Literature Review: Comparing Transformer-Based and Traditional Models Enhanced with Explainable AI for Interpreting Informal Sentiment in Student Feedback

Sreenivas X23326603

National College of Ireland

**Abstract**

This research examines the performance of XAI-explained transformer approaches when they ana- lyze sentiment expressions found within informal online student feedback. Support Vector Machines (SVM) maintain clear explanations but their capability to understand complex emotional educational feedback falls inadequate. Compatibility issues arise because BERT-type transformers master con- textual understanding while functioning as non-disclosure systems which restrict their educational suitability. The proposed research conducts a comparative assessment between traditionalmodels and transformer models through the usage of XAI tools LIME and SHAP to produce explainable results. This research fills the accuracy-interpretability gap to develop trustable sentiment analysis systems which provide explanations in educational environments.

**Keywords:** Sentiment Analysis, Explainable AI, Transformer Models, Student Feedback, Natural Language Processing

# Introduction

Students use Coursera and Udemy educational platforms as central tools for modern learning and provide thousands of evaluations that document their educational journeys. The informal feedback contains many abbreviations together with slang phrases and emojis and emotional signs. The complex language structure of such content creates obstacles for Natural Language Processing techniques that try to detect sentiment signals in this type of information.

Traditional sentiment analysis models Na¨ıve Bayes and Support Vector Machines (SVM) are val- ued because they provide straightforward understanding in addition to ease of interpretation. These processing models exhibit acceptable performance on structured systems yet experience issues when pro- cessing casual student comments which exhibit spontaneous verbalization. They struggle to comprehend subtle contextual hints and strong emotional signals which appear throughout casual speaking.

BERT along with RoBERTa demonstrate extensive contextual comprehension due to their self- attention ability which distinguishes them from traditional models. The models demonstrate superi- ority in every state-of-the-art sentiment analysis evaluation. Black-box status is a common criticism of these models because they obscure their inner mechanisms to end users – a drawback that impedes accountability in learning environments which require explanation.

The gap in interpretable artificial intelligence has been filled through LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) tools. The tools trace com- plex model choices to make both processes and systems transparent thus building user trust. The addition of XAI tools to NLP pipelines enables educational institutions to understand the reasoning behind sentiment analysis outcomes in addition to automation functionality.

The research evaluates different approaches to sentiment analysis of informal student feedback by researching traditional and transformer-based models with integrated XAI explanations. The research has the goal to discover the optimal accuracy and interpretability pairing for educational applications.

**Research Question:** How do transformer models augmented with explainable AI compare to tra- ditional models in interpreting sentiment from noisy, informal student feedback in online learning plat- forms?

# Literature Review

## Traditional Approaches to Sentiment Classification

Traditional sentiment classification has been powered by Na¨ıve Bayes algorihm alongside Support Vector Machines (SVM). Linear boundaries enable speedy and easy interpretation of these models that work in high-dimensional spaces. MOOCs educational settings featuring student feedback containing slang terms and abbreviations with emotion dimensions pose significant challenges for traditional sentiment detection systems as most previous systems used binary sentiment detection methods according to [(1).](#_bookmark0) The models provide clear transparency but lack the ability to detect vital contextual details which education environments require for correct interpretation.

## CNN and RNN-Based Sentiment Analysis

The deep contextual analysis of sentiment analysis became possible through the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) combined with syntactic and temporal syntactic structure preservation. [(2)](#_bookmark1) introduced a BERT-CNN hybrid that improved emotion detection on ISEAR and SemEval datasets. Performance improvements were noted while the models kept their unexplainable nature because they included no built-in method to clarify which predictions were gen- erated. Education requires explainable models due to the necessity for teachers to understand system decisions that determine curriculum alterations so the models’ difficult-to-determine processes restrict their use in educational settings.

## Transformer-Based Models

The natural language processing field has experienced a transformation due to BERT and RoBERTa transformers which allow input sequences to receive simultaneous attention from all directions. [(3)](#_bookmark2) cre- ated a model employing transformer embeddings together with attention-based RNNs to handle student feedback with remarkable accuracy results. Transformer-based solutions work as unexplained informa- tion systems. The predictive success of these models remains excellent yet their nontransparent nature has become a concern thus driving a necessary convergence between these models and Explainable AI (XAI) systems.

