

Case Study 1

Information Extraction from Biographical Interviews

Project Plan Blueprint

Team Members

Mr. Ramu Adusumalli(11036082)

Mr. Guna Selvaraj(11035757)

Mr. Srujan Prakash Gowda(11036104)

Mr. Bala Gurram(11036053)

Prof. Dr. Binh Vu

Applied Data Science and Analytics

Document Approvals

|  |  |  |  |
| --- | --- | --- | --- |
| **Role** | **Name** | **Signature** | **Date** |
| External Stakeholder | Dr. Dennis Möbus |  | 02-11-2023 |
| Internal Stakeholders | Prof. Dr. Binh Vu  Prof. Dr. Swati Chandna |  | 02-11-2023 |
| Technical Advisors | Sina Mehraeen  Dr. Kamelia Reshadi |  | 02-11-2023 |
| Product Owner | Mr. Ramu Adusumalli  Mr. Guna Selvaraj  Mr. Srujan Prakash Gowda  Mr. Bala Gurram |  | 02-11-2023 |
| Scrum Master | Mr. Ramu Adusumalli  Mr. Guna Selvaraj  Mr. Srujan Prakash Gowda  Mr. Bala Gurram |  | 02-11-2023 |
| Scrum Team | Mr. Ramu Adusumalli  Mr. Guna Selvaraj  Mr. Srujan Prakash Gowda  Mr. Bala Gurram |  | 02-11-2023 |

Roles:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprints** | **Product Owner** | **Scrum Master** | **Scrum Team** |
| Sprint 1 | Mr. Bala Gurram | Mr. Ramu Adusumalli | Mr. Srujan Prakash Gowda  Mr. Guna Selvaraj |
| Sprint 2 | Mr. Srujan Prakash Gowda | Mr. Guna Selvaraj | Mr. Ramu Adusumalli  Mr. Bala Gurram |
| Sprint 3 | Mr. Guna Selvaraj | Mr. Srujan Prakash Gowda | Mr. Ramu Adusumalli  Mr. Bala Gurram |
| Sprint 4 | Mr. Ramu Adusumalli | Mr. Bala Gurram | Mr. Srujan Prakash Gowda  Mr. Guna Selvaraj |
| Sprint 5 | Mr. Bala Gurram | Mr. Ramu Adusumalli | Mr. Srujan Prakash Gowda  Mr. Guna Selvaraj |
| Sprint 6 | Mr. Srujan Prakash Gowda | Mr. Guna Selvaraj | Mr. Ramu Adusumalli  Mr. Bala Gurram |

Planning Basics

Scope:

The Institut für Geschichte und Biographie (IGB) at the FernUniversität in Hagen is a pioneering institution in Germany for the field of Oral History. It has had a significant impact on how biographical interviews are conducted and understood in the country. The IGB's research data archive, known as the Archiv “Deutsches Gedächtnis” (ADG), houses a substantial collection of interview materials, including digital transcripts, audio recordings, and video files.

The primary aim of this project is to develop an efficient and user-friendly NLP model with optimal parameters. This model will be used to extract valuable information from the interview transcripts that are accessible to archive users. By doing so, we can gain a deeper understanding of the perspectives, opinions, and experiences shared by the interviewees. This extracted information can have various applications, including research analysis, decision-making, report generation, and obtaining insights into specific topics or issues discussed during the interviews.

Scrum Project Plan:

***Product Backlog:***

1. As a data scientist, I would collect the given dataset and existing work, and study it, so that I have a clear understanding of the problem statement and the approach to be considered.
2. As a data scientist, I would like to study the state-of-the-art literature, so that I can identify key knowledge gaps and the scope for improvement.
3. As a data scientist, I would like to develop a machine learning model, so that useful and relevant information can be extracted from the given dataset.
4. As a data scientist, I would like to test the developed prototypes, so that I can check and obtain the optimized accuracy.
5. As a data scientist, I would like to visualize or generate thematically clustered keywords, so that it will be easy for the historian to look for the specific information with keywords.

