

**Individual Assignment: Social media analytics pipeline
with Python (Part 6. Reporting and Reflection)**

Lim Jun Yi Keith

Faculty of Arts and Social Sciences, Department of Communications and New Media

NMC5342: Introduction to Applied Social Media Analytics

Chen Zhicong

November 3, 2025

Word Count: 1,491

Introduction

I chose to scrape eight subreddits (threads between 2021-2025) relating to immigration in Singapore — totalling 1,073 posts (Lim, 2025a; Lim, 2025b).

Immigration is sociopolitically important because: 1. Singapore remains highly reliant on non-citizens for economic growth; foreigners account for around 40% of Singapore's population. (NPTD, 2025). 2. Greater immigration has led to more unhappiness, with some attributing job scarcity to foreigners 'stealing' jobs (Sin, 2021). 3. There is growing anxiety about the erosion of cultural identity (Mathews & Tay, 2021). From 2021 to 2025, the non-citizen population grew by 25% (1.96 million-2.45 million), while the citizen population grew by a meagre 4.6% (3.5 million-3.66 million).

This dataset is a microcosm of broader public sentiment toward immigration in Singapore. Analysing it provides early insights into emerging xenophobia, allowing for proactive interventions and communications from governmental agencies.

Methods

The workflow (Lim, 2025c), is as follows.

Python Setup & Refresher

I imported `os` and `warnings` to provide clarity due to runtime warnings. Then, I refreshed my Python knowledge (loop and function tasks).

Data Familiarisation

I inspected the dataset through `df.head()`, `df.info()`, and `df.describe()`. Afterwards, I identified key columns, then checked for null values and duplicates. Lastly, I generated simple descriptive summaries by finding the total comments, date range, and most active users.

Data Cleaning & Keyword Analysis

I removed nulls, duplicates, bot comments, and deleted comments. I imported `re` to remove URLs, extra spaces, and unnecessary punctuation. Later, I imported `nltk` and `spaCy` for text preprocessing. The dataset was tokenised and lemmatised, with stopwords removed — creating a new column, `cleaned_text`. For keyword analysis, I installed `collections.Counter` and `WordCloud` libraries to check the frequency of the most used terms while visualising a `WordCloud`. Next, I used `scikit-learn`'s `TfidfVectorizer` to compute TF-IDF scores while creating a TF-IDF `WordCloud`.

Sentiment and Topic Modeling

I imported SentimentIntensityAnalyzer from vaderSentiment.vaderSentiment and TextBlob. I used TextBlob for computing polarity and subjectivity while VADER was used for labelling sentiment. Then, I visualised the sentiment distribution through matplotlib. Afterwards, I used gensim's LDA for bigrams and stopwords filtering — with five topics surfaced. Lastly, pyLDAvis was used to visualise the topics.

Network Analysis

I used pandas, numpy, networkx, matplotlib, and collections. I matched the columns (IDs to authors), then constructed a weighted graph (nodes = users, edges = replies). Afterwards, a human-only subgraph was made to show users interacting with one another, rather than users interacting with the post itself.

Later, key network metrics (weighted degree, in-degree, out-degree, betweenness centrality, network density) were shown. Visualisations exhibited overall network and top nodes by degree centrality. Greedy modularity was also used to show clusters and frequent interactions. The graph was then exported in GraphML format.

I also visualised a WordCloud based on TF-IDF. Other terms like “delete” and “ns” (referring to compulsory National Service) emerged. They likely appeared disproportionately more in certain subreddits even though they were not the most frequently used terms overall.

Possibly, “delete” could be tied to certain Reddit slang such as Redditors posting “delete this” sarcastically upon seeing controversial posts. However, another explanation could be that some deleted comments were missed during the cleaning process and hence appeared in the WordCloud. Regarding “ns” — it is a contentious topic concerning Singaporean citizenship and identity. Some Singaporeans may feel that it is unfair that immigrants may ‘enjoy’ the same benefits as locals while not having to serve Singapore.

Sentiment Analysis

Polarity and subjectivity were based on TextBlob’s sentiment analysis. Polarity: -1 (negative), 0 (neutral), +1 (very positive); subjectivity: 0 (very objective) and 1 (very subjective). The mean reveals that the scraped texts were slightly positive but close to neutral (0.0526); they were not very subjective (0.348). The median showed that half the scraped texts were neutral or less (0.00); there was some subjectivity (0.369). Based on min/max, some posts were extremely negative (-1) but none were extremely positive (0.80).

