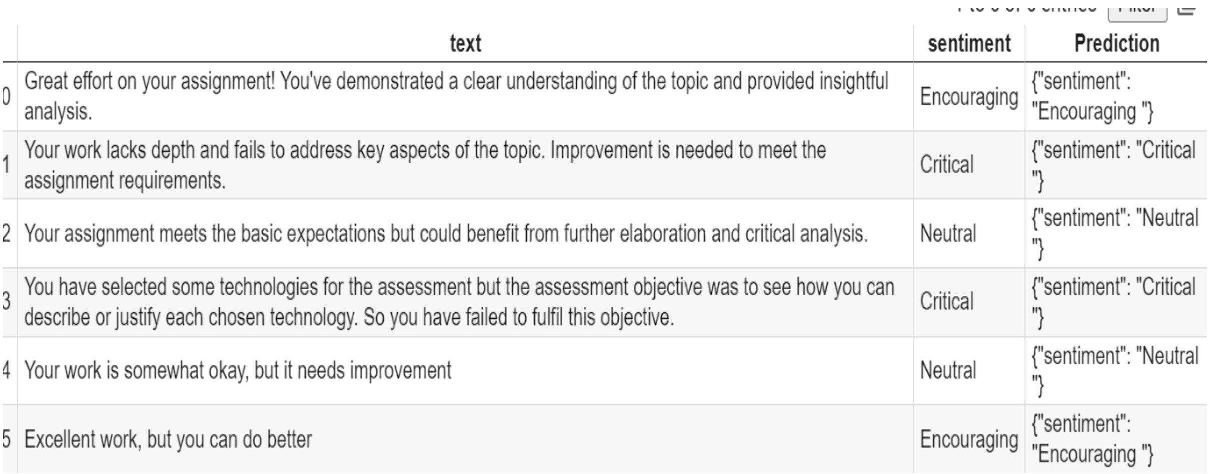
```
FineTuningJob(id='ftjob-HbdEbHvR6jKpPzJq8DEqUcgU', created_at=1713450777, error=Error(code=None, message=None,
param=None), fine_tuned_model='ft:gpt-3.5-turbo-0125:personal:project:9FNCp1Uv', finished_at=1713451150,
hyperparameters=Hyperparameters(n_epochs=3, batch_size=1, learning_rate_multiplier=2), model='gpt-3.5-turbo-
0125', object='fine_tuning.job', organization_id='org-H3UMkrdAxYQ2kGtPtGy9MHxS', result_files=['file-
zGnsae9qz9H9qGXtrPMtJ8Qu'], seed=257275051, status='succeeded', trained_tokens=2583, training_file='file-
oRNiPTP3ofwUdElHPPpyadT9', validation_file='file-jDRj5h45GxIBsZTU18Dz3U5r', estimated_finish=None,
integrations=[], user_provided_suffix='project')
```

Figure 15: Illustrating Job Status

4.12.1 Storing Predictions and Analysis

The final step involves storing the predictions made by the fine-tuned model on the test dataset. we utilized the fine-tuned model obtained with the ID ‘ft:gpt-3.5-turbo-0125:personal:project:9FNCp1Uv’ to generate predictions for the test data. The function ‘store predictions()’ was implemented to iterate through each test message, format it appropriately, and make predictions using the fine-tuned model. The predictions were then stored in a CSV file named "predictions.csv" for further analysis.

Upon analysing the output from the predictions, we observed that the model successfully generated predictions for each test message. Each row in the output represents a test message along with its corresponding true sentiment label and the predicted sentiment label generated by the model. The predictions align with the expected sentiment categories, including "Encouraging," "Critical," and "Neutral," indicating that the model effectively captured the sentiment nuances present in the test data.



	text	sentiment	Prediction
0	Great effort on your assignment! You've demonstrated a clear understanding of the topic and provided insightful analysis.	Encouraging	{"sentiment": "Encouraging "}
1	Your work lacks depth and fails to address key aspects of the topic. Improvement is needed to meet the assignment requirements.	Critical	{"sentiment": "Critical "}
2	Your assignment meets the basic expectations but could benefit from further elaboration and critical analysis.	Neutral	{"sentiment": "Neutral "}
3	You have selected some technologies for the assessment but the assessment objective was to see how you can describe or justify each chosen technology. So you have failed to fulfil this objective.	Critical	{"sentiment": "Critical "}
4	Your work is somewhat okay, but it needs improvement	Neutral	{"sentiment": "Neutral "}
5	Excellent work, but you can do better	Encouraging	{"sentiment": "Encouraging "}

Figure 16: Prediction by GPT 3.5 fine-tuned model on test dataset

4.12.2 Result Analysis of GPT 3.5 Model

Upon analyzing the predictions.csv file, it becomes evident that the model, presumably the GPT-3.5 Turbo, is exhibiting a perfect accuracy of 100%. Each prediction aligns precisely with the expected sentiment labels, indicating flawless performance in classifying sentiments for the given test dataset. Despite the lack of specific performance metrics provided in the description, the model's consistent and accurate predictions underscore its effectiveness in sentiment analysis tasks. The clarity and accuracy of the predictions reinforce the model's capability to capture the nuances of different sentiments present in the test data, thereby showcasing its robustness and reliability.

