University of Hertfordshire UH

Project and Data Management Plan

Student Name: Jhansi Bajepalli

Student Number: 22019746

1 Project Overview

1. Project Title

Text Sentiment Analysis and Classification Based on Large Language Models and Advanced Neural Network Techniques.

2. A Short summary of the project topic and background

An area of AI called " NLP " studies how computers understand and utilise human languages (Devlin *et al.*, 2019). Utilising neural network models such as Transformers, text summarisation pipeline may include operations such as segmenting sentences, extracting features, and applying both abstractive and extractive summarisation methods (Baqach and Battou, 2023)(Redhu, 2018)(Taha *et al.*, 2024).

The present work is particularly focused on text sentiment analysis and classification that employ LLMs and neural network algorithms. In this research, the IMDB dataset is used, which is collected from Kaggle. The Data preprocessing includes data cleaning, feature extraction, data splitting, classification models, and model assessment procedures will all be employed in this system development. The suggested model will be assessed using a variety of performance assessment measures, including the confusion matrix, F1 score, recall, sensitivity, specificity, accuracy, and precision in classification.

3. Research Question

The research questions are as follows:

- **RQ1:** What is the effectiveness of large language models combined with advanced techniques in accurately classifying and analyzing sentiment in textual data?
- **RQ2:** How do different feature extraction methods impact the performance of sentiment classification models on the IMDB dataset?
- **RQ3:** What are the comparative strengths and weaknesses of the developed models based on various performance metrics?

4. Project Aim and Objectives

The aim of this research is to develop an advanced sentiment analysis model for movie reviews using LLMs and sophisticated neural network techniques, enabling accurate classification of sentiments expressed in textual data. Project objectives are provided below:

- To conduct data pre-processing and cleaning on the IMDB movie review dataset to ensure a quality and relevance of an input data for analysis.
- To extract meaningful features from the pre-processed data that will enhance the model's ability to classify sentiments accurately.
- To implement various classification models using LLMs and advanced NN techniques to categorize movie reviews based on their sentiment.
- To assess how well the suggested model performs in terms of classification F1 score, confusion matrix, recall, sensitivity, accuracy, and precision.

- To demonstrate model's adaptability for handling large datasets and its applicability in different domains with minimal retraining.

2 Project Plan: Task's list and/or Project Timeline

1. Task List and Timeline

Project Timeline & Task Overview

1. **Topic Selection** (27-Sep to 28-Sep): Finalize project topic.
2. **Abstract** (Complete by 29-Sep): Write a project summary.
3. **Question Reply** (Complete by 30-Sep): Answer the questions.
4. **Project Development Plan** (Complete by 01-Oct): Outline project phases.
5. **Detailed Project Plan** (Upcoming): Map detailed timeline and tasks and detail the project plan.
6. **Introduction** (27-Sep to 10-Oct): Provide background and objectives.
7. **Literature Review** (27-Sep to 10-Oct): Research existing studies and methods.
8. **Methodology** (27-Sep to 10-Oct): Define approach and techniques.
9. **Results and Discussion** (Post-Methodology): Analyze and interpret findings.
10. **Conclusion and Future Work** (After Results): Summarize findings and suggest future steps.
11. **Final Submission**: Prepare and submit completed report.

Table 1: **Timeline of the project**

Activity	27-Sep	29-Sep	30-Sep	01-Oct	02-Oct	03-Oct	04-Oct	05-Oct	06-Oct	07-Oct	08-Oct	09-Oct	10-Oct
Topic Selection	🟧												
Abstract							Complete						
Question Reply							Complete						
Project Development Plan							Complete						
Detailed Project Plan							Waiting to proceed						
Introduction							🟧						
Literature Review								⬛	⬛				
Methodology										🟦			
Result And Discussion											🟩	🟩	
Conclusion And Future Work													🟪
Final submission							Waiting to proceed						

3 Data Management Plan

The dataset Management Plan is a summary of the IMBD dataset:

1. Overview of the dataset

The dataset chosen for this project is the IMDB Movie Reviews Dataset, available on Kaggle. This is a database of fifty thousand movie reviews, all of which have been tagged with a favourable or negative attitude. Each set of reviews has 25,000 reviews; 25,000 are used for training and 25,000 are used for testing. Originally gathered and pre-processed by Stanford University, this dataset is commonly utilized for NLP tasks.

2. Data collection

The IMDB dataset(IMDB Dataset of 50K Movie Reviews | Kaggle) are sourced from Stanford University, where it was compiled and prepared for public use in sentiment analysis research. The data is available on Kaggle, and it contains reviews from various users, likely collected from the IMDB platform itself.

3. Metadata

- **Format**: The dataset is in text format, with each review labeled as either "positive" or "negative."
- **Records**: It contains 50,000 reviews, split equally into training and testing sets.
- **Type**: A data is categorical, as each review is labeled according to sentiment classification.
- **Size**: Approximately 84 MB in compressed format.

4. Document control

NOT START YET

5. ReadMe File

This file will detail dataset characteristics, pre-processing steps, code structure, and any dependencies required to run the project.

6. Security and storage

The dataset is stored on Google Drive, ensuring secure access with encryption and restricted permissions. Regular backups are maintained to protect data integrity and continuity.

7. Ethical Requirements

- Does the data come under GDPR requirements?

No, as a data is anonymized and does not contain personal information.

- Does the project conform to UH ethical policies?

An initiative complies with UH's ethical standards, yes.

- Do you have permission to use the data for your proposed research project?

Yes, a dataset is accessible to the general public for research purposes on Kaggle.

- Are you assured that the data was collected ethically?

Yes, a data was ethically gathered and provided by Stanford University.

REFERENCES

Baqach, A. and Battou, A. (2023) ‘Text-Based Sentiment Analysis’, in *Lecture Notes in Networks and Systems*. Available at: https://doi.org/10.1007/978-3-031-26384-2_10.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Redhu, S. (2018) ‘Sentiment Analysis Using Text Mining: A Review’, *International Journal on Data Science and Technology* [Preprint]. Available at: https://doi.org/10.11648/j.ijdst.20180402.12.

Taha, K. *et al.* (2024) ‘A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights’, *Computer Science Review*, 54, p. 100664. Available at: https://doi.org/https://doi.org/10.1016/j.cosrev.2024.100664.