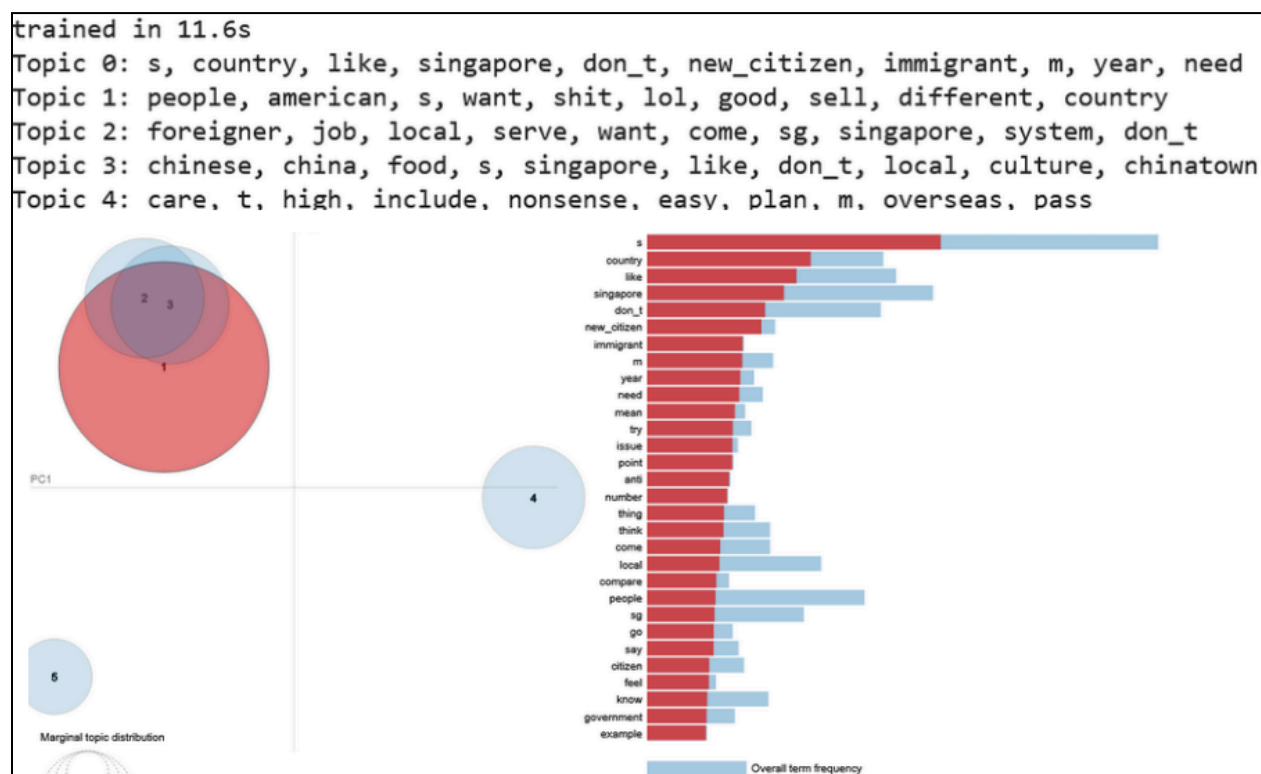
Sentiment analysis was then conducted. The majority of the text was positive or neutral (82.8%). A large minority was negative (17.2%) — expected, considering that immigration is provocative.

Topic Modeling

Table 2: 5 topics were identified with Genism, with pyLDAvis helping to visualise said topics.

Table 2

Topic Visualisation



Interpreting the topics, we can see that: 1. Topic 0: Singapore and citizenship ("new_citizen", "immigrant") 2. Topic 1: Attitudes (perhaps concerning informal posts: "Shit", "lol") 3. Topic 2: Jobs ("foreigner", "job") 4. Topic 3: Chinese immigrants ("Chinese", "China") 5. Topic 4: Attitudes toward migration ("nonsense", "easy", "overseas").

These topics informed the aforementioned sentiment distribution. Topic 1 likely contained informal and sarcastic language that TextBlob possibly misclassified as neutral. Topics 0 and 2 were likely polarised, involving sensitive issues like immigration, government policy, and foreigners. Topic 3 likely also showed mixed sentiments, reflecting tensions surrounding China-Chinese migrants and other ethnic groups like Indians. Topic 4 probably showed varied opinions due to its focus on migration and life in Singapore.

