Using NLP to Understand Self Descriptions of Frequent Drinker and Non-Frequent Drink On OKCupid

Phuong Thao Nguyen
Vanderbilt University
phuong.thao.h.nguyen@vanderbilt.edu

1. INTRODUCTION

Alcohol plays a significant role in social settings but poses substantial health and behavioral risks. The World Health Organization notes that harmful alcohol use causes around 3 million deaths yearly, accounting for 5.3% of global deaths [18]. Excessive drinking is linked to an increased risk of various cancers, including those of the oral cavity, pharynx, larynx, esophagus, and liver [3]. Additionally, alcohol consumption can lead to behavioral issues such as impulsivity, violence, and impaired decision-making, as it may inhibit the suppression of impulsive emotions, often resulting in impulsive decisions in social situations [17]. These factors point to the complex challenges associated with alcohol consumption.

Despite the known risks, alcohol consumption remains prevalent due to social dynamics in both personal and professional spheres. Research by Allen, Loeb, Kansky, and Davis examined the link between positive peer relationships and enhanced quality of life, emphasizing the role of social bonds in shaping behaviors, including drinking patterns [2]. These relationships often promote conformity to group norms, influencing individual drinking choices as a means to fit in and maintain social standing within a group [5]. Additionally, in professional settings, alcohol serves as a tool for relaxation and socialization, helping to build connections that can advance careers. It enhances emotional processing and reduces anxiety, functioning as a social lubricant that fosters more extroverted interactions [4]. Notably, research indicates that drinking is financially incentivized, with male and female drinkers earning significantly more than their non-drinking peers [14].

Alcohol's integration into social and workplace norms shapes identities, prompting research into its impacts on health, behavior, and society. Traditional research methods, while insightful, often face challenges like time-consuming data collection and limited sample sizes [12]. Recent advances in Natural Language Processing (NLP) have improved research efficiency by automating data extraction and analysis, saving over 120 hours and reducing costs by approximately $1,500 [1]. NLP effectively analyzes self-descriptions on social platforms, aiding healthcare providers in understanding patient needs and implementing preventative measures for alcohol-related issues.

NLP has also been used to identify patient safety concerns by analyzing discussions in healthcare forums. Using topic modeling, researchers have identified key concerns among recovering alcoholics, such as the urge to drink during the day, anxiety, and cravings for sugar among 2000 identified themes[10]. NLP enables longitudinal studies, revealing correlations between alcohol-related discussions on social media and excessive drinking patterns across various regions [10]. Researchers have extracted topics that distinguish drinkers from non-drinkers on platforms like Facebook, focusing on language use and social interactions[12].

Limited research exists on how alcohol influences social behaviors in dating contexts, with only preliminary findings suggesting that heavy users of dating apps are less likely to regularly consume alcohol [6]. This area remains underexplored, especially regarding the application of NLP techniques to analyze behaviors on dating platforms.

While Scannell [17] highlights the risks of online dating platforms, particularly among college students, such as sexual assault, and Huang et al. [9] underscores the susceptibility of users, especially Taiwanese adolescents, to online victimization, little attention has been given to exploring alcohol-related self-descriptions on these platforms. These findings suggest the importance of investigating how such self-descriptions correlate with vulnerability and risk behaviors.

This study aims to bridge this gap by identifying themes in user descriptions on online dating platforms that differentiate between drinkers and non-drinkers, shedding light on broader behavioral impacts and potential risks associated with online dating experiences. By leveraging NLP techniques, this research contributes to the development of secure online environments, informed by evidence from victimization and alcohol use, ultimately enhancing safety in the dating world.

2. PURPOSE STATEMENT AND RESEARCH QUESTION

This research employs natural language processing (NLP) and machine learning to investigate how alcohol consumption influences user profiles on the dating app OkCupid. This platform matches users based on their self-descriptions and answers to essay prompts, along with data like age and recreational drug use. The study aims to analyze these descriptions to identify how users portray themselves and their preferences, particularly in relation to alcohol consumption. By applying NLP, it is anticipated that patterns linking alcohol use with other profile features will be uncovered, enhancing the understanding of user dynamics and preferences within the context of online dating.

Primary Research Question: Is it possible to differentiate between frequent vs. infrequent drinkers based on their OkCupid profiles? What terms differentiate these groups?

Secondary Research: How do topics, extracted from OkCupid profiles using the topic modeling method MNF, vary across the frequent and infrequent drinker groups?

