

The University of Hong Kong
Department of Computer Science

COMP4801 Final Year Project
Interim Report

Automated Social Media Research and Sentiment Analysis using LLMs

Group: FYP23047

Students:

Hung Jasper (3035787671)

Ko Berrick Eldridge (3035784710)

Supervisor:

Dr. Kong Lingpeng

21 January 2024

I. Abstract

Social media offers organizations valuable insights into user feedback and public sentiment. Manual social media research can be impractical, and existing approaches to sentiment analysis have shown vulnerabilities. This project aims to address these challenges by developing an automated system to perform social media research and conduct aspect-based sentiment analysis utilizing state-of-the-art LLMs, such as GPT and Gemini. The system consists of four main procedures: data collection, data pre-processing, data analysis, and result visualization. At present, the implementation for data collection and the first stage of data analysis has been completed. The successful implementation of the system will not only provide organizations with a valuable tool but also contribute to the broader understanding of LLMs and their capabilities.

II. Acknowledgment

We would like to express our sincere gratitude to our supervisor, Dr. Kong Lingpeng, for his guidance and support throughout the project. We would also like to thank our CAES9542 instructor, Mr. Matthew Anderson, for his guidance in technical writing and presentation.

III. Table of Contents

I. Abstract.....	2
II. Acknowledgment.....	3
III. Table of Contents.....	4
IV. List of Figures.....	6
V. List of Tables.....	7
VI. Abbreviations.....	8
1. Introduction.....	9
1.1 Background.....	9
1.1.1 User Feedback.....	9
1.1.2 Social Media.....	9
1.1.3 Sentiment Analysis.....	10
1.1.4 Large Language Models.....	10
1.2 Project Motivations.....	11
1.2.1 Automated Social Media Research.....	11
1.2.2 Sentiment Analysis using LLMs.....	11
1.3 Related Work and Significance of Project.....	11
1.4 Objectives.....	11
1.5 Deliverables.....	12
1.6 Report Outline.....	12
2. Methodology.....	13
2.1 System Design.....	13
2.2 Procedures.....	14
2.2.1 Data Collection.....	14
2.2.2 Data Pre-processing.....	15
2.2.3 Data Analysis.....	15
2.2.3.1 Aspect Identification and Classification.....	16
2.2.3.2 Aspect-based Sentiment Analysis.....	16
2.2.3.3 Aspect-based Summarization.....	17
2.2.4 Result Visualization.....	17
3. Work Completed and Results.....	18
3.1 Data Collection.....	18
3.1.1 Data Collection from Reddit.....	18
3.1.2 Data Collection from YouTube.....	18
3.1.3 CSV Files.....	19
3.2 Data Analysis.....	20
3.2.1 Aspect Identification and Classification.....	20
3.2.2 Performance Evaluation.....	20
4. Difficulties and Mitigations.....	22
4.1 Data Collection.....	22
4.1.1 Data Collection from X, Facebook, and Instagram.....	22

4.2 Data Analysis.....	23
4.2.1 Restricted Access to Online Services.....	23
4.2.2 Instability of Online Services.....	24
5. Project Status and Proposed Schedule.....	25
6. Conclusion.....	26
VII. References.....	27

IV. List of Figures

Figure 1. An overview of the system workflow.....	13
Figure 2. An overview of the workflow for aspect identification and classification.....	16
Figure 3. A sample CSV file with data collected from Reddit.....	19
Figure 4. A sample CSV file with data classified into identified aspects.....	20
Figure 5. API plans for accessing X's data.....	22
Figure 6. "Page Public Content Access" permission required to access Meta's data.....	23
Figure 7. An internal server error occurred when using Gemini.....	24

V. List of Tables

Table 1. Targeted social media platforms for data collection.....14

Table 2. Project status and proposed schedule.....25

VI. Abbreviations

LLM	Large Language Model
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformers
API	Application Programming Interface
CSV	Comma-Separated Values
RoBERTa	Robustly optimized BERT approach
PRAW	Python Reddit API Wrapper
VPN	Virtual Private Network

1. Introduction

This section introduces the project. Section 1.1 provides the project's background; section 1.2 explains the project's motivations; section 1.3 presents related work and justifies the project's significance; section 1.4 outlines the project's objectives; section 1.5 details the project's deliverables; section 1.6 provides an overview of the remaining content of the report.

1.1 Background

1.1.1 User Feedback

User feedback is a valuable source of information, helping organizations understand the needs and preferences of their users. This understanding allows them to identify areas for improvement and make informed decisions when developing new products. Addressing user feedback effectively can also make users feel heard and lead to improved customer loyalty.