## Emotion Label Diversity with GoEmotions

The GoEmotions dataset which [(4)](#_bookmark3) established contains 58,000 Reddit comments that were labeled with 27 specific emotional categories. This dataset enables detailed emotional classification which goes well beyond basic positive and negative sentiments thus enabling enhanced emotional perceptiveness. The re- searching domain of Reddit restricts direct implementation of its approach in evaluating student feedback because such commentary contains educational messaging and teaching-focused content. GoEmotions remains a standard test for measuring the emotion detection capabilities of transformer models.

## Explainable AI in NLP

Research reviews XAI tools for model interpretation especially for high-stakes domain education. Model reasoning steps become accessible through these explanation tools which permit users to track predictive reasoning. When NLP model frameworks receive applications they permit educators to observe the linguistic indicators that produce specific sentiment predictions thus building user trust and system usefulness.

## Explainability in Educational NLP

Educational feedback evaluation by [(5)](#_bookmark4) revealed that NLP applications commonly maintain opaque model outputs. The research supports both accuracy improvements and explainability additions due to interpreter functions which allow educators to respond constructively to student concerns about artifi- cial intelligence systems. When information systems lack transparency institutions often fail to notice important feedback despite its existence.

## Emotion and Engagement in AI Learning Systems

The core element for personalized learning systems consists of identifying emotional states and engage- ment through techniques described by [(6).](#_bookmark5) The main objective in sentiment analysis requires more than classification outcomes because actionable insights serve as the ultimate goal. The combination of XAI with emotion modeling makes the delivered insights both understandable and educationally meaningful for human educators.

The application of Explainable Transformers deals with Aspect-Based Sentiment Analysis. Trans- formers within aspect-based sentiment analysis (ABSA) demonstrate efficient management of nuanced sentiment classification operations. RoBERTa obtained the best performance level through evaluation of transformer-based models BERT, RoBERTa, and XLNet when operating on MAMS and SemEval ABSA benchmark datasets. The research improved interpretability through the integration of LIME, SHAP, integrated gradients, attention weight visualization, and Grad-CAM as explainable AI techniques. The analytical tools allowed researchers to uncover both context-based relations and vital elements that affected prediction results. These added techniques enable tracking how predictions relate to specific language patterns specifically important in educational sentiment analysis because of its essential need for transparency [(7).](#_bookmark6)

## Systematic Review of Transformer-Based Sentiment Analysis in Educa- tion

The research community has recently conducted a systematic review of transformer architecture applic- ations in educational sentiment analysis. The research reviewed forty-one peer-reviewed publications from 2017 until now and grouped their data sources into MOOCs and social media and audio surveys. Transformer-based models demonstrate exceptional performance at analyzing informal and emotionally charged feedback which characterizes online learning feedback processes according to the review results. Research pointed out that computational complexity and inadequate cultural adaptability were shown to be major limitations. The review endorses the implementation of Explainable AI tools to enhance user trust and interpretability in educational system applications of these models [(8).](#_bookmark7)

## Interpretable Transformers in Virtual Learning Environments

The research presented a transformer-based hierarchical model which interpreted student review senti- ment through identifying sentence summaries with maximum sentiment information. The model achieved sentiment classification with an additional benefit of providing extractive summary explanations through an attention-based method. Through their dual-task implementation educators gained immediate access to sentiment polarity along with key supporting text from the input. These models produce better trust relationships with users along with decreased bias which indicates their effectiveness for virtual learning student feedback scenarios. The proposed solution resolves performance-quality conflicts which exist between sentiment analysis applications [(9).](#_bookmark8)

## Explainability Challenges in Deep Learning-Based Sentiment Analysis

The broad survey has studied sentiment analysis along with Explainable Artificial Intelligence (XAI) by finding deep neural networks to be the main framework for sentiment categorization. The advanced predictive accuracy of models such as BERT and LSTM creates issues because educational settings need transparent explanations for necessary prediction validations.

The paper organizes explainability methods under two categories consisting of intrinsic (ante-hoc) and post-hoc approaches. Interpretable linear classifiers and decision trees form part of the ante-hoc methods while post-hoc models like SHAP and LIME work with already trained black-box platforms. The authors highlight local interpretability since it delivers single prediction explanations that prove both more achievable and practical to use in NLP tasks which include student feedback analysis.

Transformer models feature attention mechanisms that researchers use for making interpretability improvements through token-based visualizations of decision-influencing input tokens. The studied ap- proaches provide significant value to this research because they make predictions on informal student feedback more transparent.

