Meeting Minutes - 22 Feb 2022

THE LISTENING SQUAD

AI SOCIAL LISTENING TOOL

Meeting Minutes for 22 February 2022

Date:	📅 **22 Feb 2022**
Time:	🕔 5:00PM
Location:	MS Teams
Present:	@ Ow Ling Jia @ Tian Le Cheow @ Chen Jian Yu @ Joshua Wong @ Sarah Chin
Absent:	@ ONG JHIN YEE _

🥅 Goals

Updates for Prof Jisun on what’s happening with the project.

🗒️ Agenda

No.	**Agenda Item**	**Remarks**
	Updates on setting up scrapers on AWS a. Testing of the scrapers on the server b. Confirmed methods for scraping	@ Chen Jian Yu
2.	Machine Learning Methodology a. Trend Analysis b. Keyword Analysis c. Sentiment Analysis	@ Ow Ling Jia @ Joshua Wong
3.	Frontend Progress a. Showcase current progress for the webpage	@ Tian Le Cheow

🗣️ Discussion topics

Agenda	**Action By**
Updates on setting up scrapers on the server	

- Scraper - Backend
    - created a few servers (3 in total)
    - 1st one: Reddit and twitter data - Reddit can only get the most recent 7 days
        - Reddit Historical: Working but most recent 3 months cannot get comments; no error, just blank
        - Twitter Historical: Cannot get anything past 7 days
            - tried using snscrape; plans to try again some time this well
        - Instagram Historical: the data cannot be scraped and data cannot be retrieved, might not want to use IG anymore
        - Facebook: For now we will use Facepager, since it is the only one that can manually scrape data
        - YouTube: We are using Selenium, might have to try the code again (daily)
            - Historical: might have to give up as we run into the same issue
    - Prof Jisun: Twitter might be okay to forgo since Singaporeans don't use this social media a lot
        - Instagram: If not possible, don't need to pursue
        - First, use historical data to visualise how it will appear on the dashboard
        - YouTube: Selenium method is slow so might not be scalable
            - Prof suggested using this YouTube API: https://developers.google.com/youtube/v3
        - **Action 1**: Data collection process to be sent to Jisun
    - Issues with Facepager on different social media sites: YouTube have to use the playlist to get the videos but not every channel have multiple playlists
        - On top of that, not every video is included in the playlists
- AWS Server: Can close the server and transfer all on the local server
    - Will the local server crash, seeing that the AWS one crashed?
    - Prof: local server should be powerful enough to handle it; loop back to the server either this week or next week since most people wont be running it during this period
    - **Action 2**: Transfer all the data to the local server
    - Running MongoDB on the local server should be okay; extracted data is stored on MongoDB

Reddit, Twitter and Youtube daily scrapers are more or less fixed and there are no pressing issues

- **Action 3**: Let Prof Jisun know about the size of the historical data

Cost of the server: roughly US$12

No updates from Amazon's side

Jian Yu

Machine Learning Methodology

Graphs

- The team tested the trend analysis module to see how it would look like over time
- Engagement metrics will change according to the social media platform
    - e.g. Twitter - number of tweets, Facebook - number of posts

Trending Topics

- Previously discussed: we will use pre-defined topics to allow the users to choose
- Scraped articles from CNA and got all the article tags from there
- Tags are the combined and further refined by removing stopwords (NLTK, Gensim, spaCY)
- After the dictionary was created, the team added more topic words; top 500 words for, eg education
    - The dictionary had to be further refined as some words did not belong to the topic
- The code will compare the text with the predefined data and choose the topic that has the longest dictionary (best match) and define the data as that topic
    - will have to refine the topic dictionaries even more as testing the data showed complications in matching the text to the correct topic
    - **Action 1**: Refine the topic dictionaries
    - Prof Jisun: would it be better to build a simple NLP library and compare it against the text instead of using the keywords
        - noticed the overlap with the words over the topics
        - need labelled data from the social media itself ideally
        - build an ML model on the article and see how it performs, if its not working well, whatever words used on the article may not cross over to social media
        - the problem might also be too many words in the dictionary
    - Jian Yu: does it make sense to label the tread itself instead of the comment itself?
        - Comments may not have a lot of information but the post itself might be more definitive
        - Prof Jisun agrees; labelling the posts itself might be better and can use weights to see which topic the post belongs to

