Public Perception and Use of Hookah on Twitter in the US

Yiwei Han
Data Science
yhan32@u.rochester.edu

Puhua Ye
Data Science
pye3@u.rochester.edu

Mengwei Wu
Data Science
mwu52@ur.rochester.edu

Yuka Shimazaki
Data Science
yshimaza@u.rochester.edu

I. Introduction/Motivation

Nowadays, Hookah cafes have gained popularity over the whole world, including Middle East, Britain, France, Russia, and the United States. Hookah bars, cafes and restaurants are increasingly popular among young people, especially near universities [1]. According to American Lung Association, 79.6 % of current hookah users aged 12-17 say that they like using hookah because they can socialize in hookah lounges [2]. Furthermore, another main reason that hookah smoking is more popular is that people generally think hookah is safer way of tobacco than cigarettes [3].

Interested in understanding how people perceive hookah/waterpipe in general, we specifically looked at tweets published from March 2021 to March 2023. We want to look at deeper look at the tweets published in Twitter to understand the hookah perception, which will provide guidelines for future hookah regulation. We preprocessed and filtered the twitter data into commercial and noncommercial tweets using keywords. For both commercial and non-commercial tweets, we did topic modelling using BERTopic and LDA modelling to gain a comprehensive understanding of our tweets. Also, we did hookah brand frequency analysis on commercial and non-commercial tweets to figure out which hookah brands are most often mentioned in tweets. For non-commercial tweets, we first human labeled attitude and user or not columns to train our models of RoBERTa and Llama 2. Based on the results of the models, we visualized key insights to finding interesting stories with times series analysis, US maps, heatmaps, emoji analysis and topic modelling using BERTopic for positive attitude and negative attitude in non-commercial tweets. With our effective visualizations, we are able to understand any anomalies in the number of tweets, attitude rate for each state, the day and time when people write tweets about hookah, and popular emojis people use for different attitudes toward hookah.

In this project, we try to understand the public perception of Hookah and the use of Hookah on Twitter in the US and promotion strategies of Hookah products on Twitter. By understanding the hookah use, we can provide useful guidance for future hookah regulation and possible tobacco prevention campaigns.

II. Data Set Description

A. Description

The dataset is provided by the University of Rochester's Clinical & Translational Science Institute (CTSI) with the tweets coming from March 2021 – March 2023. And we mainly focus on the following variables: created_at, tweet text, user_id, and location of the tweet. Some other variables we also look at are quote, reply, retweet and favorite count, user_profile_image_url.

B. Data Cleaning

Our first step is to do data preprocessing, including removing duplicated tweets and irrelevant tweets. To filter out non-commercial tweets related to hookah in the US, we did data filtering based on the keywords related to US, hookah, and promotion. The result we got is that the total number of tweets in the US is 927,290, the total number of Hookah-related tweets is 323,347, the total number of non-commercial tweets is 299,544, and the total number of commercial tweets is 23,803, as seen in the table.

Total Tweets in the US	Total Tweets of Hookah-related	Total Tweets of Non-commercial	Total Tweets of Commercial
927,290	323,347	299,544	23,803

TABLE I
Data After Data Preprocessing

III. Exploratory Analysis

In order to have a deeper understanding of hookah on tweets, we analyzed the hookah brand on commercial tweets and non-commercial tweets, topic modeling of commercial tweets, and topic modeling of non-commercial tweets.

A. Hookah Brands

We first did hookah brand analysis based on the brand names and the associated flavors of each brand.

The bar charts show the frequency of hookah brands in both commercial and non-commercial tweets. We can notice that top 2 brands in both commercial and noncommercial tweets are Adalya and Starbuzz, with Adalya takes over half of the market share. The Top 3 brand for commercial Tweets is

