#MugLife

"Which beverage reigns superior? An analysis of the great Tea vs Coffee debate."

Milindi Kodikara, Syeda Shabnam Khan, Mahawattage Perera

1 Introduction 4
2 Method 4
2.1 Data collection 5
2.1.1 Reddit 5
2.1.2 YouTube 5
2.1.3 Google Reviews 5
2.2 Pre-processing 6
2.2.1 Exploration of data using distribution of term frequency 8
2.3 Approach 8
2.3.1 N-grams 9
2.3.2 Sentiment analysis 9
2.3.3 Topic modelling 9
2.3.4 Graphs and Networks 10
2.3.4.1 Egonet 10
2.3.4.2 Reply graphs and centrality 10
2.3.4.3 Community detection 11
2.3.5 Bag-of-words model 12
3 Analysis 12
3.1 Reddit 12
3.1.1 *Who* are the key influencers in the tea and coffee communities 12
Top players in the tea and coffee community 12
Who gets influenced by whom? 13
Extent of their influence across reddit 15
3.1.2 *What* they love to talk about 17
N-grams 18
Topic modelling and word cloud 18
3.1.3 *When* redditors show interest and *why* 22
3.1.4 *Where* their interests lay 24
Leaf and Bean type 25

Tea and Coffee style 25
Brewing style 26
Origin of leaves and beans 26
Countries 27
3.1.5 *How* they feel about tea and coffee 28
3.2 YouTube 29
3.2.1 *What* do they love to talk about 29
3.2.2 *When viewers* show interest and *why* 32
3.2.3 *How* they feel about tea and coffee 34
3.3 Google Reviews 34
3.3.1 *What* do they love to talk about 34
3.3.2 *Where* their interests lay 36
4 Conclusion 37
References 38
Appendix A 39
Appendix B 40

1 Introduction

In this age of being chronically online, one of the hottest debates on social media is **"Tea vs Coffee"** [1-4]. This topic has always been an interest of ours as we are international students from South Asia, one of the largest areas of tea and coffee consumption and production in the world. Naturally, this led us to explore the research question: **"Tea vs Coffee: Which beverage reigns superior?"**.

In this paper, we attempt to put a stop to this debate once and for all using Natural Language Processing techniques on popular social media. Posts and comments from social media sites Reddit, YouTube and Google Reviews were explored and analysed to find influencers and communities surrounding these topics. Furthermore, we determine the trending concepts, topics and feelings discussed within these communities.

We employ sentiment analysis, topic modelling, and social media and network analysis techniques to provide insights into **who** the active influencers are, **what** they think and care about, **when** their interest is peaked and **why**, **where** there is interest and **how** they feel about all things related to tea and coffee.

Tea and coffee both have their respective communities in social media. While tea has a large number of communities with diverse interests, coffee has a smaller number of communities with powerful influencers. Tea drinkers love talking about their leaves and tea-ware. They have a consistent social media presence over time. On the other hand, coffee drinkers are passionate about their brews and grinders. They have been getting more active in social media recently.

2 Method

The experiments involved the creation of a natural language processing (NLP) and analysing system, MugLife[1]. MugLife includes means for collection, pre-processing, exploration and analysis of data from social media platforms Reddit, YouTube and Google Reviews.

The statistics of the data collected can be found in Table 1.

[1] https://github.com/Milindi-Kodikara/MugLife

2.1 Data collection

2.1.1 Reddit

Top posts and their associated comments have been extracted as these posts have garnered the most upvotes over time. The subreddits from which data related to tea were collected from ***tea*** and ***TeaPorn*** while coffee data was collected from ***coffee*** and ***pourover*** as these subreddits were the largest subreddits related to tea and coffee and contained the largest number of active users.

The PRAW API, the python reddit API wrapper was used for data collection. This is as MugLife was written in Python and PRAW is a widely used, simple and efficient API wrapper to build using Reddit data.

2.1.2 YouTube

By developing specific keywords and requirements relating to tea and coffee, a systematic approach was utilized to collect videos, ensuring the data's relevance and scope. To be able to automate the retrieval process and collect viewer statistics like likes and comments along with video titles and channel names, our approach made use of YouTube's API. After a thorough preparation step, the gathered data was sorted to remove unnecessary information and organize the most significant points for further analysis.

2.1.3 Google Reviews

15 big cities from countries that can be associated with either beverage whether in terms of culture or origin of the drink were selected. Reviews for all places that serve either tea or coffee within a radius of 25 KM from the city center were collected from google reviews. The data was extracted using keywords tea, coffee, cafe, bubble tea, coffee shop, tea house. The Google Places API [Places Detail, n.d.] was used for gathering data, a popular service from Google with limited free API calls.

Social media platform	Data type	Tea	Coffee	Total
Reddit	Posts	1,996	1,970	3,966
	Comments	58,968	159,226	218,194
	Total	**60,964**	**161,196**	**222,160**
YouTube	Videos	4896	4783	9679

	Comments	2500	2500	5000
	Total	**9869**	**9783**	**14,679**
Google Reviews	Reviews	23395		23395
	Total	**23395**		**23395**

Table 1: Summary of collected data from social media platforms

2.2 Pre-processing

Upon exploration of a subset of the data extracted, pre-processing was conducted to improve the accuracy and reliability of the data to be analysed. The pre-processing techniques included the following:

1. Conversion of all text to lowercase
2. Removal of digits
3. Removal of punctuation marks, emojis and special characters
4. Removal of username tags, mentions and links
5. Tokenization: The process of splitting a sequence of words to individual tokens utilizing the NLTK library.
6. Stripping whitespace
7. Removal of stop words (redundant words such as 'a', 'the' etc.)

An example of a reddit post and the associated pre-processed tokens are depicted in Figures 1 and 2, respectively.

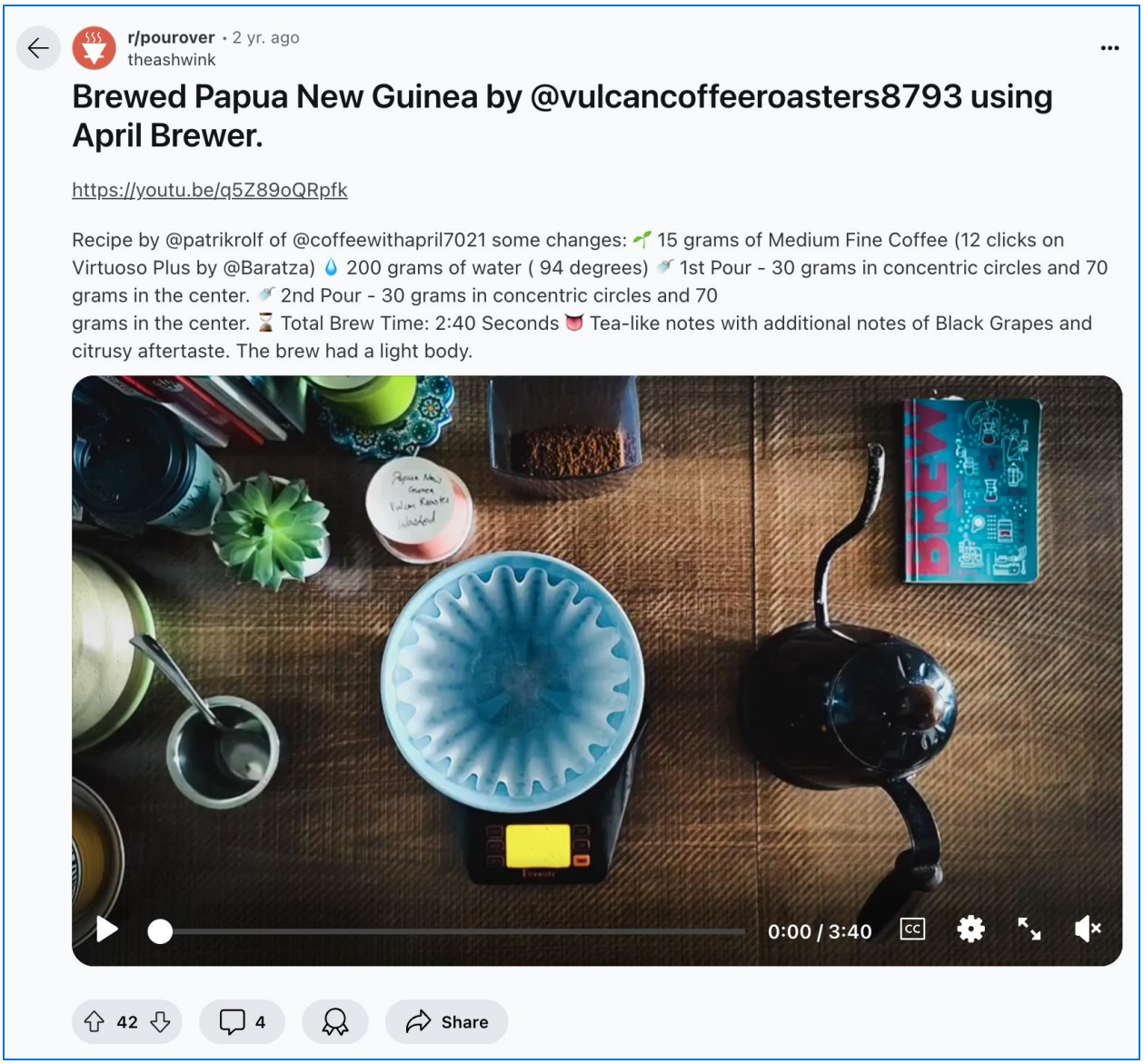


*Figure 1: Example reddit post from the **pourover** subreddit*

['brewed', 'papua', 'new', 'guinea', 'using', 'april', 'brewer', 'recipe', 'changes', 'grams', 'medium', 'fine', 'coffee', 'clicks', 'virtuoso', 'plus', 'grams', 'water', 'degrees', 'pour', 'grams', 'concentric', 'circles', 'grams', 'center', 'pour', 'grams', 'concentric', 'circles', 'grams', 'center', 'total', 'brew', 'time', 'seconds', 'tea-like', 'notes', 'additional', 'notes', 'black', 'grapes', 'citrusy', 'aftertaste', 'brew', 'light', 'body']

Figure 2: Pre-processed reddit post from Figure 1

2.2.1 Exploration of data using distribution of term frequency

A distinct difference between the un-processed and pre-processed data could be seen upon the exploration of the frequency distributions of the top 50 unigrams. This cements the necessity for the application of the pre-processing techniques for analysis as the pre-processed data is void of redundant and meaningless tokens. Figures 3-6 show the term frequency distribution before and after pre-processing of the Reddit data for tea and coffee datasets.

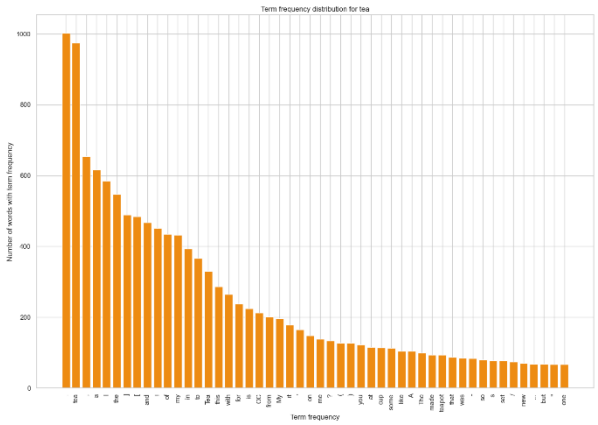


Figure 3: Term frequency unprocessed reddit data, tea

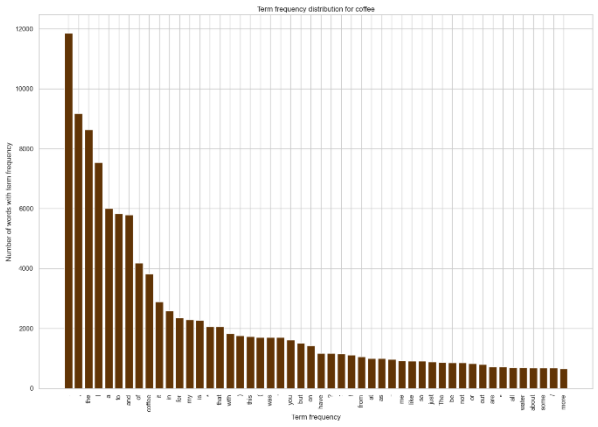


Figure 4: Term frequency after pre-processing reddit data, coffee

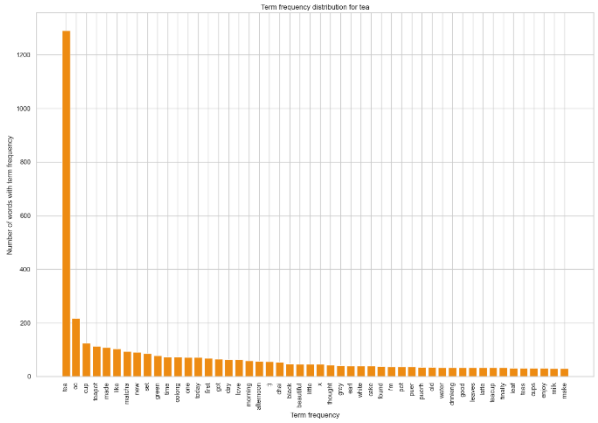


Figure 5: Term frequency pre-processed reddit data, tea

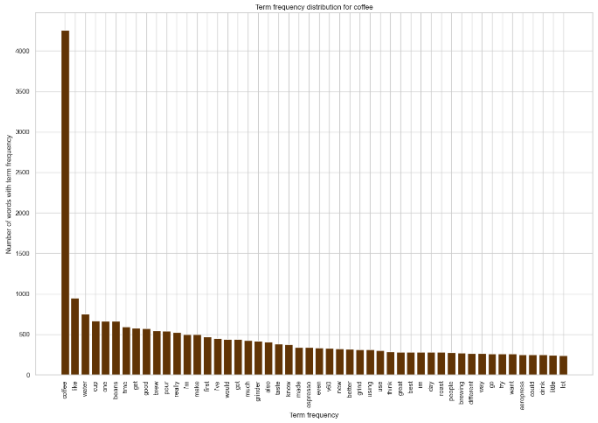


Figure 6: Term frequency after pre-processing reddit data, coffee

2.3 Approach

2.3.1 N-grams

To determine the key words and themes in the social media sites unigrams, bigrams and trigrams were utilized given that these methods take context of the words into account.