**Phase 1: Project Initialization**

Sprint 1:

User Stories:

1. As a researcher, I want to find all recent papers on word embeddings for qualitative research data, so that I understand the state-of-the-art.

2. As a data scientist, I would like to learn about the state-of-the-art literature to identify key knowledge gaps and areas of improvement, so that I have a clear understanding of the problem statement.

Tasks:

1. Project Understanding and Kick-off Meeting
2. Understanding the Problem Statement and identifying the knowledge gaps.
3. Find and study all the recent related papers.

**Phase 2: Development and Evaluation of Machine Learning Model**

Epic 1: Development of Machine Learning Model

Sprint 2:

User Stories:

1: As a data scientist, I want to understand existing work and provided datasets, so that I can decide how it should be pre-processed for the next tasks.

2: As a data scientist, I want to preprocess the data, so that I have proper input data to train my models.

Tasks:

1. Understanding existing CoLabs and datasets.
2. Preprocessing the data.

Sprint 3:

User Stories:

1. As a data scientist, I want to develop the NLP model for training the word embeddings, so that useful information can be extracted.

Tasks:

1. Develop the NLP model for training the word embeddings.
2. Create CoLab with all parameters set to an optimized default, to train a new model.

Sprint 4:

User Stories:

1. As a data scientist, I want to implement the most purposeful evaluation- and exploration methods, so that the data can be analyzed by distant reading and parameters can be optimized.

Tasks:

1: Implement the most purposeful evaluation and exploration methods

2: Create a CoLab, with all evaluation and exploration tools implemented, documented, and explained

**Phase 3: Optimization and Visualization**

Epic 3: Optimization and Visualization

Sprint 5:

User Stories:

1. As a data scientist, I want to find the ideal representation of the interviews through evaluating and parameter tuning, so that stakeholders can rely on the output.

Tasks:

1. Create an ideal representation of the interviews through evaluating and parameter tuning.
2. Visualize the results from the developed model in the form of clusters.

Sprint 6:

User Stories:

1. As a data scientist, I want to track the process of evaluation and optimization, so that the stakeholders can rely on the output and document their methodology.

Tasks:

1. Track the process of evaluation and optimization.
2. Overall testing and further improvements according to feedback.

**Phase 4: Report**

Sprint 7:

User Stories:

1. As a data scientist, I want to integrate the feedback given by the stakeholders, so that I can update the development accordingly and document the results.

Tasks:

1. Integrate the feedback and update the developments.
2. Document the case study results.
3. Create the Presentation.

Important Dates

* Kick-off Meeting: 20-10-2023
* Sprint Planning Meeting: 22-10-2023
* Daily: Daily standup within the team.
* Weekly Meeting with Technical Advisors: Every Friday till project completion
* Bi-weekly with internal stakeholders/Sprint Review Meeting: TBD
* Sprint Retrospective: The end date of every sprint given in the Project Plan

TimePlan

Please find below the time plan for the project:



Meeting Notes:

|  |  |  |
| --- | --- | --- |
| Meeting 1: Project Kick-Off Meeting with Stakeholders | | |
| Date: 20/10/2023 | Time: 15:00-16:00 | Location: MS Teams |
| Participants: | Meeting Head:  Dr. Dennis Möbus  Prof. Dr. Binh Vu  Prof. Dr. Swati Chandna Sina Mehraeen  Dr. Kamelia Reshadi | Attendees:  Mr. Ramu Adusumalli  Mr. Guna Selvaraj  Mr. Srujan Prakash Gowda  Mr. Bala Gurram |
| Agenda: | * Introduction of Project * Introduction of Stakeholders, technical advisors, and team | |
| Discussions: | * Briefly discussed the project details and problem statement. * Introduced to datasets. * Got to know the background of the given project. * Briefly discussed the expected result. * Got to know the Stakeholders and Technical Advisors. | |
| Prepared by: | Srujan Prakash Gowda | |

Appendix