Subtle Asian Data

Naitian Zhou
University of Michigan
naitian@umich.edu

David Zhao
University of Virginia
dz6hu@virginia.edu

Abstract

Recent *subtle asian* Facebook groups have gained incredible popularity among young adult Asian diaspora. We focus on one of these groups – *subtle asian dating* – and conduct a content analysis on its posts to examine what it can tell us about the members of Asian youth subculture, and how their identity and culture affect their view of and approach to dating and relationships.

1 Introduction

The *subtle asian dating* (SAD) Facebook group is a spin-off from *subtle asian traits*, a group started in September 2018 which has since amassed over a million members. The group consists primarily of young Asian first and second generation immigrants [4]. While *traits* is a veritable treasure trove of Gen Z Asian angst and frustrations, we chose to conduct our analysis on *dating* instead due to the relatively higher proportion of text posts and the relatively lower proportion of poorly executed Pikachu memes.

We use our dataset of posts to address some of the most interesting and pressing questions about the Asian immigrant youth and their relation to social media. We examine first the language used in SAD posts: the distribution of usage of "oppa", "baba", and "daddy", whether "bubble tea" is preferred nomenclature over "boba", and how emojis are used in SAD posts. We then take a closer look at the SAD demographic: the breakdown of occupations / educations and ethnicities.

It is our hope that, in addressing this large range of questions, we can paint a better overall picture of the typical SAD user, and the SAD subculture in general. Hopefully, this also opens the door to addressing deeper questions in the SAD community, such as "Where do all of these attractive people come from?" and "Why am I still single?"

2 Related Work

2.1 The Subtle Asian Franchise

To our knowledge, there has been no previous research on SAD posts, given that the group was created in November 2018. That said, there have been other applications of the SAD group. The website `subtleasianmatches.com`, launched in early December 2018, provides a matchmaking service based on responses to a questionnaire site users fill out. In correspondence with the subtleasianmatches team, it seems they currently use a naive similarity measurement for matches, but with the cleaner and more uniform data available from their questionnaire, there is plenty of potential research to be done on that front [2].

2.2 Social Media Content Analyses

There exist many content analyses of social media pages in literature. Oiarzabal conducted a survey of the Basque diaspora and its presence in social media. However, most of his research came from survey results gathered from the online communities (web pages, blogs, and social media pages) he was investigating [6]. A study on the Facebook Band Directors Group (BDG) investigated posting habits and popular topics within the community [1]. This type of content analysis through aggregating user contributions is more similar to the present study, but the BDG content analysis still involved a considerable amount of manual labeling, which we tried to avoid.

2.3 Emerging Asian Immigrant Adults

There is extensive literature on Asian American emerging adults (approximately college age – the same demographic as the average SAD user), the majority of which focuses on the relationship between emerging adult Asian Americans and their parents [3] [7]. These are generally small scale studies with individual interviews with fewer than 50 subjects.

3 Dataset

We compiled our corpus of SAD posts by crawling the Facebook group using Selenium. While Facebook has an API endpoint for getting posts in groups, it is only accessible by group admins, ruling it out as a valid method of data collection. The corpus was compiled mid-November 2018, and posts were acquired in reverse-chronological order. The data collection was done in two stages: first, on the group page, links to each individual post were acquired. Then, the actual post content and metadata was retrieved from those links. This served to provide more uniform and complete data.

For each post, we recorded the author, the timestamp, the post body, and the number of comments. We discarded any posts which lacked all of those pieces of information, as well as any deleted posts.

The data collection was cut short after Facebook temporarily blocked the account used for crawling. This resulted in a significantly smaller dataset than previously expected: of the 3876 links to posts, we were only able to get the post data for 786 of them, of which 749 were usable. Because the script used a personal account, and the author did not want to permanently miss out on subtle asian memes, we decided to stop crawling after the ban.

We also had to process the comment counts, because Facebook renders comment counts of over 1,000 differently. For example, a post with over 1,200 comments becomes 1.2K. This means we only have approximations of comment counts for very popular posts.

To preserve the anonymity of users, we have included only aggregate and non-identifying data in this paper. For more information or access to the complete dataset, please contact the authors.

4 Language of a SAD Post

We performed this study on the 749 valid posts. Most of the analysis was done on a concatenated corpus of all the posts, which was then lowercased and word- and emoji-tokenized.

4.1 Oppa vs Baba vs Daddy

Daddy is a slang term of endearment, used for (usually older or wealthier) men with a sexual connotation. It has risen in popularity recently – recently, the New York Times called 2018 the "year of the daddy" [10]. The term "baba" presumably arises from the Mandarin Chinese translation of "daddy". The Korean term "oppa" can have a flirtatious connotation, but it means older brother and can often be used platonically to describe an older male that you are close to or trust [11].