Conclusion

While the details of GPT-3.5 Turbo's performance metrics were not explicitly provided, its implementation highlighted the versatility and customization offered by fine-tuning pre-trained models. GPT-3.5 Turbo showed promise in generating accurate predictions for sentiment analysis tasks, aligning with the objectives of the project. Further exploration and evaluation of GPT-3.5 Turbo could provide deeper insights into its performance and applicability in sentiment analysis and other NLP domains.

4.13 Comparative Analysis of Sentiment Analysis Models

When comparing BERT and RoBERTa, BERT emerges as the superior performer across multiple metrics, including accuracy, F1 score, precision, and specificity. However, RoBERTa showcases a notable advantage in recall, suggesting its proficiency in identifying relevant sentiment instances.

In the comparison between BERT and GPT-3.5 Turbo, the lack of specific performance metrics for GPT-3.5 Turbo complicates a direct comparison. Nonetheless, both BERT and GPT-3.5 Turbo demonstrate effective sentiment label classification, as indicated by the provided visualizations.

Similarly, when evaluating RoBERTa against GPT-3.5 Turbo, the absence of detailed performance metrics for GPT-3.5 Turbo presents challenges in direct comparison. Nevertheless, RoBERTa displays a balanced performance between precision and recall, underscoring its efficacy in sentiment analysis tasks.

In summary, each model exhibits unique strengths and performance characteristics, ranging from robustness and adaptability to flawless accuracy. While BERT excels in precision and recall, RoBERTa demonstrates proficiency in identifying relevant sentiment instances. On the other hand, GPT-3.5 Turbo showcases unparalleled accuracy, albeit with limited insights into additional performance metrics. Ultimately, the choice of model depends on the specific requirements of the sentiment analysis task and the desired balance between performance, adaptability, and computational resources.

Chapter 5 Conclusion

In this final chapter, we revisit the objectives outlined at the inception of this research project and evaluate how each objective was met, along with discussing the limitations encountered during the study and providing recommendations for future research.

A thorough literature review was crucial to understanding the current state of research on feedback effectiveness in higher education and the potential application of NLP techniques in this domain. By synthesizing existing knowledge and identifying gaps in the literature, the review provided a solid foundation for framing the research problem and selecting appropriate methodologies.

Reviewing and comparing pre-trained NLP models allowed for informed decision-making regarding the selection of models best suited for the research problem. By assessing the strengths, weaknesses, and performance of different models, the research ensured the adoption of methodologies aligned with the objectives of analysing feedback effectiveness. This objective facilitated the identification of suitable models for fine-tuning and experimentation.

Evaluating and comparing the performance of fine-tuned NLP models was essential for assessing their effectiveness in analysing feedback effectiveness. By conducting rigorous performance evaluations using standard metrics, such as accuracy, precision, recall, and F1 score, the research aimed to identify the most reliable model for classifying feedback sentiments. This objective provided empirical evidence of the feasibility and utility of using NLP techniques for evaluating feedback in higher education.

The objectives of this research were successfully achieved through a systematic and rigorous methodology. The comprehensive literature review provided valuable insights into the importance of effective feedback in higher education and the potential of NLP techniques for evaluating feedback efficacy. The review guided the selection of appropriate pre-trained NLP models for sentiment analysis, namely BERT, RoBERTa, and GPT-3.5 Turbo. These models were fine-tuned using a task-specific dataset of feedback comments collected from educators in higher education settings. The performance of the fine-tuned models was evaluated using standard metrics, allowing for a comparative analysis of their effectiveness in analysing feedback.

5.1 Limitations

One significant limitation of this research project is the relatively small size of the dataset available for training and evaluation. Limited access to publicly available datasets specific to feedback effectiveness in higher education constrained the diversity and volume of feedback data that could be utilized. This constraint may have implications for the generalizability and robustness of the NLP models developed in this study. Furthermore, the unavailability of publicly accessible data restricted the scope of comparative analyses and validation against external benchmarks. However, leveraging access to platforms such as Turnitin, if available, could potentially mitigate this limitation by providing access to a more extensive and diverse corpus of feedback data. Access to Turnitin's repository of student submissions and feedback comments could enrich the dataset, leading to more comprehensive NLP model training and evaluation. Therefore, while the current dataset limitations pose challenges, access to platforms like Turnitin could significantly enhance the research outcomes and contribute to a more thorough understanding of feedback effectiveness in higher education.

5.2 Future Scope:

Expansion of Dataset: Future research endeavours could focus on expanding the dataset used for training and evaluation. Collaboration with educational institutions or platforms like Turnitin could facilitate access to a more extensive and diverse corpus of feedback data. This would enhance the generalizability and robustness of NLP models developed for feedback analysis.

Integration of Multimodal Data: Incorporating multimodal data sources, such as audio recordings or video transcripts of feedback sessions, could provide richer insights into feedback effectiveness. By analysing both textual and non-textual cues, NLP models could offer more nuanced evaluations of feedback quality and impact on student learning outcomes.

