Monitoring Twitter chatter to assess the beliefs, attitudes, and knowledge of Thai youth and detect temporal patterns of alcohol use and alcohol-related risky behaviors.

Dr. Francois R. Lamy
francois.lam@mahidol.edu

Table of Contents

A. Project title 4

B. Project duration 4

C. Total budget 4

D. Principal investigator 4

E. Background and significance 5

F. Main Objectives and Outcomes: 9

G. Method: 9

1. **Data collection** 9
2. **Data curation** 10
3. **Qualitative content analysis** 11
4. **Sentiment Machine Learning Classifier** 11
5. **Temporal patterns of alcohol-related tweets** 12
6. **Comparing Twitter data to secondary dataError! Bookmark not defined.**

H. Results 14

1. **Manual Qualitative Content Analysis: Creating the codebook.** 14
2. **Manual Qualitative Content Analysis: Intercoder reliability agreement results** 17
3. **Manual Qualitative Content Analysis: Results** 18

4. Computational Data Analysis Results 21

a. Supervised Machine Learning algorithm f-score. 21

b. Category-labeling results 21

c. Sentiment analysis results. 22

d. Terms frequency analysis 22

e. Temporal heatmaps of Personal Communication tweets 25

f. Temporal heatmaps of Negative Personal Communication tweets 26

g. Temporal heatmaps of Positive Personal Communication tweets 28

I. Discussion 29

J. Suggestions/Future Research/Policy Proposals 31

K. Challenges and Limitations 32

L. References 34

M. Annexes 44

Table of Figures

Figure 1. Example of a temporal heatmap based on Facebook engagement 13

Figure 2. Project Overall Architecture...14

Figure 3. Qualitative Content Analysis Dataflow 17

Figure 4. Temporal Heatmap of Personal Communication tweets per day and month....26

Figure 5. Temporal Heatmap of Negative Personal Communication tweets per day and month 27

Figure 6. Temporal Heatmap of Negative tweets per weekday and hour 28

Figure 7. Temporal Heatmap of Positive tweets per weekday and hour 29

Table of Tables

Table 1. Themes emerging from Personal Communication tweets coded as Positive 18

Table 2. Themes emerging from Personal Communication tweets coded as Negative.....19

Table 3. Most common words in tweets labeled as Negative 23

Table 4. Most common words in tweets labeled as Positive..24

Abstract:

Background: Thailand has the highest total alcohol consumption per capita of the Southeast Asian Region and alcohol is considered to be responsible of 2.1% of yearly death in Thailand. This exploratory research aimed to describe the behaviors, sentiments, attitude, and knowledge related to alcohol consumption in Thai tweets.

Method: The Twitter API was called using the tweepy Python library to collect alcohol-related tweets in the Thai Language using 12 search terms. Fifteen thousand random tweets were manually coded by three Thai native speakers to assess the beliefs, knowledge and sentiments associated to collected tweets. Supervised machine learning algorithms were trained to classify tweets per category and sentiment based on the results of the qualitative content analysis using the scikit-learn Python library. Standard Term Frequency analysis of tweets labeled as positive and negative by the sentiment machine learning algorithms were achieved using the nltk Python library. Temporal heatmaps of tweets classified as Personal communications and tweets labeled as positive and negative by the machine learning algorithms were designed using the calplot and seaborn Python libraries.

Results: A total of 10,888,196 unique tweets were collected from the 01/01/2023 to the 12/07/2023. "Personal Communications" represented 49.1% of the coded random sample, 15.3% were "Pornography-related" tweets and 35.6% were coded as Irrelevant. Among Personal communication tweets, 81.1% were coded as Neutral, 4.9% as Positive and 14.0% as Negative. The open coding of both positive and negative tweets labeled as Personal communication described well-known alcohol-effects and experiences (e.g., hangover, aggressivity, prosocial behaviors, euphoria). The Term Frequency analysis revealed the presence of words potentially related to harmful events (e.g., "ขับ," "อันตราย," "ชน," "อาวุธ") in the tweets labeled as negative. The temporal heatmaps suggest that negative Personal communication tweets were posted most frequently on Saturday evening.

Conclusion: Despite changes in Twitter/X data collection policy, over 10 million tweets were collected over the study period. Nevertheless, Twitter/X is not a viable data source to collect digital information related to alcohol or health in general. Twitter/X users tend to display a rather negative attitude toward alcohol. No alcohol-related prevention messages were found during the data analysis calling for designing and disseminating health information through social media. Several potential future research are also formulated.

บทคัดย่อ

ที่มาและความสำคัญ: ประเทศไทยมีอัตราการบริโภคแอลกอฮอล์ต่อหัวประชากรสูงที่สุดในภูมิภาคเอเชียตะวันออกเฉียงใต้ โดยแอลกอฮอล์เป็นสาเหตุของการเสียชีวิตในแต่ละปีมากถึงร้อยละ 2.1 การวิจัยชิ้นนี้เป็นการศึกษาเชิงสำรวจ การวิจัยเชิงสำรวจนี้มีวัตถุประสงค์เพื่อบรรยายพฤติกรรม ความรู้สึก ทัศนคติ และความรู้ที่เกี่ยวข้องกับการบริโภคเครื่องดื่มแอลกอฮอล์ในทวีตภาษาไทย