Traditionally, user feedback is collected through surveys, interviews, and focus groups. However, these methods come with several disadvantages. First, they rely on limited sample sizes, resulting in some user segments being underrepresented. Moreover, both designing and conducting them requires significant time and effort. Additionally, these methods typically provide predefined questions and lack open-ended discussions, thereby limiting the depth of the feedback collected.

1.1.2 Social Media

In the current digital era, social media has become an integral part of our lives, providing platforms for individuals to freely express opinions and engage in discussions. As of 2023, it is estimated that around 60.5% of the global population are active social media users [1]. The surge in social media usage has resulted in a substantial volume of user-generated content across such platforms, making them valuable resources for organizations to gain insight into user feedback and public sentiment.

The social media landscape is diverse, including various types that suit different content and communication styles. The main categories encompass social networking sites like Facebook that promote connections, discussion forums like Reddit where users answer each other's questions, and video-sharing platforms like YouTube for sharing videos [1].

1.1.3 Sentiment Analysis

Sentiment analysis is the process of analyzing emotions from text and classifying them into positive, neutral, or negative sentiments [2]. It has found broad applications, particularly in brand monitoring, market research, and the analysis of user feedback.

The analysis can be categorized into different types based on scope and complexity. The prevalent type is document-level sentiment analysis, which analyzes the sentiment of an entire text [3]. Another type, known as aspect-based sentiment analysis, identifies specific aspects within a text and analyzes the sentiment associated with each identified aspect [3].

There are two main approaches to conducting sentiment analysis. The simpler lexicon-based approach assigns sentiment scores to keywords based on a predefined dictionary [4].

However, this method is prone to inaccuracies, as the same word can convey both positive and negative meanings depending on the context. The machine learning approach is more sophisticated, which predicts sentiments using trained models. Nevertheless, this approach still falls short of fully comprehending human language, especially in areas like irony and sarcasm [4].

1.1.4 Large Language Models

Large Language Models (LLMs) are machine learning models designed for Natural Language Processing (NLP) tasks. Leveraging transformer architectures, these models demonstrate proficiency in understanding and generating human-like language, making them valuable tools in applications like translation and text generation. Prominent examples of state-of-the-art LLMs include Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformers (GPT) and Gemini.

Prompt engineering is an essential process in interacting with LLMs. The design and development of well-optimized prompts can significantly enhance the capabilities of LLMs and improve the quality of their outputs [5]. The elements of a well-crafted prompt include clear instructions, relevant context, input data, and an output indicator [5].

1.2 Project Motivations

1.2.1 Automated Social Media Research

The significant amount of user-generated content on social media presents a valuable resource for organizations to collect user feedback and perform sentiment analysis. While manual methods of social media research are available, the vast number of posts and comments across various platforms makes this approach labor-intensive, time-consuming, and often impractical. Therefore, this project aims to develop a tool to automate the research process.

1.2.2 Sentiment Analysis using LLMs

Existing approaches to sentiment analysis have demonstrated vulnerabilities, particularly in understating human language accurately. Considering that LLMs possess the capability to comprehend human language, this project aims to explore the use of LLMs as an improved approach to sentiment analysis.

1.3 Related Work and Significance of Project

Similar studies have been conducted previously. Kheiri and Karimi explored the use of GPT models for sentiment analysis, and a similar study was conducted by Carneos et al [4, 6]. However, their studies focused on performing document-level sentiment analysis using GPT models specifically. In contrast, this project seeks to explore the potential of various LLMs in the context of aspect-based sentiment analysis. Moreover, the studies will be implemented through a real-world system, which can serve as a valuable tool for organizations to gather user feedback and understand areas for improvement.

1.4 Objectives

This project aims to develop an automated system that performs social media research and conducts sentiment analysis using state-of-the-art LLMs. With the above agenda, this project has the following objectives:

1. Collect relevant textual data from social media based on user-specified keywords.
2. Clean and transform the collected data appropriately for further analysis.
3. Conduct aspect-based sentiment analysis on the preprocessed data using LLMs.
4. Generate reports with appropriate visualizations to present the analysis results.

1.5 Deliverables

The project aims to deliver a command-line program that achieves the aforementioned objectives. The program should encompass the following procedures: data collection, data pre-processing, data analysis, and result visualization.

