Broadsight Tech Project

Yuyang Peng, Shiyi Chen, Yahan Cong, Rui Mao

Introduction

Effective managing and maintaining the organizational image is essential for large public and private sectors. This requires competent handling of media information and responses. Our company's Broadsight Tracker is an online collaboration platform tailored for media teams, enabling them to communicate, document, and measure communications efficiently. The platform allows real-time sharing of media work and offers data management, analysis, and reporting features to gauge team value. Understanding user preferences and needs is vital for enhancing the platform and attracting users. Our project aims to answer the following research questions:

1. What needs have users stated?
2. What are the interest hotspots of our users?
3. What functional preferences do our users have?
4. What specific needs or preferences do different user groups have?

To address these questions, we plan to analyze both customer demands and platform data. Direct customer needs are primarily reflected in their communication emails with the Broadsight Tracker team, while the Broadsight Tracker database captures user focus areas, preferences, and other relevant information. Our project is strategically divided into two main segments: Usage Analysis and Email Scraping. While these segments operate independently, they both contribute to the overarching goal of enhancing the firm's ability to analyze media interactions.

The Usage Analysis team focuses on identifying users' preferences for Broadsight Tracker's features through platform data. Our goal is to utilize visual tools to provide managers with a clear, intuitive understanding of user engagement across different functionalities. We also aim to correlate services with media interactions on the platform, exploring user interest hotspots between the two. Finally, by examining usage variations among different users, we want to

find a rule to segment customers effectively. It will enhance the company's understanding of user preferences and improve platform features to offer more customized services.

Simultaneously, the Email Scraping team concentrates on enhancing data extraction methods to increase the accuracy and efficiency of collecting relevant information from various media sources. This data is essential for developing a more robust analytical framework capable of handling extensive volumes of media interactions, categorizing them by impact level, and incorporating this information into the firm's strategic planning tools.

Literature Review and Related Work

1. Email Scrapping:

The integration of Natural Language Processing (NLP) technologies into email management systems exemplifies significant advancements in the field of computational linguistics and artificial intelligence. This section reviews the technical underpinnings and applications of NLP in email management as demonstrated by Stanford NLP, Expert.ai, and Twinword, providing a framework for the proposed project.

1.1. Stanford NLP's Sentiment Analysis

Stanford's NLP framework includes sophisticated linguistic analysis tools that are essential for extracting and interpreting the sentiment from textual data. This technology employs a combination of machine learning algorithms and linguistic rules to analyze the structure and content of language. In the context of email sentiment analysis, Stanford NLP utilizes tools like the Stanford CoreNLP, which integrates various annotators such as the sentiment annotator. This annotator applies a deep neural network to classify sentences into categories such as "very negative", "negative", "neutral", "positive", and "very positive". This classification helps in assessing the emotional tone of emails, which is crucial for prioritizing responses in business communications (Stanford NLP Group).

1.2. Expert.ai's Email Management Solution

Expert.ai offers an AI-based framework that incorporates a custom NLP model capable of classifying text and uncovering the sender's intentions. The technical process involves natural language understanding (NLU) tasks that analyze the syntax and semantics of the text, extracting entities such as names, companies, and products mentioned within the emails. Moreover, Expert.ai's framework includes sentiment analysis to evaluate the emotional tone of the messages. This system uses a proprietary blend of machine learning models and linguistic algorithms to provide a detailed analysis of emails, enabling automated prioritization based on content relevance and urgency. This automation is particularly beneficial for managing large volumes of emails in organizational settings, enhancing productivity by reducing manual sorting (Expert.ai).

1.3. Twinword's Text Classification and Sentiment Analysis

Twinword utilizes NLP technologies to offer APIs for text classification and sentiment analysis, which are directly applicable to the automation of email categorization. Their system picks out keywords and categories through machine learning models that have been trained on large datasets. The sentiment analysis API analyzes the tone of the email context, classifying it into predefined emotional categories. These technologies combine to sort emails automatically, setting priorities and streamlining responses based on content analysis. Twinword's APIs highlight the practical application of text analysis in real-world scenarios, showcasing how NLP can significantly improve the efficiency of email management by automating tasks that traditionally required substantial human intervention (Twinword).

1.4. Integration into Project Objectives

These technological insights inform the project's approach by highlighting the potential of NLP to automate and enhance the efficiency of email management systems. By adopting similar technologies, this project aims to develop an advanced system for Broadsight Tech Data that automates the categorization and prioritization of emails. This will not only improve response times but also ensure that critical communications are addressed promptly, leveraging NLP to enhance decision-making processes in corporate environments.