วิธีวิจัย: ผู้วิจัยได้ดึงข้อมูลทวีตที่เกี่ยวข้องกับแอลกอฮอล์ในภาษาไทย โดยเรียกใช้ API ของ Twitter ผ่านไลบรารี Tweepy ในภาษา Python ด้วยคำค้นหาจำนวน 12 คำ จากนั้นได้ให้นักวิจัยชาวไทยจำนวน 3 คนทำการวิเคราะห์เนื้อหาจาก 15,000 ข้อความที่สุ่มมาจากคลังข้อความทั้งหมด เพื่อจำแนกประเภทข้อความ เช่น ความเชื่อ ข้อมูลข่าวสาร และความรู้สึกเกี่ยวกับแอลกอฮอล์ อัลกอริทึมการเรียนรู้ด้วยเครื่องแบบมีผู้สอน (Supervised machine learning) ถูกฝึกฝนเพื่อจำแนกทวีตตามประเภทและความรู้สึกตามที่จัดหมวดหมู่ได้จากการวิเคราะห์เนื้อหาโดยใช้ไลบรารี scikit-learn ในภาษา Python และใช้ไลบรารี nltk ในภาษา Python วิเคราะห์ความถี่ของคำศัพท์พื้นฐานที่แสดงความรู้สึกเชิงบวกและเชิงลบ ทั้งนี้ ผู้วิจัยได้ใช้ calplot และ seaborn ในภาษา Python สร้างแผนภูมิ Heat Map เชิงเวลาเพื่อแสดงผลความถี่ของทวีตในหมวดการสื่อสารส่วนบุคคล และหมวดย่อยที่ระบุความรู้สึกเชิงบวกและเชิงลบเกี่ยวกับแอลกอฮอล์

ผลการศึกษา: มีข้อความทวีตที่ไม่ซ้ำซ้อนกันที่เกี่ยวข้องกับแอลกอฮอล์ในช่วงระหว่างวันที่ 1 มกราคม – 12 กรกฎาคม 2566 จำนวนทั้งสิ้น 10,888,196 ข้อความ เมื่อพิจารณาจากข้อความที่สุ่มเลือกสำหรับการวิเคราะห์เชิงเนื้อหา พบว่าร้อยละ 49.1 เป็นการสื่อสารส่วนบุคคล ร้อยละ 15.3 เป็นการกล่าวถึงสื่อลามก และร้อยละ 35.6 ไม่เกี่ยวข้องกับแอลกอฮอล์ ในกลุ่มข้อความที่อยู่ในหมวดการสื่อสารส่วนบุคคล 81.1 เป็นข้อความที่ไม่แสดงความรู้สึกหรือความคิดเห็นต่อแอลกอฮอล์ หรือเป็นกลาง ในขณะที่ร้อยละ 4.9 แสดงความรู้สึกหรือความคิดเห็น

เชิงบวก และร้อยละ 14.0 แสดงความรู้สึกหรือความคิดเห็นเชิงลบ ผลจากการเข้ารหัสแบบเปิดสำหรับทวีตทั้งเชิงบวกและเชิงลบในหมวดหมู่การสื่อสารส่วนบุคคล สามารถสื่อถึงผลกระทบและประสบการณ์จากการดื่มแอลกอฮอล์ เช่น เมาค้าง ก้าวร้าว เป็นมิตร เคลิบเคลิ้ม เป็นต้น จากการวิเคราะห์ความถี่ของคำศัพท์ พบว่ามีคำที่เกี่ยวข้องกับเหตุการณ์ที่อาจเป็นอันตรายเช่น “ขับ” “อันตราย” “ชน” และ “อาวุธ” ในทวีตที่ถูกจัดอยู่ในกลุ่มความรู้สึกหรือความคิดเห็นเชิงลบ นอกจากนี้ แผนภูมิ Heat Map เชิงเวลาระบุว่ามีการโพสท์ข้อความสื่อสารส่วนบุคคลที่เป็นเชิงลบบ่อยที่สุดในช่วงวันเสาร์เวลากลางคืน

สรุป: แม้จะมีการเปลี่ยนแปลงนโยบายเกี่ยวกับการเก็บข้อมูลจาก Twitter/X แต่มีทวีตมากกว่า 10 ล้านข้อความที่ถูกเก็บรวบรวมได้ในช่วงเวลาของการศึกษา อย่างไรก็ตาม แพลตฟอร์มดังกล่าวอาจไม่สามารถเป็นแหล่งข้อมูลดิจิทัลเกี่ยวกับแอลกอฮอล์หรือสุขภาพโดยทั่วไป จะเห็นได้ว่า ผู้ใช้ Twitter/X มักแสดงทัศนคติที่ค่อนข้างลบต่อแอลกอฮอล์ นอกจากนี้ ยังไม่พบข้อความที่เกี่ยวกับการป้องกันอันตรายจากการดื่มแอลกอฮอล์ในระหว่างการวิเคราะห์ข้อมูล ดังนั้น ควรมีการส่งเสริมให้มีการออกแบบและเผยแพร่ข้อมูลด้านสุขภาพผ่านสื่อสังคมออนไลน์ ทั้งนี้ ข้อค้นพบจากการศึกษาครั้งนี้สามารถนำไปต่อยอดการวิจัยได้อีกหลากหลายประเด็นในอนาคต

A. Project title

Monitoring Twitter chatter to assess the beliefs, attitudes, and knowledge of Thai youth and detect temporal patterns of alcohol use and alcohol-related risky behaviors.

B. Project duration

1st October 2022-31th May 2024

C. Total budget

276,100THB

D. Principal investigator

(PI) Assistant Professor Francois R. Lamy / Faculty of Social Sciences and Humanities / Department of Society and Health / Health Solutions Research Unit (HSRU) / Mahidol University (MU)

E. Background and significance

According to WHO, alcohol, or ethanol, is “a toxic and psychoactive substance with dependence producing properties” and harmful consumption of alcohol is a causal factor in over 200 diseases and injury and is responsible for up to 5.1% of yearly deaths worldwide (WHO, 2018). Heavy Episodic Drinking (HED) is the most common form of harmful drinking and can be defined as the consumption of 6 or more standard alcoholic beverages in a single occasion over the last 30 days (WHO, 2022). Alcohol consumption remains the main risk factor for premature death and disability among those aged 15 to 49 years, accounting for 10% of all deaths in this age group. In addition, disadvantaged and especially vulnerable populations have higher rates of alcohol-related death and hospitalization. Although long-term alcohol consumption causes a variety of non-communicable diseases such as cirrhosis (Rehm et al., 2010), liver cancer (Turati et al., 2014) or diabetes (Baliunas et al., 2009), alcohol use is also linked to several potentially adverse behaviors such as, pedestrian accidents (Öström & Eriksson, 2001), polysubstance use (Staines et al., 2001), unsafe sex (Chersich & Rees, 2010; Staines et al., 2001), unwanted pregnancy (Rassool & Villar-Luis, 2006), but also domestic violence (Bègue et al., 2012; Curtis et al., 2019) criminal behaviors (Dingwall, 2013) and drunk driving, which can lead to car/motorbike (Tulloh & Collopy, 1994).