Keyword Analysis

The team will be using NER to get the keywords

Action 1, 2 & 3: Ling Jia

- But the results aren't very good; doesn't seem to get the results that we want
- The same word can also reflect different entity
- '#' represents tokenizing; LinkedIn might not be in the library so it breaks it down into different attributes to be analysed
  - **Action 2**: Find out what’s happening with the code
  - Prof Jisun: Library might have different ways of storing the word
    - Look at aggregated level and compare the methods and choose which one would be the best
    - Some models are used for research instead of real world applications and therefore may not be the best model
    - Split data by month or year and try the different models to see which one works best
    - **Action 3**: Try the proposed method and update Prof Jisun

Sentiment Analysis

Polarity and emotions were looked into

Emotion classifier - captures 7 different emotions (Hugging Face)

- There’s no neutral emotion, so the team manually inserted a threshold to get a neutral emotion (7.5)
- Calculated the average number of Facebook comments per post and the length of the comments
  - Limitations to long and short comments: Unable to tell whether the entirety of the comment is relevant to the topic at hand
- Tested the accuracy of the code
  - Results were not good
  - Data was trained on the twitter dataset but doubt that it is the training data’s fault
  - Could be the model’s fault
  - Prof Jisun: more first person based data and might be affecting the model
    - Hugging face zero shot classification: pretrained data where you give the label and the model will just find the best label; in classroom setting it works well so can try it out
    - Have backup plan: choose the ones that works the best at this moment
    - Problem with the hugging face: Works relatively slow
    - Prof Jisun shared some codes: https://github.com/anjisun221/css_codes/blob/main/ay21t1/Lab05_text_classification/Lab05_text_classification%20-%20Students.ipynb
  - Problem to address: the more labels there are the worse it works
  - Positive and negative might work better as compared to searching for emotion
  - Prof Jisun: Reasonable accuracy should be 70%
    - For emotions the accuracy might not be as high

Looking into Singlish words

- A lot of the data is using Singlish but a lot of models are not trained to classify Singlish words
- Still need to label the data even if the tokenizer is working

Jian Yu: Core features better or more features?

- Prof Jisun: focus on core first, but from professors' side, project might look a little too barren and basic
  - Don’t have to focus on everything on secondary features but can include some that makes us stand out
  - Create at least one special feature that makes us stand out to justify to the stakeholders
- Prof Jisun to update us abt linking the GPU server with the local server

Jian Yu: The app itself or the machine learning models?

- If the app can work on historical data it should be fine as well
- Even if we show until last November on the app itself it should be okay
- Visualisation would be very important to tie everything together

Frontend Progress

Build using Vue2 instead of Vue3 because some of the applications still do not work, but it can still be migrated over in the future **Search Feature** - In the search box, entering a query would then have a dropdown showing the related topics - Date period can also be selected or customised - Sentiment and platform filter has also been included Tooltip - shows the user how the component can be used - Intend of showing the sentiment related to the topics in the Trending Topics segment **Trending Analysis** - The graph sown will be affected by the main filters that the user can customise **Top Keywords** - Intend on inserting more data - Will insert the legend also - **Action 1**: Work on the attributes above Graphs are using 2 different packages - since some of the features are only supported in Vue3, therefore need to use 2 different packages	Tian Le

:note: Meeting Notes - Updates

Update	Risk Level	Mitigation Plan
School server has been restarted and can be used		Transfer all the data to the local server

Decisions

Administrative Matters

Date of Next Meeting:	**24 Feb 2022**
Time of Next Meeting:	5PM