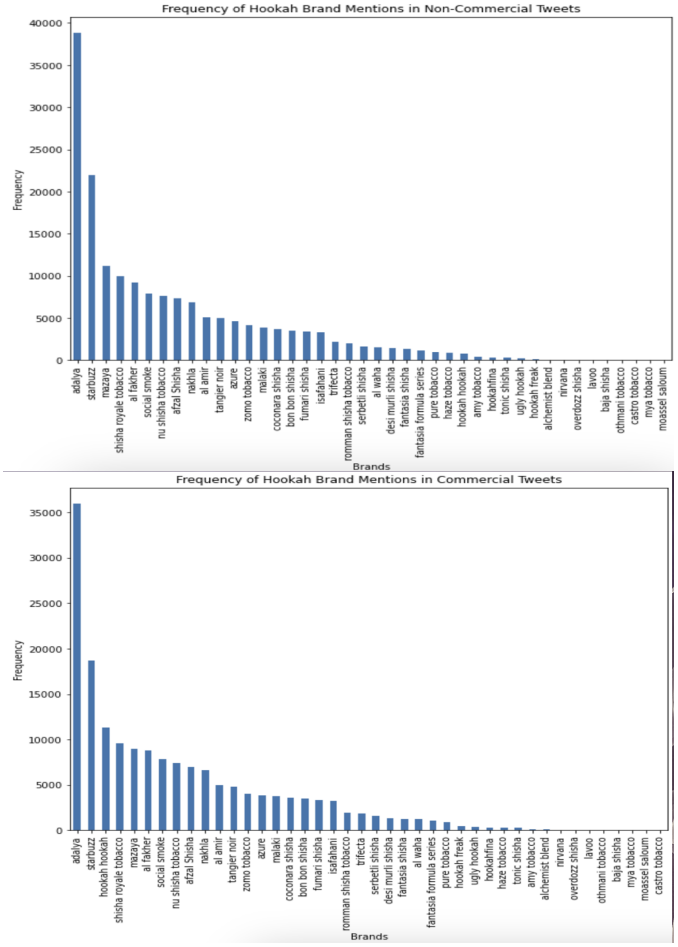

Fig. 1. Hookah Brand Frequency in Commercial and Non-Commercial Tweets

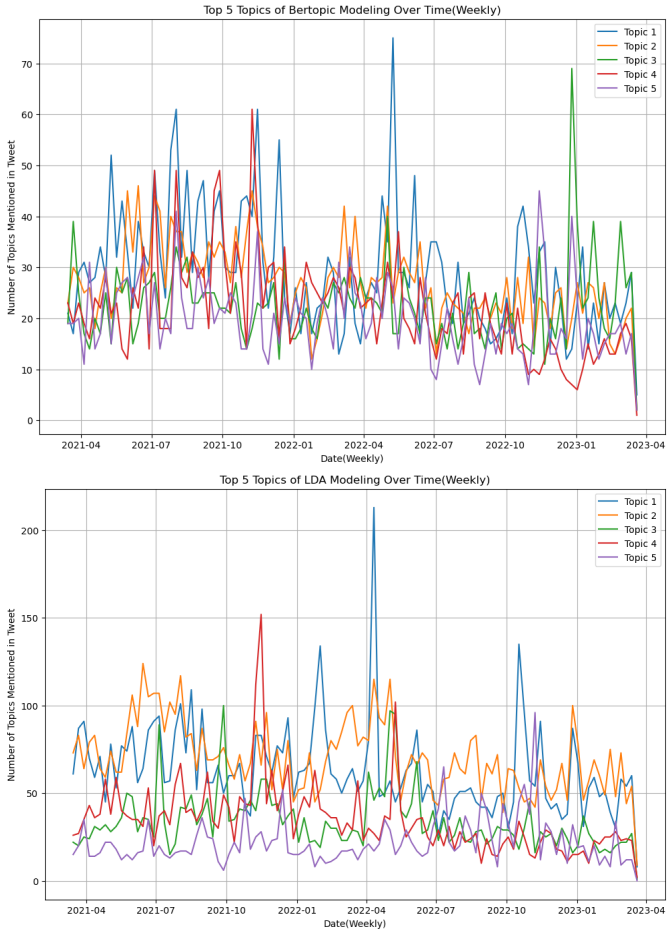

Fig. 2. Top 5 Topics of BERTopic and LDA Modelling for Commercial Tweets Time Series

Hookah Hookah, while this brand falls lot behind in the non-commercial Tweets, indicating Hookah Hookah does a lot of advertisements and promotions to attract customers.

B. Topic Modelling of Commercial Tweets

We used BERTopic and LDA models for topic modeling and analysis of commercial tweets, and we used Kmeans to clustering first before using BERTopic model.

1) Top 5 Topics of BERTopic and LDA Modelling Time Series Analysis: The time series plots depicts the number of tweets covering the top five topics over time each week after BERTopic modeling and LDA modeling respectively.

It can be seen from the two time series plots that it peaked in April 2022, and the topic with the peak in both plots is Topic1, which the keywords in both topics are about hookah, tonight, drinks, and food. This indicates a significant increase in commercial tweets about nightlife events in April 2022.

2) Top 5 Topics of BERTopic and LDA Modelling Analysis: Two tables in Figure.3 and Figure.4 list summary of each topic, the keywords of each topic, and the examples listed for each topic the top five topics after the BERTopic modeling and the LDA modeling. Each topic summary is summarized based on keywords and examples, indicating the topic classification included in each commercial tweets.

Topic Summary	Topic keywords	Example tweets in topic
The topic is related to nightlife events, which providing a combination of entertainment and dining.	free, night, tonight, bottles, hookah, shots, tuesday, food, tacos, drinks	Tonight ladies free by 11pm Ladies free hookah in groups $125 bottles till 11pm 🍾🍾 https://t.co/9ea3ab8gw1
This topic is clearly related to the sale of hookah-related products.	hookah, store, selling, niggas, sell, like, dollars, prices, nigga, paying	Dallas! Selling hookah cups and candy tips this weekend... #Dallas #hookah https://t.co/5Zwg07gRaa
This topic is associated with parties and night entertainment at Thursday, Friday, and Saturday, especially mentioned ladies have free admission.	party, ladies, happy, night, thursday, free, tonight, therapy, friday, saturday	📸 EVERY THURSDAY it's Hookah / Ladies night, PLEASE HONOR ALL CDC RULES AND GUIDELINES RELATED TO COVID-19 ————————... https://t.co/76pz1hz3Ec
This topic is about lounges and clubs that offer hookah, drinks and food, especially in cities like Atlanta and Miami.	lounge, hookah, atlanta, drinks, liquor, miami, food, lounges, club, like	Stop by the hookah lounge tonight.. Then slide for bit to eat $20 Smoke & Eat $50 All you can EDS https://t.co/Q3ndvcJtqd
This topic focuses on smoking-related products, including hookah, weed, and tobacco, as well as the sale or promotion of various flavored products.	smoke, smoking, tobacco, hookah, vape, flavored, products, weed, flavor, flavors	Smoke Shop (Black) Swooper Feather Advertising Flag & Pole Kit – Perfect for Smoke Shops, Vape Stores, Hookah Spots https://t.co/RWPiWOepnO

Fig. 3. Table for Top 5 Topics of BERTopic Modelling for Commercial Tweets

C. Non-Commercial Tweets

We also did topic modelling by using BERTopic and LDA models in Non-Commecial Tweets.

Topic Summary	Topic keywords	Example tweets in topic
This topic revolves around night events with Hookah, especially with free admission for ladies.	free, hookah, tonight, night, ladies, open, specials, drink, drinks, entry	Ladies free hookah in groups $125 bottles till 11pm 🍾🍾 https://t.co/9ea3ab8gw1
This topic is related to selling hookah stores.	hookah, smoke, like, store, shisha, lounge, smoking, need, sell, good	Selling my hookah, coals (box of 100), hookah coal burner, and shisha. Make me an offer https://t.co/zb52dVChOm
This topic is related to the promotion of waterpipe types and new products. It's about the innovation of new flavors and the material of water pipes.	hookah, waterpipe, available, party, glass, durable, engineered, smoothest, revolutionary, pipe	@QAce420 Engineered to be the smoothest, most durable, and revolutionary waterpipe available!... https://t.co/EYarDApxxb
The topic focuses on the customer experience, using posted photos to emphasize the love for hookah and the flavor of hookah to promote hookah and its flavors.	hookah, come, happy, flavor, posted, photo, love, perfect, restaurant, shop	Just posted a photo @ Mangos Bar, Hookah & Restaurant https://t.co/qFi8Mwj7DV
This topic focuses on the promotion of hookah lounge party events.	hookah, sunday, lounge, come, food, drinks, music, back, party, link	$20 mint hookah ALL night $5 shots til 11pm! @HollywoodLounge Tonight!!! https://t.co/mmTRimlvL0

Fig. 4. Table for Top 5 Topics of LDA Modelling for Commercial Tweets

Figure.5 and Figure.6 provides the keywords in each topic of top five topics after BERTopic modeling and LDA modeling, then we can know descriptions for each topic.

```
Topic 1: smoke, smoking, weed, hookah, like, smoked, cigarettes, tobacco, shit, people
Topic 2: music, food, drinks, vibes, free, hookah, good, vibe, party, night
Topic 3: hookah, like, need, love, make, pass, lmao, know, shit, really
Topic 4: hookah, spot, house, tonight, time, home, need, night, bought, today
Topic 5: hookah, light, girl, head, like, tips, broke, hate, bitch, hitting
```