Thailand has the highest total alcohol per capita consumption of the South East Asian Region (SEAR) (WHO, 2018) and alcohol is considered to be responsible of 2.1% of yearly death in Thailand (Nontarak et al., 2022). In addition, Thailand has the second highest HED rate in the SEAR region with 42.8% of the respondents of the Smoking and Drinking Behavior Survey 2017 having such an episode yearly and 10.3% binge drinking one or several days per week (Vichitkunakorn et al., 2021). HEDs, or binge drinking episodes, are linked nationwide to intimate partner violence (Wichaidit & Assanangkornchai, 2020) and to lethal traffic injuries, especially among the 15-24 years old age range, where alcohol is involved in a quarter of lethal traffic injuries (WHO, 2019). Although 13.6% of the 15-19 years old and 33.5% of the 20-24 years old

Thai citizens are regularly consuming alcohol (Office, 2017), HED is more frequent among these age ranges, especially among university students with national studies indicating that over 50% of students engaged in HED (Assanangkornchai et al., 2009; Kitchua et al., 2012; Tonkuriman et al., 2019). Understanding the reasons and patterns of alcohol usage and HED among the youth and young adults is paramount to limit the adverse consequences linked to frequent and/or heavy usage of alcohol for the future of the Thai society.

Over the past twenty years, numerous studies have investigated the prevalence and patterns of alcohol use (Assanangkornchai et al., 2010; Chaveepojnkamjorn, 2012; Chaveepojnkamjorn et al., 2009; Chaveepojnkamjorn & Pichainarong, 2011; Hongthong et al., 2012; McNeil et al., 2016; Office, 2017; Pichainarong & Chaveepojnkamjorn, 2010; Saingam et al., 2012), risk/protective factors influencing drinking behavior (Assanangkornchai et al., 2018; Assanangkornchai et al., 2010; Boonchooduang et al., 2017; Chaveepojnkamjorn & Pichainarong, 2010; Khondok et al., 2012; Luecha et al., 2020; Pengpid & Peltzer, 2012; Tantirangsee et al., 2014; Tonpornkrang et al., 2015; Vantamay, 2012; Vantamay, 2009), attitudes toward alcohol (McNeil et al., 2016; Siviroj et al., 2012), as well as the beliefs that youth associated to alcohol (Vantamay, 2009) among the Thai population. Most of these studies relied on large scale surveys and expensive direct, human facilitated data collection. In the modern context, youth and young adults are importantly exposed to social media and sometimes rely heavily on these media to find information, which can, in turn, modify their perception and attitudes toward alcohol and alcohol usage.

Analyzing data from social media could bring additional insights to the existing monitoring systems: through their systematic review of combined usage of surveys and social media data (SMD), Reveilhac et al. (2022), suggest that these two methods are generally combined to 1) predict any forms of social opinions or political outcomes; 2) comparing data sources studying a given social phenomenon; 3) using survey measures to design social media research; 4) enriching survey results with SMD; 5) recruiting individuals on social media to conduct a targeted survey, and; 6) generating new insights based on prior or “under-investigated” topics using SMD (Reveilhac et al., 2022).

Despite the benefits of combining these two types of data, several discrepancies exist. The main difference remains in the type of data collected: surveys are designed to answer questions at hand, while social media-based research collects raw data that needs to be further processed to answer a research question (Callegaro & Yang, 2018). Moreover, the type of population studied varies as well: SMD are collected from individuals who have access to such media and who are generally 18-49-year-old, female and tend to be more educated (PewResearch, 2021), while surveys can specifically target a subtype or the general population. Furthermore, SMD represent only data from users who are willing to publicly voice their opinion: some social media users may not post about specific behaviors, while others might be very vocal about similar behaviors creating imbalance in the representation of one particular attitude, behavior or sentiment (Al Baghal et al., 2021). In the field of substance use, web-based self-disclosures may possibly produce more accurate and valid reports of substance use behaviors than conventional methods, where under-reporting and social desirability biases present significant challenges (Richardson et al., 2003). Overall, combining both data would benefit research as they cover different aspects of a same phenomenon: SMD being large, immediate, inexpensive, able to reach hidden population and are easy to collect, while survey data being of higher quality, well-structured, representing the general or a targeted population and encompassing sociodemographic characteristics (Eck et al., 2021; Wang et al., 2018). Despite the existing benefits inherent in combining these two methods and to the best of our knowledge, there is no research trying to compare and contrast SMD to survey data, and potentially enriching existing data, in the field of substance use in general, and alcohol/HED in particular.

The present research took the exploratory steps to assess the capacity to use social media data (Twitter/X) to collect indirectly longitudinal alcohol-related data in Thai language. Collecting information regarding alcohol-related behaviors, attitudes, knowledge and beliefs through this type of media could provide timely additional data at a lower cost and potentially enrich existing survey results. This type of research belongs to the field of infodemiology that can be defined, according to Eysenbach, as "the study of the determinants and distribution of health information and misinformation" (Eysenbach, 2002). More precisely, infodemiology is "the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy." (Eysenbach, 2011). Because they harness Big Data to identify and monitor emergent health trends in a timely manner (Beaunoyer et al., 2017; Paul et al., 2016), infodemiologists tend to heavily rely on computer sciences techniques such as Data Mining, Natural Language Processing or Machine Learning. Correlatively, this branch of epidemiology has grown substantially since the advent of the semantic web and the development of social media, and has been fruitfully applied to the surveillance of influenza (Alessa & Faezipour, 2019; Nagar et al., 2014; Pawelek et al., 2014), depression (Leis et al., 2019; Reavley & Pilkington, 2014), suicide (Jashinsky et al., 2014; Sueki, 2014), eating disorders (Beguerisse-Díaz et al., 2017; Cavazos-Rehg et al., 2019; Karami et al., 2018; Lapinski, 2006) and diabetes (Beguerisse-Díaz et al., 2017; Karami et al., 2018). This approach has also been used in the field of drug use, abuse and addiction with studies focusing on alcohol (Cavazos-Rehg et al., 2015; Krauss et al., 2017), tobacco and e-cigarettes (Zhan et al., 2017), stimulants (Hanson et al., 2013), opioids (Daniulaityte et al., 2015; Shutler et al., 2015), or drug supply (Pinyopornpanish et al., 2018; Rosenbaum et al., 2012; Schmidt et al., 2011).

