TODO TITLE

A report submitted to the University of Manchester for the degree of Bachelor of Science in the Faculty of Science and Engineering

Author: Sai Putravu
Student id: 10829976
Supervisor: TODO

2025

School of Computer Science

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

Chapter 1

Introduction

1.1 Background

Fill this section out

Introduce the research topic.
The things in this section will include

- Talk about the ISIS research facility
- Talk about the Operational Cycle for ISIS (graph too)
- Talk about the Datasets, operalog

1.2 Motivation

Motivate the research project.
The things in this section will include

- Introduce the problem: Auto-categorisation and label inference
- Identify the data input, expected output, data shape and explain why this motivates the project
- Natural Language Processing
- Semantic Similarity
- Need for clustering

Chapter 2

Literature Review

The things in this section will include

- Looking at general predictive maintenance
- Looking at general predictive maintenance in industrial applications
- Similar pairwise sentence similarity literature
- Similar literature in text clustering
- Similar literature in specifically sentence clustering in industrial applications

Chapter 3

Technical Background

Describe the various technical factors required before attempting to understand the methodology

The things in this section will include

- Discuss sentence embedding, similarity measures: BERT, RoBERTA, MPNET, XLM, NOMIC.
- Dimensional reduction techniques and need for them (UMAP, PCA, t-SNE).
- Clustering methods: kmedoids, DBSCAN, DBSCAN*/HDBCAN
- Clustering evaluation methods:
- Maybe briefly touch on Optuna?

I don't remember these off the top of my head

Chapter 4

Methodology

Describe the methods and procedures used.

The things in this section will include.

- Explaining data format and data visualisation: wordcloud.
- Data cleaning steps, including removing key words such as Ion Source.
- Text preprocessing steps (cleaning) and computational challenges (tensorflow).
- Choosing the best sentence embedding transformer: MPNET, NOMIC.
- Data visualisation (before and after sentence embedding): similarity visualisation, explain unique sentences, token length distribution.
- Motivate why clustering in higher dimensions performs worse
- UMAP, PCA, t-SNE comparison. Motivate using UMAP.
- UMAP hyperparameter optimisation.
- Performing clustering with kmedoids, dbscan, hdbscan.
- Using optuna.
- Evaluation of results and choosing the best model (and arguing why hdbscan is the best by looking at the variance of dbscan and inflexibility of kmedoids)

- Touch on the production of a CLI application that allows you to mix and match various parts of the pipeline. Motivate the need for command line tool.

Chapter 5

Results and Discussion

Describe the results and analyse the results

- Analyse the word cloud
- Analyse the sentence embedding results.
- Analyse UMAP vs. PCA vs. t-SNE qualitatively and later quantitatively (compared to the clustering)
- Anaylse the UMAP hyperparameter optimisation qualitatively, mention that we use Optuna

Chapter 6

Conclusion

Summarize your findings and suggest areas for future work.

Bibliography