Social Media Comparison Requirements

Social Media Comparison

Social Media Team 1
Andres Dicochea, Peter Mansbacher, James Mortensen, Eric Slick

Software Requirements Specification

Version: (1.0) **Date: (09/16/2021)**
Version: (1.1) **Date: (10/24/2021)**

Table of Contents

1. Introduction

1.1 Purpose

The purpose of this document is to give a comprehensive overview of the project requirements for "Evaluating Social Media Across Competitors". This SRS document will serve as a reference to the engineers designing and developing the project. It is also a proposal for the client.

1.2 Scope

The problem this project is aiming to solve: given a valid social media profile name and platform, can we scrape various social media platforms in order to gather data from posts and return a useful comparison of metrics. Capabilities of the software include taking a platform and brand as inputs, and then comparing at least three brands at once. Five or more brands would be better though. Also, the program needs time frames for collecting the posts. This way the comparisons can be compared over time. The goal is to have an output of, "Post records in a database for brands that we wanted to get; and a spreadsheet export with the brand, its posts over time and certain data metrics for each post". Benefits will be that with this software, the user will be able to compare the content, impressions, and engagement of brands on specific social media platforms. The software being produced for the "Social Media Evaluator" project will be a web interface-based program. The constraints are no API data collection, limited information on posts, legality of having to utilize open-source programs, needing an account, and possibly the collection rate.

1.3 Definitions, Acronyms, and Abbreviations.

Table 1- Definitions

Web scrape	Fetching data from various sources to capture uniform structured data elements
GUI	Graphic User Interface
API	Application Programming Interface
Open-Source	Code that is freely available for modification and redistribution
Social Media Platform	A medium for content, containing pictures, videos and user impressions, and engagement for that content
HTML	Hypertext Markup Language
CSV	Comma Separated Value
Brand Handles	The account name used by brands on social media platforms

1.4 References

[1] IEEE 830-1993 - IEEE Recommended Practice for Software Requirements Specifications

[2] Debra Parcheta 9/14/2021 - Social Media Competition- Program to Evaluate Social Media Across Competitors.

1.5 Overview

The rest of the SRS contains more detailed information pertaining to the requirements of the project. Customers/potential users care about section two, developers care about sections two and three.

2. The Overall Description

The Social Media Comparison system will be used to compile information from competing brand's social medias. Generally, this system will require the user to specify which social media platform they would like to analyze, the timeframe in which they would like posts to be collected, and the specific brand handles. Typically, social media platforms feed information to programs using APIs, which require account login information. Our program will collect all the required information without the login information for specific accounts.

2.1 Product Perspective

Our tool for "Evaluating Social Media Across Competitors" must interact with several different services, including multiple websites and a database. It will essentially serve as a messenger between social media platforms and our database, relaying all the requested information by the user so that different brand's social media presences can be compared quantifiably.
Our software will need to interact with the following social media networks:

- Facebook
- Twitter
- Instagram
- TikTok
- YouTube
- Pinterest

From these applications, the following information will be collected for all posts within the user's designated timeframe:

- Post URL (This URL is unique for every post, and will always get one back to that post even if it is no longer accessible through the platform)
- Post Date
- Post Text

- Image URL (Provides the image location, if applicable to the post)
- Impressions (Number of times the post has been viewed/visited by users of the platform)
- Engagements (Number of times users interacted with the post. The ways in which users can interact with posts varies by platforms, see **2.1.1 System Interfaces** for specific ways interactions are measured for each platform)

This information, along with the user's designated platform and brand handles, will be entered into our database. Also, all information requested to be scraped by the user will be downloadable via CSV file.

2.1.1 System Interfaces

Our program will collect the following information each time it is used:

- Social Media Platform
- Brand Handle
- Post URL
- Post Date
- Post Text
- Image URL
- Impressions
- Engagements

This applies to all aforementioned social media platforms, and it will only collect this information for posts *within* the user's designated timeframe. What qualifies as an engagement may vary depending on the social media platform being scraped.