2. Usage analysis

In the context of usage analysis, since Broadsight Tracker has a diverse user base, understanding user behaviour will help our client company understand how customers respond to services. By analyzing users' behaviours and preferences, our company can accurately segment their customer base (Alfian, Ijaz, Syafrudin, Syaekhoni, Fitriyani, & Rhee, 2019). This enables them to customer their products and services effectively and helps them to provide unique preferences and behaviours to various user groups. And for customer segmentation, K-means and hierarchical clustering are current common practices. Additionally, Kuo and Zulvia proposed combining the Artificial Bee Colony(ABC) algorithm with K-means to reduce the K-means' sensitivity of initial centroids (Kuo & Zulvia, 2018). And Li and her group suggested using an adaptive particle swarm optimization (PSO) algorithm to improve K-means. It will enhance the global clustering optimization, and avoid getting stuck in local optima(Li et al., 2021). We believe these algorithms could be insightful for future segmentation of user groups based on different preferences.

Dataset

The main dataset for the Email Scraping group is the authorized emails (in .msg and .eml format) for targeted model training and testing.

For Broadsight Tracker usage analysis, we have three datasets. The first one contains the website's interactive response information. This dataset contains all user interactions with the webpage over the past month, including user submissions, contact content records, and team contributions integration. The information in this dataset about users' usage of the website can help us analyze user preferences for web page functions.

Our other two datasets: ServiceLog and MediaInteraction, record past user service records of Broadsight Tracker. ServiceLog includes details of past media/issue work followed up by Broadsight Tracker, including its type, topic, company, etc. The MediaInteraction dataset records information used in these media/issue works, such as story links, background, and key messages. These two databases help us analyze the main characteristics of the company's (potential) user groups and explore differences in usage patterns and needs among different users.

Objectives and Aims

1. Email Scraping Group

The aim of the Email Scraping group within Broadsight Technologies' project is to apply NLP (Natural Language Processing) techniques to extract key information from a plethora of email communications. This involves developing sophisticated algorithms that can identify and parse relevant data from unstructured text, allowing for the automatic categorization of content based on predefined criteria such as sentiment, topic relevance, and urgency. The primary objective is to streamline the process of data handling, reducing manual labor and enhancing the accuracy and speed of data retrieval.

Key research questions:

- How can NLP tools be optimized to accurately extract relevant information from diverse email formats?
- What criteria can be developed to determine the relevance of extracted information to specific research or business needs?
- How can the email scraping system be scaled to handle large volumes of data without compromising processing speed or accuracy?

2. Usage Analysis Group

The aim of group Usage Analysis is to gain a comprehensive understanding of user interaction patterns and preferences within a system by analyzing detailed API call data.

Key research questions:

- How do user interaction patterns with analytics features vary among different client groups, and can these variations be systematically classified according to industry?
- What are the user preferences and behaviors in the use of Broadsight Technology's system?

Goals:

- Identify the most popular services offered by the company and determine the specific customer segments these services cater to.
- Develop clustering models to group clients based on their usage patterns, such as frequency of use, number of active users, and types of services or media interactions engaged.
- Investigate how users interact with different sections of the analytics features, specifically comparing preferences for team analytics versus custom analytics.
- Identifying preferred types of API endpoint being tracked and common issue topics, highlighting the users' tendency to use the site.

Methodology

1. Email Scraping

- Data Preparation:

A. Unstructured Data Acquisition: This stage involves collecting a broad range of unstructured data sources, such as emails, textual documents, and other digital communications, which are essential for the data conversion process.

B. Schema Specification: Flexible JSON schemas are defined to serve as blueprints for the desired structured outputs. These schemas ensure that the converted JSON files meet diverse application requirements.

C. Preprocessing: Data preprocessing is essential for optimizing model performance. This process includes cleaning the data to remove extraneous elements, correcting errors, and segmenting text to suit the model processing needs better.

- Model Mechanism:

A. Jsonformer Class Development: A utility class, termed 'Jsonformer,' is developed to facilitate interaction with machine learning models to generate structured JSON outputs. This class includes several key functionalities:

B. Prompt Extension: It modifies base prompts to guide the output of machine learning models toward specific JSON formats, aligning with predefined schemas.

C. String and Object Generation: The class handles the conversion of model outputs from raw strings to structured JSON objects and values, iteratively populating the schema.
D. Model Selection and Integration: Various pre-trained large language models available through cloud-based APIs are used. These models are selected based on their capability to understand and generate text effectively, and their performance is evaluated by the relevance and completeness of output.