2.3.2 Sentiment analysis

Sentiment analysis was conducted via two methods, Count sentiment analysis and Vader sentiment analysis. These unsupervised models were utilized given that the ground truth of the sentiments for the datasets was not available.

Sentiment analysis was conducted to explore how people feel towards coffee and tea and explore engagement over time.

Count sentiment analysis determines the average sentiment by counting the number of positive and negative words [Figure 7] while Vader computes the degree of positive and negative sentiment using a lexicon set specifically built from social media [Figure 8].

```
use, coffee, press, grind, beans, put, directly, coffee, press, would, use, coffee, press, want, coffee, use, milk, instead,
water, sorry, questions, really, basic, i'm, year, old, trying, make, mom, feel, special, birthday, edit, thanks, awards, ill,
give, update, whether, mom, like, coffee, celebrate,

------------Count sentiment value------------
1
-----------------------------------
```

Figure 7: Count sentiment of a reddit post, count sentiment value of 1

```
use, coffee, press, grind, beans, put, directly, coffee, press, would, use, coffee, press, want, coffee, use, milk, instead,
water, sorry, questions, really, basic, i'm, year, old, trying, make, mom, feel, special, birthday, edit, thanks, awards, ill,
give, update, whether, mom, like, coffee, celebrate


------------Count sentiment value------------
compound: 0.9001
-----------------------------------
```

Figure 8: Vader sentiment of a reddit post, compound vader sentiment value of 0.9001

2.3.3 Topic modelling

Topic modelling was conducted to determine what concepts, topics and themes were discussed within the collected datasets for tea and coffee. The unsupervised approach,

topic model Latent Dirichlet Allocation (LDA), was utilized for topic modelling and topics were visualized using pyLDAvis and word clouds.

LDA functions by finding clusters of co-occurring words by encouraging topics to have a few prominent words and the documents to have a few prominent topics.

2.3.4 Graphs and Networks

The popular Python network analysing tool networkx was utilised to construct graphs for analysis.

2.3.4.1 Egonet

Ego-centric networks of the top users (from here on referred to as *influencers*) were explored to determine their influence on the tea and coffee communities.

Influencer influence was determined by looking at who replies to their posts and comments and who they reply to in return with the influencer at the center of the network. As the reply links are directional a DiGraph is used to construct the graph where inward directed edges represent replies from other users. The reply count was stored to indicate the popularity of each neighbours.

2.3.4.2 Reply graphs and centrality

Reply graphs were employed to understand the users of the tea and coffee subreddits by analysing their communities and behaviours.

The nodes represents the users. The directed edges represent the replies to another redditor's posts or comments, and the replies to their own posts and comments.

Centrality is computed to determine important users within the community by creating nodes out of the posts' authors and creating edges of the replies between them. Degree (ranks users with more connections higher in terms of centrality), eigenvector (generalise degree centrality by incorporating the importance of neighbours) and katz (adding bias to the computed eigenvector centrality values) centralities were computed and embedded in the nodes of the network. The statistics of the centralities are depicted in Figures 9 and 10.

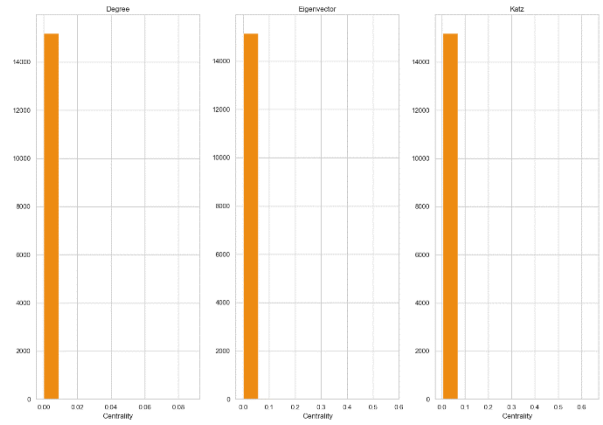


Figure 9: Centrality statistics, tea

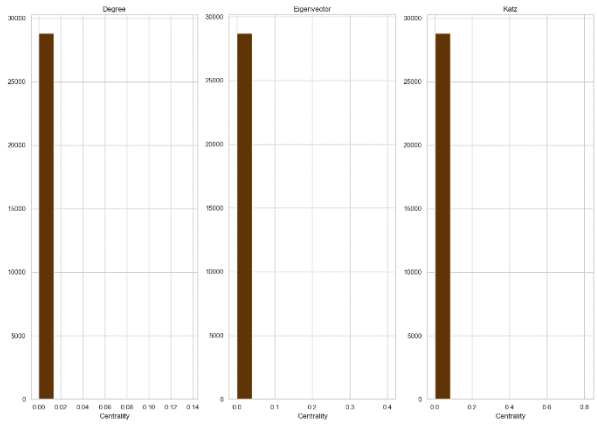


Figure 10: Centrality statistics, coffee

Additionally, the global clustering coefficient was computed to measure transitivity of the reply graphs, constructed based on triplets of nodes. The global clustering coefficient for tea and coffee reply graphs were 0.003262778311738053 and 0.014468822054633144 respectively.

2.3.4.3 Community detection

Communities surrounding users of the tea and coffee datasets were identified based on similar characteristics utilizing built-in networkx community detection algorithms to determine their interests and behaviours.

Community detection was conducted on the reply graphs created based on the tea and coffee data. The reply graphs were converted to be undirected graphs as the selected community detection algorithms to be implemented could only be implemented on undirected graphs.

Clique Percolation Method (CPM) and Louvain algorithms were implemented as the former is a node based community detection algorithm whilst the latter is a group based, global community detection algorithm.

CPM uses cliques as seeds to find larger, overlapping communities. For our explorations, the clique size was set to 3. As CPM does not have a modularity maximisation implementation, Louvain algorithm, a greedy modularity optimisation method, has been implemented to maximise modularity to gain clearer and distinct communities.

The visualization platform Gephi was used to explore the graphs.

Given that there is no ground truth to compare against, the graphs were analysed by comparing against the users' interests.

To determine the influence of redditors from the tea and coffee subreddits on all subreddits in reddit, an influencer network was constructed upon the subreddits tea and coffee redditors are active in. The influencer network was created to distinguish patterns and similarities between the users. 221 and 192 unique, top voted authors from the tea and coffee subreddits and their influence on all subreddits is analysed in Section 3.

2.3.5 Bag-of-words model

In addition to the afore mentioned methods, bag-of-words model was implemented for the extraction of unique features in reddit to determine where users' interest lay as a community using CountVectorizer.

3 Analysis

Key points from the analysis can be found on the MugLife website[2].

3.1 Reddit

3.1.1 *Who* are the key influencers in the tea and coffee communities

Top players in the tea and coffee community

221 and 192 top voted and most commented unique authors, and their posts and comments in subreddits they are active in are explored in this section to identify the top influencers and the extent of their influence.

Zishateapot was identified as the top influencer in the tea subreddits with 125 out degrees and 7 in degrees [Figure 11] while Pilot_Maven was identified as the top influencer in the coffee subreddits with 486 in degrees and 74 out degrees.

[2] https://muglife.github.io/ (Only compatible with desktop view, currently)

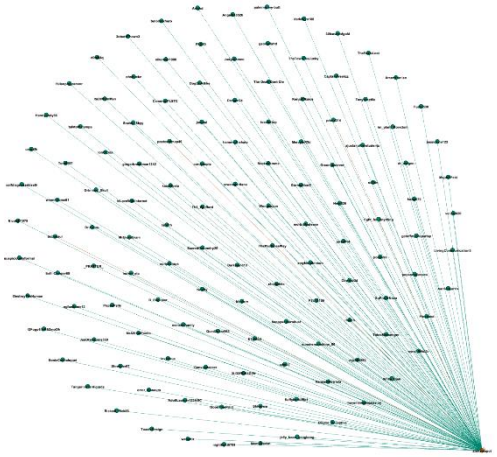


Figure 11: Top influencer, tea: Zishateapot

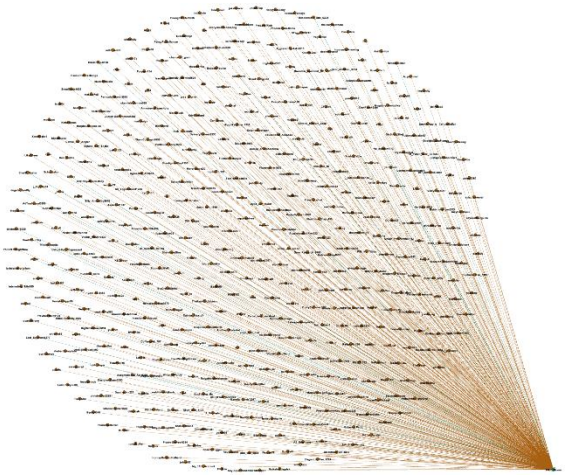


Figure 12: Top influencer, coffee: Pilot_Maven

Who gets influenced by whom?

Based on centrality statistics, the global clustering coefficient for tea and coffee reply graphs were 0.003262778311738053 and 0.014468822054633144 respectively. We could infer from these figures that the reddit community does not comprise of tightly interconnected groups of users.

This was further corroborated via Figures 9 and 10 which depicts a significantly low centrality across degree, eigenvector and katz centralities showing that the average connectedness is low within these two communities thereby leading us to the conclusion that the influence one user has over the other is low.