Although there is a large variety of social media platforms, Twitter/X is particularly useful for substance abuse surveillance because Twitter/X data are publicly accessible and reflect casual, unedited disclosures and communications from large numbers of people. At the time the project started, Twitter/X grew to over 12.8 million active users in April 2022 (Kemp, 2022) and was mostly used by the younger subpopulation (the average age of Twitter user in Thailand was 18.6 years old) (Insider Intelligence, 2022). Overall, this project mined Twitter/X data to continuously capture opinion and health-related data as well as to identify risky behaviors in the textual content of collected alcohol-related tweets.

To demonstrate the feasibility of this study, alcohol-related tweets were collected using the X Application Programming Interface (API) with several "search terms" for sixteen days from the 29th of June 2022 to the 16th of July 2022 (retweets were filtered out). 557,064 tweets that contained the entity “alcohol” and some of the commonly used terms, slang terms and hashtags associated to alcohol and drunkenness (e.g., 'เหล้า', 'ดื่มเหล้า แดกเหล้า', 'เบียร์', 'ดื่มเบียร์ แดกเบียร์', 'ไวน์', 'ดื่มไวน์', 'นักดื่ม', 'ขี้เหล้า ขี้เมา', 'มึนเมา', 'เมาเหล้า', 'เมา') were collected and stored. Out of that number, 241,279 (43.3%) contained geolocation information. Depending on the search terms, extracted tweets were relevant to health research as they depicted intentions as well as actual behaviors, attitudes and knowledge regarding alcohol (e.g., “10วันมานี่หมดค่าเหล้าค่าเที่ยวไปเกือบหมื่นได้อะ ชีวิตล่มจมชิบหัย 5555555” or “อยากกินเหล้าจริงๆ แม่ง” or “เศร้าอ่ะ อยากกินเหล้าเลยเนี่ย” or “กุเก่งมาก เลิกเหล้ามาได้อาทิตย์นึงละ”) and HED (e.g., “เมื่อคืนเหมือนเมา คิดถึง ละก็ร้องไห้ออกมา” or “เบื่อจิง เมาละต้องตื่นมาขอโทษทุกคนที่หลุดปากพูดความลับตัวเอง งอแงห่าเหวไรไม่รู้ โอ้ยถ้าเมาแล้วจะเป็นแบบนี้อะนะ” or “เมาจัดๆ หัวจะทิ่ม”). This preliminary exploration confirmed that Twitter/X data contained large volumes of relevant information for alcohol research in Thailand.

F. Main Objectives and Outcomes:

This project developed an exploratory platform to process real-time social media data (Twitter/X) and semi-automated information extraction about beliefs, knowledge, attitudes and harms related to alcohol consumption. The specific Research Objective of this exploratory research was therefore to describe the behaviors, sentiments, attitude, and knowledge on alcohol consumption in Thai tweets.

G. Method:

1. Data collection

Twitter (now X) is a micro-blogging service provider and social network platform that was launched in 2006. X reported 330 million monthly active users generating over 700 million tweets per day (https://www.internetlivestats.com/twitter-statistics/#trend) at the time this research was submitted to the granter.

Tweets are limited to no more than 280 characters and thus contain very brief information. However, because of the large volume of data generated by Twitter users, analysis of tweets can provide valuable population level metrics as well as geolocation of Twitter users. Tweets were collected continuously via the Twitter's streaming API using the "tweepy" Python library based on twelve generic search terms (i.e., 'ดื่มเหล้า', 'แดกเหล้า', 'ดื่มเบียร์ แดกเบียร์', 'ไวน์', 'ดื่มไวน์', 'นักดื่ม', 'ขี้เหล้า ขี้เมา', 'มึนเมา', 'เมาเหล้า', 'เมา', 'เบียร์', 'เหล้า'). Tweet unique ID, timestamp, location, text and Twitter/X user unique "screen name" were collected and stored.

Despite being able to collect tweets as intended at the beginning of the project, data collection was stopped earlier than expected following several changes in Twitter/X policies. Twitter developers with "Academic" account access were originally able to retroactively mine up to 10 million tweets per month. However, this type of access was revoked on the 23th of May 2023 and turn into a "Unlimited" account access, allowing the researcher to collect a sample of tweets of interest daily. This type of account was then terminated on the 12th of July 2023 and change to "Basic" account allowing the monthly mining of 1,500 tweets. This last change negated the ability to continue data collection as initially intended and the data collection was stopped on that date.

2. Data curation

Text from collected tweets were further curated using a combination of the nltk and pythainlp Natural Language Processing (NLP) Python libraries. To palliate the ambiguity regarding the term "alcohol," which can also be employed as a cleaning agent, we developed a supervised machine learning algorithm that automatically categorized tweets based on their relevance, limiting the impact of amphibolic terms on the data analysis. The supervised ML was trained based on the qualitative content analysis coding results (See 3).

3. Qualitative content analysis

Qualitative content analysis is designed to analyze and group data into topics/themes generated inductively. In other words, themes and categories emerge from the raw data examine by a team of researchers constantly comparing and contrasting their interpretation of the raw data. Data can be texts, images, videos or recordings. Qualitative content analysis generally followed a seven-steps process:

1) formulate the research questions;

2) select a sample to be analyzed;

3) define categories/themes to be applied;

4) generate a coding scheme outlining the coding rules;

5) implement the coding process;

6) determine intercoder reliability, and;

7) analyze the outputs inherent to the coding process (Kaid, 1989).