The survey data validates the primary research assumption which states that BERT-type models com- bined with XAI tools enable accurate and easily understood sentiment analysis solutions for education- based applications.

## Research Niche

A major problem arises from transformer models achieving better accuracy than traditional models since educational situations demand students to understand model decisions. The goal of this study is to unite XAI techniques with transformer-based models and analyze them against standard interpretable models such as SVM. The research introduces a method which assesses both interpretation accuracy and quality to suggest appropriate tools that explain sentiment in student feedback.

# Research Method & Specification

## Research Method

The research employs experimental comparison to measure how effective and explainable two different sentiment analysis methods are in the processing of unstructured student feedback. Student comments gathered from Udemy and Coursera learning platforms make up the research domain. The large text data available on these platforms presents an issue for standard NLP systems because users often utilize informal language, slang, short forms, emojis and contextual sentiment cues that make classification difficult.

Two research pipelines will be built and assessed to tackle the main inquiry on how transformer models with explainable AI perform relative to traditional models when analyzing sentiment within informal student feedback data on online learning platforms.

### Model 1: Traditional NLP Pipeline

This pipeline will consist of:

A pipeline includes Text Preprocessing steps which consist of Tokenization followed by stop- word removal then lemmatization and vectorization through Term Frequency-Inverse Docu- ment Frequency (TF-IDF). A Support Vector Machine classifier relies on the TF-IDF vectors for classification purposes. Simple interpretability stems from ranking features based on linear SVM coefficients that provide a baseline explainability method.

The method has gained popularity because it provides straightforward explanations with simple application. The method demonstrates low performance levels when applied to contextual and informal language situations.

### Model 2: Transformer-Based Pipeline with XAI

This pipeline includes:

The model utilizes a BERT (Bidirectional Encoder Representations from Transformers) ver- sion which received training from the same information base as Model 1. BERT will undergo fine-tuning on sentiment-labeled datasets during which early stopping and learning rate op- timization techniques will optimize performance. Programs like LIME and SHAP will serve as post-hoc explanation tools for creating model-agnostic explanations during the integration stage. The study will conduct attention weight visualization to analyze which words in the text drove sentiment determinations.

Organizations will achieve a balance between predictive power and interpretability through present- day modeling innovation coupled with interpretability methods.

### Data Collection and Labeling Strategy:

Student feedback acquired from Coursera and Udemy course discussion areas and reviews becomes the data source according to an ethical collection method. Each piece of information will receive full anonymization treatment as a privacy protection measure. The annotation process combines two approaches for labeling student feedback through which researchers will:The GoEmotions dataset and similar resources serve as existing datasets to share inform- ation for the project. The annotation process involves prompting GPT-4 to create sentiment labels which will be validated by humans for maintaining accuracy levels and consistency.

### Experimental Design:

Different split groups within the dataset will serve both pipelines for training and testing procedures in order to permit comparative evaluation. The data will be separated into training and validation

amounts of 70Both evaluation methods will measure quantitative elements like F1 scores while simultaneously testing model explanation abilities to provide interpretability insights.

The research employs two predictive models to measure both prediction accuracy and observation of explanation transparency as well as user-friendly explanation generation among both models. The research findings will help determine how AI systems should be used for educational sentiment analysis tasks that require interpretable models.

## Research Resources

The research implementation depends on contemporary open-source tools that prioritize platform ac- cessibility and scientific transparent operation and repeatable experimentation. The chosen tools and resources offer both quick experimentation possibilities along with scaleable model training in addition to rapid deployment capabilities for Explainable AI methods. The project employs the following complete list of technical resources:

* + - **Programming Language:** Python 3.11

Python stands out since its provides an extensive set of machine learning along with NLP libraries in addition to integration capabilities with visualization tools and explainability frameworks.

* + - **Development Environment:** Jupyter Notebook and Google Colab

The project relies on Jupyter for visualizing programming tasks and document creation with integ- ration of the free GPUs in Google Colab (utilizing NVIDIA Tesla T4 performance). The research team benefits from Google Drive storage by receiving both expandable and accessible research development capabilities.