Fig. 5. Top 5 Topics of BERTopic Modelling for Non-Commecial Tweets

```
Topic 1: shisha, hookah, lounge, tobacco, shooting, flavor, shot, mint, dead, pipe
Topic 2: hookah, smoking, like, smoke, niggas, love, shit, people, really, nigga
Topic 3: hookah, brunch, drink, specials, night, bottle, last, mimosas, stop, available
Topic 4: hookah, smoke, like, need, wine, good, home, time, make, drink
Topic 5: hookah, lounge, food, hooka, music, drinks, tonight, free, good, like
```

Fig. 6. Top 5 Topics of LDA Modelling for Non-Commecial Tweets

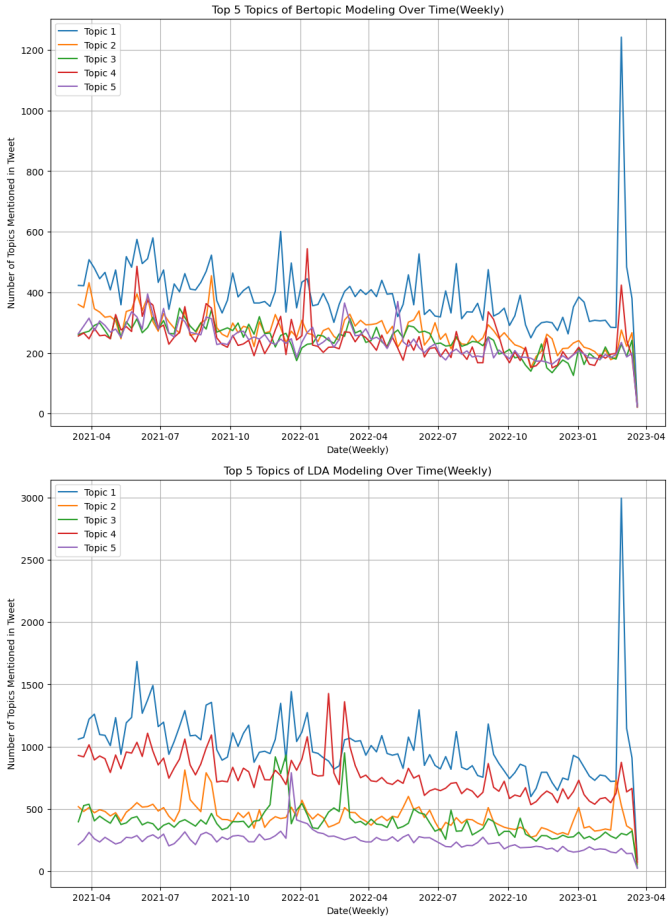

Fig. 7. Top 5 Topics of BERTopic and LDA Modelling for Non-Commercial Tweets Time Series

The time series plots portray the volume of non-commercial tweets per week containing the top five topics that's topic modeled by the BERTopic model and the LDA model. Among them, it is clearly noticed from both figures that there was a significant surge in the number of non-commercial tweets covering all topics, which also led to a corresponding peak in March 2023. It can be noticed that topic1 remains dominant over time in these two plots, so topic1 obviously have peaks in both plots, which contains keywords related to smoke, hookah, tobacco, and shisha. This shows that during this time, there was an increase in non-commercial activities related to shisha and hookah.

IV. Model Development

A. Human Labelling

The initial step in our model development was the creation of a high-quality human labeling training dataset. The accuracy and reliability of the labeled data directly influence the performance and validity of our models. We initiated the process with working together to label a set of 100 tweets. This primary objective was to develop a mutual understanding of the labeling criteria and ensure consistency in identifying attitude towards hookah and user status. Following the initial labeling, we independently labelled 300 tweets. However, the Kappa statistic, a measure of inter-rater agreement that assess the consistency of our labeling process, was below 0.7, indicating a need for further calibration. So, we discussed collectively, focusing on the differences in those 300 tweets, ensuring a common understanding among us. After that, we independently labelled different 300 tweets again and the Kappa score was above 0.7, which represents a high degree of agreement between us on the attitude and user status of tweets towards hookah. Then, each of us labelled 400 different tweets. Finally, our training dataset consisted of 2300 human-labelled tweets and each tweet was categorized based on its attitude towards hookah (positive, neutral, or negative) and the user status (user or non-user).

B. Data Engineering

In the context of our project, data engineering was a critical phase. After we trained the model using a dataset consisting of 2300 tweets, we realized that the model was not as good as we expected. So we hope that we can expand the dataset by means of data augmentation to address the limitations of our initial dataset. We used synonym replacement technique, which involved altering certain words in tweets with their synonyms, and back translation through French, aimed at generating linguistically varied but semantically consistent ver-

sions of the original tweets. These steps were instrumental in enhancing the dataset's diversity, enabling our models to learn from a wider range of language use. After data augmentation, our dataset has a total of 14,365 pieces of data, of which 9,046 are POSITIVE, 2,916 are NEUTRAL, and 2,403 are NEGATIVE as shown in TABLE II.

Dataset	Total Tweets	POSITIVE Count	NEUTRAL Count	NEGATIVE Count
Original	2,300	1,128	856	316
After Augmentation	14,365	9,046	2,916	2,403

TABLE II
DATA AUGMENTATION COMPARISON

Next, recognizing the inherent class imbalance in our dataset, we employed under-sampling techniques. This approach was critical to ensure that the model did not develop a bias towards the more frequently occurring classes. The reason why oversampling was not chosen is that we think that the difference between the number of POSITIVE and the number of NEUTRAL and NEGATIVE categories is too large. Using oversampling would have resulted in having too many duplicates of NEUTRAL and NEGATIVE data, which would have resulted in overfitting the model. By creating a more balanced training dataset, we enhanced the model's ability to generalize and accurately classify tweets across different categories. After under-sampling, our training dataset has a total of 7303 pieces of data, of which 2,500 are POSITIVE, 2,634 are NEUTRAL, and 2,169 are NEGATIVE and the evaluation dataset has a total of 1437 pieces of data, of which 921 are POSITIVE, 282 are NEUTRAL, and 234 are NEGATIVE as shown in TABLE III.

Dataset	Total Tweets	POSITIVE Count	NEUTRAL Count	NEGATIVE Count
Augmented Dataset	14,365	9,046	2,916	2,403
After Under-sampling	7,303	2,500	2,634	2,169

TABLE III
DATA UNDER-SAMPLING COMPARISON

C. Roberta Model

In our project, we firstly developed a model based on RoBERTa which known for its proficiency in handling complex language data. RoBERTa is a pre-trained language model, which structure is based on the encoder part of transformer model, that can be fine-tuned on specific tasks, where it generates embedding that can contained all the information in the sentence and can be utilized in further model layers for prediction or classification. The structure of our RoBERTa model was intricately designed to effectively classify tweets. It included an input layer accept the tokenized tweet through specific RoBERTa tokenizer, a core RoBERTa layer utilizing for context comprehension, a pre-classifier for preliminary processing to optimize and prepare the data features so that the final classification layer can make more effective decisions, ReLU activation for introducing non-linearity, a dropout layer to prevent overfitting, and a final classifier that determined the attitude and user status based on the processed data.