1.6 Report Outline

The remainder of this report consists of six sections. Section 2 provides an explanation and justification of the methodology employed in the project; section 3 discusses the work completed and the results available thus far; section 4 addresses difficulties encountered along with their respective mitigations; section 5 outlines the schedule for the remaining work; section 6 summarizes the essential components of the report.

2. Methodology

This section explains and justifies the approach taken for the project. Section 2.1 describes the overall design of the system; section 2.2 details the procedures involved in conducting social media research and sentiment analysis.

2.1 System Design

The system will be implemented as a command-line program rather than a fully-fledged application. By avoiding the need to design and develop dedicated frontends or backends, more effort can be directed toward developing the core capabilities for conducting social media research and sentiment analysis, which is the primary focus of the project.

Additionally, command-line programs offer high flexibility for integration into existing applications and high extensibility for future developments.

The system will be developed in Python because it offers numerous libraries, such as Pandas, that align well with the project's requirements. Moreover, Python is widely recognized as the language of choice for data analysis and machine learning, with abundant online resources available for reference.

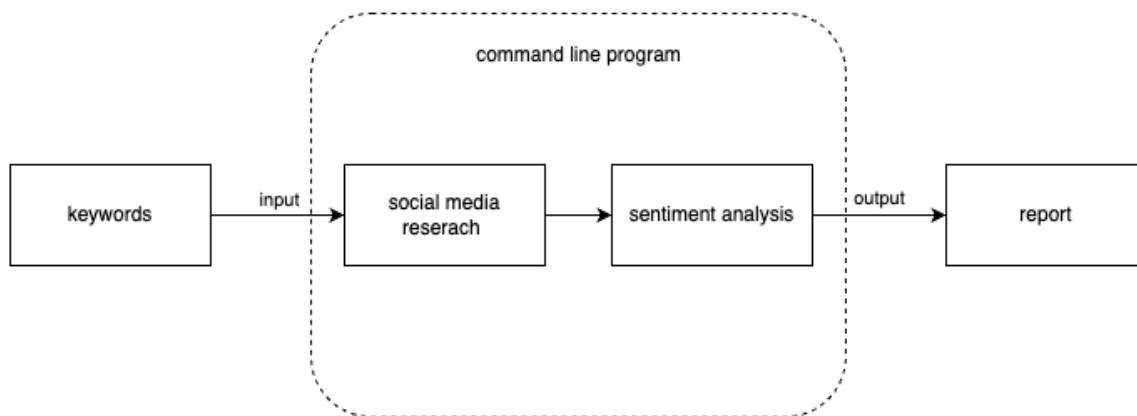


Figure 1. An overview of the system workflow

The proposed workflow of the system is illustrated in Figure 1. First, users will be prompted to input keywords of interest. Then, the system will perform social media research by collecting related data and proceed to conduct sentiment analysis on them. Finally, a report will be generated to present the findings.

2.2 Procedures

The system consists of four main procedures for performing social media research and sentiment analysis: data collection, data pre-processing, data analysis, and result visualization.

2.2.1 Data Collection

Based on the provided keywords, the system will collect relevant posts and corresponding comments from various social media platforms. This will be accomplished through the official Application Programming Interfaces (APIs) whenever accessible, ensuring reliability and efficiency. If APIs are not available, web scraping techniques will be employed to crawl and extract relevant data. In either case, the process will strictly adhere to each platform's terms of service to ensure legal and ethical practices.

	X	Reddit	YouTube	Facebook	Instagram
rank [1]	10th	15th	3rd	1st	4th
type	discussion forums		video-sharing platform	social networking sites	
official API	yes	yes	yes	yes	yes

Table 1. Targeted social media platforms for data collection

Table 1 displays the social media platforms targeted for data collection, including X, Reddit, YouTube, Facebook, and Instagram. These platforms are selected for several reasons. First, they are among the most used social media platforms in the world. Second, the selection encompasses major types of social media. Third, each of them provides official APIs for accessing their data.

While various types of data, such as images or videos, are available on these platforms, the system will exclusively collect textual data. This is because the subsequent analysis will focus solely on text. In addition to the textual content of the post or comment, attributes such

as the corresponding rating (number of likes minus dislikes) and the date of posting will be collected, depending on the depth of the subsequent analysis.

Collecting all available data relevant to the provided keywords is impractical. To strike a balance between computational feasibility and data comprehensiveness, the system will collect a maximum of 1,000 data instances from each social media platform. Subsequently, the collected data will be stored in Comma-Separated Values (CSV) files for easy retrieval and to facilitate later procedures.