### Core Libraries and Frameworks:

The data processing tasks utilize pandas, numpy, NLTK, along with spaCy to normalize and tokenize and preprocess sentiment values. The text scraping process uses BeautifulSoup in combination with GPT-based prompts for automatic annotation of the collected data. The project uses scikit-learn for SVM and Na¨ıve Bayes basic models and employs HuggingFace Transformers for BERT-based model development and trials. Local (per prediction) explana-

tion generation and global explanations emerge from the use of LIME, SHAP, and transformers-interpret.

**Visualization:** matplotlib, seaborn, and plotly for data analysis and interpretation of model behavior through graphs and heatmaps.

### –• Datasets:

The research uses sentiment-labeled student reviews which were ethically obtained through scraping publicly accessible Coursera and Udemy sections. The GoEmotions Dataset serves as a strong emotion classification framework featuring 27 emotion tags for model assessment through transformer-based emotion detection methods.**IMDb/Stanford Sentiment Tree- bank (optional):** For cross-domain sentiment robustness testing.

### –• Hardware and Compute Resources:

The **Cloud Compute** component relies on Google Colab Pro as either a paid or unpaid version that utilizes the NVIDIA Tesla T4 or P100 GPUs when performing transformer per- formance optimization. The installation features an optional backup feature which maintains synchronization between Anaconda-based Python environment on local devices for offline val- idation purposes and reproducible results.

### Version Control and Reproducibility:

Every code file together with models and documentation will reside inside a dedicated private GitHub repository and implement commit tracking alongside environment logging and readme documentation to support replication. The sharing of annotated experiments plus visual insights utilizes either the Jupyter Book or Markdown Export feature.

The selected resources enable both the research’s solid technical foundation and its ability to scale and repeat experiments. This system provides standardized evaluation through both traditional and deep learning NLP frameworks while streamlining Explainable AI integration processes for educational applications.

## Evaluation

Evaluation of the models will include multiple quantitative metrics for measuring their effectiveness while simultaneously conducting qualitative assessments of their interpretability features. A dual eval- uation process verifies that the models deliver both precise results and deliver information that enables educational stakeholders to act on them.

### Model Performance:

Standard evaluation metrics Accuracy, Precision, Recall and F1-Score will help measure how well the classifiers identify positive and negative and neutral sentiment categories. The as- sessment of F1-Score remains crucial when dealing with data that is unbalanced between classes. Sytematic classification errors will become apparent through confusion matrix ana- lysis which helps understand how well the model recognizes different sentiment categories. The K-Fold cross-validation process will be implemented for assessing generalizability while ensuring the model does not accept false relationships. The configuration protects consist- ent model behavior when working across various training-testing partitions. A paired t-test and Wilcoxon signed-rank test will evaluate metric distribution results between models for determining statistically meaningful performance differences.

### Interpretability of Predictions:

The term *Fidelity* represents the ability of local surrogate explanations derived from LIME or SHAP to reproduce original model operations. Higher fidelity indicates that the explanation accurately depicts the actual decision boundary of the model. The evaluation determines if the selected features from the explanation function as an independent unit capable of generating identical predictions. The quality of explanation directly relates to having a comprehensive approach. Educational experts will perform human evaluation of explanations to determine their clarity as well as trustworthiness and useful usability. Feedback acquisition happens through surveys and Likert-scale questionnaires which measure subjective interpretability. The observation of Cognitive Load will evaluate explanations based on their cognitive com- plexity to determine their understanding difficulty. These assessment methods will guarantee stakeholders who lack technical training can still get meaningful value from them. The assess- ment includes analyzing heatmaps and token-level importance visualizations especially from attention-based models to establish their visual effectiveness and human expectability.

Multiple methods of data triangulation including quantitative and qualitative measures will serve to confirm the research results. The experimental assessment will evaluate which model pipeline between a traditional SVM system and a transformer-based BERT pipeline integrates XAI produces the most actionable and educationally relevant and interpretable insights. The investigation will produce re- commendations about choosing the most effective model type for real-life academic feedback analysis systems.