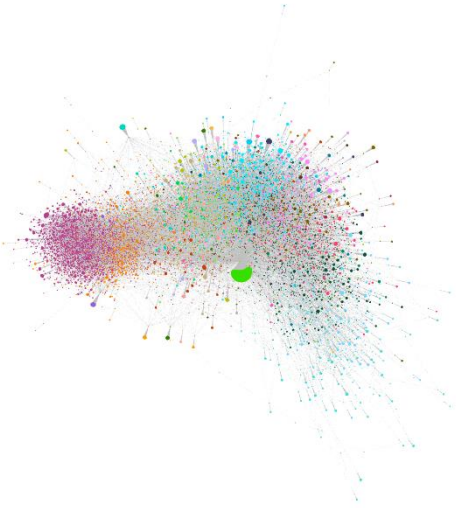


Figure 13: Coffee community partitioned with louvain algorithm and modularity resolution 1.0

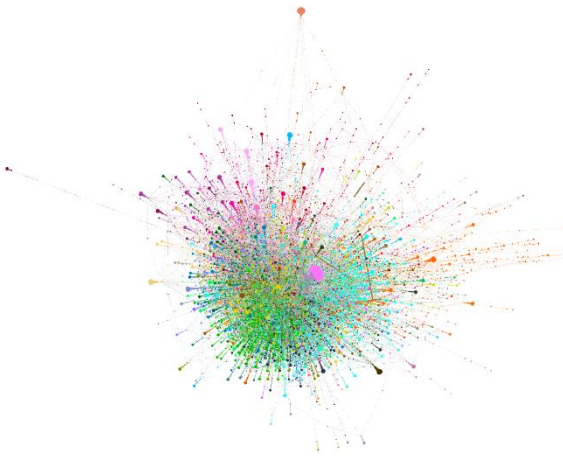


Figure 14: Tea community partitioned with louvain algorithm and modularity resolution 1.0

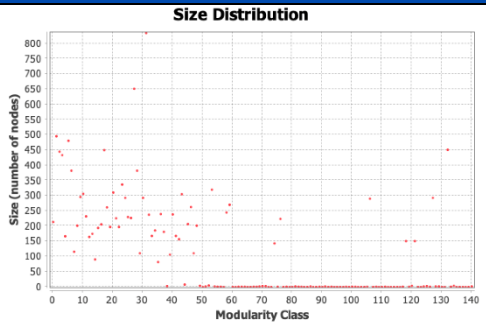
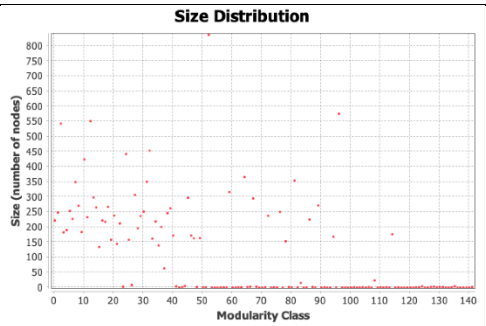
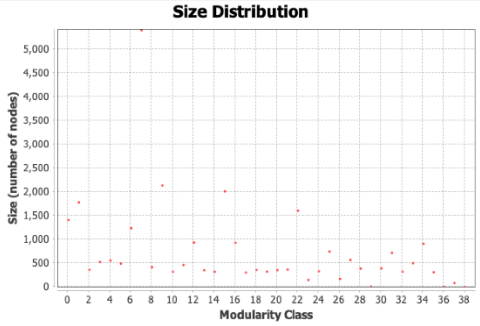
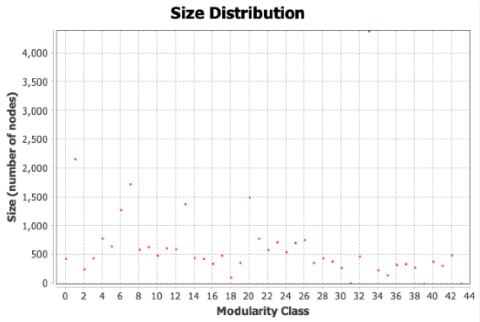
Community type	Algorithm	Modularity	Number of communities	Size distribution
Tea	CPM	0.690	141	
	Louvain	0.689	142	
Coffee	CPM	0.645	39	
	Louvain	0.645	44	

Table 2: Summary of community statistics, reddit

It can be observed from Figures 13 and 14, and Table 2 that there is a very high number of communities detected while depicting high modularity scores for the tea and coffee communities at resolution one.

While distinct and large communities could be observed, the influence of individual redditors is low. This could be due to Reddit not being a platform where a majority of redditors 'follow' each other akin to social media platforms such as Twitter and Instagram.

Extent of their influence across reddit

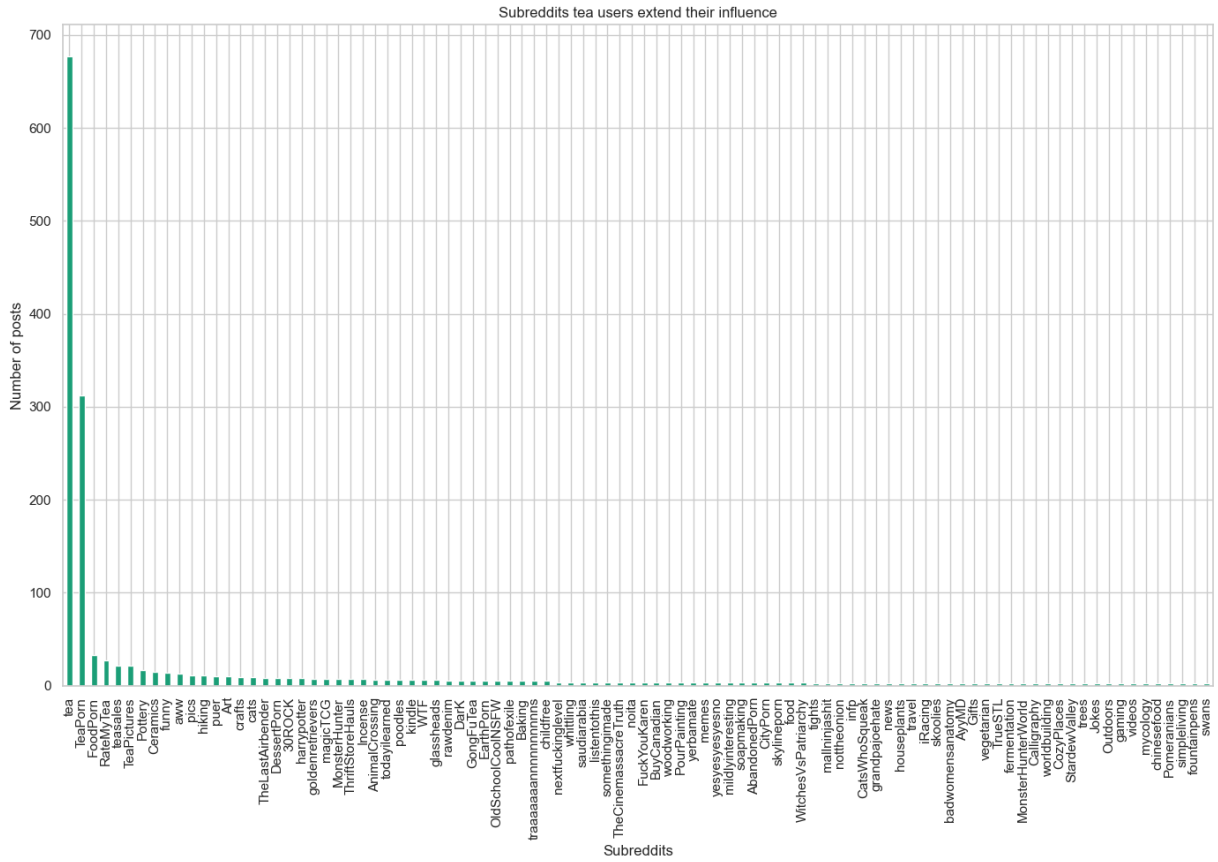


Figure 15: Influence of authors from tea subreddits on other subreddits

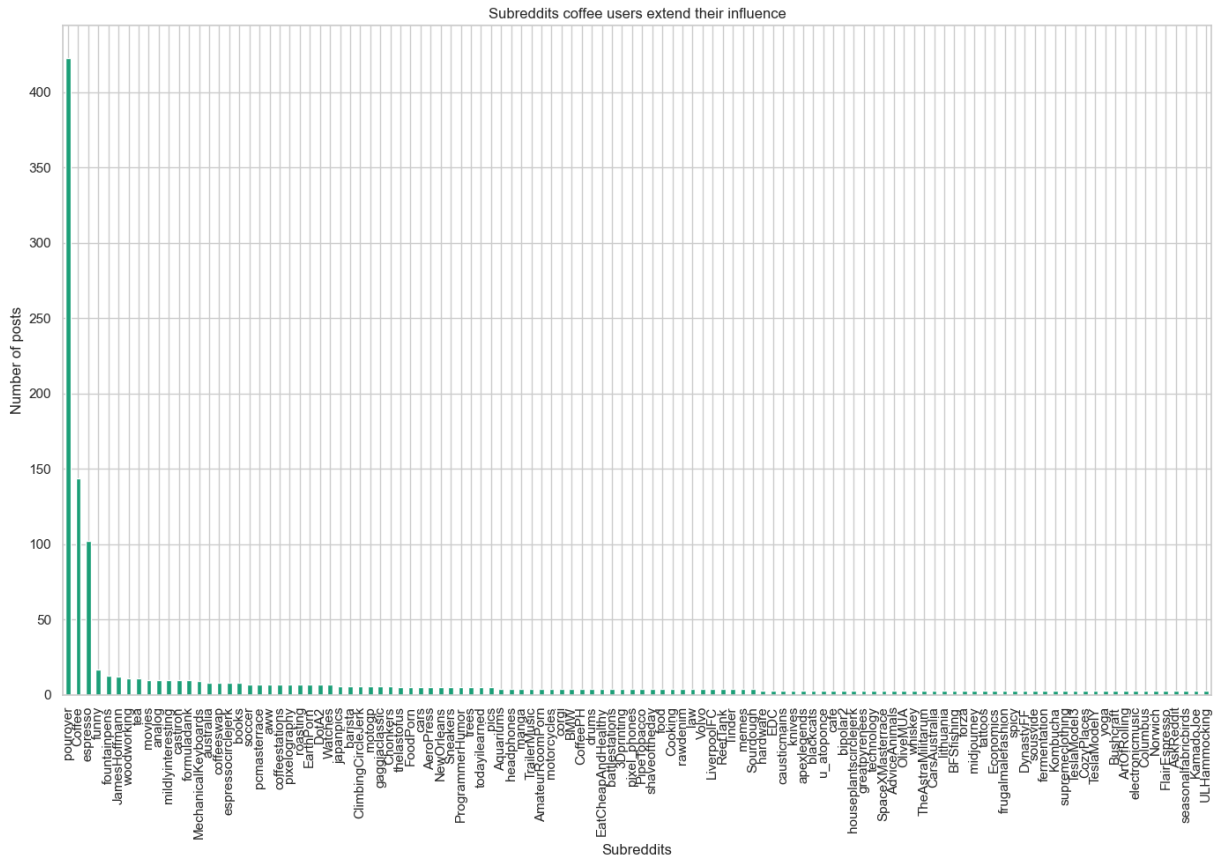


Figure 16: Influence of authors from coffee subreddits on other subreddits

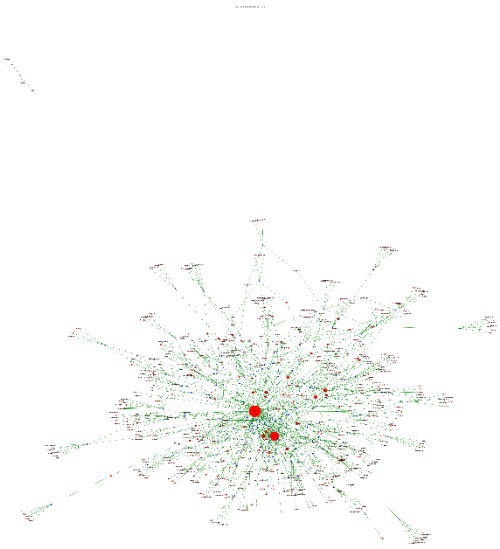


Figure 17: Network graph of extension of

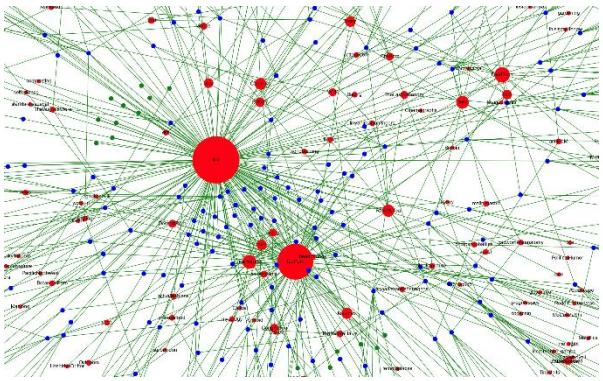


Figure 18: Network graph of extension of influence of authors from tea subreddits on other subreddit, zoomed-in

influence of authors from tea subreddits on other subreddit, zoomed-out

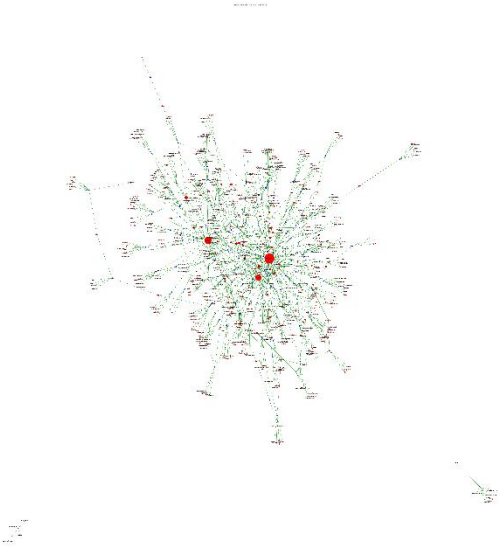


Figure 19: Network graph of extension of influence of authors from coffee subreddits on other subreddit, zoomed-out

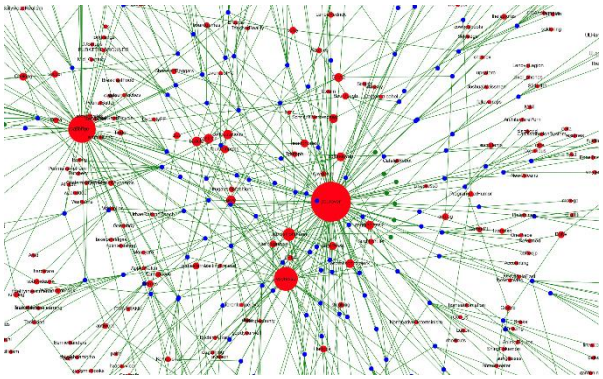


Figure 20: Network graph of extension of influence of authors from coffee subreddits on other subreddit, zoomed-in

Figures 15 to 20 depict the subreddits the tea and coffee communities extend their influence on.