Primary Hypothesis: Based on the self-descriptions of users on OKCupid, it is possible to differentiate between them. Users who frequently drink are likely to have descriptions that reflect social outings and activities, just as those who drink less or not at all will exhibit similar traits in their profiles. Whereas infrequent drinkers or non-drinkers might emphasize interests in lifestyle or hobbies. It is expected that the content of these descriptions will gravitate towards themes reflective of their respective drinking habits.

Secondary Hypothesis: In addition, it is hypothesized that frequent drinkers will express topics related more to social events, such as large gatherings, to express their extroverted personality. While infrequent drinkers will express topics in regards to more one-on-one interactions or their interpersonal relationships.

3. METHOD

In this investigation, the analysis will be divided into two steps. Initially, logistic regression with TF-IDF as the features will be employed to ascertain whether it is possible to distinguish between frequent and infrequent alcohol users based solely on text descriptions. Should this approach prove successful, it will be acknowledged that text-based methods can effectively differentiate between the two groups. Subsequently, the analysis will delve into the topics within each essay or prompt and utilize Non-Negative Matrix Factorization (NMF) techniques to extract important words, thereby facilitating an investigation into the themes associated with the retrieved topics.

3.1 Data

3.1.1 BACKGROUND & ORIGINAL SOURCE

Data was obtained from Kaggle but originally scraped from OkCupid, with usernames and dates removed for de-identification purposes[15]. OkCupid users are categorized by sex and sexual orientation, with 20,533 straight females, 1,996 bisexual females, 1,588 gay females, 31,073 straight males, 3,985 gay males, and 771 bisexual males[15]. Other information in the dataset includes typical user attributes such as age and ethnicity, as well as lifestyle variables like diet, drinking habits, and smoking habits. The dataset comprises ten essays per user, each capturing different facets of their personal lives, such as self-perception, activities, talents, social habits, cultural preferences, essentials in life, contemplative thoughts, typical Friday night activities, private admissions, and criteria for messaging potential matches[15].

3.1.2 DATA CLEANING

To train the model, the first step is to engineer a single column that consolidates all possible essays. However, it is crucial to examine whether there is a disproportionate number of essays answered by users, as not all users respond to them. Each essay is counted to gauge the response rates. Despite variations in response rates across these essays—ranging from 54,458 responses for the most completed essay to 40,721 for the least completed—the differences are considered minor enough to warrant merging all essays into a single column. Subsequently, all users who did not answer all the questions are removed, streamlining the analysis for modeling purposes.

This consolidated column will be utilized solely for modeling the differentiation between the two groups. However, during topic modeling, the most relevant essays will be selected and the technique applied to investigate potential topics. To prepare the text for analysis, HTML tags are initially removed using BeautifulSoup, followed by the employment of spaCy to lemmatize the words, remove stopwords, and filter out non-alphabetic characters. This process yields a cleaned version of the text, ready for further analysis or processing. This cleaning process is repeated for both the single column containing combined essays and for all individual essays. Subsequently, the drinking label for each user is examined.

3.1.3 CREATE FREQUENT vs INFREQUENT GROUP

When examining drinking habits, many users (20,665) identified themselves as 'social' drinkers, a category that does not specify the frequency of their drinking. To clarify the analysis, this group was excluded, and the focus was placed on users who drink more regularly and those who drink less often. These groups were labeled as 'frequent' drinkers (who drink 'often' or 'very often') and 'infrequent' drinkers (who drink 'rarely' or 'not at all'). Excluding 'social' drinkers and categorizing the remaining users allowed for a separate analysis of their essay responses, providing a clearer understanding of how lifestyle choices correlate with personal characteristics in the dataset.

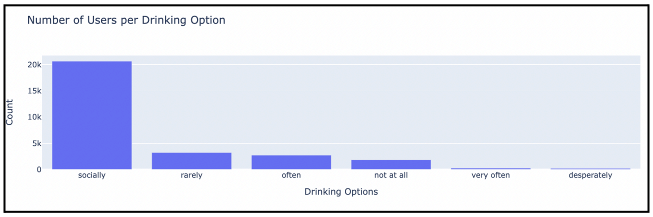


Figure 1 Count of Each Drinking Responses

3.1.4 *DOWNSAMPLE*

Segmenting the data into two groups based on drinking habits focuses the analysis but also introduces limitations. Initially, there were 3,153 users who drank frequently and 5,094 who drank infrequently. To enhance the fairness and accuracy of the analysis, especially for predictive modeling such as logistic regression, the size of the infrequent drinkers group was reduced. This balancing improves the results from machine learning algorithms, but it also results in a smaller dataset, potentially omitting significant variations among users. Ultimately, 3,153 users were set for the infrequent drinking group and 3,153 for the frequent drinking group along with their essay responses.