Network Analysis

The human-only interaction network had few edges per user (Table 3); users like “OkAdministration7880” and “leavingSg” were the most active, driving conversations while users like “jinhong91” attracted the most responses, likely because their comments attracted discussions. The user “leavingSg” was both active and controversy-sparking.

Table 3

Human-only Network


```

Human-only network:
Nodes: 507 Edges: 483

Top-10 human authors by weighted degree (total replies sent+received):
bukitbukit: 25
leavingSg: 25
OkAdministration7880: 20
sengleong88: 20
Historical_Drama_525: 19
KagenTheGOATcuck: 18
Jiakkantan: 16
Spiritual_Contest194: 14
hatboyslim: 12
Ehehehe090: 12

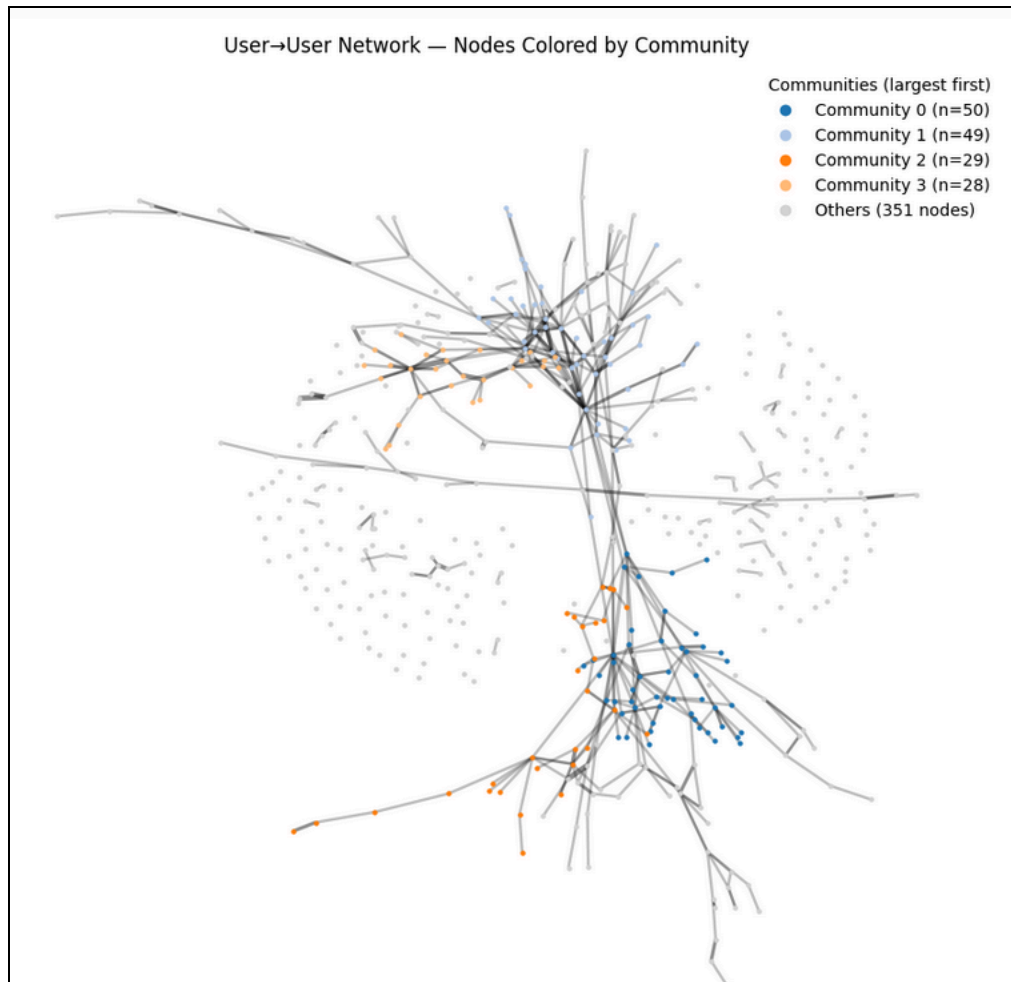
Top-10 most replied-to human authors (weighted in-degree):
leavingSg: 14
jinhong91: 9
Historical_Drama_525: 9
Fair-Second-642: 8
KagenTheGOATcuck: 8
minisoo: 8
boliaostuff: 8
nooobmaster23: 8
drwackadoodles: 7
hatboyslim: 7

```

Interestingly, the full Reddit user–user reply network (Table 4) showed fragmented discourse. Most discussions were siloed and alienated — happening within individual threads.