Table 1: Keyword Appearances in Concatenated Corpus

Oppa	**Baba**	**Daddy**
57	10	71

In tabulating the number of keyword appearances, we also factored in the native language translations (so for "baba", we searched for both "baba" and the Chinese equivalent). "Daddy" appears to remain the most popular, but in this Asian-influenced facebook group, "oppa" is a popular alternative.

Table 2: Common Qualifiers in Concatenated Corpus

Term	Qualifier (Count)
oppa	church (6), korean (5), an (5), boba (3), kpop (2)
baba	boba (5)
daddy	sugar (9), a (2), - (2), boba (2), and (2)

It is also useful to note that these terms are often accompanied by a qualifier. The most obvious example is when "daddy" actually refers to "sugar daddy". In Table 2, we take a look at common qualifiers for these terms. For these, we looked only at the English versions of each term.

An interesting feature is that "baba" appears exclusively in the context of "boba baba", in contrast to the other two keywords, which all have multiple qualifier prefixes. This implies that "baba" does not have any connotation attached to itself, unlike "oppa" or "daddy"; rather, the playful or flirtatious meaning comes from the alliterative "boba baba".

Another notable bigram is "church oppa". Many qualifiers for "oppa" are, unsurprisingly, related to Korean culture ("korean oppa", "kpop oppa", etc), and "church oppa" is no exception. Its prominence demonstrates the relatively large presence of Christianity in Korean culture. Over 70% of Korean US immigrants are involved in Korean ethnic churches [8].

4.2 Boba vs Bubble?

One of the most highly contested questions in *subtle asian* groups is the proper name for pearl milk tea, which is frequently referred to as *boba* or *bubble* tea. These beverages are popular among the SAD demographic: Asian American Pacific Islander youth and young adults [5]. "Boba" refers to the tapioca balls, and all three names refer to a beverage which contains these balls. This study is concerned only with the variant of tea which contains tapioca balls.

Table 3: Common Successive Words in Concatenated Corpus

Term	Word (Count)
bubble	tea (33), pop (2), milk (2), boys (1)
boba	**,** (7), date (6), **.** (5), and (5), **-** (4)

The word "bubble" occurs in the corpus 38 times, as opposed to 213 times for "boba". By this naive metric, it is obvious boba is a more popular term than bubble. Observing the contexts in which these words appear, it becomes clearer why this is (see Table 3). "Bubble" is almost used exclusively in the context of bubble tea, but "boba" is much more versatile. Not only is it frequently used by itself (note the prevalence of punctuation in successive words, showing it often does not act as a qualifier for another noun), it occurs also in "boba baba", "boba time", "boba bae", etc.

Interestingly, "boba tea" only occurs once. This can be again explained by the fact that generally, "boba" is used as a standalone noun to refer to pearl milk tea, as opposed to "bubble tea" or "bubble milk tea", in which bubble acts only as a descriptor for the noun (tea).

4.3 Emoji Use

A common theme among SAD posts is the use of emojis. In fact, over two thirds of the posts analyzed contained at least one emoji.

There are a total of 10461 emojis in the corpus. The 10 most popular emojis can be seen in Table 4.

We measured word count by splitting the text on non-alphanumeric characters and taking the length of the resulting list. On average, close to 5% of the "words" in a post were emojis. The post with the highest proportion of emojis consisted of almost 35% emojis. It contained 119 emojis (29 distinct emojis). The post with the most emojis contained 168 emojis (which made up 24% of words in the post). That post contained 68 distinct emojis, which is the highest number of distinct emojis.

Table 4: Ten Most Popular Emojis

Emoji	**# Occurrences**
🔥	392
💦	339
‼️	285
😉	261
😍	259
🚨	244
🍆	209
👅	202
😤	163
👀	154

In fact, much like lexical diversity, it is interesting to look at the emoji diversity in a post – that is, what is the ratio of distinct emojis to the total number of emojis? How creative are users with their posts? Initial observation of the data shows an average emoji diversity of .65 (STD=0.25), but this is skewed by the large number of posts with few emojis. This is made apparent in Table 6, which shows that, of the posts with a diversity of 1, they had on average fewer than three total emojis, and the majority had only one.

Table 5: Emoji Counts for Posts With Emoji Diversity = 1

mean	2.562500
std	2.697283
min	1.000000
25%	1.000000
50%	1.000000
75%	3.000000
max	15.000000

If we disregard all posts with emoji diversity of 1, we find the average emoji diversity to be 0.55 (STD=0.21). That means posts generally contain lots of repeated emojis. Few posts contain a nonneglible (>1) number of emojis that are all different (N=55). The 75th percentile is 3 distinct emojis.