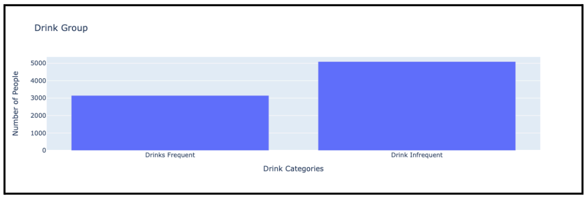


Figure 2 Infrequent vs Frequent Drinker Before Balance

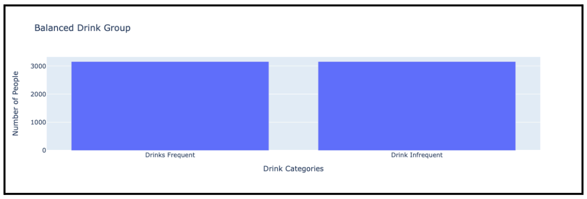


Figure 2 Infrequent vs Frequent Drinker After Balance

3.2 Logistic Regression Model

3.2.1 TF-IDF AS Features

To construct a model that classifies users as frequent or infrequent drinkers, the analysis combined essays and utilized TF-IDF (Term Frequency-Inverse Document Frequency) to transform text data into an appropriate format. TF-IDF converts essays into weighted numerical vectors, thereby enhancing the logistic regression model's capability to distinguish between the two groups based on their drinking habits. By weighting words according to their frequency and inversely by their commonness across all documents, TF-IDF emphasizes unique terms associated with either frequent or infrequent drinking behaviors. This method ensures more accurate classification by highlighting words that strongly differentiate between user categories. Adjusting TF-IDF parameters, such as setting the minimum document frequency (min_df) to 0.01 and maximum document frequency (max_df) to 0.95, aids in filtering out terms that are too rare or too common, resulting in a matrix of relevant words for training the logistic regression model.

3.2.2 Logistic Regression

For this logistic regression model, the target variable is classified as either 'frequent' or 'infrequent', based on the data from the drinks column. The input features are represented by a TF-IDF matrix. The dataset is partitioned into an 80/20 split for training and testing, respectively, with a random state of 42 to ensure consistency. The iteration limit for the model is set at 1000 to allow adequate adjustments and convergence. Model performance is assessed using various metrics including accuracy, which measures overall correctness, along with precision and recall for class identification effectiveness. Additionally, a confusion matrix is employed to provide a visual representation of the classification outcomes.

3.2.3 Determining Distinct Words Separating the Two Groups

Using NLP to Understand Self Descriptions of Frequent Drinker and Non-Frequent Drink On OKCupid

Phuong Thao Nguyen
Vanderbilt University
phuong.thao.h.nguyen@vanderbilt.edu

To discern how specific words affect predictions, the analysis involved examining coefficients from logistic regression, which indicate the strength and direction of the relationship between each word and the likelihood of classification as a frequent or infrequent drinker. These coefficients and their corresponding words were organized into a dataframe and sorted the top 1000 words with the lowest coefficients, which are strongly associated with infrequent drinkers, and labeled them as negative indicators. Conversely, the top 1000 words with the highest coefficients, associated with frequent drinkers, were sorted and labeled as positive indicators. By combining these sets into a single dataframe, a direct comparison was facilitated, unveiling the most polarizing terms between frequent and infrequent drinkers and illustrating how specific words in dating profiles can signal different lifestyle choices.

3.3 Topic Modeling

The analysis was then advanced by applying topic modeling to individual essays provided by users, rather than aggregating text as done in logistic regression. Specifically, focus was placed on seven of the ten essays from the dataset, chosen because they offer significant insights into personal behaviors and characteristics. This approach allows for a more nuanced understanding of how users express themselves in different contexts of their online profiles, enhancing the ability to draw distinct correlations between textual expressions and user behaviors.

essay0: My self summary

essay2: I'm really good at

essay3: The first thing people usually notice about me

essay5: The six things I could never do without

essay6: I spend a lot of time thinking about

essay7: On a typical Friday night I am

essay8: The most private thing I am willing to admit [15]

This selection excludes essays mainly related to occupational and leisure activities, allowing us to concentrate on those that provide deeper insights into lifestyle choices and behavioral patterns. Our targeted approach, utilizing Matrix Factorization (MNF) and the Silhouette score, enhances our understanding of how these behaviors correlate with broader personal traits.

3.3.1 Silhouette Score (NMF)

After obtaining the essays, the analysis aimed to determine the optimal number of topics for each one. Traditionally, the elbow test is used for this purpose by plotting variation explained against the number of clusters. When adapting this method for Non-negative Matrix Factorization (NMF), the focus shifted to reconstruction error, with tests conducted for up to 15 topics. However, the absence of a clear 'elbow' in the reconstruction error graph indicated that increasing the number of topics could lead to a generalized model.