Table 4

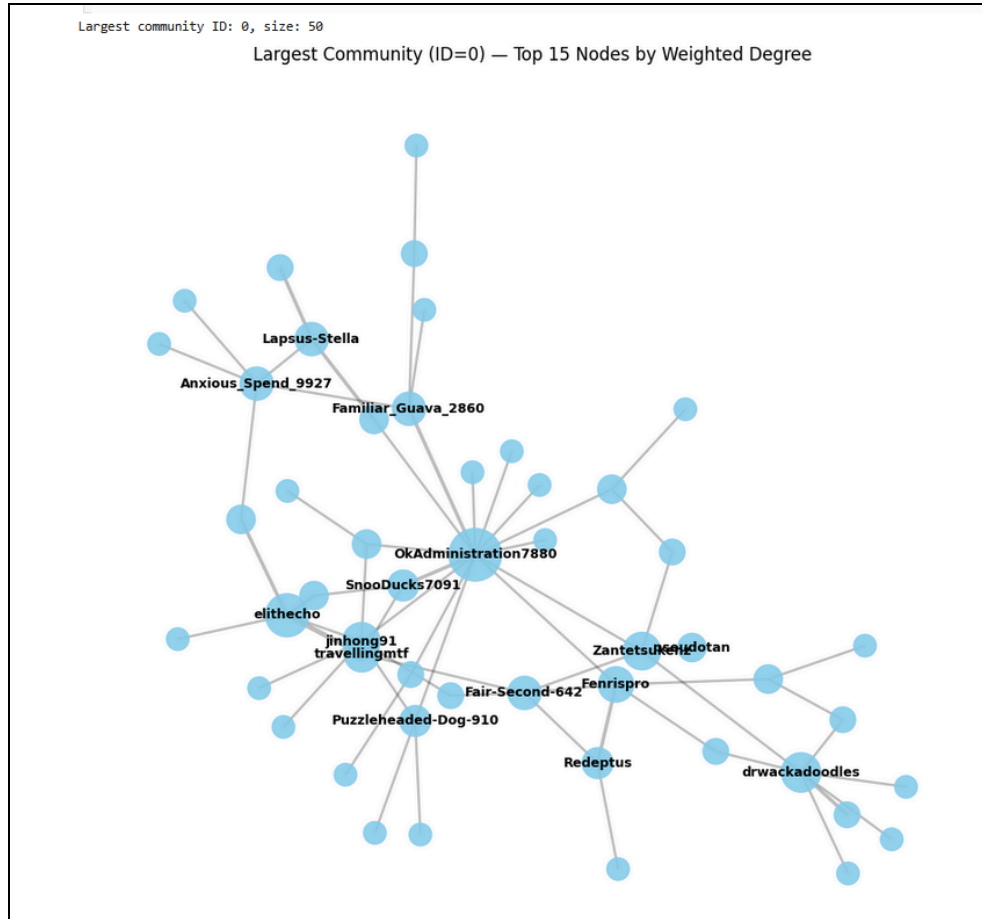
User-User Reply Network



Lastly, The cluster revolved around “OkAdministration7880” (Table 5), who was the most active engager. The structure showed a core-periphery pattern, where there were a few highly active nodes (users) who anchored discussions, surrounded by less-active users.

Table 5

Top 15 Nodes



Discussion & Recommendations

1. 17.2% of posts being negative suggests a very real resentment that a large minority of Singaporeans have against immigrants — tensions are expected to escalate as immigration rises in the future.

2. Since ‘China’ and ‘Chinese’ appeared more often than ‘India’ and ‘Indian’, this suggests that negativity is directed more towards Chinese immigrants as compared to Indians (or at least, more conversations focused on Chinese immigrants).

3. Majority of the text data was labelled neutral — most locals may thus be undecided when it comes to immigration.

4. Most discourse was by a few select users. The network has low density and many disconnected conversations — suggesting limited cohesion. Therefore, the presence of a prevailing anti-immigration force is unlikely. Though relieving, many of these vocal users can still instill anti-immigration attitudes due to predominant neutrality.

Ultimately, the Singapore government must act. Around the world, anti-immigration sentiments have taken root. Locals see immigrants as the ‘other’ and blame them for societal issues (Jones & Sengul, 2025; Keane, 2025; Yamaguchi, 2025). Hence, the Singapore government must balance between protecting locals’ interests while also ensuring safety for immigrants by:

1. Keeping its word: Minister Indranee Rajah stated that the population was likely to be significantly below 6.9 million by 2030 (Rajah, 2023; as cited in NPTD, 2024); if

immigration skyrockets past that by 2030, then the government will lose Singaporeans' trust, leading to discord.

