PROJECT REPURPOST

Prototype

Abstract

The prototype document will provide status of the deliverables and snapshots of work-in-progress deliverables and report any potential risks.

Yashwanth Balan Arumugam
Pragya Avinash Mishra
Shruti Sham Kotwal
Atharva Shantanu Kulkarni
Saju Chacko Rajan

Table of Contents

Project Deliverables 2

Artifacts – Work Completed 2

Solution deployment and workflow diagram 2

User Flow Diagram 3

Wireframe for end state goal for tag suggestions 3

Artifacts – Work in Progress 5

Working ML Model for tag suggestions 5

GitLab repository of model code 5

Risks – Emerging & Retired 6

Works Cited 7

Repurpost Prototype Document

Project Deliverables

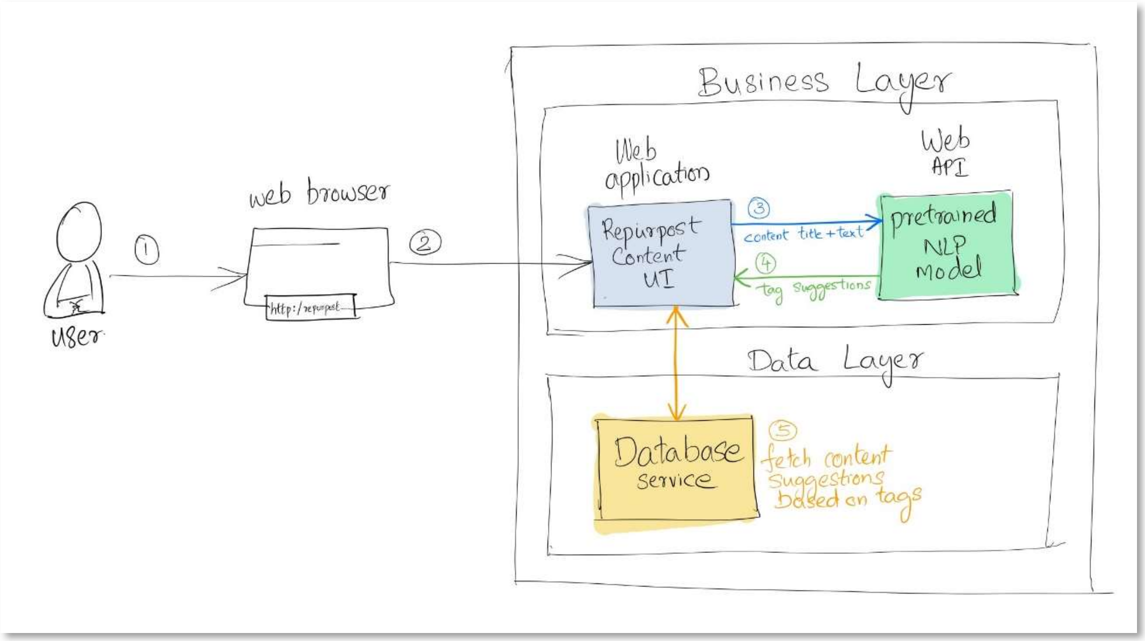
Following are the main project deliverables and its statuses:

Si#	Deliverable	Status & Details
1	Solution deployment and workflow diagram	Completed
2	User Flow diagram	Completed
3	Wireframe for end state goal for tag suggestions	Completed
4	Working ML Model for tag suggestions	In Progress
5	API endpoint to invoke ML model	Not Started
6	GitLab repository of model code	In Progress
7	Provide Web API usage guide and documentation	Not Started