Consequently, the Silhouette Score was utilized as a more refined measure of similarity within clusters. This approach acknowledged that each essay might require a different number of topics, reflecting its unique content, thereby allowing for more precise modeling of the thematic structure within the data.

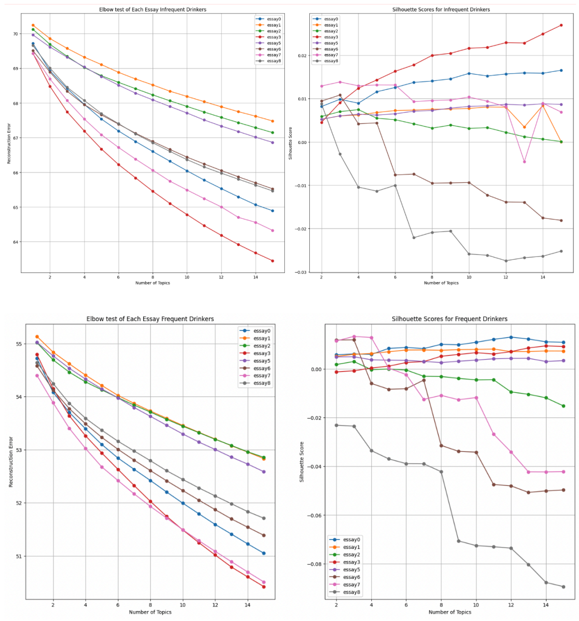


Figure 4 Elbow Test vs Silhouette Score (Frequent vs Infrequent drinkers)

For example, "My self summary" had 12 topics for both groups, indicating a deeper exploration of self-identity. Essays with lighter themes needed fewer topics, such as 3 for "I'm really good at," suggesting a narrower range of themes."The first thing people usually notice about me" required 14 topics for infrequent drinkers and 13 for frequent drinkers, showing diverse perceptions. Finally, "The most private thing I am willing to admit" was

Phuong Thao Nguyen
Vanderbilt University
phuong.thao.h.nguyen@vanderbilt.edu

analyzed with 2 topics for both groups, indicating a unified approach to discussing private matters.

After determining the optimal number of topics using the Silhouette Score, Non-negative Matrix Factorization (NMF) will be employed using the number of optimum topics outputted by silhouette scores and the user's response for each of the essays. This analysis will generate the top words for each topic, aiding in the understanding of prevalent themes within the data.

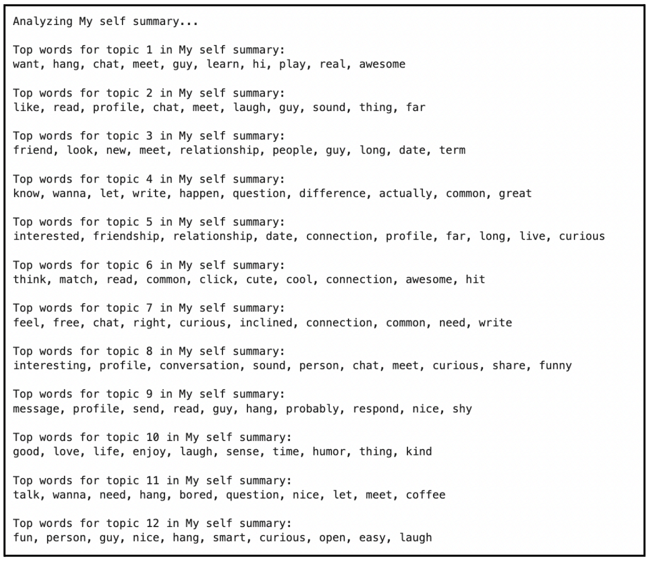
3.3.2 Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) is an unsupervised learning technique included in the Sklearn package, designed to break down high-dimensional, non-negative data matrices—like TF-IDF weighted document-term matrices from text data—into more manageable, lower-dimensional forms. This method is particularly useful in text mining and topic modeling because it identifies the key words defining each topic.

In this process start with the cleaned text responses from frequent and infrequent users to selected essays, using the optimal number of topics as determined by silhouette scores for each essay. The NMF algorithm processes these responses by first converting them into TF-IDF vectors. This conversion adjusts the weight of each word based on its frequency and its ubiquity across all documents, ensuring that the importance of each word is scaled accurately within the dataset.