2.2.2 Data Pre-processing

The collected data will undergo pre-processing to ensure its suitability for subsequent analysis. First, regular expressions will be employed to remove noise, such as account tags and hyperlinks. Next, irrelevant data, such as comments generated by bots, will be filtered out. The identification of such content requires data classification using machine learning. In this context, BERT models will be employed for their bidirectional contextual understanding, thus allowing accurate classification of the collected data into relevant and irrelevant segments.

2.2.3 Data Analysis

The pre-processed data will undergo analysis using LLMs. This procedure can be divided into three stages: aspect identification and classification, aspect-based sentiment analysis, and aspect-based summarization.

2.2.3.1 Aspect Identification and Classification

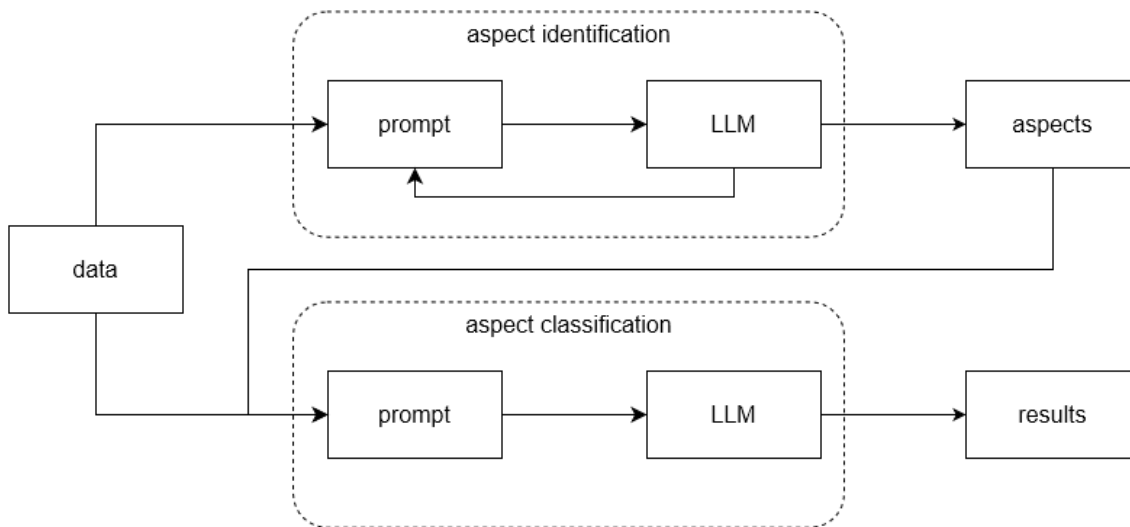


Figure 2. An overview of the workflow for aspect identification and classification

Aspects within the pre-processed data will be identified, and the data will be classified into these identified aspects. Figure 2 illustrates the overall flow for this stage. First, the data is structured into a suitable prompt and input into an LLM for aspect identification. The result will then be utilized to structure another prompt, subsequently fed into a separate instance of the LLM for aspect classification.

Given the limitations on input prompt length for LLMs, the described process will be executed in batches of 50 data instances. During the aspect identification phase, the identified aspects will serve as references for subsequent batches, preventing different names for the same aspect throughout various batches of data.

Various LLMs will be tested and compared to determine the most effective model for the final implementation. In addition, different prompts will be experimented with to identify the most effective approach for the tasks mentioned above.

2.2.3.2 Aspect-based Sentiment Analysis

With the data categorized into the identified aspects, the next stage involves classifying them into positive, neutral, and negative sentiments specific to their respective aspects, and determining the corresponding sentiment percentages. Studies have suggested approaches

involving different LLMs, such as the Robustly optimized BERT approach (RoBERTa) and GPT-based approaches [4]. These approaches vary in accuracy, model sizes, and computational complexities. Nevertheless, the utilization of different LLMs as well as prompts will be explored and evaluated to strike a balance between accuracy and efficiency.

2.2.3.3 Aspect-based Summarization

In addition to determining the sentiment percentages, key points will be extracted from the data to provide a summary for each of the identified aspects. This entails identifying the frequently occurring words, prevailing opinions, and any advice suggested for improvement. LLMs will be utilized in this stage for their summarization capabilities. Similar to previous stages, a rigorous testing and comparison process will be conducted, considering various LLMs and prompts to ensure optimal performance.