Artifacts – Work Completed

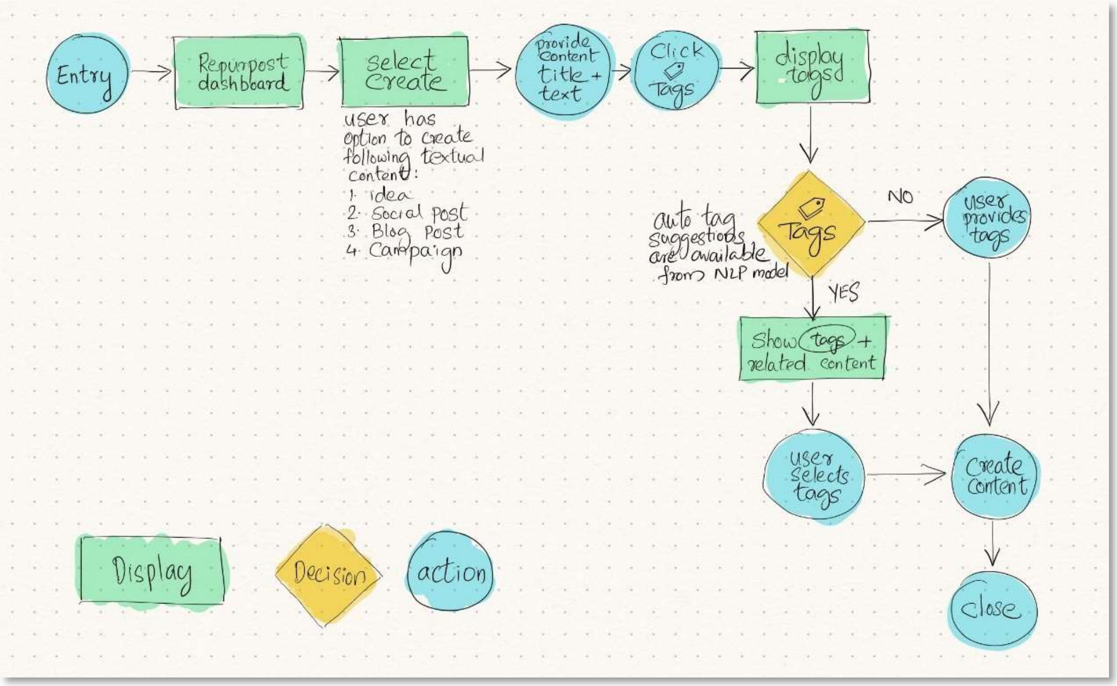
Solution deployment and workflow diagram

The solution deployment and workflow diagram display the major deployment components that are required for Repurpost to use the NLP model. For the purpose of this project, the focus is more on creating the web API NLP model is the main business logic that supplies the tag suggestions to the Repurpost platform UI.



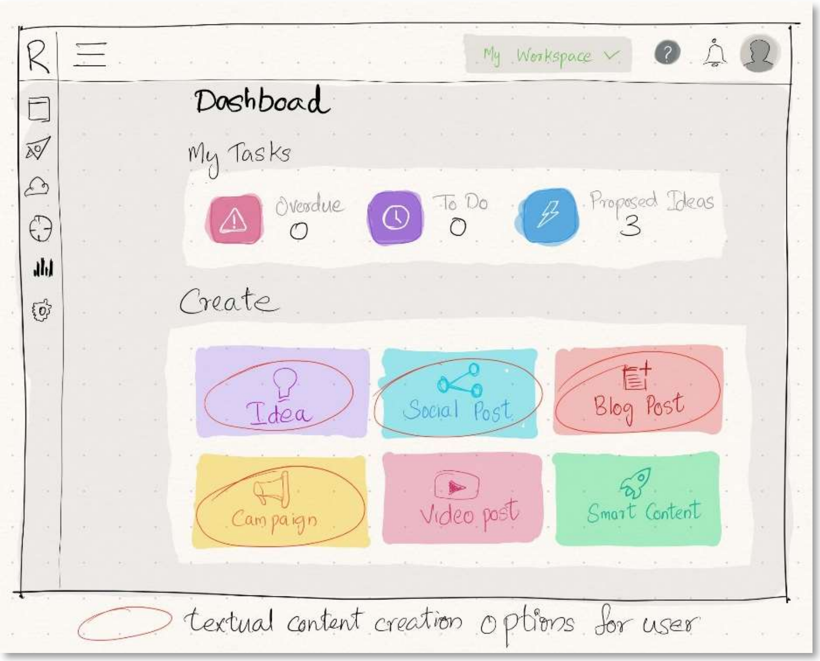
User Flow Diagram

The project team has performed some research on the Repurpost platform and has been able to identify the user flow that the user will go through for getting the content tag suggestions while they focus on content creation.



Wireframe for end state goal for tag suggestions

The end state goal for Repurpost is documented in these wireframes. Each wireframe screen is focused on using the existing Repurpost platform architecture and style and utilizing the available screen real estate on existing screen to avoid any major screen changes for users already using the platform.