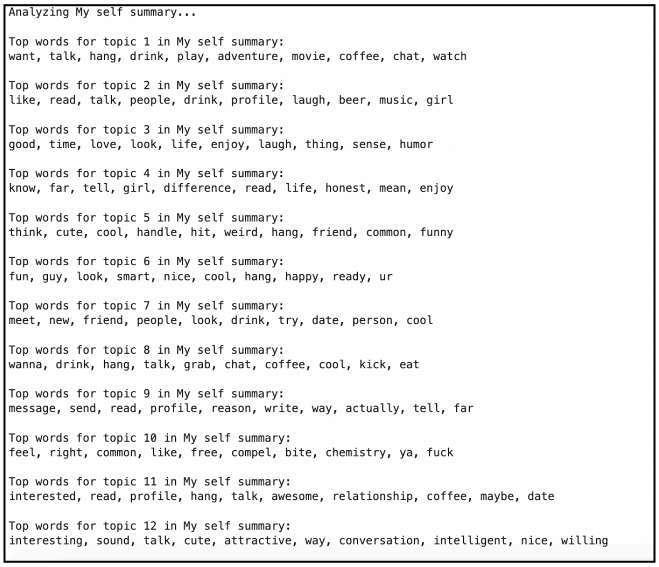
Following this, NMF decomposes the TF-IDF matrix into two matrices: the document-topic matrix (W) and the topic-term matrix (H). The document-topic matrix shows how each document correlates with various topics, while the topic-term matrix highlights the most significant terms for each topic. This separation facilitates a clear visualization and understanding of how documents relate to their respective topics.

This method was applied consistently to analyze selected essays for both the frequent and infrequent groups, specifically targeting essays such as "My self summary," "I'm really good at," "The first thing people usually notice about me," "The six things I could never do without," "I spend a lot of time thinking about," "On a typical Friday night I am," and "The most private thing I am willing to admit." This systematic approach allowed for a thorough analysis of each essay, producing a comprehensive set of key terms for each topic within the essays. This detailed examination helps deepen our understanding of the themes and subjects discussed by users, effectively distinguishing between the narratives of frequent and infrequent drinkers.



```
Analyzing My self summary...

Top words for topic 1 in My self summary:
want, hang, chat, meet, guy, learn, hi, play, real, awesome

Top words for topic 2 in My self summary:
like, read, profile, chat, meet, laugh, guy, sound, thing, far

Top words for topic 3 in My self summary:
friend, look, new, meet, relationship, people, guy, long, date, term

Top words for topic 4 in My self summary:
know, wanna, let, write, happen, question, difference, actually, common, great

Top words for topic 5 in My self summary:
interested, friendship, relationship, date, connection, profile, far, long, live, curious

Top words for topic 6 in My self summary:
think, match, read, common, click, cute, cool, connection, awesome, hit

Top words for topic 7 in My self summary:
feel, free, chat, right, curious, inclined, connection, common, need, write

Top words for topic 8 in My self summary:
interesting, profile, conversation, sound, person, chat, meet, curious, share, funny

Top words for topic 9 in My self summary:
message, profile, send, read, guy, hang, probably, respond, nice, shy

Top words for topic 10 in My self summary:
good, love, life, enjoy, laugh, sense, time, humor, thing, kind

Top words for topic 11 in My self summary:
talk, wanna, need, hang, bored, question, nice, let, meet, coffee

Top words for topic 12 in My self summary:
fun, person, guy, nice, hang, smart, curious, open, easy, laugh
```

Figure 6 Outputs from MNF for Inrequent Drinkers for Essay 1 (My self Summary)



```
Analyzing My self summary...

Top words for topic 1 in My self summary:
want, talk, hang, drink, play, adventure, movie, coffee, chat, watch

Top words for topic 2 in My self summary:
like, read, talk, people, drink, profile, laugh, beer, music, girl

Top words for topic 3 in My self summary:
good, time, love, look, life, enjoy, laugh, thing, sense, humor

Top words for topic 4 in My self summary:
know, far, tell, girl, difference, read, life, honest, mean, enjoy

Top words for topic 5 in My self summary:
think, cute, cool, handle, hit, weird, hang, friend, common, funny

Top words for topic 6 in My self summary:
fun, guy, look, smart, nice, cool, hang, happy, ready, ur

Top words for topic 7 in My self summary:
meet, new, friend, people, look, drink, try, date, person, cool

Top words for topic 8 in My self summary:
wanna, drink, hang, talk, grab, chat, coffee, cool, kick, eat

Top words for topic 9 in My self summary:
message, send, read, profile, reason, write, way, actually, tell, far

Top words for topic 10 in My self summary:
feel, right, common, like, free, compel, bite, chemistry, ya, fuck

Top words for topic 11 in My self summary:
interested, read, profile, hang, talk, awesome, relationship, coffee, maybe, date

Top words for topic 12 in My self summary:
interesting, sound, talk, cute, attractive, way, conversation, intelligent, nice, willing
```