2.2.4 Result Visualization

The system will present the analysis results through appropriate visualizations in the form of a report. For each identified aspect in the data, the report will incorporate the following proposed elements:

1. A pie chart showing proportions of positive, neutral, and negative sentiments.
2. A word cloud presenting frequently occurring words.
3. A list of prevailing opinions.
4. A list of suggestions for improvement.

Python libraries such as Matplotlib will be employed for creating visualizations like pie charts and word clouds. Additional elements such as sentiment fluctuations may be integrated to provide a more in-depth report, and the preceding procedures will be adjusted accordingly to generate the desired output.

3. Work Completed and Results

This section presents the work completed and the available results of the project. Section 3.1 presents those regarding data collection; section 3.2 presents those regarding data analysis.

3.1 Data Collection

The data collection procedure, as outlined in section 2.2.1, has been implemented for Reddit and YouTube.

3.1.1 Data Collection from Reddit

A script application was established on Reddit, generating a client ID and client secret necessary for interacting with the official Reddit API. The Python Reddit API Wrapper (PRAW) library was utilized to streamline the interaction.

The detailed steps employed for collecting data from Reddit are as follows:

1. Create a CSV file named *{provided-keywords}-reddit.csv*
2. Search for posts (known as submissions) relevant to the provided keywords
3. For each relevant post, record the textual content and the rating received in the CSV file
4. For each comment of the post and its corresponding replies, record the textual contents and the rating received in the CSV file
5. Stop when 1000 data instances have been recorded, or all relevant posts and comments have been recorded

The API grants read access to all public posts and comments. However, it imposes a rate limit of 100 queries per minute per client ID [7]. PRAW manages these rate limits by appropriately spacing out requests to the Reddit API.

3.1.2 Data Collection from YouTube

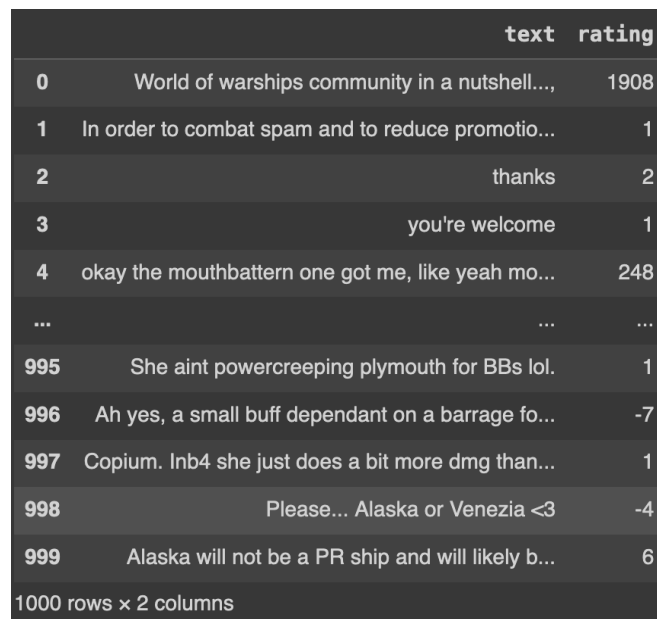
A project was set up on Google Cloud Console, with the official YouTube API enabled. An API key was generated to interact with the API. The Google Client Library was utilized to streamline the interaction.

The detailed steps employed for collecting data from YouTube are as follows:

1. Create a CSV file named *{provided-keywords}-youtube.csv*
2. Search for videos relevant to the provided keywords
3. For each relevant video,
 - i. Record the title and the rating received in the CSV file
 - ii. Search for its comments
4. For each comment,
 - i. Record the textual content and the rating received in the CSV file
 - ii. Search for its replies
5. For each reply, record the textual content and the rating received in the CSV file
6. Stop when 1000 data instances have been recorded, or all relevant videos and comments have been recorded

The API grants read access to all public videos and comments. However, it imposes a quota limit of 10,000 units per day [8]. Searching for videos costs 100 units, searching for a video's comments costs 1 unit, and searching for a comment's replies costs 1 unit.

3.1.3 CSV Files



	text	rating
0	World of warships community in a nutshell...	1908
1	In order to combat spam and to reduce promotio...	1
2	thanks	2
3	you're welcome	1
4	okay the mouthbattern one got me, like yeah mo...	248
...
995	She aint powercreeping plymouth for BBs lol.	1
996	Ah yes, a small buff dependant on a barrage fo...	-7
997	Copium. Inb4 she just does a bit more dmg than...	1
998	Please... Alaska or Venezia <3	-4
999	Alaska will not be a PR ship and will likely b...	6

1000 rows x 2 columns

Figure 3. A sample CSV file with data collected from Reddit

Figure 3 shows an example of a CSV file containing the collected data. It consists of two columns: text, representing the textual content of the collected posts or comments; rating, indicating the rating of the corresponding post or comment received.