For the classification of attitude, we used the augmented dataset to ensure the best performance, but for the classification of user status, because it is an easier task than attitude classification, we used the original dataset to reduce the cost of training. Through fine-tuned hyperparameters of our model through grid-search, the optimizing key parameters of our RoBERTa model are learning rate of 1e-5, train batch size of 4, valid batch size of 4, train epoch for attitude classification task of 15, and train epoch for user status classification task of 5.

Despite employing an augmented dataset and modifying the initial model by incorporating L2 regularization into the loss function, as well as revising the overall architecture through the integration of diverse network structures, the performance of the model still falls short of our anticipated objectives. The best performance of our model is reflected in an accuracy of 71.79% and an F1 score of 0.69, which does not meet our requirement for an accuracy rate exceeding 80%. This led us to explore the Llama2 model, one of the strongest open-source large language models, a category of tools that are highly potent in handling NLP tasks, of today.

Result	Accuracy	F1 Score
Attitude	71.79%	0.62
User_or_Not	79.78%	0.8

TABLE IV
RESULT OF ROBERTA MODEL

D. Llama2 Model

In our quest to analyze social media texts, the development of the Llama2 model marked a significant leap forward. Llama2 is an auto-regressive transformer-based language model, which is specifically designed for complex, context-rich language data, making it an ideal fit for our project. It comes in different versions with varying parameters: 7 billion, 13 billion, and 70 billion. We used the smallest version with 7 billion parameters. Llama 2 excels in various benchmarks like AGIEval, MMLU, and Winogrande, surpassing other models in reasoning and comprehension.

For fine-tuning the Llama2 model, we employed LoRA (Low-Rank Adaptation) in Parameter-Efficient Fine-Tuning (PEFT) methods. PEFT is a technique used in NLP to improve the performance of pre-trained language models on specific downstream tasks by fine-tuning a small subset of the model's parameters. This approach is beneficial for adapting models to specific tasks with less computational overhead and fewer

labeled examples. LoRA, as a PEFT technique, works by approximating the weight matrix updates in a network using a product of two smaller matrices as shown in Fig.8, thereby significantly reducing the number of parameters that need to be updated. [4] This method allows the original pre-trained weights of the model to remain unchanged while adapting the model to specific tasks, making the process more efficient and resource-friendly.

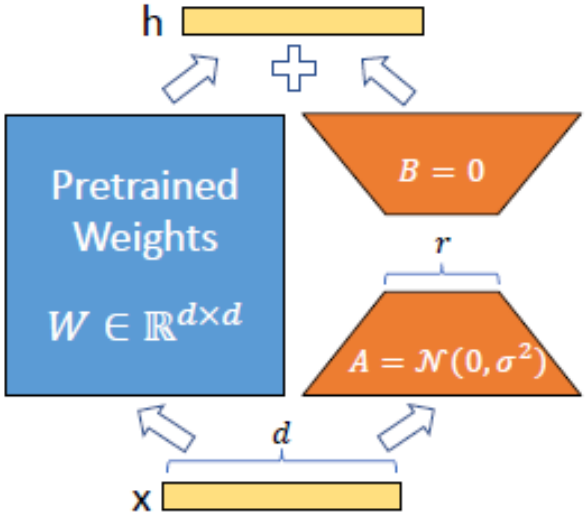

Fig. 8. reparametrization of LoRA (Source: [4])

In order to further improve the model performance, we integrated NEFTune, a novel approach to improving the fine-tuning of language models, which has shown significant improvements in model performance. By introducing noise to the embedding vectors during training, NEFTune enhances conversational ability and answer quality of models. As shown in Fig.9, in evaluations like AlpacaEval, NEFTune demonstrated an average increase of 15.1% across various datasets, indicating a substantial impact on the quality of language model outputs. [5] In our project, the application of NEFTune resulted in a remarkable improvement on accuracy of classifying attitude from 74.4% to 81.8%.

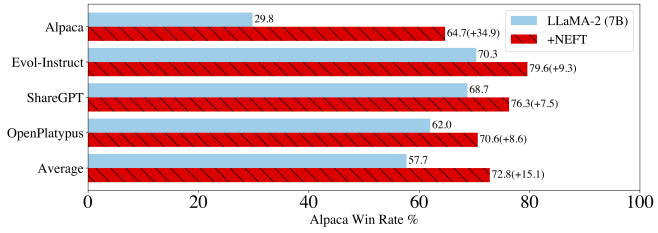

Fig. 9. Improvement by NEFTune (source: [5])

For the classification of attitude, we still used the augmented dataset to ensure the best performance. For the classification of user status, we used the original dataset to reduce the cost of training. Through tried different combinations of hyperparameters, the optimizing key parameters of our Llama 2 model are lora_alpha of 16, lora_dropout of 0.1, num_train_epochs of 20 for attitude classification and 10 for user status classification, gradient_accumulation_steps of 8, Weight_decay of 0.001, lr_scheduler_type of "cosine", and neftune_noise_alpha of 10.

Our Llama 2 model demonstrates commendable performance across various metrics. With an accuracy of xx% on attitude and xx% on user status, it efficiently identifies the correct outcomes in a significant majority of cases, indicating its overall reliability. The model's precision rate stands at xx% on attitude and xx% on user status, reflecting its effectiveness in producing a low rate of false positives. This is complemented by a recall rate of xx% on attitude and xx% on user status, illustrating the model's capability in correctly identifying all relevant instances. Furthermore, the F1 score, a balanced measure of precision and recall, is recorded at xx% on attitude and xx% on user status. This score signifies the model's robustness, combining precision and recall into a single metric. Overall, these metrics collectively affirm the model's competence in handling the classification tasks, balancing accuracy with a mindful consideration of both type I and type II errors.