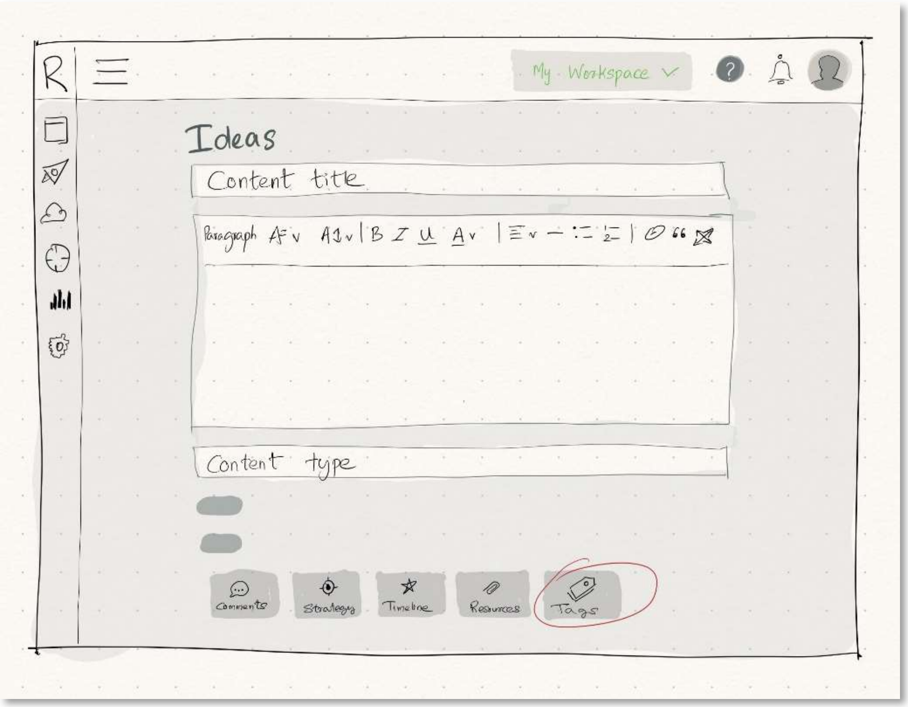


Screen 1

The landing page of the Repurpost Dashboard. The user has 4 different textual content creation options.

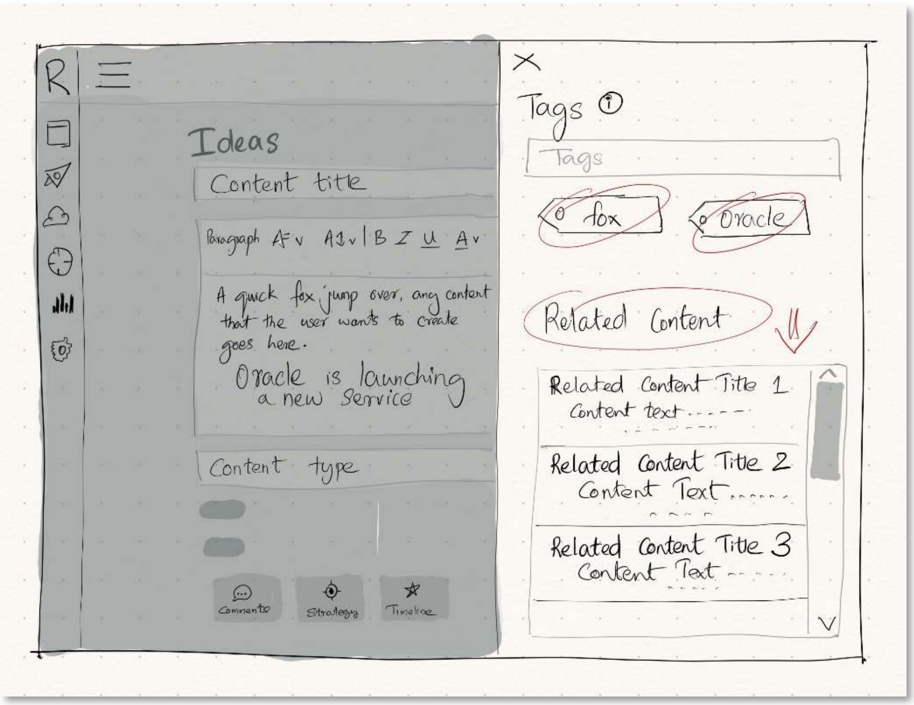
Screen 2

The 4 different textual content creation screens have most the elements that are common to each other. The main commonality is the content title and context text sections which will be the inputs for model suggestions. The tags section is accessible through a button on the lower section of the page.