3.2 Data Analysis

The data analysis procedure has been completed for the aspect identification and classification stage, as outlined in section 2.2.3.1.

3.2.1 Aspect Identification and Classification

Using the proposed framework, the system was able to first identify aspects within the input data and then classify the data into these identified aspects. The results were stored in CSV files.

▲ ▼	Text ▼	Upvotes ▼	Topics
995	Right, because the rest of the ships irl had planes they could instantly materialize out of thin air to drop charges.	1	['historical accuracy']
996	Just saying it had an absolutely bullshit reason of wargaming pulling out the historical accuracy card	1	['historical accuracy']
997	Yeah, it's hilarious when they do that because practically nothing about the game is realistic or historical to begin with.	1	['historical accuracy']
998	Mmm, yes, because the ranged submarines (Gato, I-56) are good enough at hitting targets past 8km without calculus t	3	['submarines', 'game mechanics', 'counterplay']
999	Honestly, subs don't bother me too much. The only thing I dislike are CVs when I play torp DD.	2	['submarines', 'CVs']
1000	Thrasher and U4501 are pretty trash for their respective standard, U4501 is only barely better than the non-premium B	0	['submarines']
1001	Fire the homing torps like normal torps at 30m, and then use the ping to set their depth and last minute corrections w	2	['homing torpedoes', 'shotgunning']
1002	yeah and then because you pinged, they will bomb you in response, not to mention the U4501 has an odd launcher co	2	['submarines', 'shotgunning']
1003	Shotgun isn't everything. 4501 has fantastic DPM, concealment and mobility. Besides, she has almost the same homing	3	['submarines']
1004	> This will nerf their ability to last stand and shotgun before they are overrun. What exactly? Certainly not the sub hydr	7	['submarines', 'shotgunning']
1005	when a submarine is at maximum depth and knows it is being searched by someone using hydro thanks to the comma	9	['submarines', 'hydroacoustic search', 'shotgunning']
1006	Oh right so the fact that it only lasts 6 seconds now is what makes shotgunning harder? Yeah that makes sense.	1	['shotgunning']
1007	no, what makes it harder is the scaling nerf, the fact some cruisers may have sub radar, and that several DC cruisers are	5	['shotgunning', 'submarines', 'radar']
1008	So shotgunning isn't actually directly any more difficult, just more difficult against some specific target due to subs ove	0	['shotgunning', 'submarines']
1009	Dude they nerfed Torps, Or changes them, they Deal much less damage within 3km	4	['torpedoes', 'damage']
1010	Yeah that's in the "upcoming changes after this" part, but the shotgunning is claimed to be nerfed in the earlier section	1	['shotgunning']
1011	Thats right, shotgunning will remain until they completely overhaul all dumbfires for all submarines in addition to mak	1	['shotgunning']
1012	Giving surface ships tools to detect them inherently will make it way harder to randomly surface next to one and oblite	1	['detection']
1013	> Submarine dumbfire torpedoes will be very slow and deal low damage until about 3 km distance travelled. But it cou	2	['torpedoes']

Figure 4. A sample CSV file with data classified into identified aspects

Figure 4 shows an example of a CSV file, with data collected from the previous procedure classified into the identified aspects. It consists of an additional column "Topics", which is a list of aspects identified from the corresponding post or comment.

3.2.2 Performance Evaluation

To assess the performance of different models used for data analysis, data collected from Reddit relevant to *World of Warships* was used for testing. It is a player-versus-player game that has sparked significant disputes around the concept of "game balance," referring to the equilibrium of strength between different warships in the game. The aspects (warships under

discussion) and stance (whether the warship is perceived as too strong or weak) are clearly defined in discussions found on Reddit. Thus, the dataset allows for a straightforward quantification and assessment of the model's performance. In each test, 100 data instances were randomly drawn from the resulting CSV files and evaluated.

In the current implementation, Gemini was chosen as the LLM. The system scored an accuracy of 0.8649. Most of the inaccuracies originated from the model generating a generic output. For instance, when a comment mentioned multiple ships in the game, the model classified the comment under the broad aspect 'warships' instead of each individually mentioned warship. While these outputs may be considered correct, there is a risk of losing insights if the classification is too general. For instance, it might miss specific balancing issues related to frequently mentioned individual warships. Therefore, adjustments to the prompts could be made to optimize the system's performance.