Result	Accuracy	F1 Score	Precision	Recall Rate
Attitude	81.1%	0.816	76%	81%
User_or_Not	86.3%	0.863	86%	86%

TABLE V
RESULT OF LLAMA2 MODEL

V. PERFORMANCE AND RESULTS

A. User and Attitude Distribution Analysis

The pie charts presented compare the distribution of attitudes labeled as user and not user across three categories: positive, neutral, and negative. In the 'User Attitudes' chart, the overwhelming majority, 82.0%, exhibit a positive sentiment, while neutral and negative attitudes account for only 10.5% and 7.5%, respectively. This suggests a highly favorable perception among users. In contrast, the 'Not User Attitudes' chart depicts a more balanced sentiment distribution, with 43.4% neutral, 29.6% positive, and 27.0% negative. This balance indicates a more varied perception from non-users, with less pronounced positive sentiment and significantly more neutral and negative sentiments than users. The significant neutral and negative sentiments among non-users could reflect a lack of information, preconceived notions, health concerns, or social disapproval related to hookah. The positive sentiment present among nearly a third of non-users might be influenced by cultural perceptions, social acceptance in peer groups, or curiosity about hookah.

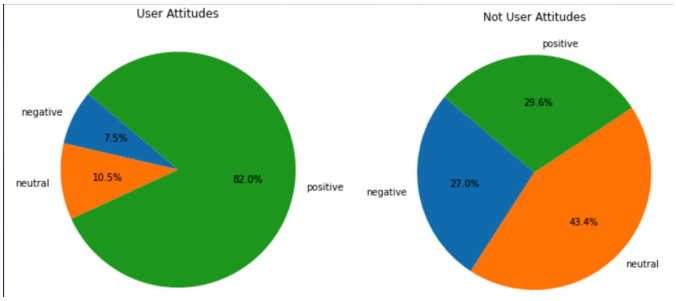

Fig. 10. User and Attitude Distribution Pie Chart

B. Number of Tweets Time Series Analysis

The time series plots depict weekly tweet volumes, categorizing them into total, non-commercial, and commercial tweets. There is a notable spike in the volume of non-commercial tweets in March 2023, which also causes a corresponding peak in the total number of tweets. This anomaly suggests a significant event or series of events that captured public attention, leading to increased Twitter activity, particularly within the non-commercial sector.

The surge in non-commercial tweets during this period could be attributed to public incidents and health discussions that put hookah use in the spotlight. The predominance of non-commercial tweets in the discourse suggests that the conversation was likely driven by individual users rather than corporate entities or commercial interests. For example, a notable incident in Detroit involving carbon monoxide poisoning from hookah use could have spurred widespread concern and discourse, as the dangers of carbon monoxide are well-documented and such events can lead to a public re-evaluation of the risks associated with hookah smoking [6]. Additionally, reports of violence and criminal activities in hookah lounges might have raised questions about the safety and regulation of such establishments. This could lead to a broader discussion on social media about the societal impact of hookah lounges, contributing to the tweet volume. Another factor possibly influencing the peak could be a heightened awareness and discussion of the general health hazards of hookah smoking, particularly the implications of second-hand smoke. Such public health discussions can resonate widely on social media platforms, as individuals share personal experiences, warnings, and health information.

C. Weekly Sentiment Time Series Analysis

The time series plots provided offer a comprehensive view of weekly sentiment trends related to hookah usage on Twitter, delineated by positive, negative, and neutral sentiments both in proportion and absolute count. The data reveals a consistent distribution of sentiment over time, with positive sentiments generally dominating. However, a significant peak in negative sentiment is observed in March 2023, both in terms of proportion and count, corresponding with a surge in the overall number of tweets.

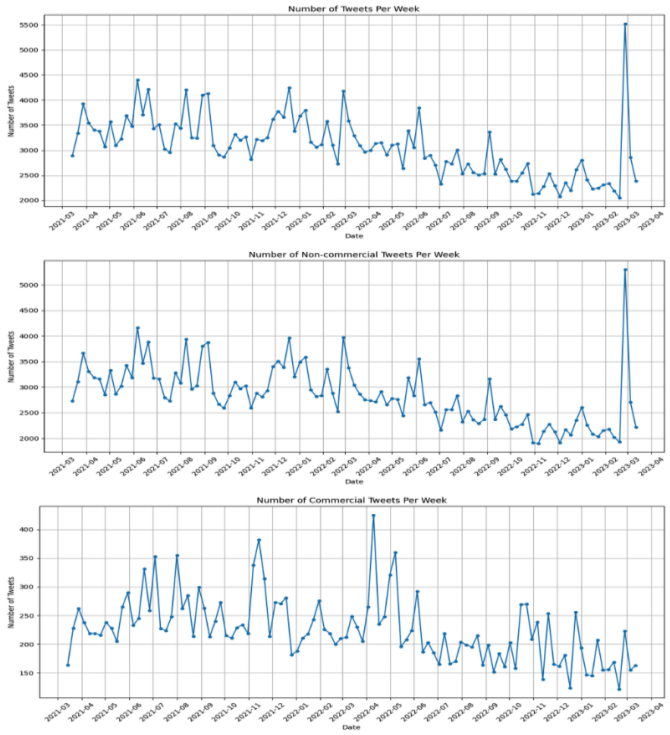

Fig. 11. Time Series Plots of Number of Tweets per Week (Top: Total Number of Tweets; Middle: Total Number of Non-Commercial Tweets; Bottom: Total Number of Commercial Tweets)

This anomalous spike in negative sentiment could be linked to specific incidents reported during this time, which aligns perfectly with the analysis of the previous plot.

D. Number of Unique User Time Series Analysis

The time series plots chart the weekly counts of unique Twitter users engaged in discussions about hookah, differentiated into total users (top plot), users who are considered 'users' of hookah (middle plot), and those classified as 'non-users' (bottom plot). The data trends are relatively stable over time with noticeable fluctuations. However, all categories exhibit a pronounced peak in March 2023, which signifies an anomalous increase in engagement from both users and non-users.

This spike in unique users corresponds with the previously noted surge in tweet volume and negative sentiment during the same period. Such a synchronicity in the data suggests that the events reported in March 2023—such as the hospitalization incident in Detroit due to carbon monoxide poisoning, the violence linked to hookah lounges, and the amplified discussions on health hazards—had a widespread impact, drawing in a larger and more diverse group of Twitter users into the conversation.