## Ethical Considerations of the Research

Research operations at the National College of Ireland abide by ethical standards defined within their research integrity guidelines. Strong emphasis exists in ethical data management of educational feedback since its content frequently contains subjective or emotionally sensitive information. Ethical considera- tions include:

* + - **Data Privacy:** Student confidentiality remains protected by handling all received feedback through an anonymization process. The system will not retain any personal information consisting of stu- dent names together with email addresses or Internet Protocol IP addresses or pseudonyms.
    - **Transparency in Labeling:**Task documentation will specify all steps regarding the application of GPT for sentiment marking alongside prompt development techniques. A validation process will run to decrease biased labels and confirm uniformity amongst multiple samples.
    - **Platform Compliance:** Only public data sources supporting scraping and academic research reuse will provide data for the study. Applications’ service terms (among them Coursera and Udemy) must be respected through strict implementation.
    - **Informed Consent (where applicable):** The study requires students to explicitly agree to survey analysis and academic data collection procedures to provide assurance about data usage and storage practices.
    - **Bias Mitigation:** The study evaluates and presents evidence about biases within models which mainly trained using English text from Western educational settings together with an acknowledge- ment of their cultural and linguistic boundaries.
    - **Model Fairness and Explainability:** XAI techniques including LIME and SHAP serve to uphold the ethical requirement of transparency. Educational organizations together with their teaching staff need to grasp AI model decision processes particularly during labeling errors and unpredictable outcomes.
    - **Reproducibility:** A public repository hosts the sharing of all code alongside datasets where authorization allows in addition to model configurations and evaluation outcomes. The academic principles embracing transparency together with peer verification and open science gain support through this practice.
    - **Responsible Reporting:**The findings will be presented with transparency while restraining ex- cessive claims regarding model accuracy or its applicability especially when dealing with students from culturally diverse backgrounds.
    - **Use of Generative AI:** The paper will identify and provide detailed information about the GPT and other large language models utilized during analysis to prevent confusion about human- generated outputs.
    - **Long-term Impact Awareness:** The research analyzes prospective influences on educational application of automated sentiment systems while examining howAlgorithmic decision systems might unfairly modify educational curriculums or student assessment processes.

Use of Explainable AI tools creates its own ethical requirement because they provide transparent interpretation and accountability especially in education where prediction reasoning needs understanding.

## Project Plan

The Capstone project continues for three months extending from May 20 until August 15 during the academic year 2025. The project schedule divides into four distinct periods with specific work assignments that result in both the end report and presentation section.

### Phase 1: Research Foundation (May 20 – June 10)

The first stage requires conducting an extensive literature review followed by collecting ethical datasets from Coursera and Udemy and applying both manual and semi-automated methods to label the received feedback.

The project moved into model development from June 3 until July 5. A Support Vector Machines (SVM) model serves as the base for this NLP development followed by transformer model fine- tuning based on BERT. The integration of Explainable AI tools LIME and SHAP will occur during this stage for interpretability purposes.

### Phase 3: Evaluation and Analysis (July 6 – July 20)

The evaluation phase incorporates standard performance metric assessment using Accuracy and F1 scores together with a qualitative assessment of XAI explanations that involves educator reviews.

### Phase 4: Finalization (July 16 – August 15)

The last stage which starts along with Phase 3 includes project report preparation followed by peer evaluation while documenting results until the final project submission.