Figure 5 Outputs from MNF for Frequent Drinkers for Essay 1 (My self Summary)

Using NLP to Understand Self Descriptions of Frequent Drinker and Non-Frequent Drink On OKCupid

Phuong Thao Nguyen
Vanderbilt University
phuong.thao.h.nguyen@vanderbilt.edu

4. RESULTS

4.1 Logistic Regression and TF-IDF

The classification report presents metrics for the model's performance in classifying users as frequent or infrequent drinkers. Both classes achieve identical precision, recall, and F1-score of 0.76, indicating balanced performance. The overall accuracy of the model is also 0.76, with macro and weighted averages reflecting consistent results across both classes, suggesting a well-performing model with equal effectiveness in predicting both frequent and infrequent drinkers.

Figure 8 Results from Training

Classification Report:	**precision**	**recall**	**f1-score**	**support**
frequently	**0.76**	**0.76**	**0.76**	**631**
infrequently	**0.76**	**0.76**	**0.76**	**631**
accuracy	**0.76**	**0.76**	**0.76**	**1262**
macro avg	**0.76**	**0.76**	**0.76**	**1262**
weighted avg	**0.76**	**0.76**	**0.76**	**1262**

The coefficients from the logistic regression model offer insights into vocabulary differences between frequent and infrequent drinkers in dating profiles. Words with negative coefficients, such as "beer," "wine," "drink," "bar," and "whiskey," suggest that mentions of these alcoholic beverages and related places are strong indicators of infrequent drinking habits. Conversely, positive coefficients associated with words like "healthy," "computer," "learn," "yoga," and "drug" indicate their prevalence in profiles of frequent drinkers.

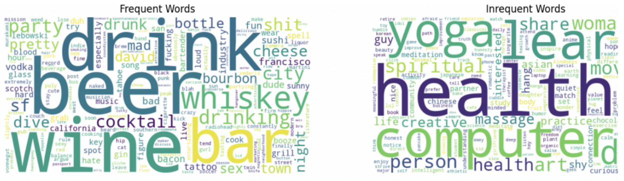


Figure 7 Word Cloud of Words that Measures distinct Words differentiating the Groups

4.2 MNF Topic Modeling

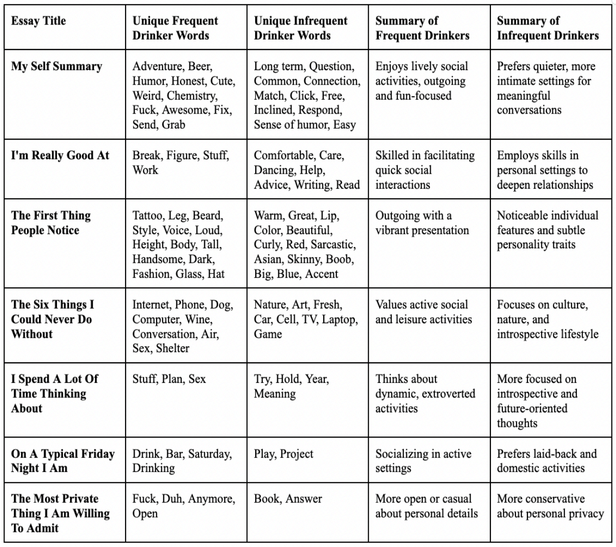
4.2.1 TOPICS

The summary of topics generated for each essay reveals interesting insights. The essays ``My Self Summary" and "The First Thing People Notice About Me" garnered the most topics, with 12 for both frequent and infrequent drinkers, and 13 for frequent and 14 for infrequent drinkers, respectively. These essays focus on users' personality traits and physical characteristics, explaining the variation in topics due to users' distinct characteristics.

Across all essays, there's a clear indication of more extroverted descriptions for frequent drinkers and a focus on relationship-building for infrequent drinkers, particularly evident in essays like "My Self Summary," "I'm Really Good At," "The Six Things I Could Never Do Without," and "On A Typical Friday Night I Am."

Interestingly, in the essay "What I Often Think About," infrequent drinkers exhibit a more future-oriented outlook, focusing on life, career building, or long-term planning compared to frequent drinkers. In "The Most Private Thing I Am Willing To Admit," frequent drinkers tend to be more direct in their topics, while infrequent drinkers are more reserved.

