# Overview of SCRUM

## Roles and Responsibilities

We adopted the SCRUM methodology for the execution for this project. We assigned clear roles to ensure a smooth and efficient work management process.

### Product Owner

The product owner (Victor) owned the product backlog and prioritised which work needed to be completed first, ensuring that the product backlog is structured, organised, and sorted based on which tasks should be worked on first. In addition, he is also responsible for adding necessary details before each sprint planning so that the scrum team can accurately estimate the level of effort. Thereafter, the scrum team can accurately forecast the amount of work they are able to accomplish for each sprint.

### SCRUM Master

The SCRUM master (Zi Ping) ensured the quality of the code by documenting and updating the notebooks regularly. He monitored and tracked the team’s progress, and removed any roadblocks for team members to move forward. He collaborated closely with the scrum team and facilitated efficient discussions for sprint plannings.

### SCRUM Development Team

Shi Jun and Keith planned on how to complete the weekly sprint tasks. Specifically, Shi Jun’s primary role is to develop the baseline models and handle the pre-processing steps, whereas Keith’s role is to prepare weekly presentations and demonstrations for the dashboards.

## SCRUM Execution

Once the project’s vision has been set and the product backlog has been completed, the SCRUM sprint cycle begins. Each sprint cycle lasted for a week and the following process occurs for every sprint cycle:

Firstly, a sprint planning ceremony was conducted before each sprint cycle. The product owner, scrum master and entire scrum team must be present for the sprint planning ceremony. During this ceremony, the product owner and the team reviewed and discussed the product backlog, deciding on what to work on in this sprint cycle. After the discussion, the team created a sprint backlog consisting of all the items that they have commited to complete by the end of the sprint. After the ceremony, the sprint will then officially commence, and the implementation phase begins.

Due to other commitments and responsibilities, the group was unable to execute daily scrum. Hence, adjustments were made such that scrum meetings are held once every two days for team members to share the progress of their tasks, so that the team can progress towards the sprint goal(s). This way, if there were any issues sounded out by any members, the other members can help or provide suggestions for him/her. These scrum meetings ensure that the team stays focused and committed towards achieving its sprint goal(s). This goes on until the end of the sprint.

At the end of the sprint, a sprint review ceremony is conducted for the team to demonstrate the work done and gather feedback. The product owner then evaluates each result against the respective items in the sprint backlog, and accepts or rejects the result depending on whether the result fulfils the objectives. If the result is accepted, the status on the product backlog can be updated to be complete. However, if the result is rejected, the team would have to look into the matter and discuss what went wrong, or what could be improved so that the objective can be met. When the review and retrospectives are completed, this marks the end of a sprint cycle, and this continues for every subsequent sprint cycle.

# Sprint 1 (Week 3, 24th Aug - 28th Aug)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | In progress |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | To be started |
| 4 | Review data pre-processing steps (code) | To be started |
| 5 | Run baseline model (on the manually tagged dataset) | To be started |
| 6 | Backtesting of model against manual tags | To be started |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | To be started |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 24/8 Mon | * Met up with client to discuss our project’s scope and deliverables * Come up with project charter |
| 28/8 Fri | * Commence manual labelling of dataset * Literature review to shortlist models good at text classification |

Sprint Review and Retrospectives:

This sprint cycle allowed the group to brainstorm on the project charter and come up with methodologies on how to answer the business objectives of the project.

In the retrospective, a suggestion was raised regarding the possible inclusion of other models which have proved feasible for text classification in some use cases.

# Sprint 2 (Week 4, 31th Aug - 4th Sep)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | In progress |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | To be started |
| 4 | Review data pre-processing steps (code) | To be started |
| 5 | Run baseline model (on the manually tagged dataset) | To be started |
| 6 | Backtesting of model against manual tags | To be started |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | To be started |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 31/8 Mon | * Prepare for requirements presentation |
| 4/9 Fri | * Finished tagging ‘G’, ‘I’, ‘K’, ‘N’, ‘O’, ‘S’ |

Sprint Review and Retrospectives:

The main objective of this sprint cycle is to prepare for the requirements presentation. The group had to lay out clearly on what is to be presented and how it should be presented.