It can be observed that most of the tea community is mainly active on tea related subreddits followed by tv show and creative subreddits. It can be inferred that the tea community consists of hobbyists who'd love a good, warm cup of tea in the morning.

Similarly, most of the coffee community is mainly active on coffee related subreddits followed by movie, games and technology related subreddits. Unsurprisingly, one of the top subreddits the coffee community is active on is Australia which could be because Australia and Melbourne especially are renowned for their coffee culture [5]. It can be inferred that the coffee community consists of tech enthusiasts and gamers who burn the midnight oil with hot coffee.

3.1.2 *What* they love to talk about

Most discussed topics, core themes and interests of the redditors were gathered using n-grams, topic modelling and word clouds.

N-grams

Beverage Type	Unigram	Frequency	Bigram	Frequency	Trigram	Frequency
Tea	tea	1291	tea, set	58	Loose, leaf, tea	9
	oc	217	green, tea	57	earl, grey, tea	9
	cup	124	early, grey	39	butterfly, pea, flower	6
Coffee	coffee	4254	french, press	145	tl, ;d, r	24
	like	950	cup, coffee	127	local, coffee, shop	16
	water	755	coffee, shop	95	good, cup, coffee	16

Table 3: Top n-grams for reddit data

According to the top n-grams and their frequencies depicted in Table 3, the tea subreddits consist of original content from the users showcasing various types of tea and tea sets while the coffee subreddits shows appreciation for good coffee with high quality whilst discussing brewing styles and local coffee shops.

Topic modelling and word cloud

10 topics were modelled and analysed to determine the trending themes and topics related to tea and coffee. The general theme and topics being discussed can be concluded as shown in Tables 4 and 5.

The topmost discussed topics in the tea community are as follows:

1. Various types of tea (Oolong, Earl grey, Chai, Matcha etc.) and their associated brewing times
2. Different times of the day to appreciate a good cup of tea
3. Showcase of tea setups

Three core themes in the tea community are as follows:

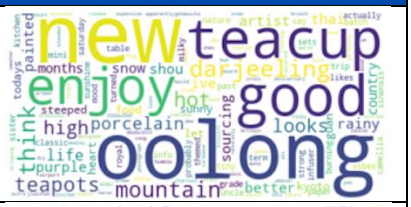
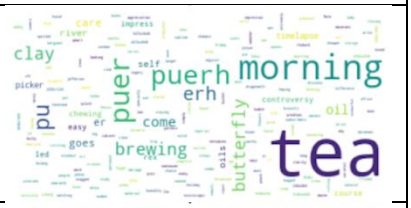
1. Sharing teatime moments
2. Enjoyment and appreciation of tea and teapots
3. Exploration of new tea

The topmost discussed topics in the coffee community are as follows:

1. Details on brewing coffee methods (V60, grinder, pour over etc.)
2. Various types of coffee styles (Eg: Espresso)
3. Discussions on roasts and beans

Three core themes in the coffee community are as follows:

1. Taste
2. Quality
3. Methods on making coffee

Topic	Inferred summary	Word cloud
new oolong good teacup enjoy darjeeling think mountain hot teapots looks high porcelain life better	The reditors discuss oolong and Darjeeling tea frequently. Moreover, they care about the atmosphere and utensils they are using.	
time leaves pot latte work make shop perfect making share thrift took chamomile appreciate long	Tea lovers care about the brewing time and tea leaves. Camomile, a floral tea is often mentioned along with leaves and time.	
tea morning puer puerh pu erh clay brewing butterfly come oil goes care er oils	Discussions on puer tea, involving the time and process of brewing.	
cake white jasmine milk im know nice blend mom makes rose pretty assam color red	Talks about jasmine, rose and assam tea with emphasis on color and quality. Cake and milk suggest add-on or food that complements tea.	
oc today little grey iced favorite year fresh china breakfast right mug great teaware handmade	The tea lovers are passionate about their tea ware. They care about the origins. The importance of time and freshness of leaves also stand out.	

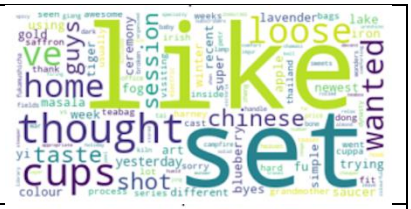
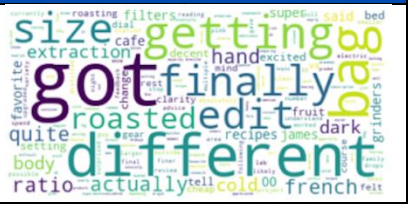
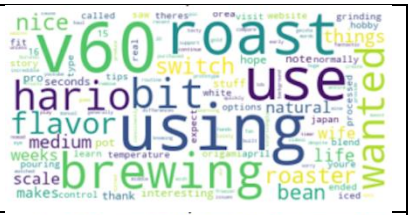
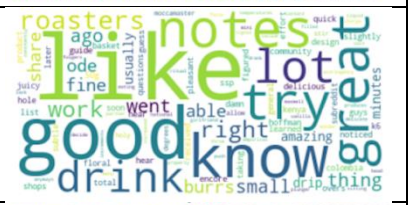
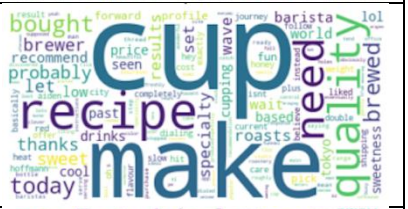
like set thought cups loose ve home wanted chinese session guys shot taste yesterday masala	The redditors talk about tea preparation, loose leaf tea and masala tea, as well as tea ware suggested by set, cup and chinese.	
love day chai beautiful japanese yunnan friend party happy delicious sweet store aged room house	Tea lovers discuss the flavor of tea. Passionate about age of leaves, storing method as well as origin, indicated by japanese and yunnan.	
got earl old drinking finally garden taiwan collection really bought hand flush os teatime late	The origin of tea, method of harvest, age of leaves, and type of leaves specifically Earl grey are key topics here.	
afternoon black raw leaf enjoying post best sencha spring yixing years photo flower lovely homemade	The commentors discuss about specific tea leaves, method of brewing. They also talk about the seasons and place to best enjoy their brew.	
cup teapot matcha green teas water gaiwan japan way pet mother pea ginger coffee orange	Discussion around location of tea and brewing ingredients. Orange and green indicates color of brew. Gaiwan, suggest discussion on tea ware. Coffee suggests comparison with coffee.	

Table 4: Topic analysis of tea community based on reddit data

Topic	Inferred summary	Word cloud
got different bag getting finally size edit roasted actually hand quite french ratio extraction dark	The coffee lovers talk about coffee roast and extraction method given by french, extraction and ration.	

v60 using use roast brewing bit wanted hario flavor roaster nice switch bean things medium	The redditors are passionate about type of roast, beans and brewing utensils.	
taste grind better im day people way kettle pretty thought machine tasting method drinking setup	The commentors discuss ways of grinding beans and brewing setup, emphasis on improving taste.	
time new espresso used ive tried filter shop days sure start zp6 second difference temp	This topic involves espresso. The commentors discuss about grinders and filters and shops to buy new coffee hardware.	
like good know great notes try lot drink roasters right work thing went ago small	The discussion is taste of coffee, possibly from different roasting method, indicated by notes and roasters.	
pour brew aeropress light pourover years fellow dont black going making press acidity washed week	The redditors speak about the type of brew, deep discussion on the taste given by washed and acidity.	
coffee really love morning experience trying feel milk looking buy order case easy carafe appreciate	Coffee lovers are interested in when and how they enjoy coffee. Their ways of serving, with milk and in a carafe.	
cup make recipe quality need bought brewed today probably sweet let thanks brewer specialty set	This topic revolves around brewing recipes and brewing utensils.	
ve grinder best coffees cups home post local started long maybe single recently bad away	Redditors talk about grinders and coffee cup, sharing ring some personal experience, indicated by started, bad and away.	

water beans think little want say year dripper happy far times add couple pours starbucks	Coffee lovers share their brewing method, indicated by water and dipper. The conversations also include Starbucks, a popular coffee place.	

Table 5: Topic analysis of coffee community based on reddit data

3.1.3 *When* redditors show interest and *why*

Key time periods of user engagement in tea and coffee subreddits since the conception of Reddit were analysed using the number of posts per date and posts per authors.

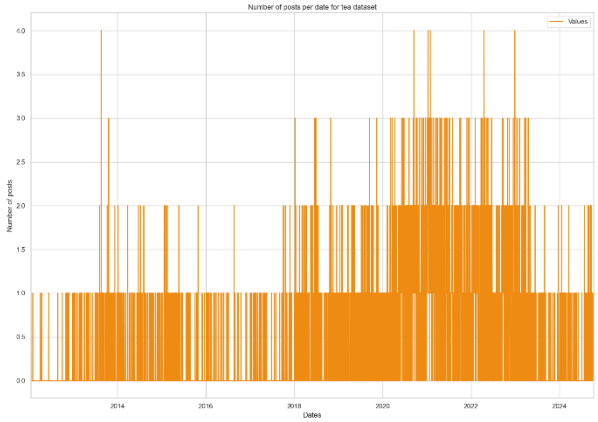


Figure 23: Posts per date reddit data, tea

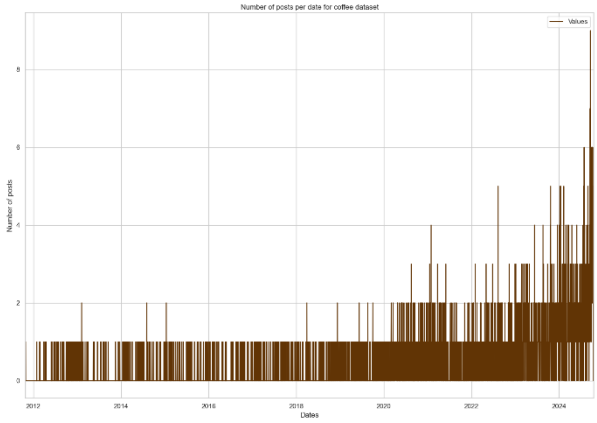


Figure 24: Posts per date reddit data, coffee

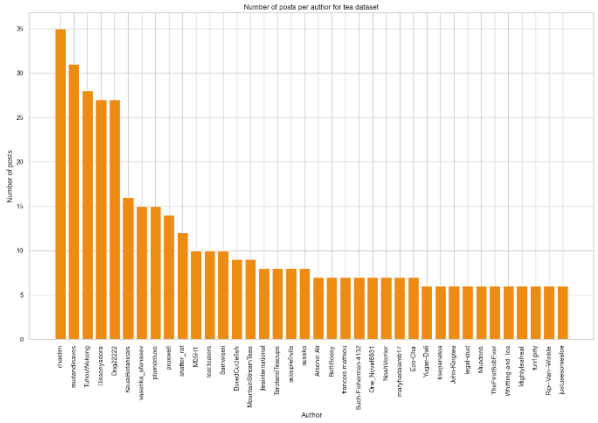


Figure 25: Posts per author reddit data, tea

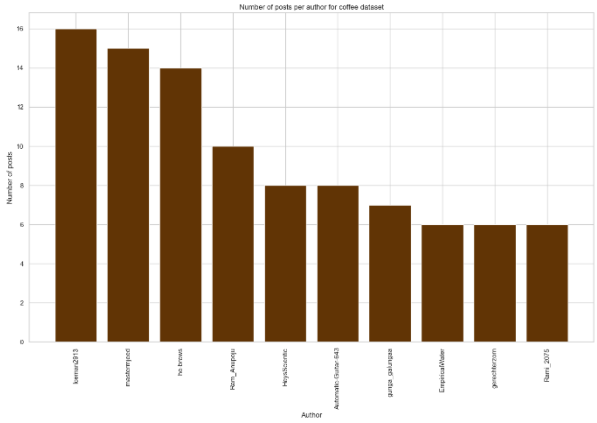


Figure 26: Posts per author reddit data, coffee

Upon analysis of the community engagement in the subreddits over time it can be observed that the period from 2018 to 2022 saw the highest engagement within the

tea community [Figures 23, 25] whereas there is a steep increase in engagement from late 2020 onwards in the coffee community [Figure 24, 26].