The steady number of unique users and non-users over time suggests a consistent level of baseline interest and engagement with hookah-related topics on social media. However, the decreasing trend in the number of unique users who are identified as 'users' might indicate a gradual decline in hookah usage or in the discussion thereof among this group, possibly

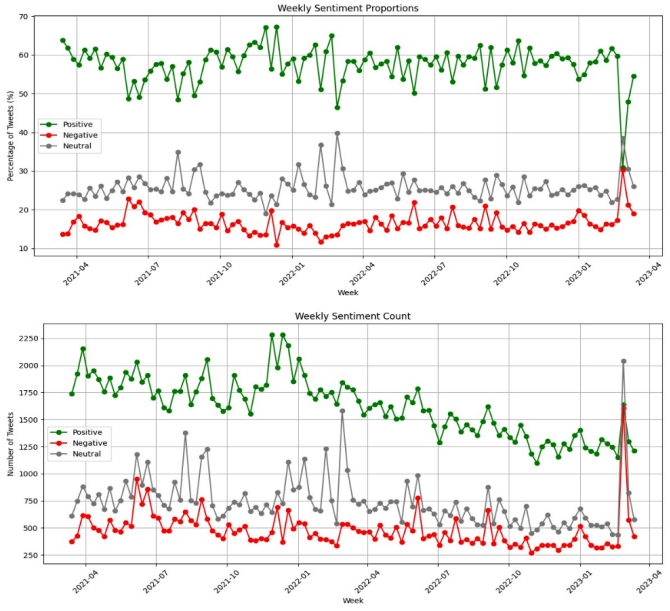

Fig. 12. Time Series Plots of Weekly Sentiment (Top: Proportion Plot; Bottom: Count Plot)

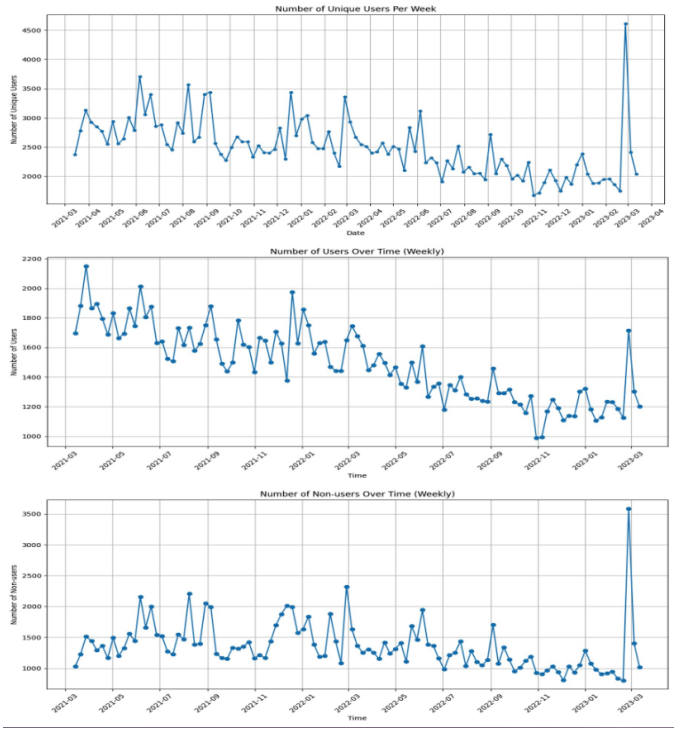

Fig. 13. Time Series Plots of Number of Unique Users per Week (Top: Number of Total Unique Users (Users + Non-users); Middle: Total Number of Unique Users Only; Bottom: Total Number of Unique Non-users only)

due to changing social habits, increased awareness of health risks, or other social dynamics not captured in this dataset.

E. Positive Attitude Rate per State

The map visualizes the positive rate towards hookah use as expressed in tweets from various states in the United States. The data represented by varying shades of blue indicates the percentage of positive sentiment in each state's Twitter discourse concerning hookah. States like Arkansas (0.6886) show the highest positive rates, suggesting a more favorable or accepting attitude towards hookah in these regions. Conversely, states such as New Hampshire (0.3056) exhibit the lowest positive rate, which could reflect a more critical or less enthusiastic perspective on hookah use.

It is noteworthy that the states with higher urban populations, like New York (0.5394) and California (0.5056), have positive rates that are not among the highest. This might indicate that while hookah might be popular in urban areas, the sentiment in conversations about it on social media does not necessarily skew overwhelmingly positive. Moreover, we can see that states in the south and east coast part of US tend to have a higher positive rate than the west and north part.

The varying positive rates across the states could be influenced by cultural, social, and legislative factors. For instance, states with stricter public health policies regarding smoking might see lower positive sentiment. In contrast, states with a vibrant nightlife and cultural scenes that include hookah bars and lounges may have higher positive rates.

F. Normalized Hookah User Count per State

The map presents a normalized representation of hookah users across the United States, adjusted for the population size

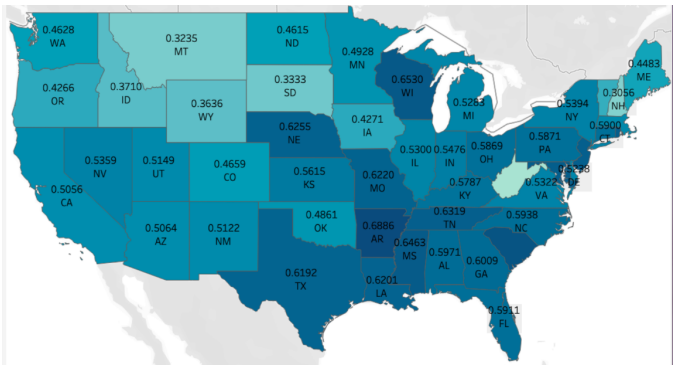

Fig. 14. Map Visualization of Positive Attitude Rate per State

of each state. This normalization allows for an equitable comparison by accounting for state-by-state population variance. Remarkably, New York stands out with the highest normalized user count of 9.866, which could suggest a significantly higher prevalence or cultural acceptance of hookah use within its population. In stark contrast, several states, such as Idaho (0.003) and New Mexico (0.003), show minimal normalized user counts, indicating that hookah use is comparatively uncommon or less culturally embedded in these regions.

The data reveals a heterogeneous distribution of hookah users, with some states like Florida (4.000) and Wisconsin (0.789) also exhibiting relatively high normalized counts. This may reflect regional cultural nuances, the presence of hookah-friendly laws, or demographic trends that favor hookah con-

sumption. Conversely, the lower normalized counts in many midwestern and northwestern states could be attributed to cultural preferences, stricter regulations, or less exposure to hookah as a social activity.

Align with the map visualization of positive attitude rate per state, we can generally see the pattern that as the south and east part of US has a higher positive rate toward hookah, it also has a generally higher unique hookah user count, indicating a positive relationship between the two.

