

Weekly Mentor Meeting: Every Friday 5:00-6:00 PM

Our github repo: [acmucsd-projects/sp23-ai-team-1 \(github.com\)](https://github.com/acmucsd-projects/sp23-ai-team-1)

Team Meeting Time: Every Monday 5:00 - 6:00 PM

Attendees: Kevin, Vanessa, Yashil, Samuel, Chi, Vincent

Important Links:

MBTI personality type using tweets : [MBTI Personality Type Twitter Dataset | Kaggle](#)

Notebook: <https://www.kaggle.com/samuelllee1/acm-sp23-team-1-project-mbti>

Action Items

Data Cleaning

- Explore what percentage of data contain emojis- remove emojis (Yashil)
- Run the detect function to see what percentage of non-English in the data- look for distribution of languages (Vanessa)
- Strip the text of all foreign language (Chi)
- Check for gibberish data (Kevin)
- Summarize notebook EDA (All)
- Make one clean function

Preprocessing

- One person make a custom dataset
 - <https://www.youtube.com/watch?v=PXOzkKB5eH0&pp=ygUPcHI0b3JjaCBkYXRhc2V0>
- one person follow the tutorial
 - Preprocessing follow the steps in the link ([Sentiment Analysis with Pytorch — Part 1 — Data Preprocessing | by Gal Hever](#))

Summary of Meeting

- Discussion of data cleaning and pre-processing done

- Implementing the detect_lang function to find distribution of languages
- Pytorch (because tensorflow sucks) → torch text

Timeline

- Step 1: Data Cleaning (28th April, Friday of Week 4)
- Step 2: Preprocessing data (1st May, Monday of Week 5)
- Step 3: Model training and validation (8th May, Monday of Week 6)
- Step 4: Decide what kind application to do and start (15th May, Monday of Week 7)
- Step 5: Continuation of the step 4- (Weeks 8,9)
- Have an app ready with Streamlit (Week 9)
- Step 6: Polish it up + presentation + showcase (Week 10)