- Deployment:

A. Model Hosting and Management: Models are hosted on cloud platforms that provide API endpoints for machine learning inference. This serverless approach allows for scalable and maintenance-free operations.
B. Integration with Cloud Services: The Jsonformer class interacts with these cloud services to manage data flow, converting unstructured inputs into JSON outputs without the need for extensive local infrastructure.
C. Operational Workflow Implementation: The deployment phase involves setting up a systematic workflow where:
D. Data inputs are processed through the Jsonformer to apply JSON schemas dynamically.
E. The Jsonformer makes multiple API calls to generate necessary JSON fields, integrating each part into a cohesive output.

2. Usage Analysis

- Data Preparation:

For our usage analysis, we plan to clean up the URL portion of the website interaction dataset. We will filter out any security information and environment access data, retaining only the elements related to user interactions, such as API interactions. And, we aim to extract parameters from the URL path as much as possible to understand better how Broadsight Tracker users engage with the website. We will utilize these requests and response information to analyze user preferences and behaviors within the system.

For the ServiceLog and Media Interaction datasets, we plan to import both datasets as JSON files into MongoDB for integrated analysis. This will allow us to comprehensively analyze user business needs and preferences.

- Analysis of User Experience:

For usage analysis, we will primarily utilize exploratory data analysis (EDA). We plan to import users' web interactions into pandas for a comprehensive analysis of user behavior, focusing on the main uses of web features, and organize our workflow into reusable functions. We will use MongoDB to integrate the ServiceLog and Media Interaction datasets to analyze users' primary concerns and common issues. The analysis results will be visualized properly to help our clients intuitively understand the overall user situation and web usage. Finally, we plan to use methods such as K-means to cluster our customers into different groups based on their needs and preferences, to explore how various user groups have differing needs when using Broadsight Tracker.

Schedule

1. Deliverables

The deliverables for this project are designed to enhance the operational capabilities and strategic insights of Broadsight Technologies. The expected deliverables are:

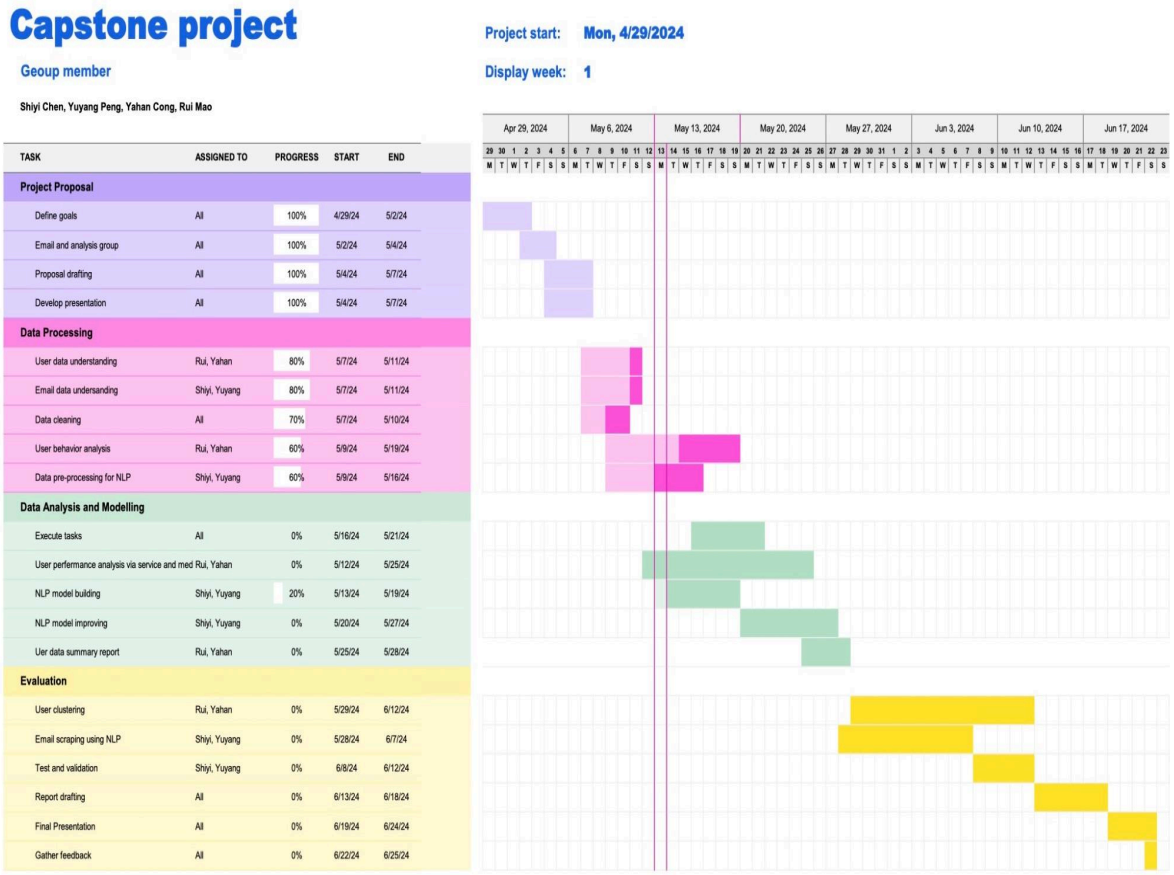
- A detailed final report summarizing the project's findings, including analyses, user interaction insights, and practical recommendations for system enhancements.
- A Natural Language Processing (NLP) tool to automate the extraction and categorization of information from emails. This tool will simplify data processing, reduce labor, and increase the accuracy and speed of data retrieval.