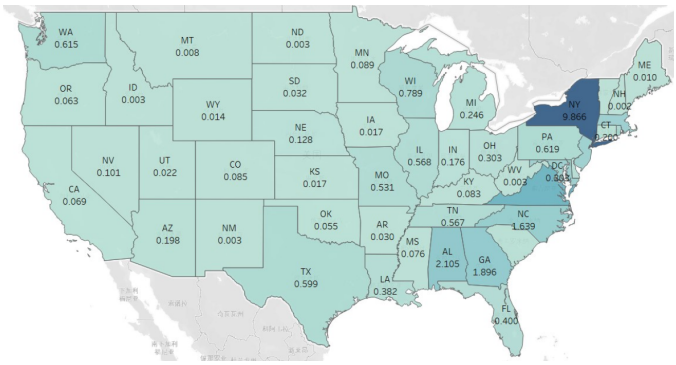

Fig. 15. Map Visualization of Normalized Hookah User Count per State

G. Positive Attitude Rate Across Day and Time

The provided heat map visualizes the variation in positive attitudes towards hookah across different days of the week and hours of the day, segregated into overall positive rates, non-user only positive rates, and user only positive rates. From the top heatmap, we can discern that, generally, the overall positive sentiment towards hookah increases during the evening and peaks at night, with the highest positivity occurring on weekends. This pattern suggests a correlation between leisure time and favorable attitudes toward hookah use.

In the middle heatmap, which focuses on non-users, there is an interesting divergence where positive sentiment is more pronounced during the daytime of weekdays, peaking on Wednesday. It suggests that non-users might associate hookah with a daytime social activity, or perhaps there is a particular event or social discussion occurring mid-week that influences this positivity. Contrastingly, during the nights of weekends, non-users exhibit the least positive attitudes, which might be associated with a perception of hookah as a late-night activity that they do not partake in or possibly due to negative experiences or exposure to second-hand smoke in social settings.

The bottom heatmap shows that users' positive sentiments towards hookah are strongest during the night-time of weekends, with Saturday being particularly notable. This trend aligns with the social habits of hookah consumption, where users are likely engaging in the activity during their free time, leading to increased enjoyment and positive attitudes. During the daytime on weekdays, users' positivity is lower, possibly because this is not a typical time for hookah use due to work or other daily responsibilities, which contrasts with their weekend behavior.

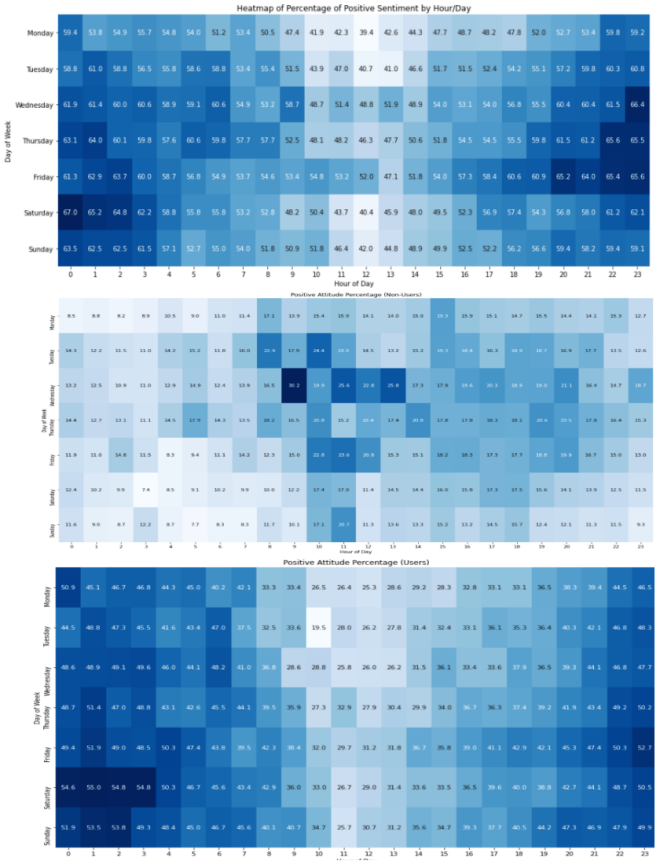

Fig. 16. Heat Map Visualization of Positive Attitude Rate Across Day and Time (Top: Total Positive Rate; Middle: Non-user Only Positive Rate; Bottom: User Only Positive Rate)

H. Top10 Emoji Count for Each Category of Tweets

The table provides a categorization of the top ten emojis used in tweets, segmented by the sentiment of the message—positive, neutral, and negative—as well as their usage in commercial tweets. Upon examination, it becomes apparent that the emojis associated with positive sentiments, such as the 'Face with Tears of Joy' and 'Loudly Crying Face', suggest expressive responses likely tied to content that evokes strong emotional reactions, with the 'Face with Tears of Joy' being the most prevalent across all categories. This indicates its versatility in communication, transcending the boundaries of sentiment to convey intensity of feeling.

The similarity in emoji usage between neutral, negative, and commercial tweets is notable, with emojis like the 'Thinking Face' and 'Red Triangle Pointed Right' appearing in all three categories. This could imply a commonality in the type of engagement these tweets are designed to provoke—perhaps a call to action or a prompting of consideration. The 'Skull' emoji, traditionally associated with dark humor or failure, features prominently in both negative and commercial contexts, which might suggest an attempt by brands to engage with trends or memes that resonate with consumers' experiences of frustration or disappointment.

In the positive category, emojis convey joy, laughter, and love, while the negative category includes symbols of sadness or discontent. Commercial tweets, interestingly, share emojis with both positive and negative connotations, perhaps reflecting a marketing strategy that seeks to engage a wide range of emotions to connect with a diverse audience.

	Positive		Neutral		Negative		Commercial	
	Emoji	Count	Emoji	Count	Emoji	Count	Emoji	Count
1	😂	22187	😂	7376	😂	20409	😂	20202
2	😭	14829	😭	5668	😭	11090	😭	10978
3	💨	7304	🤣	1590	🤣	6616	🤣	6566
4	🤣	6142	😐	1086	🚩	2275	🚩	2274
5	🔥	5286	😩	952	💀	1296	💀	1287
6	😩	4976	🚩	898	😩	1006	😩	1001
7	😍	4958	💀	883	😐	877	😐	875
8	🍾	4521	🚾	642	🙄	866	!!	858
9	😻	3222	❌	608	🚾	772	🤦	765
10	!!	3127	🤦	600	🤦	769	🚾	765

Fig. 17. Table Visualization for Top10 Emoji Count for Each Category of Tweets

I. Topic Modeling for Positive and Negative Attitude Tweets

Figure 16 illustrates the distribution of topics within negative non-commercial tweets, revealing a significant concentration on a single theme. Topic 0, which encompasses keywords such as ”hookah,” ”smoke,” ”smoking,” and terms of exasperation or negative sentiment, dominates the discourse with a staggering 96.1% of the frequency. This overwhelming focus suggests that the conversation is largely centered around the act of smoking hookah and is associated with negative experiences or viewpoints.