Our program will consider, and retrieve, the following information to measure engagements for each social media platform:

	Facebook	**Twitter**	**Instagram**	**TikTok**	**YouTube**	**Pinterest**
Likes	x	x	x	x	x	x
Comments	x	x	x	x	x	x
Shares or Retweets	x	x		x		
Saves						x
Pins						x
Closeups						x

All information will be collected without the use of login information for the requested brand's accounts.
In order for the client's existing rating system to interact with our system, all information will be entered into our database and downloadable via CSV file. From there, the client's system will handle rating each post and analyzing the posts from each brand through an ODBC connection to our database.

2.1.2 Interfaces

Users will interact with our application via a web interface.
The user-facing web interface will allow the user to specify the following information:

- The platform they would like the program to scrape (The user will only be able to scrape one platform at a time)
    - This selection can be made by the user clicking on an image of the platform's logo
- The timeframe in which the user would like posts from.
    - This timeframe cannot exceed 90 days.
    - This timeframe must be greater than 1 day.
- The brand handles the user would like to retrieve posts from.
    - Maximum of 5 brands.
    - Minimum of 1 brand.
    - Based on the selected platform, a message will appear above the inputs for brand handles which will explain to the user how to find the handle on that specific platform.

After the data is collected by our program, the web interface will be updated with a button labeled "Download as CSV file".

2.1.3 Hardware Interfaces

The hardware interfaces for this project are minimal. As it is a web-based application, it will be interacting with any devices capable of accessing a web page, primarily desktop computers and tablets. Our program does not require any input from hardware interfaces for data collection.

2.1.4 Software Interfaces

The Social Media Comparison system will receive an HTML page from the most recent version of each social media platform and parse the required information. Additionally, our system will be accessible through the most recent versions of Google Chrome, Safari, Mozilla Firefox, and Microsoft Edge.

Communication between our system and each social media platform will be performed by our system based on the BeautifulSoup framework, which will retrieve any specified HTML pages.

- Documentation can be found here: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Beautiful Soup git can be found here: https://github.com/wention/BeautifulSoup4

Selenium will be used in cases where BeautifulSoup will not retrieve the necessary information (specifically with JavaScript-heavy websites). Selenium can automate an instance of Google Chrome (or any other browser, for that matter), which is used to find the information we need across multiple webpages.

- Documentation can be found here: https://www.selenium.dev/documentation/

- Selenium git can be found here: https://github.com/SeleniumHQ/selenium

Selenium requires a driver to be installed for the specific browser it emulates. The documentation and installation information for the Google Chrome driver can be found here: https://chromedriver.chromium.org

2.1.5 Communications Interfaces

Standard TCP/IP communication will be used by our web host to communicate with the user **[RFC 793].**
No custom communication interfaces are necessary.

2.1.6 Memory Constraints

The Social Media Comparison tool will require a server to allow users to access our webpage and database. As our system will likely be retrieving and relaying a relatively large amount of data on each use, our server should run with at least 8GB of RAM. Additionally, our server will include 2 vCPU's, 160 GB Solid State Drive for storage, and 5TB of transfer storage.

2.1.7 Operations

The Social Media Comparison tool will require a server for hosting of our webpage and database. Refer to section **2.1.6** for required specifications of our server.

2.1.8 Site Adaptation Requirements

As our system is entirely web-based, there are likely no site adaptation requirements, unless the user does not have a device which can access a web page and has internet access.