Steps 1 and 2 were described previously (see E). The coding scheme (Steps 3-5) was developed by reading and re-reading a subsample of tweets. This subset was then used to explore meanings, discuss discrepancies and develop and refine the coding rules. The inductive approach in qualitative coding, which is also referred to as "open coding" (Straus & Corbin, 1990) and moves from the specific to the general (bottom-up approach) and allows examination of phenomena within their own context rather than from a predetermined conceptual basis. Step 6 was implemented by calculating the Krippendorff's Alpha coefficient (Krippendorff, 2004; Neuendorf, 2009). Krippendorff's Alpha score of 0.60–0.80 indicates moderate and above 0.8 indicates substantial agreement. Themes with score below 0.7 were redefined and retested to insure the cohesion of coder results.

4. Sentiment Machine Learning Classifier

Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text. Sentiment analysis helps convey information about the attitudes and opinions of alcohol users towards alcohol, drunkenness and other relevant issues (e.g., alcohol in society, alcoholism). Sentiment analysis consists of two main aspects: 1) identifying generic "sentiment" expressions based on a polarity "good/bad" and 2) identifying these expressions with their subjects. Sentiment analysis relies on Natural Language Processing (NLP) and ML to automatically identify and extract positive, negative, or neutral attitudes associated to some topics. Although, ML has been widely used in consumer research, its application in public health is still emerging (Kursuncu et al., 2019) with some research concerning alcohol on Twitter using such technique (Hasan et al., 2018). A supervised multi-class Machine Learning algorithm was developed using the "scikit learn" Python library to classify sentiment in tweets as negative, neutral or positive. The results from the content qualitative analysis data was used to create training datasets. We initially aimed at developing balanced datasets (e.g., 2,500 positive tweets, 2,500 negative tweets and 2,500 neutral tweets), but the sample used to train the classifier algorithms did not contain sufficient numbers of each category of tweets to be balanced (more in H.4.c). Similarly, the type of tweet sources (i.e., personal communication, news from traditional media, or advertisement for alcohol) was classified by a dedicated supervised ML classifier. This classification allowed reducing the "noise" created by traditional media (e.g., TV channel, newspaper) when tweeting news about recent events and limit the analysis to tweets sent by individuals.

Sentiment and source classifier performances were assessed using precision, recall and F-score (Powers, 2011) indicators. Precision is defined as the number of correctly classified positive examples divided by the number of examples labeled by the system as positive. Recall is defined as the number of correctly classified positive examples divided by the number of positive examples in the manually coded data. F-score is a weighted average of precision and recall measures.

5. Temporal patterns of alcohol-related tweets

Tweet metadata concerning time, date, and geographical location (when available) were collected. Frequency of alcohol-related tweets over time were calculated with an emphasis on the precise day and time within each week to degenerate temporal heatmaps of alcohol-related tweeting activity. Tweets coded as Negative were retroactively identified to create separate temporal heatmaps (see Figure 1 for an example) in order to identify when do Thai Twitter users post negative tweets about alcohol.

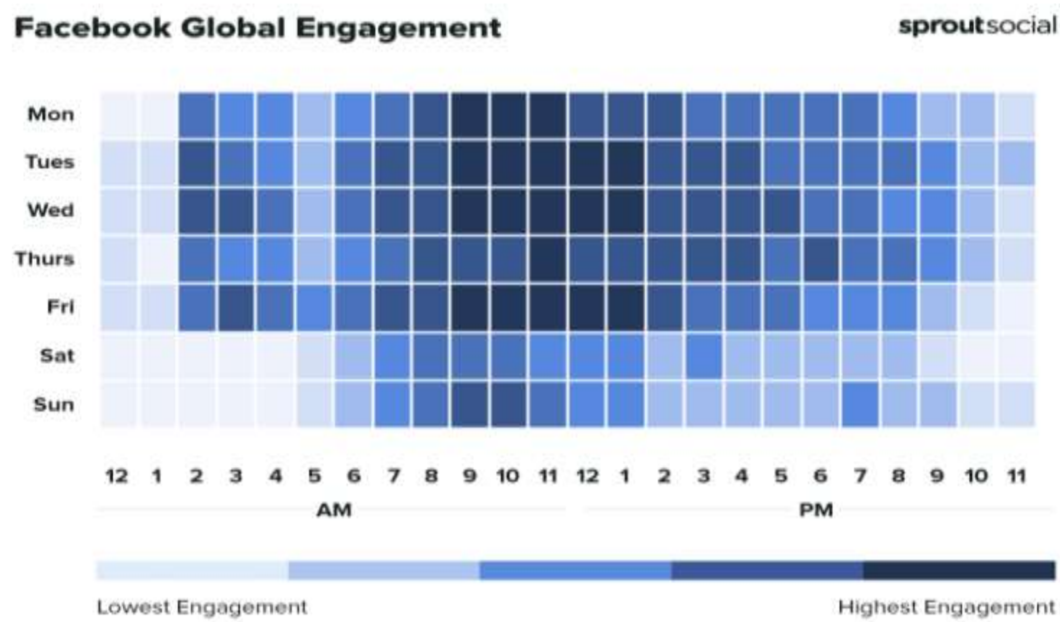

Figure 1. Example of a temporal heatmap based on Facebook engagement (Source: sproutsocial)

The overall architecture of the project is presented in Figure 2:

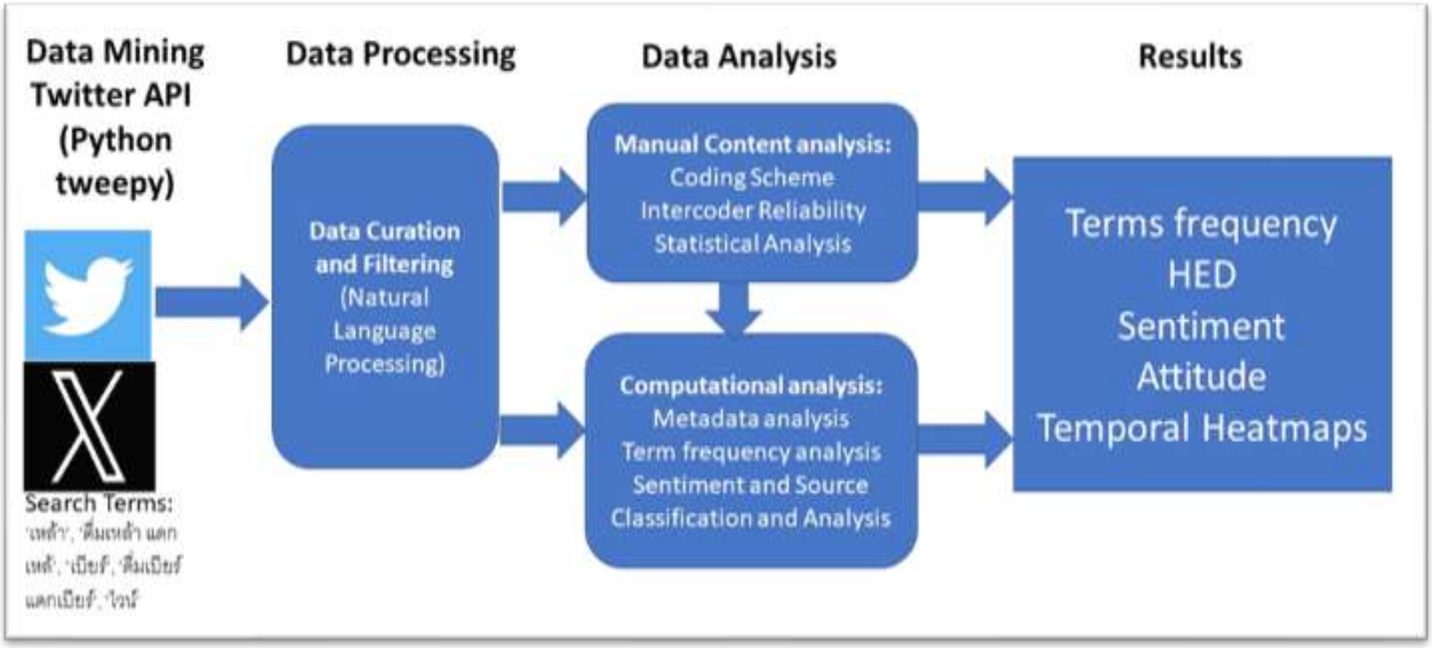

Figure 2. Project Overall Architecture

H. Results

The research protocol was considered as exempt by the Mahidol University Social Sciences IRB (MUSSIRB) (Certificate 2022/019.1912). The certificate of exemption can be found in Annex 2. Data collection started after obtention of the IRB clearance on the 1st of January 2023.

Despite the challenges linked to changes in Twitter/X policies (see G.1), 12,065,726 tweets were collected between 01/01/2023 and 12/07/2023 based on the search terms set at the beginning of the project. After data curation and filtering of duplicated tweets, the final dataset equaled 10,888,196 tweets. The results from the manual and computational analysis are presented below.

1. Manual Qualitative Content Analysis: Creating the codebook.

The initial stage of the qualitative content analysis consisted in having the three Thai native speakers reading individually 1,000 different random tweets for them to identify potential emerging topics. Then, the coding scheme was inductively developed and codes were created based on a consensual agreement among the three coders (see G.3). Second, a sub-sample of 500 similar tweets was coded by all three coders to evaluate the reliability of the coding scheme rules (see below). Third, each coder manually coded 5,000 distinct tweets from the random sample of 15,000 tweets using the developed coding scheme (see Annex 1). The coding scheme included the following codes:

1) All tweets that were not directly related to alcohol were coded Irrelevant (e.g., “เหมือนคนเมากัญชาทำเพลง”);

2) The team of coders initially differentiated four types of tweet sources: “News:” media tweets encompassed all news-related tweets, including retweets of news stories, political debates, scientific study results, and other reports (e.g., “พี่เขยใช้ 9 มม.จ่อยิงน้องเมียดับคาสวนรถไฟ ปมเขม่นกันมานานนับ 10 ปี เรื่องเมาแล้วชอบเปิดเพลงเสียงดัง หลังก่อเหตุ ไม่หนี รอมอบตัวตำรวจบางซื่อ #ไทยรัฐออนไลน์ https://t.co/8CybOGriS6”) (see also 6.); tweets that promote the sales of alcohol or events/restaurants with promotion on alcohol were categorized as “Ads” (see also 4.); tweets categorized as “Pornography-related” had to contain a mention of alcohol/drunkenness associated with a sexual act, plus a link to a book, videos or pictures (e.g., “คลิปนี้...ถ่ายตอนกลับบ้านที่ศรีสะเกษ เพื่อนชวนไปดูหมอลำแล้วไปดูผู้ชายหน้าเวทีหมอลำหมดเหล้าไปเกือบ 3 กลม เลยดึงผู้ชายออกมาขึ้นรถ ไปกินกันที่เถียงนากลางทุ่งนา(ของใครไม่รู้) สรุป..ผู้ชายเมาว่ะทำเหี้ยอะไรไม่ได้ ผมก็เลยจับเยสแม่งเลย..(มีต่อนะ) https://t.co/bUuRDGwtl7”); and “Personal communication” were tweets posted by Twitter users providing their opinions on alcohol, describing experiences of consumption, intention of use, or asking questions regarding alcohol (e.g., “ได้เจอกันนิดเดวเอง คือไปนุ่นมันไม่ได้ทำไรเยอะอยุละ เค้ามาถึงช้า1 แยกกลับ2 นอกจากตอนเมาก้ไม่ได้คุยกันละ จิงๆ แส้ด แต่ก้คุยคุ้มอยุ กุล่อยันตีสี่ คุยจนพี่เค้าเสียงแหบ555555555555”). Due to their very limited number, “News” and “Ads” tweets were combined with “Irrelevant” tweets to allow for the calculation of the reliability test;