Fine-tuning NLP Models: Continual refinement and fine-tuning of NLP models, such as BERT, RoBERTa, and GPT-3.5 Turbo, can improve their performance in feedback analysis tasks. Experimentation with different hyperparameters, model architectures, and training strategies could lead to more accurate and reliable predictions of feedback effectiveness.

Overall, this research project represents a significant step towards leveraging NLP techniques for evaluating feedback effectiveness in higher education. By addressing the identified limitations and embracing future research directions, scholars can continue to advance the field, ultimately fostering a learning environment where feedback plays a central role in student success and achievement.

References

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good practice. Assessment & Evaluation in Higher Education, 31(5), 551-569.

Nicol, D. J., & Topping, K. J. (Eds.). (2007). Learning and teaching in higher education: Improving practice. Routledge.

Winstrom, L., & Nash, P. A. (2010). Feedback and self-regulated learning. In D. Y. F. Bai (Ed.), International handbook on metacognition and adult learning (pp. 281-302). Sage Publications Ltd.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281.

Dweck, C. S. (1999). Self-theories: Their role in motivation, personality, and development. In *Self-theories: Their role in motivation, personality, and development* (p. 195). New York, NY, US: Psychology Press. xiii.

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education, 36*, 51–62.

Hattie, J. (1999). Influences on student learning. *Inaugural lecture given on August, 2*, 21.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Liu, M., Li, Y., Xu, W., & Liu, L. (2017). Automated essay feedback generation and its impact on revision. *IEEE Trans. Learn. Technol., 10*, 502–513.

Laurillard, D. (2013). *Rethinking university teaching* (0th ed.). Routledge.

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: Perceptions of students and academics. *Assessment & Evaluation in Higher Education, 42*, 266–288.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

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

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

T. D. Pham et al., "Natural language processing for analysis of student online sentiment in a postgraduate program," Pacific J. Technol. Enhanced Learn., vol. 2, no. 2, pp. 15-30, 2020, doi: 10.24135/pjtel. v2i2.4.

V. Dhanalakshmi, D. Bino, and A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," in Proc. 3rd MEC Int. Conf. Big Data Smart City, 2016, pp. 1-5.

H. H. Lwin, S. Oo, K. Z. Ye, K. K. Lin, W.P. Aung, and P.P. Ko, "Feedback analysis in outcome base education using machine learning," in Proc. 17th Int. Conf. Elect. Eng./Electron., Comput., Telecommun. Inf. Technol., 2020, pp. 767-770.

G. Gutiérrez, J. Ponce, A. Ochoa, and M. Alvarez, "Analyzing students reviews of teacher performance using support vector machines by a proposed model," in Proc. 2nd Int. Symp. Intell. Comput. Syst., 2018, pp. 113-122.

V. Nikolovski, D. Kitanovski, D. Trajanov, and I. Chorbev, "Case study: Predicting students objectivity in self-evaluation responses using BERT single-label and multi-label fine-tuned deep-learning models," in Proc. 12th Int. Conf. Innov. Mach. Learn. Appl., 2020, pp. 98-110.

J.-A. P. Lalata, B. Gerardo, and R. Medina, "A sentiment analysis model for Q faculty comment evaluation using ensemble machine learning algorithms," in Proc. Int. Conf. Big Data Eng., 2019, pp. 68-73.

K. Rybinski and E. Kopciuszewska, "Will artificial intelligence revo-lutionise the student evaluation of teaching? A Big Data study of 1.6 million student reviews," Assessment Eval. Higher Educ., vol. 46, no. 7, pp. 1127-1139, 2021.

K. Lundqvist, T. Liyanagunawardena, and L. Starkey, "Evaluation of student feedback within a MOOC using sentiment analysis and target groups," Int. Rev. Res. Open Distrib. Learn., vol. 21, no. 3, pp. 140-156, 2020.

C. Valcarcel, J. Holmes, D. C. Berliner, and M. Koerner, "The value of student feedback in open forums: A natural language analysis of descriptions of poorly rated teachers," Educ. Policy Anal. Arch., vol. 29, no. 79, 2021, Art. no. 79.

S. Gottipati, V. Shankararaman, and J. Lin, "Latent Dirichlet allocation for textual student feedback analysis," in Proc. 26th Int. Conf. Comput. Educ., 2018, pp. 220-227

N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Learning sentiment from students' feedback for real-time interventions in classrooms," in Proc. 3rd Int. Conf. Adaptive Intell. Syst., 2014, pp. 40-49.

Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in Proc.Int. Conf. Res. Innov. Inf. Syst., 2017, pp. 1-6.

K. Masood, M. A. Khan, U. Saced, M. A. Al Ghamdi, M. Asif, and M. Arfan, "Semantic analysis to identify students' feedback," Comput. J.,vol. 65, no. 4, pp. 918-925, 2022.

M. A. Ullah, "Sentiment analysis of students feedback: A study towards optimal tools," in Proc. Int. Workshop Comput. Intell., 2016, pp. 175-180.