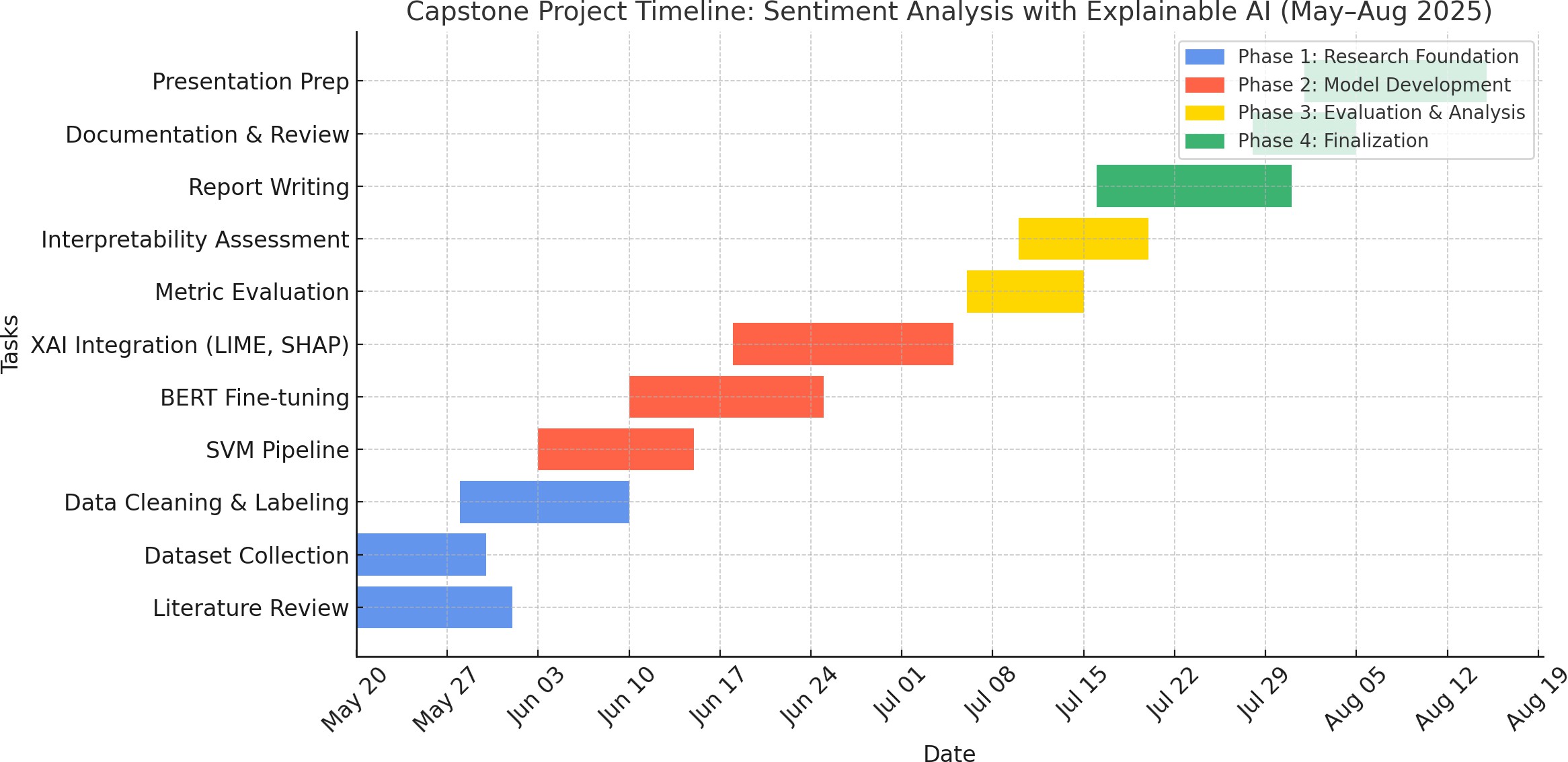


Figure 1: Capstone Project Timeline (May 20 – August 15, 2025)

# References

1. F. Dalipi, K. Zdravkova, and F. Ahlgren, “Sentiment analysis of students’ feedback in moocs: A systematic literature review,” *Frontiers in Artificial Intelligence*, vol. 4, p. 728708, 2021.
2. A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, “Bert-cnn: A deep learning model for detecting emotions from text,” *Computational Materials Continua*, vol. 71, no. 2, pp. 2943–2959, 2022.
3. I. Zyout and M. Zyout, “Sentiment analysis of student feedback using attention-based rnn and trans- former embedding,” *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 2173– 2184, 2024.
4. D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, “Goemotions: A dataset of fine-grained emotions,” *arXiv preprint arXiv:2005.00547*, 2020.
5. A. S. Sunar and M. S. Khalid, “Natural language processing of student’s feedback to instructors: A systematic review,” *IEEE Transactions on Learning Technologies*, 2023.
6. M. B. Garcia *et al.*, “Understanding student engagement in ai-powered online learning platforms: A narrative review of key theories and models,” in *Handbook of Research on Redesigning Teaching, Learning, and Assessment in the Digital Era*, ch. 1, IGI Global, 2024.
7. S. M. Aghaei *et al.*, “Explainable aspect-based sentiment analysis using transformer models,” *Big Data and Cognitive Computing*, vol. 8, no. 11, p. 141, 2023.
8. A. Pilicita-Garrido and E. Barra, “Sentiment analysis with transformers applied to education: Sys- tematic review,” *International Journal of Interactive Multimedia and Artificial Intelligence*, 2025. In press.
9. F. L. Rosa, A. Mazzeo, and G. Mazzulla, “Explainable sentiment analysis: A hierarchical transformer- based extractive summarization approach,” *Electronics*, vol. 10, no. 18, p. 2195, 2021.