3) Tweets categorized as Personal Communications were further coded by associated sentiment: Positive (e.g., “เป่าแร้ว ไม่เมา”), Neutral (e.g., “เมาจังวันนี้ เงี้ยนเลยยย #สกลนคร # สกลนครนัดเย็ด #บ้านธาตุ”) and Negative (e.g., “เขียนแท็กยังผิดอะ เมาเว่อ”);

4) The social meanings contained in the Personal Communications tweets and associated to alcohol by X users were also coded. Four main meanings were inductively discovered during the discussions between the three coders and the PI: *alcohol for socializing or cope with problems* (e.g., “#เหงาเท่าวาฬ ร้องไห้ตอนสุดท้าย ตอนเมาเมื่อคืนกับเพื่อน”); *self-identification through alcohol user or alcohol brand* (e.g., “งง ไม่เคยไม่เมา”); *belonging to a group and share its identity* (e.g., “GOT7 กับโซจู ยองแจคอแข็งที่สุด พี่มาร์คแค่แก้วเดียวตัวก็แดงแล้ว แจ็คสันจำไม่ได้เพราะเมาง่ายมาก พี่แจบอมดื่มได้เรื่อยๆแต่ก็เมา เมาทั้งๆที่บอกว่าตัวเองไม่เมา แบมแบมไม่ค่อยดื่มโซจูชอบเบียร์ คยอมดื่มโซจูได้แต่ขอไปห้องน้ำแล้วไม่กลับมาอีกเลย จินยองชอบโซแมก(โซจูผสมเบียร์)”); and, *stigmatization of alcohol drinkers/abusers* (e.g., “ชาวอักษร เลิกขี้เมาได้ยัง”);

4) During the initial discussion and based on the first 1000 tweets each coder read to obtain a sense of the alcohol-related tweets, two forms of advertisement were found: the first type are advertisements about alcohol (e.g., “สุราชุมชนไทย บ่มนานถึง 58 ปี ดีกรีเข้มข้นปนเรื่องราวขาวดำของสามัญชน และการต่อสู้เพื่อเสรีภาพของ #สุราประชาชน #PLAYREAD สัปดาห์นี้ว่าด้วยเรื่องเหล้าที่คุยยังไงก็ไม่เมา จาก #เฒ่าโลกีย์ ถึงโนนหนองลาด จากเรื่องสั้นในหนังสือเชื่อมโยงการเดินทางของผู้เขียน”), the second are advertisements not directly promoting alcohol, but rather promoting a restaurant “Happy hours” with mentions of the possibility to get drunk (e.g., “ไม่มีเพื่อนกินเพื่อนดื่มเพื่อนเมาปีใหม่เราว่างจ้างได้หรือมาที่ร้านได้นะคะ DM 😊 🙋‍♀️ #ป_ปลาคาราโอเกะ #สาย4 #อ้อมน้อย #สมุทรสาคร #นครปฐม #สามพราน #ปิ่นเกล้า #บางแค #นัดเจอ #ติดตามเค้ายัง #รีทวิต https://t.co/tSYpHNZdf7);

5) Tweets mentioning health-related consequences linked to alcohol consumption were also coded by differencing those that happened to others (e.g., “ผัวเมาขาดสติ! #สังคมวันนี้ 👉 https://t.co/AAtOOR8qoK”) from those that happened to one-self (e.g., “คือพอไม่ได้กินแอลกอฮอล์พอได้กินละกูง่วงเหลือเกิน เมาล่าสุดคือตอนพี่ออฟฟิศเลี้ยงส่งกู”);

6) News tweets were further divided into a) information from newspaper, TV channel or Internet websites (e.g., " เตือนไม่ฟัง! พี่เขยสุดทนซัดลูกโม่ระยะเผาขนดับน้องเมีย ปมเมาเปิดเพลงเสียงดัง หลังเขม่นกันมาข้ามป อ่านต่อ https://t.co/lNIW1VHIjC") and b) information aiming at sharing some knowledge (true or false) regarding alcohol and its consumption (e.g., "กินน้ำมะนาวปั่นสามารถแก้อาการเมาค้างได้ จริงหรือ? เฉลย : ไม่จริง แต่แก้อาการเมาค้างได้โดยการดื่มน้ำกล้วยปั่นกับนมและน้ำผึ้ง แต่ถ้ากินแล้วไม่เมาก็อย่ากินค่ะ เปลืองเงิน #หนูดีอรปรียา"). The final "code book" is presented in Annex 1.

Figure 3 provides a visual representation of the data flow linked to the manual content analysis.