While the community engagement in tea subreddits has been consistently buzzing, there is a clear rise of engagement in the coffee subreddits in recent years which correlates with news regarding spikes in coffee consumption in recent years [10].

Moreover, from Figures 25 and 26 it can be seen that individual redditors engage more with tea subreddits as opposed to coffee subreddits. As seen in section 3.1.2, based on the topics and themes observed, this consistent engagement could be inferred to be due to redditors inclination towards showcasing their daily cups of teas in aesthetic manners. Figure 27 shows an example post from the most frequent redditor in the tea subreddits.

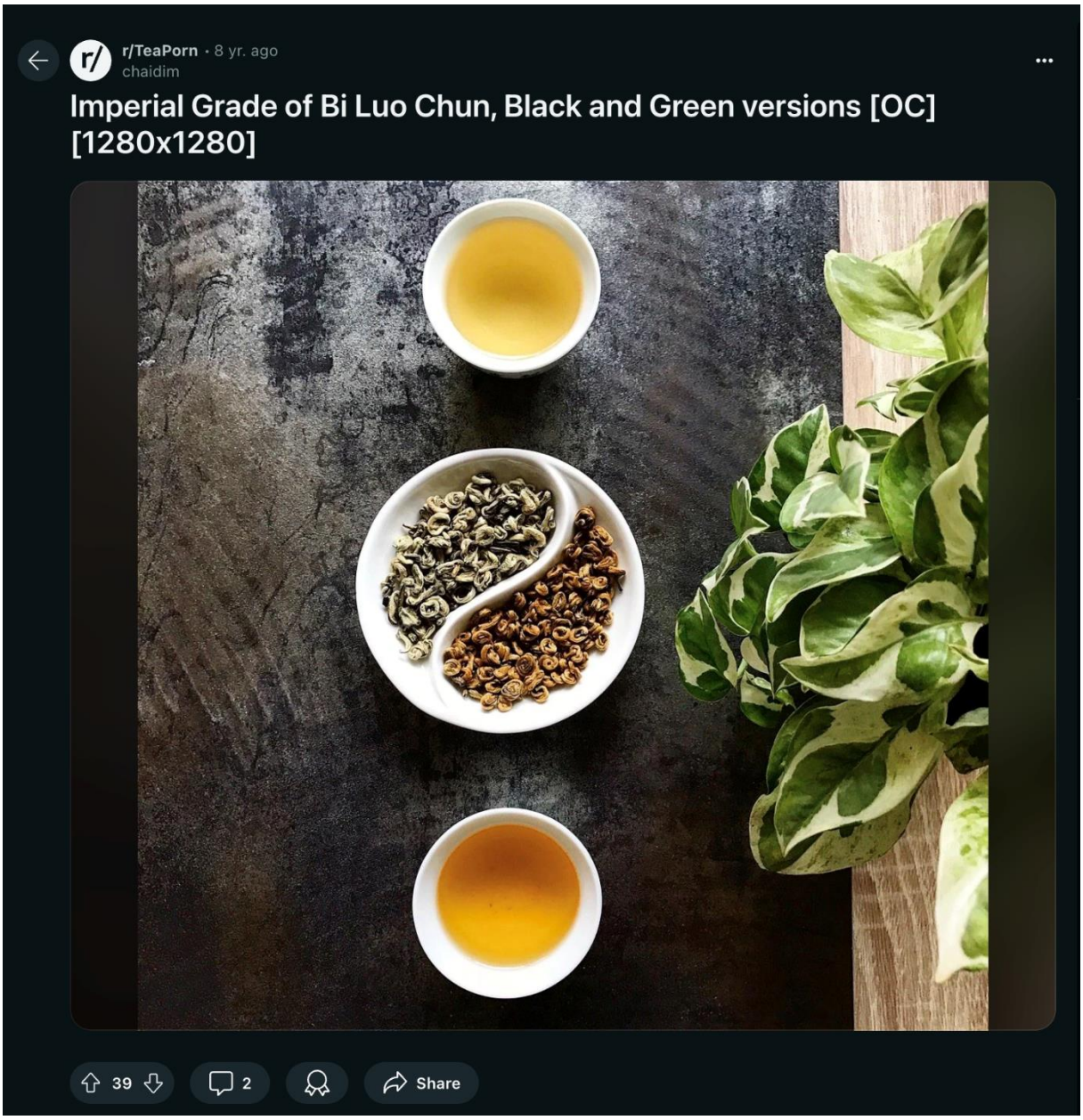


Figure 27: Post from the most frequent poster in the tea community

3.1.4 *Where* their interests lay

Interests of the tea and coffee communities were explored using frequency analysis of unique keywords to determine the popular leaf/bean types, tea/coffee styles, brewing styles, leaves/beans based on origin and countries.

Leaf and Bean type

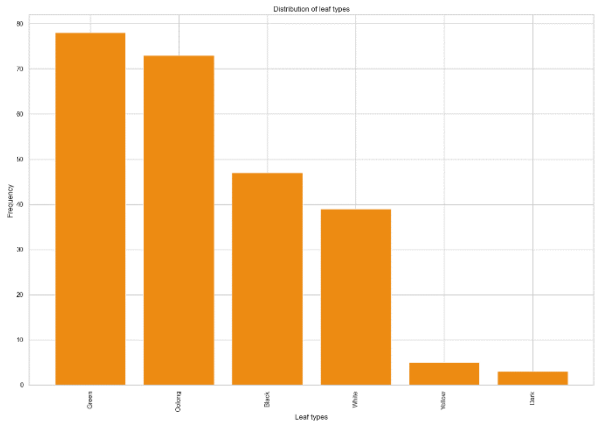


Figure 28: Leaf type

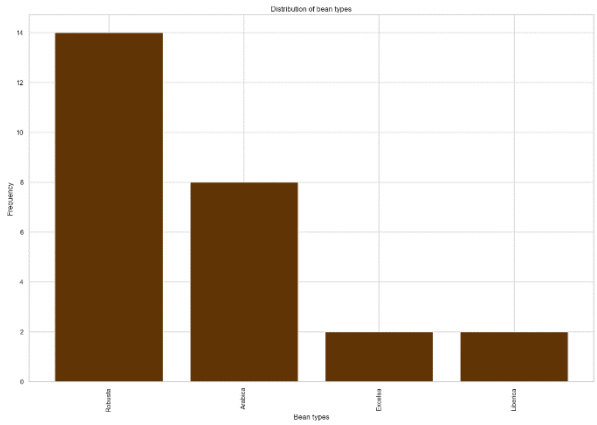


Figure 29: Bean type

Tea and Coffee style

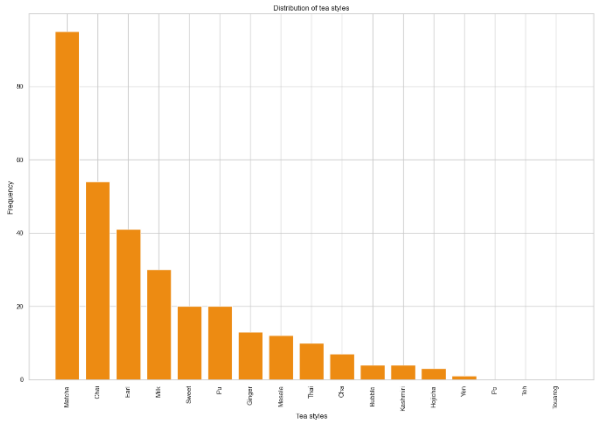


Figure 30: Tea style

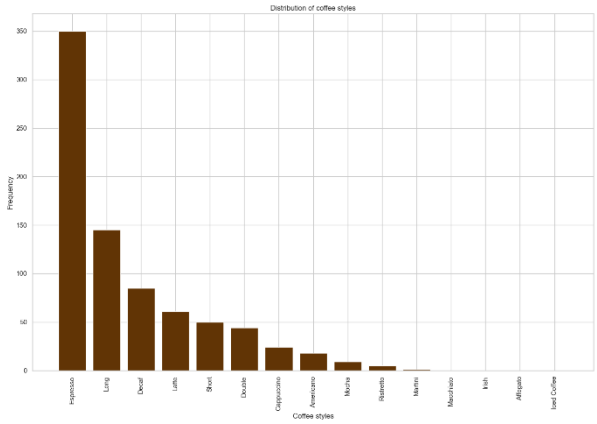


Figure 31: Coffee style

It can be observed that green tea, oolong tea and black tea dominate the taste buds of the community [Figure 28]. This could be inferred to be due to the health benefits from green and oolong tea and the liveliness that black tea brings. Moreover, Matcha, a Japanese tea style made of powdered green tea, is the reigning tea style as observed in Figure 30.

As quality of coffee and bean types were a popular discussion topic, it is no surprise to see various bean types being discussed with the dominating bean type being robusta which provides the traditional coffee flavour, albeit bitter [Figure 29]. This coincides with the top style of coffee being espresso as robusta is widely used for making this style of coffee [Figure 31].

Brewing style

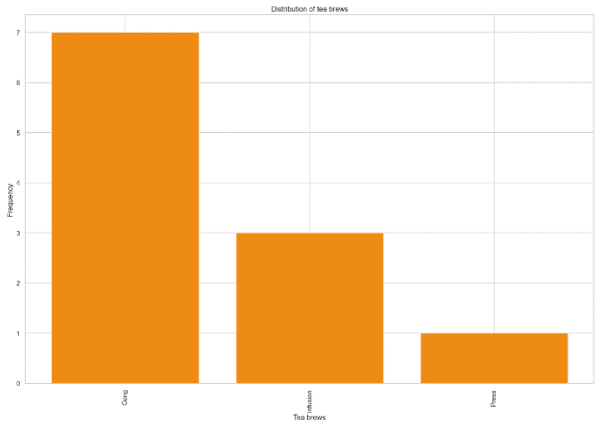


Figure 32: Brewing style, tea

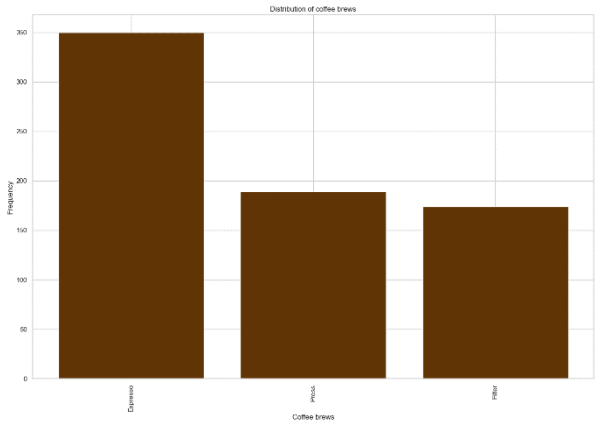


Figure 33: Brewing style, coffee

The most discussed tea brewing style was Gongfu cha which is a traditional Chinese tea preparation method [Figure 32]. While the most discussed coffee making style is Espresso, an Italian brewing style as seen in Figure 33.