5 Demographics of SAD

Information extraction techniques were used to analyze SAD posts. Pertinent details were extracted using a combination of regular expressions, parts of speech tagging, and phrase chunking. These were applied on a post level, as opposed to the corpus level used in the linguistic analysis.

5.1 Age Breakdown

We used pattern matching to extract ages from posts. We discarded ages lower than 13 and higher than 40, which yielded a total of 172 values. While the information extraction is not perfect, and there were false positives (when we incorrectly detected a number as an age), the data give us a rough view of what the age distribution is.

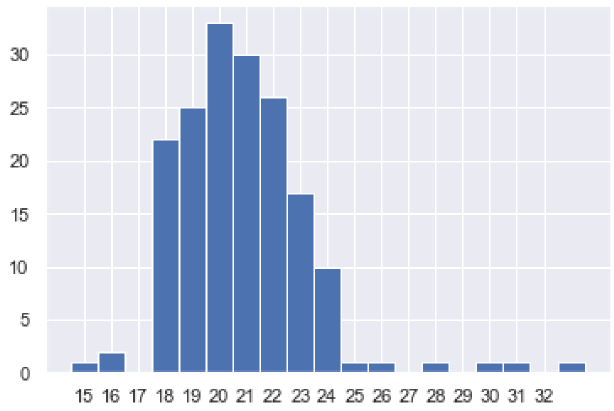

Figure 1: Age Breakdown

As visible in Figure 1, the distribution is skewed left (skew = 1.477), with a center of approximately 21 years old (mean = 20.8, median = 21). This is younger than Facebook's overall demographic, where the median is in the 35-44 age group [9]. This lends credence to the idea that SAD is representative of the Asian emerging adult demographic.

5.2 Occupation / Education

Another frequently mentioned characteristic in SAD posts is occupation. We examined five of the most popular categories of occupation (or future occupation – again, many users are college-age): medicine, computer science, engineering (of all types), finance, and law.

Table 6: Number of Post Mentioning the Occupation

Occupation	Count
medicine	45
finance	19
computer science	18
engineering	17
law	6

Medicine was the most popular category, by far. Note: the occupations of "veterinarian", "pharmacist", and "nurse", among others, were all considered to belong to the medicine category. Similarly, computer science and software engineering were both counted in the computer science category. To acquire a list of all simplifications made, please contact the authors.

One of the strategies used to extract occupation information was seeing the completions to "future . . . ", e.g. "future doctor". Interestingly, most of the occupations that arose from this context were medical or law-related. In looking at the ratios between the fields being mentioned in context and total, it seems medicine and law are more likely to be mentioned in this context.

Medicine was mentioned in this context 13 out of 45 total times, and law was mentioned 4/6 times, as opposed to engineering (1/17) and finance (1/19), which were the other two fields mentioned in the "future" context. This could be attributed to the longer time it takes to become a doctor or a lawyer, since medical school and law school are both prequisites to filling those respective positions.

5.3 Ethnic Breakdown

We used languages as a proxy for ethnic diversity among SAD users. Posts in the group are predominantly written in English, but many of them mention which Asian languages the subjects of the posts speak or are familiar with. Occasionally, non-Asian languages are also mentioned, but to a far lesser extent. Table 7 displays some of the most frequently mentioned languages.

We found these languages by examining words that appeared after “speaks” or “fluent”, which are indicators for language words. Common substitutes such as “chinese” or “canto” for Mandarin and Cantonese, respectively, were merged and interpreted as the same language.

Table 7: Number of Mentions of Popular Languages

Language	**Count**
mandarin	62
english	43
cantonese	34
korean	22
french	10

Mandarin is far and away the most mentioned language in posts, with English coming second, which makes sense, because most users are immigrants who live in English-speaking countries. This implies most SAD users are Chinese. In fact, if we consider all Chinese dialects together (including Cantonese, Shanghainese, etc), the number of Chinese language mentions rises to over 100.

There were also instances in which non-existant “languages” were mentioned to showcase another part of the subject’s personality. Some examples we found amusing include: “corgiean” or “puppynese”, “weaboo”, and “singlish”.

It is also notable that there were mentions of more specific dialects. For example, “fuzhounese” and “shanghainese” were both mentioned multiple times, with both of them being regional dialects of Mandarin. This is likely due to the fact that, though they are dialects of Mandarin, they are generally not mutually intelligible.