2. Role and Responsibility

In the project, we plan to divide into two independent groups to carry out email scraping and usage analysis respectively. Yuyang and Shiyi will work on training the NLP model to extract key information from the email dataset. Yahan and Rui will use

database analytics and visualization tools to analyze Broadsight user data, understand user behavior preferences and segment users accordingly.

3. Project Timeline



Capstone project

Geoup member

Shiyi Chen, Yuyang Peng, Yahan Cong, Rui Mao

Project start: Mon, 4/29/2024

Display week: 1

TASK	ASSIGNED TO	PROGRESS	START	END
Project Proposal				
Define goals	All	100%	4/29/24	5/2/24
Email and analysis group	All	100%	5/2/24	5/4/24
Proposal drafting	All	100%	5/4/24	5/7/24
Develop presentation	All	100%	5/4/24	5/7/24
Data Processing				
User data understanding	Rui, Yahan	80%	5/7/24	5/11/24
Email data undersanding	Shiyi, Yuyang	80%	5/7/24	5/11/24
Data cleaning	All	70%	5/7/24	5/10/24
User behavior analysis	Rui, Yahan	60%	5/9/24	5/19/24
Data pre-processing for NLP	Shiyi, Yuyang	60%	5/9/24	5/16/24
Data Analysis and Modelling				
Execute tasks	All	0%	5/16/24	5/21/24
User perfermance analysis via service and med	Rui, Yahan	0%	5/12/24	5/25/24
NLP model building	Shiyi, Yuyang	20%	5/13/24	5/19/24
NLP model improving	Shiyi, Yuyang	0%	5/20/24	5/27/24
Uer data summary report	Rui, Yahan	0%	5/25/24	5/28/24
Evaluation				
User clustering	Rui, Yahan	0%	5/29/24	6/12/24
Email scraping using NLP	Shiyi, Yuyang	0%	5/28/24	6/7/24
Test and validation	Shiyi, Yuyang	0%	6/8/24	6/12/24
Report drafting	All	0%	6/13/24	6/18/24
Final Presentation	All	0%	6/19/24	6/24/24
Gather feedback	All	0%	6/22/24	6/25/24

The project is designed considering 4 phrases, starts from April 29, 2024, and spans 8 weeks, divided into major sections: Project Proposal, Data Processing, Data Analysis and Modelling. Each task is assigned to specific team members and has an associated progress percentage. The colored blocks represent the duration of each task within the specified dates, helping visualize overlaps and project flow. The vertical line indicates the current week relative to the project timeline.

Reference

[1] Alfian, G., Ijaz, M.F., Syafrudin, M., Syaekhoni, M.A., Fitriyani, N., & Rhee, J. (2019). Customer behavior analysis using real-time data processing. *Asia Pacific Journal of Marketing and Logistics*.

[2] Expert.ai. (n.d.). *Natural Language API*. https://www.expert.ai/

[3] Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. Applied Soft Computing, 113(Part B), 107924. https://doi.org/10.1016/j.asoc.2021.107924

[4] Kuo, R.J., Zulvia, F.E. Automatic clustering using an improved artificial bee colony optimization for customer segmentation. *Knowl Inf Syst* **57**, 331–357 (2018). https://doi.org/10.1007/s10115-018-1162-5

[5] Stanford NLP Group. (n.d.). *Stanford CoreNLP – A suite of core NLP tool*s. Stanford University. https://nlp.stanford.edu/software/

[6] Twinword. (n.d.). *Twinword Ideas – Text Analysis APIs*. https://www.twinword.com/