2.2 Product Functions

The product functions have briefly been described prior to this section of the SRS document. This is a more in-depth summary of what specifically the system does. Software for the project will create a web-based user interface. The "Social Media Comparsion" product is a system with a user interface that allows the user to compare several brand handles on a platform and compare contents engagement and impressions. Then the user will have to define a start and end date for the time frame. Once the brand handles and time frame have been defined, the system will search the specified platform for each brand. Then, the following data will be scraped:

- Platform Name (text Facebook, TikTok, Instagram, YouTube, Pinterest, Twitter - could be parsed from the url)
- Brand handle (text, the account name used by the brand we want to search for on the platform)

- Post URL (text, some platforms call this the permalink or the link, it's the url that always gets you back to this particular post even if it's no longer showing on the platform because of age or censure)
- Post Date (date, a timestamp is also fine, it can be parsed)
- Post Text (Whatever the message was in the post, some messages have no text….)
- Impressions (the number of times the post has been viewed - not unique)
- Engagements (a number, a count of the actions that consumers do when they see the post, varies by platform)

In order to scrape the data, the development team will have to use frameworks and not the social media platforms' own API. The data parsed will be put into a separate system to hold the data. Once complete, the interface will show the user the list of posts and give the option to download to a spreadsheet or csv file. User will be shown a comparison of the brand handles used.

2.3 User Characteristics

As described by the client, the users of the product will not be skilled with computers. These individuals will most likely not have any high-level technical skills for computing and will need a simple interface for inputting and receiving output. These specifications will impact the design by making sure developers keep complex, confusing information behind the scenes so the user can get their results without any unnecessary complications.

2.4 Constraints

1. The "Social Media Evaluator" applicated is constrained by the developers not being able to use the social media platforms API for every platform. This may cause obstacles for the developers because creating web scrapers could prove challenging. Platforms have automation protection which requires the developers to build solutions for entering user info to view posts, get past proxies, and avoid throttling.
2. Each site has different roadblocks put in place. Figuring out how to bypass them for each social media platform will take different amounts of time most likely. Some will be easier than others.
3. Only three brands are required to be able to be compared, with 5 or more being better. Sometimes businesses may want to compare many brands at once and this would be a constraint. The more brands being searched, the longer it will take the program.
4. For this project, developers are only going to be scraping publicly found data on the social media platforms.

2.5 Assumptions and Dependencies

One assumption being made is that this application will be run on a device with internet connection. This application relies on having the internet to access the data to be scraped and to access the database that will hold data.

Another assumption being made is that the users computing device will have the similarly up-to-date technology (operating system, web browser(s)) to be compatible with the web-scraping technology that the developers of this project will implement.

Something that will be important with future social media platforms is portability. The developers of this project will either do each platform by hand or code will be portable between platforms; if portable between the assigned, it is assumed that future platforms will be able to be scraped similarly. The project will have a parent class with child classes for specific platforms. Developers will have to worry about a case where the constructor is given a platform it doesn't recognize.

2.6 Apportioning of Requirements.

In the future, it is also hoped that the application will be able to scrape from every client-mentioned social media platform as well as others not listed, already existing currently or in the future.

3. Specific Requirements

3.1.1)

- User Interface shall contain 6 different social media platforms that the user can select to compare each of the products. The following social media platforms will be an option for the user to scrape:
    - Facebook
    - TikTok
    - Instagram
    - Pinterest
    - YouTube
    - Twitter
    -

3.1.2)

- User Interface will contain 7 input fields:
    - 5 input fields for brand handles
    - 2 input fields for date ranges

3.1.4)

- The output of the software will produce a list containing:
  - Platform Name
  - Brand handle
  - Post URL
  - Post Date
  - Post Text
  - Engagements/Impressions
    - Facebook: Likes, Comments, Shares
    - Pinterest: Emoticons, Comments, Pins, Closeups, Saves
    - TikTok: Likes, Comments, Shares
    - YouTube: Likes, Comments
    - Twitter: Likes, Retweets, Comments
    - Instagram: Likes, Comments

This information will be uploaded to the client's database.

3.1.5)

- Option for user to download a CSV file to their device. This CSV file will consist of information stated in **3.1.4** from the most recent search query that the user entered.