Figure 9. Summary of Topics and Unique Top Words



Essay Title	**Unique Frequent Drinker Words**	**Unique Infrequent Drinker Words**	**Summary of Frequent Drinkers**	**Summary of Infrequent Drinkers**
My Self Summary	Adventure, Beer, Humor, Honest, Cute, Weird, Chemistry, Fuck, Awesome, Fix, Send, Grab	Long term, Question, Common, Connection, Match, Click, Free, Inclined, Respond, Sense of humor, Easy	Enjoys lively social activities, outgoing and fun-focused	Prefers quieter, more intimate settings for meaningful conversations
I'm Really Good At	Break, Figure, Stuff, Work	Comfortable, Care, Dancing, Help, Advice, Writing, Read	Skilled in facilitating quick social interactions	Employs skills in personal settings to deepen relationships
The First Thing People Notice	Tattoo, Leg, Beard, Style, Voice, Loud, Height, Body, Tall, Handsome, Dark, Fashion, Glass, Hat	Warm, Great, Lip, Color, Beautiful, Curly, Red, Sarcastic, Asian, Skinny, Boob, Big, Blue, Accent	Outgoing with a vibrant presentation	Noticeable individual features and subtle personality traits
The Six Things I Could Never Do Without	Internet, Phone, Dog, Computer, Wine, Conversation, Air, Sex, Shelter	Nature, Art, Fresh, Car, Cell, TV, Laptop, Game	Values active social and leisure activities	Focuses on culture, nature, and introspective lifestyle
I Spend A Lot Of Time Thinking About	Stuff, Plan, Sex	Try, Hold, Year, Meaning	Thinks about dynamic, extroverted activities	More focused on introspective and future-oriented thoughts
On A Typical Friday Night I Am	Drink, Bar, Saturday, Drinking	Play, Project	Socializing in active settings	Prefers laid-back and domestic activities
The Most Private Thing I Am Willing To Admit	Fuck, Duh, Anymore, Open	Book, Answer	More open or casual about personal details	More conservative about personal privacy

Using NLP to Understand Self Descriptions of Frequent Drinker and Non-Frequent Drink On OKCupid

Phuong Thao Nguyen
Vanderbilt University
phuong.thao.h.nguyen@vanderbilt.edu

5. CONCLUSION AND FUTURE WORK

This study utilizes natural language processing and machine learning to investigate the influence of alcohol consumption on behaviors and societal dynamics, paving the way for future research using NLP to explore alcohol-related topics. The analysis is conducted on OkCupid user profiles to examine how descriptions of alcohol consumption are integrated into self-descriptions.

The primary research objective is to identify linguistic distinctions between frequent and infrequent drinkers based on the content of their profiles. The findings reveal that frequent drinkers commonly use language associated with social activities and extroversion. In contrast, profiles of infrequent drinkers are more likely to include introspective and hobby-focused language.

In secondary research, the study explores the different topics that emerge from the profiles of these two groups. Profiles of frequent drinkers often feature themes of large social gatherings, whereas infrequent drinkers tend to discuss more intimate one-on-one interactions and interpersonal relationships. This suggests a more introverted disposition among infrequent drinkers. These observations are consistent with existing literature on the social behaviors associated with alcohol consumption, particularly within the context of dating platforms.

5.1 LIMITATION

Limitations include excluding the socially drinking group and downsampling infrequent drinkers, reduced sample size and may limit generalizability. This dramatically cut down our sample size. Additionally, all participants reside in California, limiting geographical diversity and cultural representation. Future research should explore demographic variables like age and include a broader demographic scope, incorporating a cross-group comparison to identify universal versus unique thematic elements across different drinking habits. Investigating language patterns related to other substances or lifestyle choices and focusing on regional differences within California could enhance understanding and applicability across diverse populations.

Our technique for silhouette score does have limitations as the best scores are closer to 0 than 1, hinting that some texts are still hard to differentiate. In addition, statistical testing for NMFis currently not applicable at the moment.

5.2 Next Step and Future Research

Looking forward, it is imperative to consider other demographic variables such as age, which might influence language style and the way individuals present themselves online. This approach can help refine our understanding of how different age groups navigate the landscape of online dating in the context of alcohol consumption.

Future research should aim to include a broader demographic scope by incorporating a cross-group comparison that includes social drinkers, to identify universal versus unique thematic elements across different drinking habits. Investigating language patterns related to different substances or lifestyle choices could offer further insights into how various facets of personality or interests influence self-description. Additionally, a focused analysis on regional differences within California could reveal whether geographic location correlates with semantic similarity or shared interests, shedding light on how cultural factors influence online self-presentation. These efforts would not only enhance the depth of our understanding but also improve the applicability of our findings across more diverse populations.

