The University of Hong Kong
Department of Computer Science

COMP4801 Final Year Project
Detailed Project Plan

Automated Social Media Research and Sentiment Analysis using NLP Models

Group: FYP23047

Students:
Hung Jasper (3035787671)
Ko Berrick Eldridge (3035784710)

Supervisor:
Dr. Kong Lingpeng

1 October 2023

Table of Contents

1. Project Background

User feedback is a valuable source of information for any business. It helps organizations understand the needs and preferences of their users, allowing 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, they have several disadvantages. First, these methods rely on a limited sample size, causing some user segments to be underrepresented. Moreover, both designing and conducting them requires significant time and effort. Besides, these methods usually provide predefined questions and lack open-ended discussions, limiting the depth of feedback collected.

To address these drawbacks, organizations are also collecting user feedback from social media. It is estimated that around 60.5% of the global population engages in social media actively [1]. Social media is such a large platform that allows organizations to access a diverse pool of user-generated content, leading to a more comprehensive understanding of user feedback. Also, instead of designing and conducting surveys or interviews, organizations can simply tap into existing conversations and content shared on social media platforms, saving both time and effort. Furthermore, social media is all about expressing opinions and engaging in discussions freely. Organizations can hence have access to authentic and in-depth user feedback. In addition, doing social media research enables organizations to understand user sentiment not only on their own offerings but also on their competitors and emerging trends. This allows them to adapt strategies accordingly and stay ahead in the market.

Doing such research manually is an option. However, analyzing such a massive volume of posts or comments across multiple platforms is time-consuming and basically impractical. Hence, this project aims to implement an automated tool for social media research and sentiment analysis using Natural Language Processing (NLP) models. While similar tools such as Brandwatch and Brand24 are already available, this project also aims to explore the capabilities of state-of-the-art NLP models, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT). They have demonstrated remarkable performance in tasks like question answering and text generation

[2, 3]. This project seeks to demonstrate the potential of these models in the context of social media research and sentiment analysis.

2. Project Objective

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

1. Collect relevant data from social media efficiently based on user-specified keywords.
2. Clean and transform the collected data for further analysis using the BERT model.
3. Perform sentiment analysis on the preprocessed data using the GPT model, with an accuracy rate of at least 80%.
4. Generate reports with appropriate visualizations to present the results clearly using the GPT model.

3. Project Methodology

This project can be divided into four phases: data collection, data preprocessing, sentiment analysis, and result visualization.

3.1 Data Collection

This project will focus on collecting and analyzing textual data for simplicity. Based on the topic of interest specified by the user, the system will collect relevant posts and comments from various social media platforms. The targeted platforms include Facebook, Instagram, X (formerly Twitter), Reddit, and YouTube due to their popularity [1]. Data will be collected through the official Application Programming Interfaces (APIs) whenever available and affordable. If not, web scraping techniques will be employed to crawl and extract relevant data. In either case, the process will be conducted in adherence to each platform's terms of use and policies to ensure legal and ethical practices. The collected data will then be stored in a Comma-Separated Values (CSV) file for further processing.

3.2 Data Preprocessing

The collected data will undergo preprocessing to ensure its suitability for subsequent analysis. The system will first filter out data that is irrelevant for sentiment analysis, such as account tagging, hyperlinks, and spam content. This can be done through text classification using a pre-trained BERT model, or using an in-house trained one if it is found necessary. Then, noises such as special characters and stopwords will be removed, while abbreviations and emojis will be appropriately handled. Finally, techniques like tokenization and lemmatization will be employed to transform the data for further processing.

3.3 Sentiment Analysis

The data will then be processed to carry out sentiment analysis. Apart from classifying the data into positive, neutral, and negative sentiments, the system will also extract and summarize useful information to provide a more comprehensive analysis. While a trained BERT model can be used for the first part since it excels in data classification [3], its label assignment technique may not be suitable for the latter part since the system will be handling unknown data. Hence, the GPT model will be used for this matter instead. It is known that the GPT model performs well in data extraction and summarization [2]. This project will also explore its capabilities in performing sentiment classification.

3.4 Result Visualization

Finally, the system will present the analysis results through appropriate visualizations in the form of a report. The report will be generated using the GPT model and will include the following proposed fields:

1. A distribution showing proportions of positive, neutral, and negative sentiments on the topic of interest.
2. A graph showing how the sentiment changes over time.
3. Top positive and negative aspects mentioned in the posts or comments collected.
4. Frequently occurring words associated with positive and negative sentiments.

Additional fields may be included in the report based on the specific analysis results. The GPT model will be fine-tuned to generate reports customized with suitable fields.

4. Project Schedule and Milestones

Schedule	**Milestones**
October 2023	● **Detailed Project Plan** ● **Project Webpage** ● System Design ● Proof of Concept
November 2023	● Data Collection ● Data Preprocessing
January 2024	● Sentiment Analysis ● **First Presentation** ● **Preliminary Implementation** ● **Interim Report**
March 2024	● Result Visualization
April 2024	● System Optimization ● **Final Presentation** ● **Finalized Tested Implementation** ● **Final Report** ● **Project Exhibition**

5. References

[1] R. Shewale. (2023, Sep. 12). *Social Media Users — Global Demographics (2023)* [Online]. Available: https://www.demandsage.com/social-media-users/

[2] J. Li and M. Zhang, “A commentary of GPT-3 in MIT Technology Review 2021”, Soochow University, Suzhou 215006, China, 2021.

[3] M. Hoang, O. A. Bihorac, and J. Rouces. “Aspect-Based Sentiment Analysis using BERT”, Linköping University, Turku, Finland, 2019.