Origin of leaves and beans

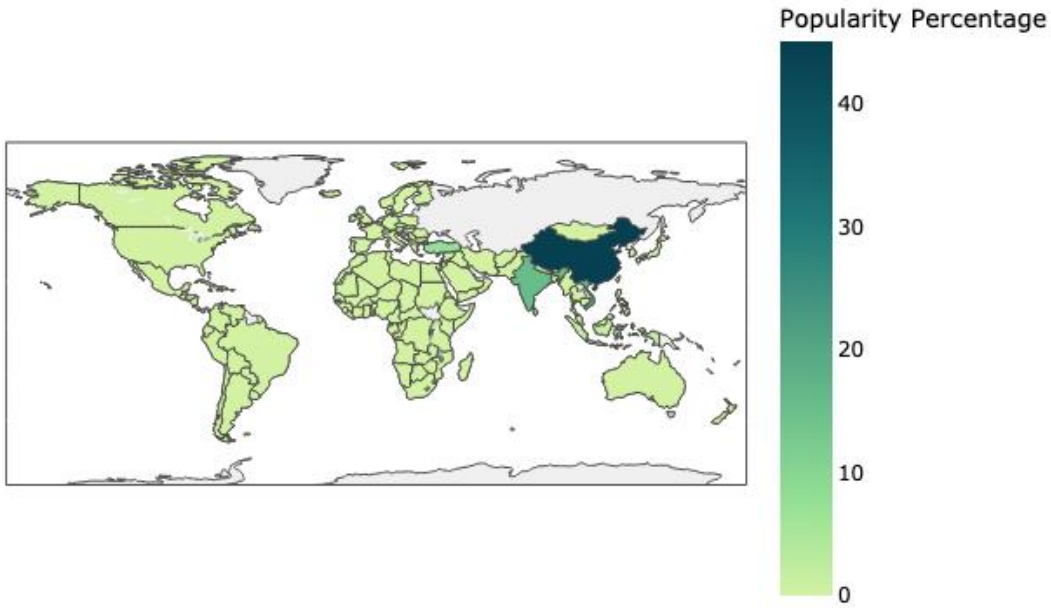


Figure 34: Origin of leaves

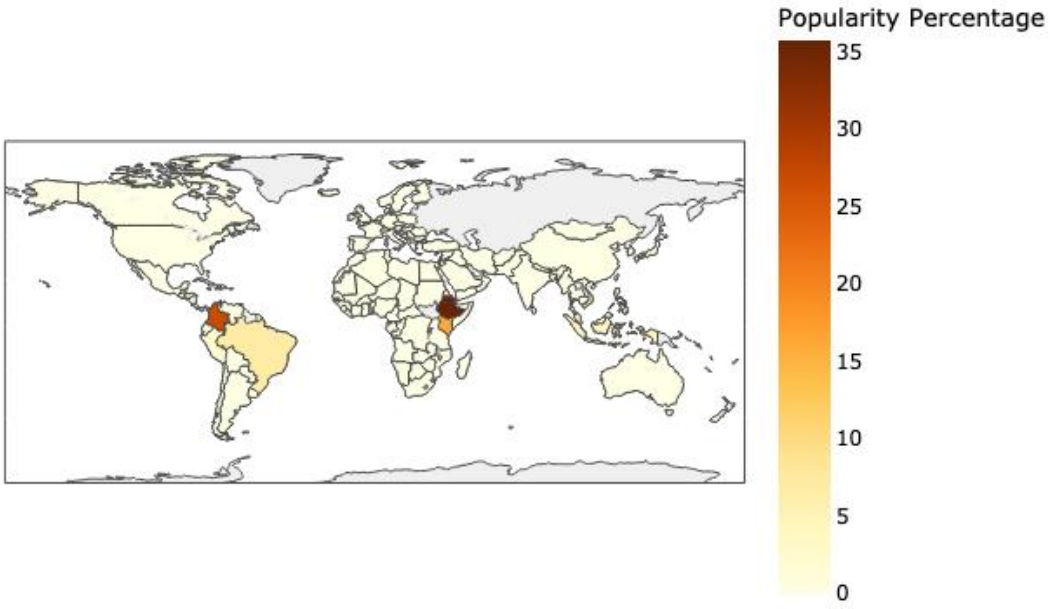


Figure 35: Origin of beans

Countries

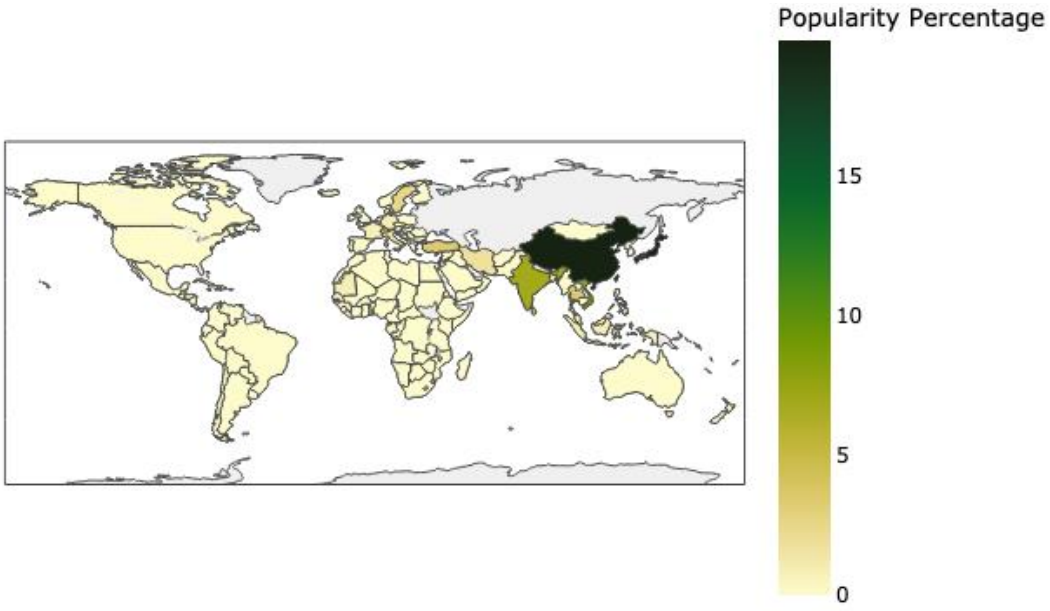


Figure 36: Popular countries in the tea subreddits

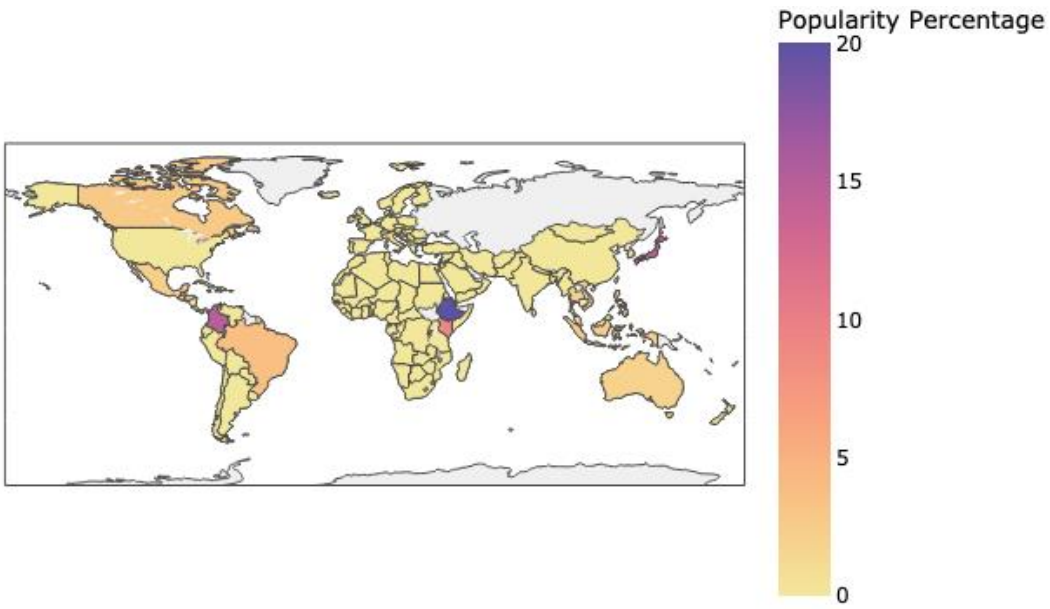


Figure 37: Popular countries in the coffee subreddits

As expected, the top tea and coffee production countries China, Vietnam, India, Sri Lanka, and Turkey are mentioned a significant amount similar to the top coffee producing countries Ethiopia, Colombia, Kenya, Brazil and Vietnam as depicted in the images from the generated interactive world maps in Figures 34 and 35.

Moreover, countries with specialty teas such as Pu-erh, Bubble tea and Matcha are shown to be mentioned frequently [Figure 36].

3.1.5 *How* they feel about tea and coffee

Sentiment analysis techniques were utilised to determine how the community feel about tea and coffee, and observe the peaks and troughs in sentiments over time.

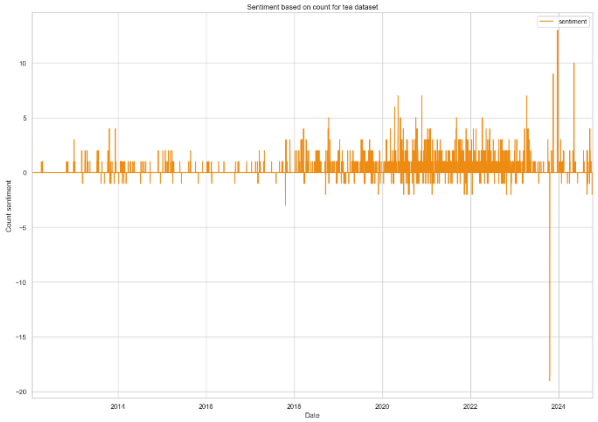


Figure 37: Count sentiment, tea

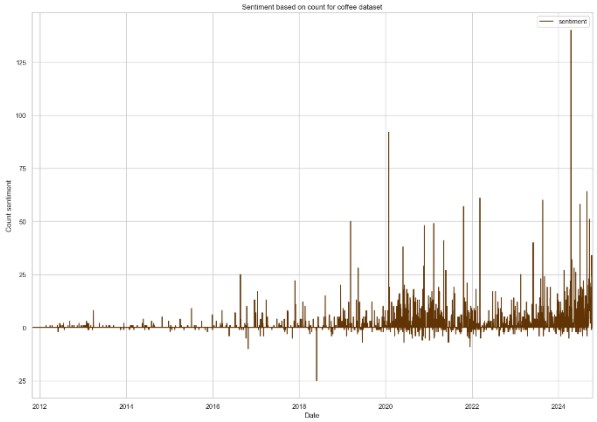


Figure 38: Count sentiment, coffee

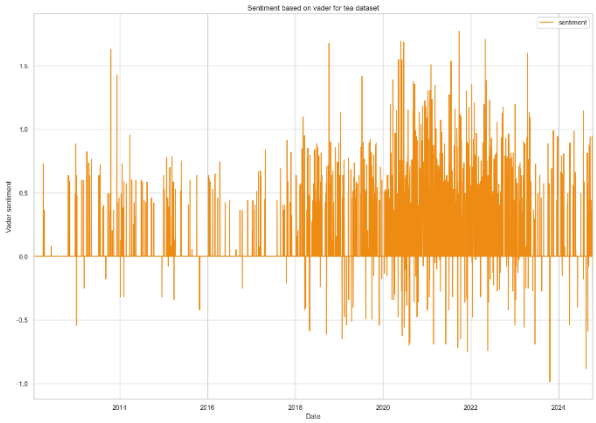


Figure 39: Vader sentiment, tea

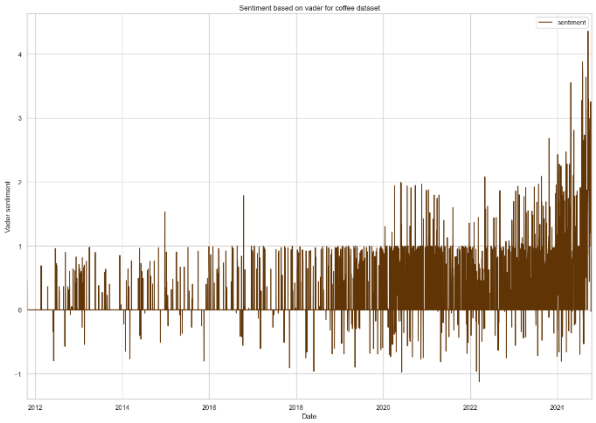


Figure 40: Vader sentiment, coffee

As seen with the frequency of posts per date, the sentiments follow a similar trend.

It can be observed that there is a contrast between the count and vader techniques with the vader sentiment analysis depicting heightened and more accurate sentiments over time due to being built specifically for analysis of social media datasets.

It can be observed that the tea community consistently expresses positive sentiments on average over time whereas the coffee community has had a steep increase in mainly positive sentiments and engagement since 2020 [Figures 39, 40].