6 Conclusion

By performing a thorough content analysis on posts in the *subtle asian dating* Facebook group, we were able to determine patterns in the local lexicon and language usage, as well as paint a clearer picture of the SAD demographic. *subtle asian dating* users are primarily college age with a large Chinese plurality. This demographic, in the context of the *subtle asian* Facebook groups, has also adopted characteristic lexicological features.

While interesting, many aspects of this research were hindered by the scarcity in available data. Having faster, increased access to posts in the group may have yielded many more insights.

This is just a high level overview of some subtle, yet interesting, traits exhibited by the emerging adult Asian community online. In future works, more specific research questions can be answered more deeply. For example, what is the most important factor in a popular post? A hottie with a body? Height? A lucrative career?

7 Acknowledgements

We would like to thank Hella Chen, Jessica Zhou, Reanne Wong, and Prabhnoor Ahuja for administering the *subtle asian dating* group, as well as all the moderators for managing the memes.

8 Disclaimer

This is a joke paper with real work behind it. It does not reflect the views, research interests, or even the educational quality of the University of Michigan or the University of Virginia. While I can’t guarantee the level of rigor requisite for an actual publication, all data and analysis done for this project was still very much real, so draw whatever conclusions you want from it, or just enjoy it for what it is: a way to keep me busy on a 12 hour flight back to the States.

References

[1] Wesley D. Brewer and David A. Rickels. “A Content Analysis of Social Media Interactions in the Facebook Band Directors Group”. In: *Bulletin of the Council for Research in Music Education* 201 (2014), pp. 7–22. ISSN: 0010-9894. DOI: 10.5406/bulcouresmusedu.201.0007. JSTOR: 10.5406/bulcouresmusedu.201.0007.

[2] Brian Gu. *Facebook Communication with Brian Gu*. Dec. 6, 2018.

[3] Hyeyoung Kang et al. “Redeeming Immigrant Parents: How Korean American Emerging Adults Reinterpret Their Childhood”. In: *Journal of Adolescent Research* 25.3 (May 1, 2010), pp. 441–464. ISSN: 0743-5584. DOI: 10.1177/0743558410361371. URL: https://doi.org/10.1177/0743558410361371 (visited on 01/07/2019).

[4] Isabella Kwai. “How ‘Subtle Asian Traits’ Became a Global Hit”. In: *The New York Times. World* (). ISSN: 0362-4331. URL: https://www.nytimes.com/2018/12/11/world/australia/subtle-asian-traits-facebook-group.html (visited on 12/29/2018).

[5] Jae Eun Min, David B. Green, and Loan Kim. “Calories and Sugars in Boba Milk Tea: Implications for Obesity Risk in Asian Pacific Islanders”. In: *Food Science & Nutrition* 5.1 (2017), pp. 38–45. ISSN: 2048-7177. DOI: 10.1002/fsn3.362. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/fsn3.362 (visited on 01/08/2019).

[6] Pedro J. Oiarzabal. “Diaspora Basques and Online Social Networks: An Analysis of Users of Basque Institutional Diaspora Groups on Facebook”. In: *Journal of Ethnic and Migration Studies* 38.9 (Nov. 1, 2012), pp. 1469–1485. ISSN: 1369-183X. DOI: 10.1080/1369183X.2012.698216. URL: https://doi.org/10.1080/1369183X.2012.698216 (visited on 01/07/2019).

[7] Laura M. Padilla-Walker et al. “Bidirectional Relations Between Parenting and Prosocial Behavior for Asian and European-American Emerging Adults”. In: *Journal of Adult Development* 25.2 (June 1, 2018), pp. 107–120. ISSN: 1573-3440. DOI: 10.1007/s10804-017-9272-y. URL: https://doi.org/10.1007/s10804-017-9272-y (visited on 01/07/2019).

[8] Kyoung Ok Seol and Richard M. Lee. “The Effects of Religious Socialization and Religious Identity on Psychosocial Functioning in Korean American Adolescents from Immigrant Families”. In: *Journal of Family Psychology* 26.3 (June 2012), pp. 371–380. DOI: 10.1037/a0028199.

[9] We Are Social. *Distribution of Facebook Users in the United States as of January 2018, by Age Group and Gender.* URL: https://www.statista.com/statistics/187041/us-user-age-distribution-on-facebook/ (visited on 01/17/2019).

[10] Bonnie Wertheim. “Year of the Daddy”. In: *The New York Times. Style* (). ISSN: 0362-4331. URL: https://www.nytimes.com/2018/12/28/style/daddy-dads-of-2018.html (visited on 12/30/2018).

[11] *What Does Oppa Mean?* Oct. 22, 2012. URL: https://lovingkorean.com/2012/10/22/what-does-oppa-mean/ (visited on 12/30/2018).