2. Leveraging integrative policies (prioritise Chinese immigrants) : immigrants should be encouraged to mix with locals (especially non-Chinese speaking groups like Malays) through dialogue sessions and community activities.
3. Strengthening job security: fresh graduates are facing a shortage of jobs, leading to blame being attributed to immigrants. Hence, the government must create more jobs via traineeships, host more job fairs and career counselling, and ensure fairer hiring processes through stricter laws and monitoring.
4. Driving educational awareness: use online influencers to debunk 'job-stealing' myths about immigrants via two-step flow of communication, adopt episodic framing to showcase immigrants' lives for greater empathy, leverage religious leaders to dialogue with disgruntled locals, and stem early signs of xenophobia through Character and Citizenship Education classes in schools.
5. Continuing to monitor online sentiment: track sentiment shifts and whether long-term policies are successful.

Methodological Limitations & Extensions

The sample size was biased, considering that Reddit threads do not represent Singapore's entirety. Additionally, the dataset was in English, which misses out Malay, Mandarin, and Tamil sentiments. The data scraped may not be comprehensive enough with only 1,073 comments taken from 2021-2025. Furthermore, human error is likely to occur since I am the sole investigator, and tools like VADER and TextBlob are prone to overlooking irony and sarcasm.

Hence, future analysis should be a joint effort between multiple agencies, focusing on covering not just Reddit, but other social media platforms. Qualitative interviews should also be conducted to provide greater sentiment accuracy, alongside larger-scale surveys that utilise stratified sampling to ensure representativeness. Finally, transformer-based models like BERT can be used to better detect humour and context.

References

- Jones, C. & Sengul, K. (2025, September 02). *How Australia's anti-immigration rallies were amplified online by the global far right*. The Conversation.
<https://theconversation.com/how-australias-anti-immigration-rallies-were-amplified-online-by-the-global-far-right-264269>
- Keane, F. (2025, September 04). *Migrant crisis: How Europe went from Merkel's 'We can do it' to pulling up the drawbridge*. BBC News.
<https://www.bbc.com/news/articles/cn5e5q7w41eo>
- Lim, K. (2025a). NMC5342_A0234542L: Dataset_(Scraping).ipynb. [Computer software].
Github.
[https://github.com/Keith-LJY001/NMC5342_A0234542L/blob/main/Dataset_\(Scraping\).ipynb](https://github.com/Keith-LJY001/NMC5342_A0234542L/blob/main/Dataset_(Scraping).ipynb)
- Lim, K. (2025b). NMC5342_A0234542L: scraped_Reddit_Dataset.csv. [Computer software].
Github.
https://github.com/Keith-LJY001/NMC5342_A0234542L/blob/main/scraped_Reddit_Dataset.csv
- Lim, K. (2025c). NMC5342_A0234542L: NMC5342_A0234542L.ipynb. [Computer software].
Github.
https://github.com/Keith-LJY001/NMC5342_A0234542L/blob/main/NMC5342_A0234542L.ipynb
- National Population and Talent Division. (2025). *Overall Population*.
<https://www.population.gov.sg/our-population/population-trends/overall-population/>

National Population and Talent Division (2024). Does the Government have a population target, for example, 10 million?

<https://www.population.gov.sg/media-centre/articles/does-the-government-have-a-population-target-eg-10-million/>

Mathews, M. & Tay, M. (2021, April 02). Understanding the ire and intricacies of immigration in Singapore. *Institute of Policy Studies*.

<https://lkyspp.nus.edu.sg/ips/publications/details/understanding-the-ire-and-intricacies-of-immigration-in-singapore>

Sin, Y. (2021, May 11). *Racism and xenophobic behaviour will become normalised if S'pore is not careful, warns Shanmugam*. The Straits Times.

<https://www.straitstimes.com/singapore/racism-and-xenophobic-behaviour-will-become-normalised-if-spore-is-not-careful-warns>

Yamaguchi, M. (2025, October 02). *Anti-foreigner sentiments and politicians are on the rise as Japan faces a population crisis*. Associated Press.

<https://apnews.com/article/japan-xenophobia-immigration-sanseito-antiglobalism-trump-foreigners-358e5eb2b9d6bfe4814cac1b35557e1f>