In the retrospective, a suggestion was raised regarding the manual labelling of the dataset. The group could create another excel sheet to consolidate the labels which have been tagged so far. This is to minimise the similarities between the labels.

# Sprint 3 (Week 5, 7th Sep - 11th Sep)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | In progress |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | In progress |
| 4 | Review data pre-processing steps (code) | In progress |
| 5 | Run baseline model (on the manually tagged dataset) | To be started |
| 6 | Backtesting of model against manual tags | To be started |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | To be started |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 7/9 Mon | * Commence baseline model testing * Finished tagging ‘H’, ‘J’, ‘L’, ‘Q’, ‘R’ |
| 11/9 Fri | * Review data-preprocessing steps |

Sprint Review and Retrospectives:

For this sprint cycle, the main focus is to run the baseline models to obtain baseline accuracies. The group has to carefully decide on the packages to be used as well as the parameters to minimise the biases of the results.

In the retrospective, a suggestion was raised regarding an improvement of the data pre-processing steps. Rather than cleaning the texts in no particular order, it is imperative to establish an order so that there would be no corner cases of unwanted words.

# Sprint 4 (Week 6, 14th Sep - 18th Sep)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | In progress |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | To be started |
| 6 | Backtesting of model against manual tags | To be started |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | To be started |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 14/9 Mon | * Finished tagging ‘M’, ‘T’ * Validate preprocessing methods |
| 18/9 Fri | * Gathered accuracies on baseline models |

Sprint Review and Retrospectives:

For this sprint cycle, 2 objectives were met. The group came up with a text cleaning algorithm and showed that the order of cleaning truly mattered. In addition, the group has run the baseline models and obtained the baseline accuracies. This will serve as a guideline to see the improvement of the manual tagging process.

# Sprint 5 (Recess Week, 21th Sep - 25th Sep)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | To be started |
| 6 | Backtesting of model against manual tags | To be started |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | In progress |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 21/9 Mon | * Finished tagging ‘P’ |
| 25/9 Fri | * Standardise and consolidate labels * Review and refine baseline models |

Sprint Review and Retrospectives:

For this sprint cycle, the group has completed the tagging of the first year dataset. However, before the group can proceed any further, they have to standardise the labels to minimise any human error made while labelling.

In the retrospective, a discussion was made and the group planned on how to standardise the labels in the shortest amount of time.

# 

# Sprint 6 (Week 7, 28th Sep - 2nd Oct)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | Completed |
| 6 | Backtesting of model against manual tags | To be started |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | Completed |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 28/9 Mon | * Standardise and consolidate labels * Run baseline models with manually tagged dataset |
| 2/10 Fri | * Obtain results of baseline models, and compare accuracies to pick best one |

Sprint Review and Retrospectives:

For this sprint cycle, the group ran the same models with the same parameters to ensure a fair comparison.

In the retrospective, discussion was made to consider also the time complexity of getting the predicted probabilities rather than looking only at the accuracy.

# 

# Sprint 7 (Week 8, 5th Oct - 9th Oct)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | Completed |
| 6 | Backtesting of model against manual tags | In progress |
| 7 | Prediction on the Y2M1 dataset | To be started |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | Completed |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 5/10 Mon | * Prepare for interim presentation |
| 9/10 Fri | * Predict same year dataset for backtesting |

Sprint Review and Retrospectives:

For this sprint cycle, the group conducted backtesting to further minimise any human errors made while manually labelling the data. This is inevitable as different people were labelling the data and could have varying opinions.

In the retrospective, discussion was made to consider the threshold that the group should set rather than looking at the entire dataset. Looking at the entire dataset is not feasible due to time constraints.