The system's performance using *facebook/bart-large-mnli*, an encoder model designed for zero-shot classification, was also assessed. The resulting accuracy was 0.3514, significantly lower than Gemini. This discrepancy can be attributed to the encoder model's strong bias toward exact wordings, leading to numerous false classifications. Additionally, the model struggled with capturing some abbreviations, resulting in the omission of many aspects during classification. While these issues could potentially be addressed through further fine-tuning of the encoder model, LLMs stand out due to their powerful generalization ability and proficiency in understanding abbreviations, providing greater flexibility and broader applications compared to classification encoder models.

4. Difficulties and Mitigations

This section addresses the encountered difficulties and the corresponding mitigation strategies. Section 4.1 details those in data collection; section 4.2 details those in data analysis.

4.1 Data Collection

4.1.1 Data Collection from X, Facebook, and Instagram

As outlined in section 2.2.1, the system is intended to collect relevant posts and comments from Reddit, YouTube, X, Facebook, and Instagram. While the implementation for the first two platforms has been completed, some obstacles prevented the implementation for the remaining platforms.

Free	Basic	Pro	Enterprise
For write-only use cases and testing the X API	For hobbyists or prototypes	For startups scaling their business	For businesses and scaled commercial projects
<ul style="list-style-type: none">Rate limited access to v2 post posting and media upload endpoints1,500 Posts per month - posting limit at the app level1 app IDLogin with XFree	<ul style="list-style-type: none">Rate limited access to suite of v2 endpoints3,000 Posts per month - posting limit at the user level50,000 Posts per month - posting limit at the app level10,000 Posts per month - read-limit rate cap2 app IDsLogin with X\$100 per month	<ul style="list-style-type: none">Rate-limited access to suite of v2 endpoints, including search and filtered stream1,000,000 Posts per month - GET at the app level300,000 Posts per month - posting limit at the app level3 app IDsLogin with X\$5,000 per month	<ul style="list-style-type: none">Commercial-level access that meets your and your customer's specific needsManaged services by a dedicated account teamComplete streams: replay, engagement metrics, backfill, and more featuresMonthly subscription tiers
Get started	Subscribe now	Subscribe now	Apply now

Figure 5. API plans for accessing X's data [9]

In the case of X, while the platform provides an official API, it requires a minimum cost of US\$ 100 per month to have read access to its data, as shown in Figure 5. Given the project's budget of HK\$ 2,000, utilizing the API is not feasible. While web scraping could be an alternative, it is stated clearly that such a practice is against the platform's terms of service [9].

Page Public Content Access

The **Page Public Content Access** feature allows an app access to the Pages Search API and to read public data for Pages for which you lack the **pages_read_engagement** permission and the **pages_read_user_content** permission. Readable data includes business metadata, public comments and posts. The allowed usage for this feature is to analyze and/or display posts and engagement on Pages.

Allowed Usage

- Analyze and/or display posts and engagement on Pages.

Common Endpoints

`/page/feed`
`/page-post`
`/page-post/comments`

Additional Details

- This permission or feature requires successful completion of the App Review process before your app can access live data. [Learn More](#)
- This permission or feature is only available with business verification. You may also need to sign additional contracts before your app can access data. [Learn More Here](#)
- While you are testing your app and before you submit it for review, your app can only access content on a Page for which the following is true: The person who holds the admin role for the Page also holds an admin, developer, or tester role on the app. If you want the app to be able to access public content on other Pages, you must submit this feature for review. Once you set your app to live mode, it will not be able to see any Page public content without this feature.

Figure 6. “Page Public Content Access” permission required to access Meta’s data [10]

For Facebook and Instagram, Meta offers official APIs for both platforms, but access to their public data requires the “Page Public Content Access” permission, as shown in Figure 6.

Unfortunately, obtaining this permission has proven challenging. Similar to X, web scraping is prohibited according to their terms of service [10].

Given the constraints posed by the above restrictions, the system has determined to limit data collection to Reddit and YouTube exclusively.

4.2 Data Analysis

4.2.1 Restricted Access to Online Services

For the implementation of data analysis, LLMs, including GPT and Gemini, were intended to be utilized and accessed through their respective online services. However, direct access to both services from Hong Kong was still not officially supported.