3.1 External Interfaces

- **Search Interface**

  Purpose:
  - Presented to user once they successfully access the web page interface
  - Enter up to 5 brand handles that the client is searching for and update a database with all information that was gathered

  Input:
  - The user will be asked to enter no less than 3 fields to begin a search query
    - Up to 5, no less than 1 brand handles
    - 2 date fields (start – end)

  Output:
    - The destination of the output will be on screen as well as the csv that is available for them to download

  Valid Input:
    - Brand handles: The brand handles need to be a perfect match to what the social media page is registered as. If there are any differences between the input and the actual brand handle, the user will receive an error message that could not be found.
    - Date ranges: The date ranges that are input need to be formatted in a chronological order

  Relationships to other inputs/outputs:

- The social media results that the user searched for in the website interface will be an input for the database

Screen Formats:

- All 6 social media choices located at top of the page
- Below social media choices will be date section with two inputs (start and end)
- Next section is where user inputs all brand handles
- Last section which is located at bottom of the page is a button to initiate the search and download the csv

Data Formats:

- Brand handles: must match brand handles exactly or error will be produced

End messages: N/A

3.2 Functions

3.2.1) Validity Checks:

- The application that we develop will check for correct handles that are inputted to search social media accounts
- The system shall check for valid dates

3.2.2) Exact Sequence of Operations:

- User will input brand handles and a date range into the GUI and the application
- Application will check if the brand handles that the user inputted are valid and are referencing existing accounts
- A validation message will be presented to the user if the brand handles that were entered were found
- Display a proper error message to the user and ask them to try again if the brand handles do not exist

3.2.3) Responses to Abnormal situations

- Incorrect brand handle:
  - Message printed to user that the brand handle field is incorrect and to check if it is correct
- Incorrect date format:
  - Message printed to user that the date fields are incorrect
- Communication to remote database failure:
  - The system shall print a message to the user that the system database cannot be reached
  - The user will be asked to contact the database administrators

3.3 Performance Requirements

- Supports simultaneous users

- Below are listed the factors that will play a key role in the amount of information to be handled:
    - Frequency of using the application
    - Amount of brand handles that are searched when using the application

Dynamic numerical requirements:

- 100% of the data scraping will be performed immediately when the user presses the scrape button
- Performance will be dependent on the amount of information that is available for each brand handle on a social media platform

3.4 Logical Database Requirements

- Data retention requirements will be that the database will hold all information that is listed in **3.1.4**
- The data that is collected will have no relationships with each other
- Accessing capabilities will be dependent on the database administrator

3.5 Design Constraints

- 20 simultaneous users will be supported through Amazon LightSail
- Amazon LightSail will provide 160GB of storage, 8GB of RAM, and 5TB transfer speed
- User will only be able to input 3 brand handles for one specific social media platform
- Only one social media platform will be able to be selected when searching for statistics

3.5.1 Standards Compliance

- Data naming will be presented in CSV file that is available to download once the search for every brand handle is completed
- Any data manipulation will be logged through AWS

3.6 Software System Attributes

- The website will provide web-scraping functionalities capable of retrieving information for the following 6 social media outlets:
    - Twitter, Pinterest, Instagram, Facebook, TikTok, Youtube

- The website will communicate with a database and store data with relevant information such as dates, brand handles, likes and comments into a CSV file and then into the database.

3.6.1 Reliability

Website:

- The website will ensure that retrieved data from social media outlets is correct.

Database:

- The database will be evaluated each day in order to validate the correct storage of retrieved information by submitting a test query to the webscraping tool and storing the retrieved information object.

3.6.2 Availability

Website:

- **Maintain session information relating to:**
    - Data inputs
        - The brand handle of a company on a social media outlet which the user would like to gather statistics on.
        - Which social media outlets the user would like to retrieve information from and compare retrieved data against.
        - Date range in which the information being retrieved must fall within.
    - User login

The website will be expected to recover from any crashing or downtime within an hour.

- **Upon crashing of website**
    - Identify data loss associated with previously queried information whether from database or prior to web-scraping
- Requested brand information
- Requested time frame
- Requested social media outlets

The web-scraping tool will recover given any bad data inputs based on the failed request

- Notify user of invalid URL
- Notify user of failure to obtain requested information
    - Notify user if information does not exist
    - Notify user if information cannot be scraped

Database:

The database will store information related to user queries that were made on the website including: brand handle queried, date range quered, and information with respect to the social media outlet that was queried such as likes, shares and comments for Facebook.