The remaining topics—Topics 1 through 4—account for a marginal proportion of the discussion, with percentages ranging from 3.1% to 0.1%. These topics include a variety of terms, some of which are colloquial or slang, indicating specific subcultures or contexts within the broader negative sentiment toward hookah. Topic 2, for example, includes ”shisha,” ”tobacco,” and ”cigarette,” suggesting discussions that may relate to comparisons or conflations between hookah smoking and other forms of tobacco use. Topic 4 includes terms such as ”doncic,” ”crazy,” and ”crying,” which might be referencing specific events or public figures that have entered the conversation in a tangential or metaphorical manner. The data suggests that the negative sentiment is not diffuse but rather concentrated on particular aspects or incidents related to hookah.

Figure 17 depicts the distribution of topics within positive non-commercial tweets, demonstrating a dominant theme among discussions. Topic 0, which includes keywords such as ”hookah,” ”smoke,” ”lounge,” and related terms indicative of social and leisure activities, such as ”tonight,” ”drinks,” and

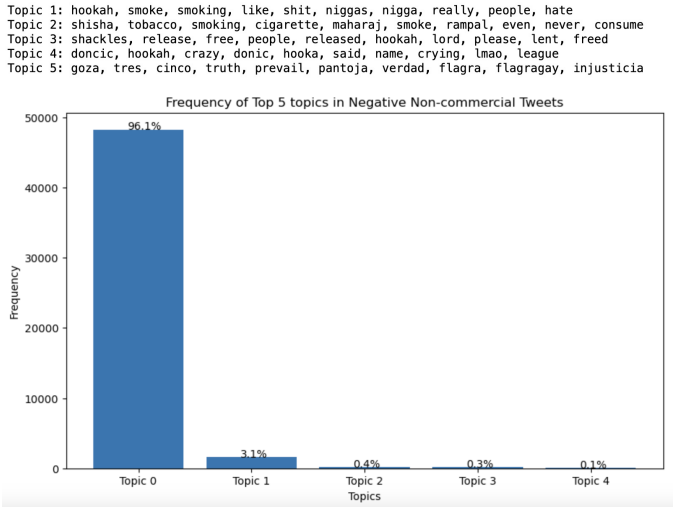

Fig. 18. Frequency of Top5 Topics for Negative Attitude Tweets

”food,” overwhelmingly represents the bulk of the conversation, accounting for 92.7% of the frequency. This significant concentration suggests that the positive sentiment is closely tied to the social aspect of hookah use, possibly reflecting its integration into social gatherings and events where food and beverages are enjoyed.

The other topics, although marginal in comparison, give additional context to the conversations. Topic 1, with words like ”shisha,” ”bitch,” ”santana,” and ”need,” may reflect a more personalized or possibly music-related discussion that associates hookah with contemporary cultural elements. Topic 3 mentions ”reservations,” ”brunch,” ”mimosas,” and ”specials,” indicating a context of hookah being discussed within the setting of social dining or brunch culture.

The minuscule percentages for Topics 1 through 4 suggest that while these subjects are present in positive discussions about hookah, they are not as pervasive as the primary theme of hookah in a social and recreational setting. This data could be particularly informative for business strategies, indicating that hookah is perceived positively when associated with social leisure and hospitality.

VI. Conclusion and Future Works

A. Conclusion

In this project, we conducted a comprehensive analysis of hookah-related tweets on Twitter between March 2021 and March 2023 to understand public perceptions of hookah. We also performed exploratory analyses, topic modeling, and machine learning models (particularly RoBERTa and Llama2) that allowed us to perform analyzes from large amounts of data.

Moreover, the Llama2 model outperformed RoBERTa in classifying, we got the accuracy of attitude is 81.8% and accuracy of user or notuser is 86.3%. Based on the results of the model, we performed time series analysis, US map,

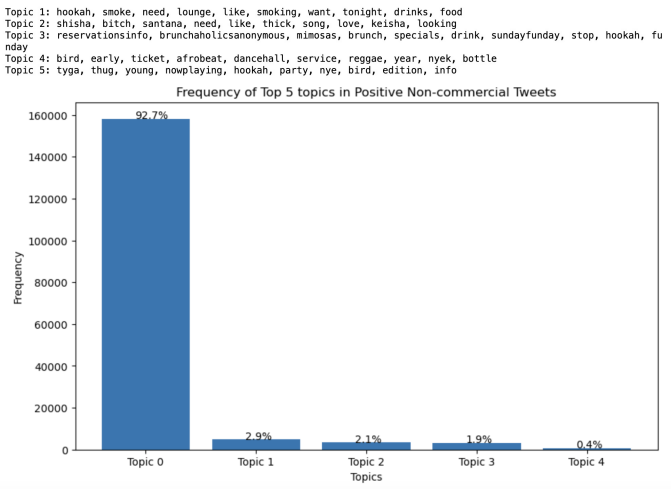

Fig. 19. Frequency of Top5 Topics for Positive Attitude Tweets

heat map and emoji analysis, as well as analysis after topic modeling of positive and negative tweets.

Finally, the analysis revealed significant spikes in tweet volume, sentiment, and user engagement in March 2023. This surge may have been triggered by specific events and discussions regarding the health hazards, safety concerns, and social aspects of hookah use. Positive attitude was common, especially on weekends and evenings, reflecting hookah's association with social activities. Geographic analysis shows regional differences in positive attitude rates and number of users. An analysis of emojis in tweets revealed the emotional expression of different symbols, and attitude topic modeling also enumerated positive and negative themes surrounding hookah, respectively.

B. Future Works

For future works, our results can be utilizeed to provide useful guidance for future hookah regulation and possible tobacco prevention campaigns, and we can also gather Hookah User image data to perform the facial recognition algorithm to identify the demographics of the users, regulate on the underages hookah usage condition and formulate policies.

REFERENCES

[1] J. B. Jukema, D. E. Bagnasco and R. A. Jukema, Waterpipe smoking: not necessarily less hazardous than cigarette smoking, December, 2013.

[2] A. S. Gentzke, M. Creamer, K. A. Cullen, B. K. Ambrose, G. Willis, A. Jamal and B. A. King, Vital Signs: Tobacco Product Use Among Middle and High School Students — United States, 2011–2018(MMWR), February 2019.

[3] M. Jawad and G. Power, Prevalence, correlates and patterns of waterpipe smoking among secondary school students in southeast London: a cross-sectional study, February 2016.

[4] E. Hu, Y. Shen, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[5] N. Jain, P. Chiang, Y. Wen, J. Kirchenbauer, H. Chu, G. Somepalli, B. R. Bartoldson, B. Kailkhura, A. Schwarzschild, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, NEFTune: Noisy Embeddings Improve Instruction Finetuning, October, 2023.

[6] J. Greene, 24-Year-Old Gets Carbon Monoxide Poisoning After Smoking Hookah: 'I Thought I Was Going to Die' (Exclusive), November, 2023.