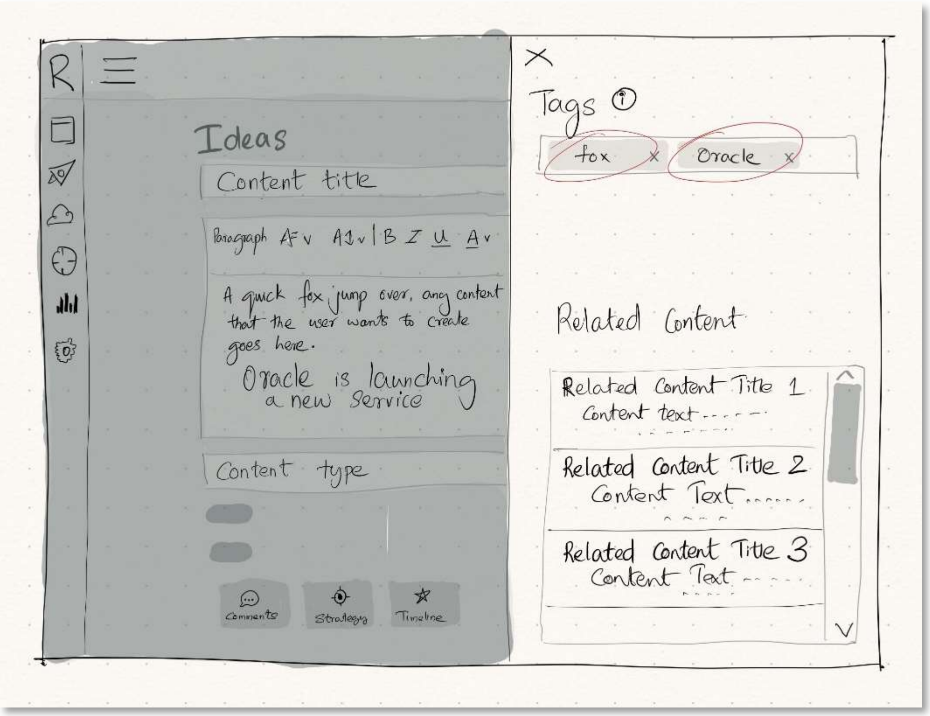
Screen 3

The tag screen will be shown when user clicks the 'Tag' button. The content title and text on the content screen will be provided as inputs to the web API and tag suggestions will be displayed on the screen for user to select, if the user chooses to after checking the related content.


A quick fox jump over, any content that the user wants to create goes here.
Oracle is launching a new Service

Screen 4

The user has an option to choose any or all of the suggested tags after reviewing related content details based on the title and text of the content that user provided on the main content creation page.



Artifacts – Work in Progress

Working ML Model for tag suggestions

As per the project plan, the dataset identification and cleanup steps are in progress. The python notebook used for dataset load and cleanup steps is provided in the embedded document provided below. The dataset load and join steps have been completed. The cleanup steps are in progress and details are as follows

1. Remove English stop words – Complete
2. Remove HTML tags – Complete
3. Convert to lowercase – Complete
4. Remove punctuations and special characters – In Progress
5. Lemmatize words – Pending

PDF

StackOverFlow_Raw_Dataset_Cleanup

GitLab repository of model code

The team plans to utilize GitLab for storing the project code and dataset that will be used for this project. Currently, only the datasets have been checked-in to the repository. The python notebook and documentation will be added as soon as we have the artifacts ready for check-in. The GitLab repository is accessible from Repurpost Auto Tag Suggestions.

Risks – Emerging & Retired

Following are the main risks that are identified in the project along with the plan to address each of them:

Identified Risks	Details and Risk Mitigation Plan	Status
Dataset Size	Processing and preparing the 1.75 GB dataset with 1.25 million rows is causing slowness in cleaning up the dataset. As a mitigation, using parallel processing packages like Pandarallel and Dask is allowing us to speed up function application on each row. The main limitations is the number of CPU cores that are available on the machines.	Retired
Model Accuracy	Unlike traditional machine learning model accuracy, accuracy, precision, recall and f1 score is not the right metrics for evaluating the performance of the trained models. Since the problem statement requires a multi-label classification model the models are going to be evaluated using the Hamming loss factor and Jaccard similarity index.	Emerging
Web API Deployment	Other than using Flask for creating an API in local machines, the team has not explored any other cloud API deployment yet. Options are being explored and render.com is an option that is being considered to deploy the NLP model as a public API for demo and testing purposes.	Emerging

Works Cited

Artifacts	Reference Documents
Solution deployment and workflow diagram	PDF Repurpose high level solution diagra
User Flow diagram	PDF Repurpost user flows.pdf
Wireframe for end state goal for tag suggestions	PDF Repurpost proposed wireframe
Working ML Model for tag suggestions	PDF StackOverFlow_Raw_Dataset_Cleanup
Updated Project Plan (28th October 2022)	W Latest Project Plan - Team Repurpost.do