Project and Data Management (PDM) Plan

# 1. Project Overview

**Project Title:**  
Building an Emotion Classification Model Using Natural Language Processing (NLP)

**Project Introduction and Background:**  
Emotion classification is a critical subfield of NLP that focuses on identifying human emotions from textual data. Applications range from sentiment analysis in customer feedback to mental health monitoring via social media. Despite advancements in NLP techniques like transformer-based models (e.g., BERT, GPT), the performance of these models on emotion-specific tasks remains understudied. This project aims to evaluate the effectiveness of various NLP methodologies in accurately classifying emotions from text, leveraging a balanced dataset to mitigate biases common in emotion-labeled data.

The significance of this work lies in its potential to enhance human-computer interaction systems, improve mental health diagnostics, and refine customer service automation. By comparing traditional machine learning algorithms (e.g., SVM, Naive Bayes) with deep learning approaches (e.g., LSTM, BERT), this project will provide insights into optimal techniques for emotion recognition tasks.

**Research Question:**  
How well do different NLP techniques perform on emotion classification tasks?

**Project Objectives:**

1. Compare the accuracy, precision, and recall of traditional machine learning models versus deep learning architectures for emotion classification.
2. Implement and fine-tune state-of-the-art NLP models (e.g., BERT) on the emotion-balanced dataset.
3. Evaluate the ethical implications of using publicly available emotional data, ensuring compliance with GDPR and institutional policies.
4. Develop a reproducible workflow using GitHub for code version control and collaboration.

**Reference List:**

1. Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
2. Mohammad, S.M. and Turney, P.D., 2013. Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3), pp.436-465.
3. Hutto, C. and Gilbert, E., 2014, May. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

# 2. Project Plan: Task List and Timeline

**Task List:**

1. Literature Review (Week 1–2):
   * Survey existing studies on emotion classification and NLP techniques.
   * Identify gaps in model performance and dataset biases.
2. Data Collection & Preprocessing (Week 3–4):
   * Download the [emotion-balanced dataset](https://huggingface.co/datasets/AdamCodd/emotion-balanced) from Hugging Face.
   * Clean and tokenize text data; balance class distributions if required.
3. Model Implementation (Week 5–7):
   * Train traditional models (SVM, Naive Bayes) using TF-IDF features.
   * Fine-tune pre-trained transformers (BERT, DistilBERT) on the dataset.
4. Evaluation & Analysis (Week 8–9):
   * Compare model performance using metrics: accuracy, F1-score, confusion matrices.
   * Interpret results to identify strengths/weaknesses of each approach.
5. Report Writing & Presentation Preparation (Week 10–11):
   * Draft the final report and create presentation slides.
   * Rehearse presentation delivery and anticipate questions.
6. Ethical Compliance Check (Ongoing):
   * Ensure data usage aligns with GDPR and UH ethical policies.

**Timeline (Gantt Chart):**

| Task | Weeks 1-2 | Weeks 3-4 | Weeks 5-7 | Weeks 8-9 | Weeks 10-11 |
| --- | --- | --- | --- | --- | --- |
| Literature Review | ██████ | ███ |  |  |  |
| Data Preprocessing |  | █████████ | █ |  |  |
| Model Implementation |  |  | ████████ | ███ |  |
| Evaluation |  |  |  | ████████ | █ |
| Report & Presentation |  |  |  |  | ██████████ |

# 3. Data Management Plan

**Dataset Overview:**

* **Source:** [AdamCodd/emotion-balanced](https://huggingface.co/datasets/AdamCodd/emotion-balanced" \t "_blank) on Hugging Face.
* **Original Purpose:** Curated for balanced emotion classification (e.g., joy, sadness, anger).
* **Data Fields:**

The data fields are:

* + text: a string feature.
  + label: a classification label, with possible values including sadness (0), joy (1), love (2), anger (3), fear (4), surprise (5).
* **Data Splits**

The dataset has 2 configurations:

* + split: with a total of 20\_000 examples split into train, validation and test
  + unsplit: with a total of 89\_754 examples in a single train split
* **Size:** ~20,000 records; estimated 2.46MB in CSV format.

**Data Collection & Storage:**

* Data will be downloaded directly from Hugging Face.
* Primary storage: GitHub repository (private) and University OneDrive.
* Backup frequency: Weekly commits to GitHub.

**Metadata:**

* File formats: CSV (text data), Python scripts (code).
* Expected codebase size: <1GB (including model checkpoints).

**Document Control:**

* **GitHub Repository:** will be shared once repo is created after project proposal approval
* **Commit Frequency:** Weekly updates with descriptive commit messages (e.g., "Added BERT fine-tuning script").
* **File Naming:** Structured as [task]\_[date]\_v[version].ext (e.g., data\_preprocess\_2025-01-10\_v1.py).

**ReadMe File Contents:**

* Dataset source and preprocessing steps.
* Instructions to replicate model training and evaluation.
* Dependencies list (Python libraries, hardware requirements if any).

**Security & Ethical Compliance:**

* **GDPR Compliance:** Dataset is anonymized and publicly available for research.
* **UH Ethical Policies:** Confirmed via UH ethics checklist (no personal/identifiable data used).
* **Permissions:** Dataset is open-access under Creative Commons license (CC-BY 4.0).
* **Original Ethics:** Data sourced from platforms where users consented to public use.