To address this issue, a virtual private network (VPN) was employed to establish access to the services. However, access to paid services of OpenAI like using GPT-4 was still blocked, as payments using credit cards issued in Hong Kong were declined.

4.2.2 Instability of Online Services

```
google.api_core.exceptions.InternalServerError:  
500 An internal error has occurred. Please retry  
or report in https://developers.generativeai.google/guide/troubleshooting
```

Figure 7. An internal server error occurred when using Gemini

Throughout the data analysis implementation, the use of GPT and Gemini through their online services proved to be unstable, experiencing frequent interruptions. Figure 7 shows one such example, where Gemini's services faced timeouts during peak usage hours. These interruptions not only impeded the development process but also raised concerns about the reliability and stability of the system.

Since these interruptions were unpredictable and unavoidable, a strategy using try blocks was implemented. When sending queries through the online services, the system would make up to two retry attempts. If all three attempts failed, the system would skip the current batch of data for analysis. This approach has proven to improve the system's reliability.

5. Project Status and Proposed Schedule

Date	Tasks	Status
October 2023	Project Research	completed
December 2023	Data Collection	completed
January 2024	Data Pre-processing	in progress
February 2024	Data Analysis - Aspect Identification and Classification	completed
	Data Analysis - Aspect-based Sentiment Analysis	in progress
	Data Analysis - Aspect-based Summarization	to be completed
March 2024	Result Visualization	to be completed
April 2024	System Optimization	to be completed

Table 2. Project status and proposed schedule

Table 2 outlines the current project status and presents the proposed schedule for the remaining tasks. As addressed in section 3, the implementation of data collection as well as aspect identification and classification has been completed. The remaining tasks include the implementation of data pre-processing, the remainder of data analysis, and result visualization. The work on data pre-processing is currently in progress. It is anticipated to conclude by January 2024. Concurrently, progress is underway for aspect-based sentiment analysis, and together with aspect-based summarization, they are projected to be completed by February 2024. Lastly, result visualization is scheduled for completion by March 2024.

6. Conclusion

This project aims to develop a system to automate the process of social media research and to perform aspect-based sentiment analysis using state-of-the-art LLMs.

At present, data collection has been implemented for Reddit and YouTube. The aspect identification and classification stage for data analysis has been completed as well. Concurrently, ongoing work is progressing on data pre-processing and aspect-based sentiment analysis, with anticipated completion in the near term.

The successful implementation of the system will not only provide organizations with a valuable tool to conduct social media research and sentiment analysis but also contribute to the broader understanding of LLMs and their capabilities.

VII. References

- [1] R. Shewale. (12 Sep. 2023). Social Media Users — Global Demographics (2023) [Online]. Available: <https://www.demandsage.com/social-media-users>. [Accessed: 30 Sep. 2023].
- [2] A. Hassan and H. Korashy. "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, pp. 1093-1113, Dec. 2014.
- [3] A. Nazir, Y. Rao, L. Wu and L. Sun. "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 845-863, Jun. 2022.
- [4] K. Kheiri and H. Karimi. "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," USU., Logan, UT, USA. Jul. 2023.
- [5] Prompt Engineering Guide. (n.d.). DAIR.AI. [Online]. Available: <https://www.promptingguide.ai/>. [Accessed: 20 Jan. 2024].
- [6] D. Carneros-Prado, L. Villa, E. Johnson, C. C. Dobrescu, A. Barragan and B. Garcia-Martinez, "Comparative Study of Large Language Models as Emotion and Sentiment Analysis Systems: A Case-Specific Analysis of GPT vs. IBM Watson," presented at the 15th International Conference on Ubiquitous Computing and Ambient Intelligence, Riviera Maya, Mexico, Nov. 2023
- [7] Reddit API Documentation. (n.d.). Reddit. [Online]. Available: <https://www.reddit.com/dev/api>. [Accessed: 20 Jan. 2024].
- [8] YouTube Dat API - Quota and Compliance Audits. (n.d.). Google. [Online]. Available: https://developers.google.com/youtube/v3/guides/quota_and_compliance_audits. [Accessed: 20 Jan. 2024].
- [9] Twitter API | Products | Twitter Developer Platform. (n.d.). X. [Online]. Available: <https://developer.twitter.com/en/products/twitter-api>. [Accessed: 20 Jan. 2024].

[10] Page Public Content Access. (n.d.). Meta. [Online]. Available:
<https://developers.facebook.com/docs/features-reference/page-public-content-access/>.
[Accessed: 20 Jan. 2024].