6. REFERENCES

[1] Abram, M. D., Mancini, K. T., & Parker, R. D. (2021). Methods to integrate natural language processing into qualitative research. All Articles. https://doi.org/10.1177/1609406920984608

[2] Allen, J. P., Loeb, E. L., Kansky, J., & Davis, A. A. (2022). Beyond susceptibility: Openness to peer influence is predicted by adaptive social relationships. International Journal of Behavioral Development, 46(3), 180-189. https://doi.org/10.1177/0165025420922616

[3] Blot, W. J. (1992). Alcohol and cancer. Cancer Research, 52(7 Suppl), 2119s-2123s. PMID: 1544150.

[4] Dunbar, R. I. M., Launay, J., Wlodarski, R., et al. (2017). Functional benefits of (modest) alcohol consumption. Adaptive Human Behavior and Physiology, 3, 118–133. https://doi.org/10.1007/s40750-016-0058-4

[5] Duell, N., Clayton, M. G., Telzer, E. H., & Prinstein, M. J. (2022). Measuring peer influence susceptibility to alcohol use: Convergent and predictive validity of a new analogue assessment. International Journal of Behavioral Development, 46(3), 190-199. https://doi.org/10.1177/0165025420965729

[6] Flesia, L., Fietta, V., Foresta, C., & Monaro, M. (2021). The association between dating apps and alcohol consumption in an Italian sample of active users, former users, and non-users. Social Sciences, 10, 249. https://doi.org/10.3390/socsci10070249

[7] George, S., Rogers, R. D., & Duka, T. (2005). The acute effect of alcohol on decision making in social drinkers. Psychopharmacology, 182(1), 160-169. https://doi.org/10.1007/s00213-005-0057-9

[8] Giorgi, S., Yaden, D. B., Eichstaedt, J. C., Ashford, R. D., Buffone, A. E. K., Schwartz, H. A., Ungar, L. H., & Curtis, B. (2020). Cultural differences in tweeting about drinking across the US. International Journal of Environmental Research and Public Health, 17(4), 1125. https://doi.org/10.3390/ijerph17041125

[9] Huang, T.-F., Hou, C.-Y., Chang, F.-C., Chiu, C.-H., Chen, P.-H., Chiang, J.-T., Miao, N.-F., Chuang, H.-Y., Chang, Y.-J., Chang, H., & Chen, H.-C. (2023). Adolescent use of dating applications and the associations with online victimization and psychological distress. Behavioral Sciences, 13(11), 903. https://doi.org/10.3390/bs13110903

[10] Jelodar, H., Wang, Y., Rabbani, M., et al. (2020). A collaborative framework based for semantic patients-behavior analysis and highlight topics discovery of alcoholic beverages in online healthcare forums. J Med Syst, 44, 101. https://doi.org/10.1007/s10916-020-01547-0

[11] Jose, R., Matero, M., Sherman, G., Curtis, B., Giorgi, S., Schwartz, H. A., & Ungar, L. H. (2022). Using Facebook language to predict and describe excessive alcohol use. Alcoholism: Clinical and Experimental Research, 46(5), 836–847. https://doi.org/10.1111/acer.14807

[12] Marengo, D., Azucar, D., Giannotta, F., Basile, V., & Settanni, M. (2019). Exploring the association between problem drinking and language use on Facebook in young adults. Heliyon, 5(10), Article e02523. https://doi.org/10.1016/j.heliyon.2019.e02523

[13] OpenAI. (2024). ChatGPT (4) [Large language model]. https://chat.openai.com

[14] Peters, B. L., & Stringham, E. (2006). No booze? You may lose: Why drinkers earn more money than nondrinkers. Journal of Labor Research, 27, 411–421. https://doi.org/10.1007/s12122-006-1031-y

[15] Rudeboybert. (n.d.). OkCupid codebook (Revised). GitHub. https://github.com/rudeboybert/JSE_OkCupid/blob/master/okcupid_codebook_revised.txt

[16] Scannell, M. J. (2019). Online Dating and the Risk of Sexual Assault to College Students. Building Healthy Academic Communities Journal, 3(1), 34–43. https://doi.org/10.18061/bhac.v3i1.6688

[17] Stappenbeck, C. A., & Fromme, K. (2014). The effects of alcohol, emotion regulation, and emotional arousal on the dating aggression intentions of men and women. Psychology of Addictive Behaviors, 28(1), 10–19. https://doi.org/10.1037/a0032204

[18] World Health Organization. (n.d.). Alcohol. Retrieved from https://www.who.int/news-room/fact-sheets/detail/alcohol