3.6.3 Security

Website:

The website front end will contain a GUI with various input boxes for the user including date ranges, brand handles and the amount of brands to compare.
The server will perform validity checks related to webscraped data in order to avoid critical vulnerabilities and failures in the database

- Input Validation relating to brand handles to avoid improper brand handle input and the subsequent return of information not pertaining to the original query.
- Input Validation relating to date ranges such that date ranges must fall within the alloted timeframe of the server IE 90 days.
- Requested information is valid IE not "404 Error" being stored

Website will utilize SSO and integrate Google login.

Database:

Database will keep a log and history of previously queried/scraped results and the user who requested them.

3.6.4 Maintainability

Website:

Website will have a simple interface for user input queries regarding brand, timeframe, social media outlets to be scraped.
Existing frameworks which run the webscraping tool will be updated accordingly.
Daily tests to test working functionality of the web scraper including validation of returned data and validation of retrieval of correct brand handle.

Database:

Validate that database is storing data correctly and that previously scraped data still exists in the database as long as it is not passed the maximum storage time of 90 days.

3.6.5 Portability

The program will be portable to any website or server stack that supports the technologies used in the project.
The program will require a web connection, a host OS, and a database to be ported to another system.

3.7 Organizing the Specific Requirements

3.7.1 System Mode

Interfaces will be website based and functionality will be back-end based still tied to the website.

3.7.2 User Class

Users will be able to query the web-scraping tool to pull brand handle information.

Admins will be able to query the web-scraping tool and modify the webserver configuration.

3.7.3 Objects

There will be a webscraping parent class which is inherited by webscraping children classes related to each of the 6 social media outlets.
Each child class will return an informational object which lists the information for any particular brand handle that has been searched with respect to the social media outlet that was queried.
The webscraping tools will be accessible to the users and informational objects will be obtained by the user once they enter a brand handle to be scraped.

3.7.4 Feature

Web-scraping tool which allows the user to obtain an informational data object pertaining to the brand handle that was scraped.
User input functions for comparison of brand handles on up to 5 different social media outlets at once.
Submit button which passes the brand handles input by the user for comparison to the web-scraping tool.
Date-range which will allow the user to select two dates in which they would like to search and compare social media outlets for information.
Database which records data object information queried from brand handles by the webscraping tool.
Functionality to port queried data object information into a CSV file to be recorded into a database.

3.7.5 Stimulus

Feature: Retrieve Brand Information
User will input up to 5 brand handles to retrieve data from. User will input a date range to search for information pertaining to brand handle. User will submit the query to the webscraping tool.

Feature: Database Storage
User will input up to 5 brand handles to retrieve data from. User will input a date range to search for information pertaining to brand handle. User will submit the query to the webscraping tool. The webscraping tool will submit the retrieved information to the database.

3.7.6 Response

Expected response from social media outlet containing scraped information related to engagement and

3.8 Additional Comments

No additional comments as of current state of project.

4. Change Management Process

Should changes need to be made to this requirements document, team members will edit the necessary information and notify the client to have it approved. Additionally, the client will have access to the development Github and has access to the development team's Slack channel. Strong communication will be maintained between the development team and the client, so any changes to the requirements document should not be a surprise to the client.
When changes are made and approved, the version number on the first page of this document will be updated.

All changes are logged below:

- 10/12/2021
    - Requirements Document feedback received from the client. Team has begun updating all necessary sections.
- 10/26/2021
    - Updates completed for sections 2.0 - 2.18, 3.0 - 3.8, based on client feedback.
- 10/28/2021
    - Updates completed to section 2.2 - 3.0, document is ready to be presented to client again

5. Document Approvals

The potential people approving of the SRS document will be listed below:

- Andres Dicochea
- James Mortensen
- Eric Slick
- Peter Mansbacher
- Debra Parcheta