3.2 YouTube

3.2.1 *What* do they love to talk about

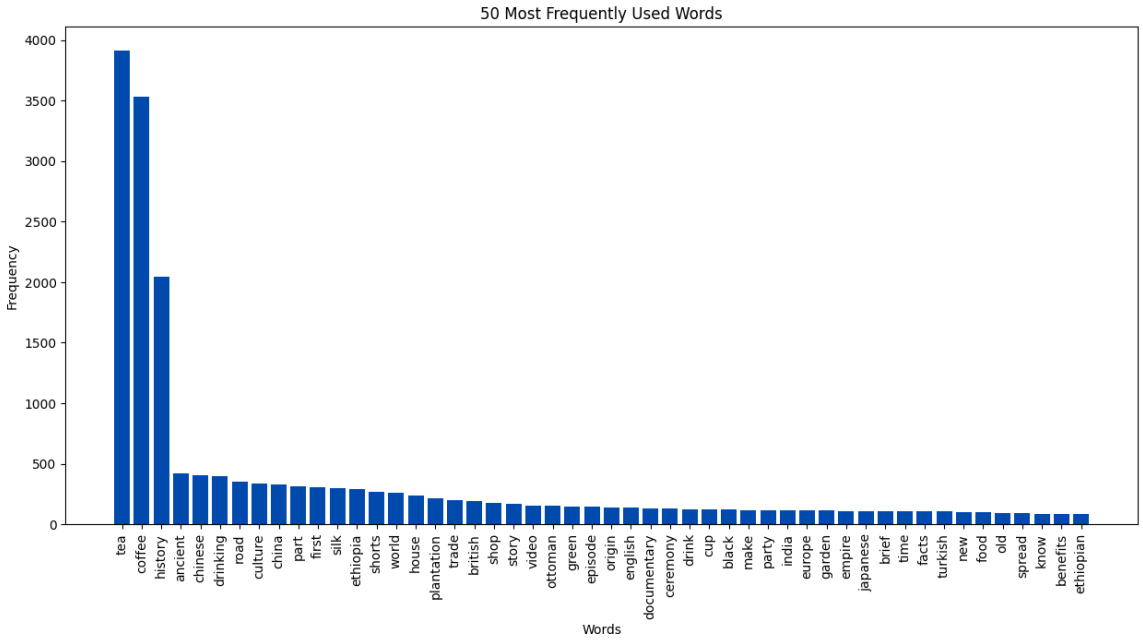


Figure 41: Word Frequency YouTube

The majority of words related to tea in the bar graph clearly show that, within the context of the YouTube data analyzed, tea is more popular than coffee. The word "tea" is referenced the most, indicating that there is more information focused on tea than coffee.

Figure 42: Topic analysis based on YouTube data

The appearance of other related names that are linked to the cultural, historical, and commercial aspects of tea, greater than any terms that might be related to coffee, further supports this dominance. This data offers strong evidence that tea is more noticeable and may even be more popular on this platform, which could be a reflection of larger patterns in social media debates between these two drinks

3.2.2 *When viewers* show interest and *why*

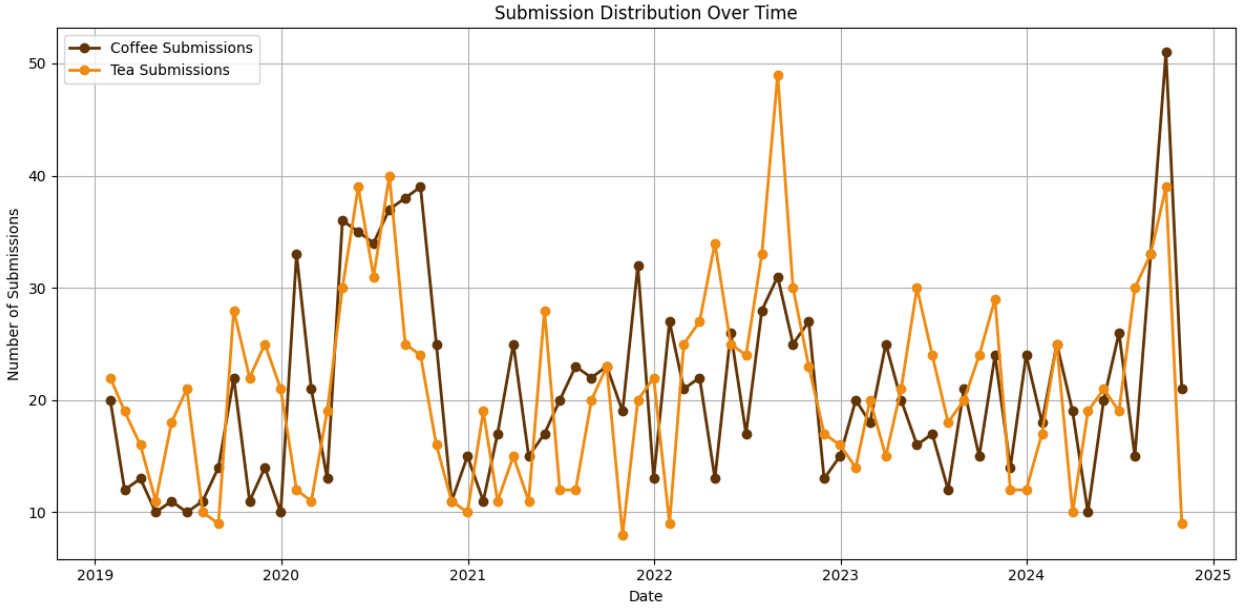


Figure 43: Submission distribution over time

The graph focuses attention to the variations and time trends in social media activity around these two beverages. While submission rates for both tea and coffee exhibit fluctuation, it is evident that tea submissions regularly outpace those for coffee in multiple periods, most notably in 2021 and 2024. This pattern might indicate a stronger or longer-lasting interest in tea during specific times, which could be impacted by commercial, seasonal, or cultural variables that encourage more content production and viewer interaction on YouTube.

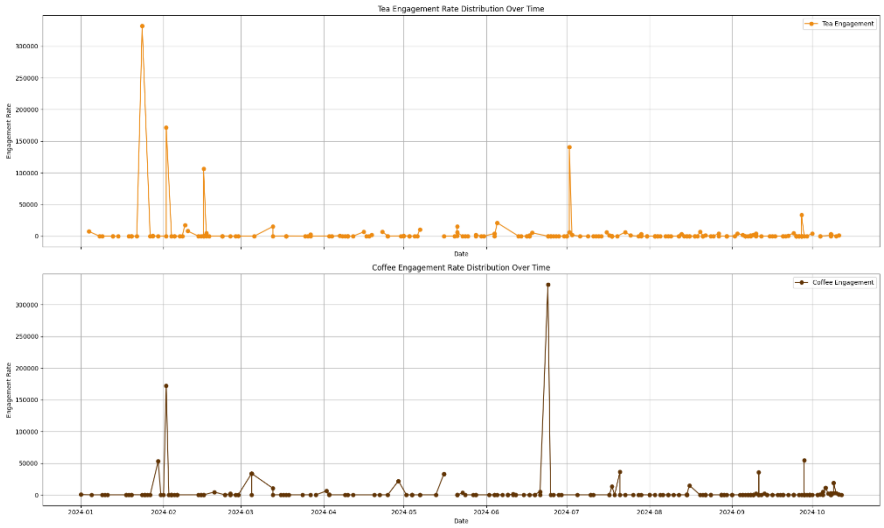


Figure 44: Engagement rate distribution over time

According to data from YouTube, the dual-line graph shows the distribution of interaction rates for tea and coffee over time over a few months in 2024. The engagement rate for tea is displayed in the upper panel; its irregular jumps indicate periodic high-interest occasions or tea-related viral content. The engagement rate for tea largely stays at a baseline with little variation, despite these sporadic jumps. As opposed to tea, the lower panel shows the engagement rate for coffee, which similarly shows periodic jumps but with fewer instances. Similar engagement patterns can be seen in both graphs, with brief spikes in activity interspersed with times of low activity. These spikes may be related to certain campaigns or chunks of content that suddenly attracted viewers' attention.

3.2.3 *How* they feel about tea and coffee

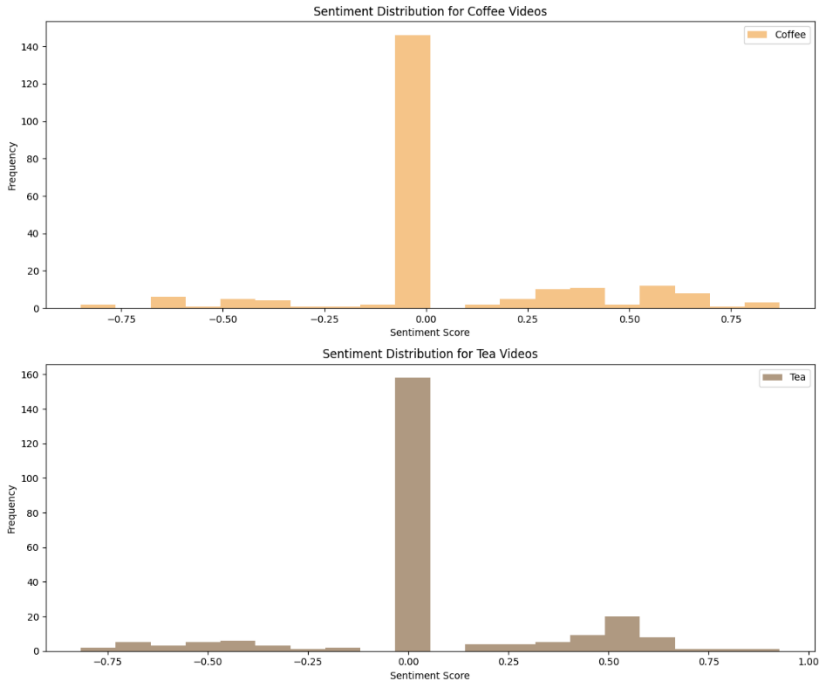


Figure 45: Sentiment distribution YouTube

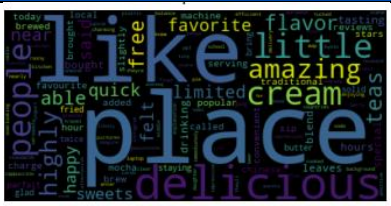
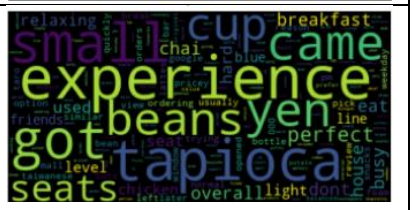
The sentiment distribution for tea and coffee, as determined by analysis of data from YouTube, is shown in the provided histograms. Most of the sentiment scores for coffee related videos are neutral, with fewer videos displaying either positive or negative extremes. These values cluster strongly around a score of zero. With a slightly wider range across positive and negative ratings, albeit still less frequently, the sentiment distribution for tea videos, in comparison, shows a significant crest at a neutral sentiment score.

These graphics are essential for showing that, while content about tea and coffee generally has a neutral tone, content about tea shows a somewhat greater range of sentiments. The fact that tea shows a wider range of emotions than coffee videos shows that tea is not only more popular than coffee, but also more treasured. These results contribute to understanding of the affective resonance of content on viewers and help us to leverage its higher impact on emotion as well as popularity than coffee.

3.3 Google Reviews

3.3.1 *What* do they love to talk about

Topic modelling results of the reviews suggest that the cafe goers care about the ambiance and service quality most. Matcha and bubble tea are popular among the reviewers when it comes to tea. Latte, espresso and black coffee are hot topics for coffee lovers. Deserts and sandwiches are also discussed in the reviews. Moreover, the results also indicate that the reviewers frequent these places during breakfast and lunch times.

Topic	Inferred summary	Word cloud
place like delicious little cream people amazing highly teas flavor free able favorite felt limited	The reviewers discuss the place, its people and taste of food. Cream, tea and flavor associated with the taste of beverage also appear together.	
staff really service taste menu best better location different quiet table thought options scones visited	Reviewers look for quality of service and staff at these places. Location and quiet suggest the ambiance are also important.	
experience tapioca small got came yen beans cup seats perfect busy chai breakfast dont eat	Reviewers talk about their experience and seating arrangements. Brea kfast places, tapioca suggest tea, beans suggest coffee. Yen indicates the topics have something to do with Japan and discussions around pricing.	
tea recommend want think price cute work cozy iced espresso excellent minutes wanted chocolate meal	Discuss both tea and coffee. Emphasis on ambiance and price. Minutes indicate service time of the places. Meal and chocolate suggest items on menu other than drinks.	
good great milk order cake love feel worth beautiful use customers times relax open stop	This topic suggests discussion on ambiance and positive customer experience. The places serve milk and cake.	

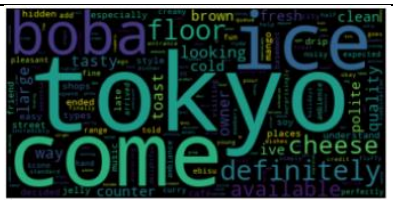
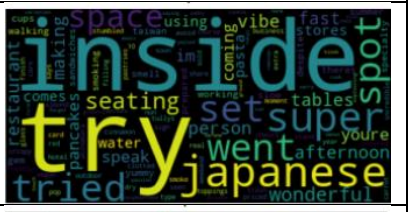
tokyo come ice boba definitely floor cheese available looking way tasty large quality owner toast	Tokyo suggests the reviews are for places in Tokyo. Boba or drinks with tapioca seems to be popular in Japan. The reviewers care about taste and quality. Iced, boba, cheese and toast also suggest menu preferences.	
nice time shop drink friendly latte ordered drinks quite area enjoy visit sugar sit lot	Reviewers care about staff attitude and associate it with positive experience. Latte and sugar come up, suggesting preference of sweet drinks.	
try inside japanese super spot tried went set space seating wonderful afternoon restaurant tables fast	The cafe interior and seating space are important. Afternoon and Japanese suggest Japanese lunch places.	
cafe store atmosphere matcha bit kind pretty hot lunch outside lovely bubble japan selection high	Preference of match and bubble tea. Hot suggesting discussion on weather. The reviewers emphasise on the atmosphere and service.	
coffee food sweet station located english make sweetness day black long recommended ve strong sandwich	In this topic, sweet and sandwich appears with strong, black, long, coffee suggesting which food is bought together with coffee. The reviewers also care about the location of the coffee place.	