# 

# Sprint 8 (Week 9, 12th Oct - 16th Oct)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | Completed |
| 6 | Backtesting of model against manual tags | Completed |
| 7 | Prediction on the Y2M1 dataset | In progress |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | To be started |
| 9 | Repeat tasks 5 - 8 for the remaining months | To be started |
| 10 | Refine and improve on baseline models | Completed |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 12/10 Mon | * Backtesting of model against manual tags after checking |
| 16/10 Fri | * Re-training of the model and proceeding with forward prediction |

Sprint Review and Retrospectives:

For this sprint cycle, the group will refit all of the year 1 data with the new labels and retrain the model again. The result will be a model which is ready for forward prediction.

In the retrospective, discussion was made to consider the length of each snowball cycle, whether it should be 1 month or 3 months. Ultimately, as the model has only learnt 1 year of data, it would be more suitable to snowball by months rather than quarters as quarterly data could prove too much for the model to handle.

# 

# Sprint 9 (Week 10, 19th Oct - 23th Oct)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | Completed |
| 6 | Backtesting of model against manual tags | Completed |
| 7 | Prediction on the Y2M1 dataset | Completed |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | In progress |
| 9 | Repeat tasks 5 - 8 for the remaining months | In progress |
| 10 | Refine and improve on baseline models | Completed |
| 11 | Association mining | To be started |
| 12 | Dashboard | To be started |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 19/10 Mon | * Forward prediction of Y2Q1 dataset |
| 23/10 Fri | * Review and correction of poorly classified text |

Sprint Review and Retrospectives:

For this sprint cycle, the group cleaned up the labels for the Y2Q1 dataset and re-trained the model with Y1 and Y2Q1 data.

# 

# Sprint 10 (Week 11, 26th Oct - 30th Oct)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | Completed |
| 6 | Backtesting of model against manual tags | Completed |
| 7 | Prediction on the Y2M1 dataset | Completed |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | Completed |
| 9 | Repeat tasks 5 - 8 for the remaining months | Completed |
| 10 | Refine and improve on baseline models | Completed |
| 11 | Association mining | In progress |
| 12 | Dashboard | In progress |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 26/10 Mon | * Meet with client to understand requirements of dashboard * Explore and experiment with dashboard design |
| 30/10 Fri | * Work on association mining code |

Sprint Review and Retrospectives:

For this sprint cycle, the group came up with a draft of the dashboard including details like what visualizations to show on the dashboards and what are the business values that the users can gain from the dashboards.

In the retrospective, discussion was made to consider the different dashboard designs and ultimately decide on the one that suits the client the most.

# 

# Sprint 11 (Week 12, 2nd Nov - 6th Nov)

|  |  |  |
| --- | --- | --- |
| **Updated Product Backlog** | | |
| **No.** | **Item** | **Status** |
| 1 | Tagging of first year dataset | Completed |
| 2 | Literature Review | Completed |
| 3 | Run baseline model (on the original dataset) | Completed |
| 4 | Review data pre-processing steps (code) | Completed |
| 5 | Run baseline model (on the manually tagged dataset) | Completed |
| 6 | Backtesting of model against manual tags | Completed |
| 7 | Prediction on the Y2M1 dataset | Completed |
| 8 | Review the predicted Y2M1 dataset and manually tag the poor classified dataset | Completed |
| 9 | Repeat tasks 5 - 8 for the remaining months | Completed |
| 10 | Refine and improve on baseline models | Completed |
| 11 | Association mining | Completed |
| 12 | Dashboard | Completed |

|  |  |
| --- | --- |
| **Date** | **Completed Task(s)/Progress Update(s)** |
| 2/11 Mon | * Meet up with client to demonstrate dashboards to gather feedback * Adjust association mining rules * Finalise dashboard design |
| 6/11 Fri | * Work on final report, presentation and consolidation of code |

Sprint Review and Retrospectives:

For this sprint cycle, the group did a last touchup on the dashboards and refined the association rules to show more meaningful association rules on the dashboards.

In the following week (Week 13), the group focused on completing the deliverables (report, presentation and code) for both the field and academic supervisors.