Table 6: Topic analysis of google reviews data

3.3.2 *Where* their interests lay

The term frequencies suggest coffee is the most recurring term followed by tea. Tea appears more time in reviews from Tokyo, Taipei, Morrocco, Colombo and London. Coffee appears more in reviews from Istanbul, Dhaka, Cape Town and Hanoi. Contrary to expectations, cities like Melbourne, San Fransisco, and New York which are known to love coffee discuss tea more. Similarly, Delhi and Beijing, from India and China talk more about coffee than tea.

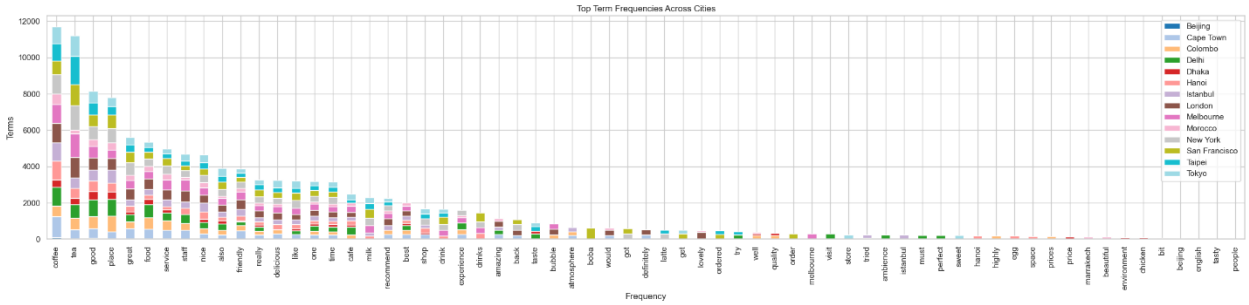


Figure 46: Term frequencies across cities

4 Conclusion

This report aims to resolve the long-standing debate on social media 'Tea or Coffee' - which is the superior drink? Using data from 3 social media sites we explore the communities, sentiments and topics around tea and coffee.

We dove deep into subreddit to identify the people that influence the tea and coffee communities. Our findings show the top influencer in coffee communities has more post engagement compared. They hold more influence in their community than the top tea influencer. This is explained by the outcome of community detection which shows tea has almost 3 times more communities than coffee. The communities are similar in strength with tea (Louvian modularity: 0.689) being slightly better than coffee (Louvian modularity: 0.645). Due to the close-knit structure, information moves faster within the community, but information spread from different communities may be slower.

Tea and coffee lovers are particularly passionate about their brewing method evidenced by the results of topic modelling on YouTube and Subreddit data. Tea drinkers are enthusiastic about their leaves and tea-ware. Coffee drinkers show similar enthusiasm about their roasts and grinders. Results from sentiment analysis on both YouTube and Reddit data show tea community expressing more positive sentiments compared to coffee community. The analysis on Google reviews shows people talk more about coffee from cafes than tea. However, based on subreddit tea community seem to have consistent engagement over time.

Coffee seems like a new*bean* in the beverage game whereas tea has had a much stronger, con*tea*nuous presence in the lives of people. So, to leave the debate to rest, we humbly put forward that, it is in fact tea that reigns superior . . . for now ;)

References

[1] (2024) *Coffee vs. tea: Which is better for your health?* Available at: https://www.washingtonpost.com/wellness/interactive/2022/coffee-vs-tea-nutrition-health/ (Accessed: 20 October 2024).

[2] Lang, A. (2019) *Coffee vs. tea: Is one healthier than the other?*, *Healthline*. Available at: https://www.healthline.com/nutrition/coffee-vs-tea (Accessed: 20 October 2024).

[3] *Coffee vs. tea: Which drink is healthier?* (2024) *Forbes*. Available at: https://www.forbes.com/health/nutrition/coffee-vs-tea/ (Accessed: 20 October 2024).

[4] *Tea vs Coffee: Which Drink Has Greater Health Benefits?* (2020) *Tea vs coffee: Which drink has greater health benefits?* Available at: https://www.teadrop.com.au/blogs/blog/tea-vs-coffee-which-drink-has-greater-health-benefits?srsltid=AfmBOooBFuVUgLzwpdgrmffQuJhZm2_M-kec6CA1arELVeF7uevNQH9I (Accessed: 20 October 2024).

[5] Admin (2023) *Australian Coffee Culture explained*, *Coffee For The People Roasting Co.* Available at: https://cftproastingco.com.au/australian-coffee-culture-explained/ (Accessed: 20 October 2024).

[6] *ARTBYMOGA*. Available at: https://artbymoga.com/ (Accessed: 20 October 2024).

[7] Samridhprasad (no date) *SAMRIDHPRASAD/Reddit-Analysis: Perform network analysis on reddit*, *GitHub*. Available at: https://github.com/samridhprasad/reddit-analysis (Accessed: 20 October 2024).

[8] vestland (2019) *Color map based on countries' frequency counts*, *Stack Overflow*. Available at: https://stackoverflow.com/questions/59297227/color-map-based-on-countries-frequency-counts (Accessed: 20 October 2024).

[9] "Place details" Google for Developers, viewed 20 October 2024, <https://developers.google.com/maps/documentation/places/web-service/details>.

[10] Published by M. Ridder and 22, M. (2024) *Global Coffee Consumption 2021/21*, *Statista*. Available at: https://www.statista.com/statistics/292595/global-coffee-consumption/ (Accessed: 20 October 2024).

Appendix A

I. Github repository (analysis code): https://github.com/Milindi-Kodikara/MugLife

II. Github repository (Github organisation and website): https://github.com/MugLife/MugLife.github.io

III. Kanban Board: https://github.com/users/Milindi-Kodikara/projects/3

IV. Reddit data sample

V. Google Reviews data sample

VI. YouTube data sample

Appendix B

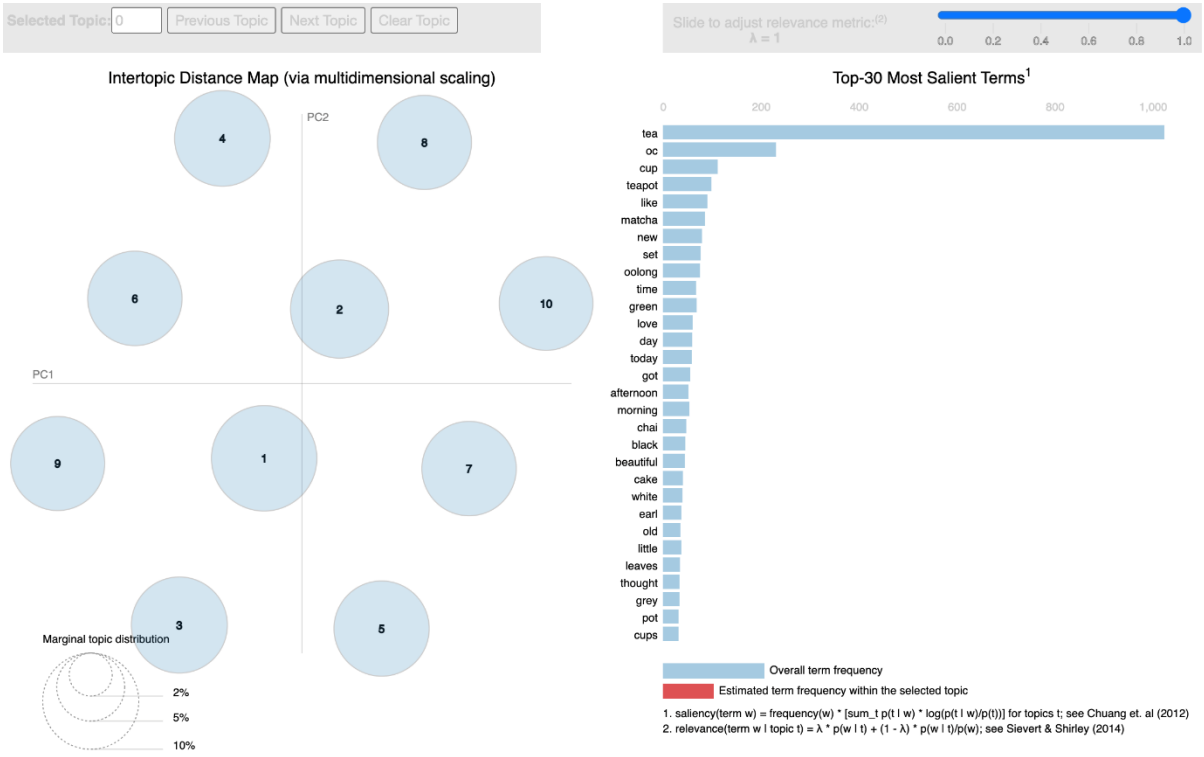


Figure B.1: Intertopic Distance Map and the top 30 most salient terms, tea

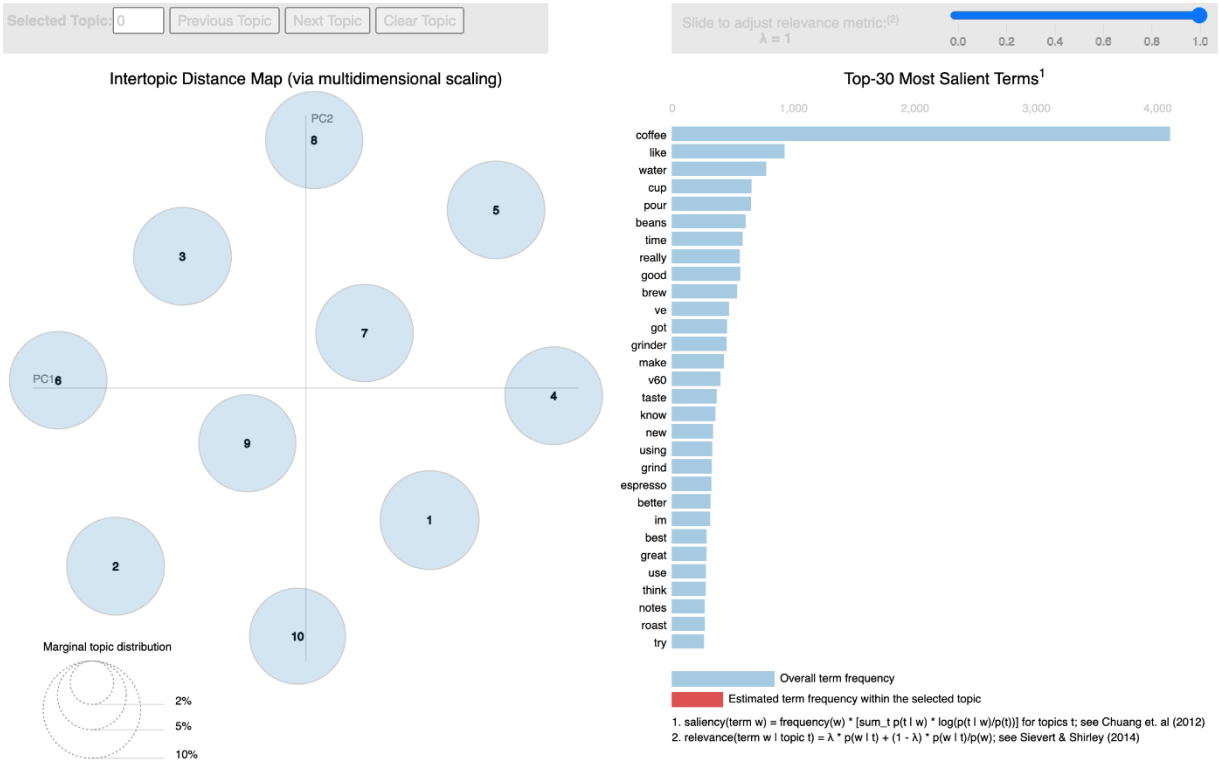


Figure B.2: Intertopic Distance Map and the top 30 most salient terms, coffee