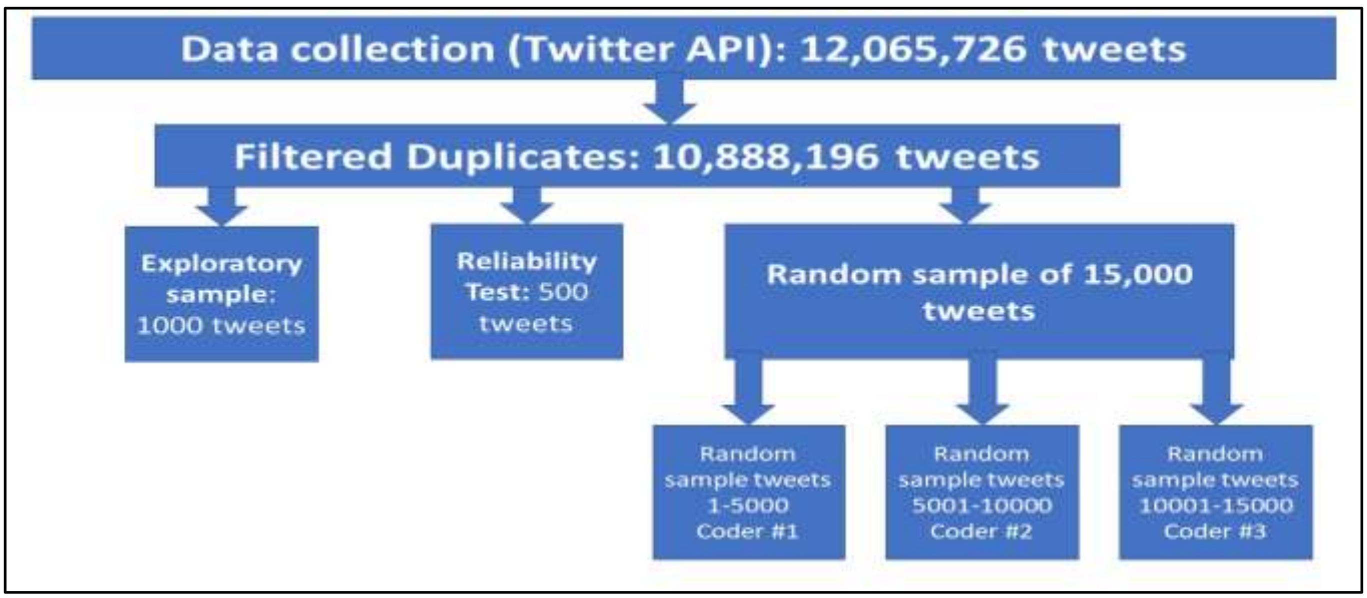

Figure 3. Qualitative Content Analysis Dataflow

2. Manual Qualitative Content Analysis: Intercoder reliability agreement results

The Krippendorff's Alpha (KA) intercoder reliability assessment of the random sample of 500 tweets show substantial agreement for "Category" (Percentage Agreement 93.2%, KA 0.90) and moderate (high) agreement for "Sentiment" (Percentage Agreement 88.4%, KA 0.78) codes. There were not enough tweets for the other codes in the 500 tweets to run a reliability test. We have therefore limited our analysis to tweets that fell under the Category and Sentiment codes.

3. Manual Qualitative Content Analysis: Results

"Personal Communications" (PC) represented 49.1% of the 15,000 random sample of tweets (n=7,368), 15.3% (n=2,296) were "Pornography-related" and 35.6% (n=5,336) Irrelevant. Concerning the sentiment of PC tweets, 81.1% were coded as Neutral (n=5,979), 4.9% (n=359) as Positive and 14.0% (n=1,030) as Negative. The open coding (see G.3) of both Negative and Positive tweets was then achieved by the three coders.

The open coding of PC tweets labeled as Positive revealed five main themes:

1) Alcohol makes persons appear relax and displaying amusing/charming behaviors while inebriated;

2) Alcohol boost self-confidence or boldness (e.g., improved sense of humor/ being more sociable/be outgoing);

3) Alcohol provides some emotional/psychological benefits (e.g., more relax, help to relieve stress, increase happiness);

4) Alcohol improves sexual experience/performance;

5) Alcohol is linked to a sense of reward and pleasure (e.g., reward after hard day of work, drink with friends after work) (examples are displayed in Table 1).

Table 1. Themes emerging from Personal Communication tweets coded as Positive

Themes	Examples
1) Perceived benefit observed on others	- นึกภาพไปเดทแล้วพี่เมาๆทำหน้างี้ กุเป็นคู่เดทคือหลงแย่

2) Boost in self-confidence or boldness	- สรุปคือเด็กสินสาดแบบกู เมาแล้วพูดแต่อิ้งค่ะ ครูต้องภูมิใจ ใครมาทักกู มาคุยด้วย ต้องสลับมาคุยอิ้งกับกูหมด เมาแล้ว fluent อะค่ะ
3) Perceived psychological benefits	- ชอบดื่มไวน์ อย่าถามว่าดื่มตัวไหนยี่ห้ออะไร เพราะผมแค่รู้สึกมีความสุขที่ได้ดื่ม ได้เสพบรรยากาศก็พอ
4) Improved sexual experiences	- หลังดื่ม ไวน์ ไปนิดหน่อย ก็จะอมควยแบบฟินๆหน่อย ชอบอมควยอยากอมหลายๆควยพร้อมกัน 💕
5) Sense of reward and pleasure	- สำคัญคือ โต๊ะข้าง ๆ เต้นสนุก และชนแก้วเก่ง 55555 ชอบมากมิตรภาพในสถานที่มึนเมา - น้องเมาในคอน สนุกมากกกกกกกก แบบมากกกกก โยกหัวสะบัดเลย 🔥 แล้วกุก็ร้องเพลงในคอนคุ้มละ55555555555

Four main themes appeared from the open coding of Personal Communication tweets labeled as Negative:

1) Difficulties in handling drunk persons (e.g., fight, unpredictable behavior);

2) Drunkenness caused undesirable behaviors a) to others: violence, rudeness, aggressiveness, burden placed on others and b) to oneself: losing belongings, fall, shameful or embarrassing behaviors;

3) Health-related problems: a) Physical: i.e., hangovers, vomiting, headaches, muscle pain and b) Mental: i.e., feeling depressed, dysphoria, anxiety;

4) Accidents – posts regarding car accidents linked to alcohol consumption (examples are displayed in Table 2).

Table 2. Themes emerging from Personal Communication tweets coded as Negative

Themes	Examples
1) Difficulty in handling drunk persons	
	- อย่าเมาแล้วโทรมานะ กุด่าจริงนะอหหหห
2) Drunkenness causing undesirable behaviors	

a. to others: violence, rudeness, aggressiveness, burden placed on others	- กูโคตรจะไม่ชอบคนเวลาเมาแล้วแหกปากเสียงดัง ออกลูกนักเลงชิบหาย - ถึงแล้วๆๆๆมะกี้เรย วันนี้เจอคนเมาจะต่อยคนบนรถ กุงงมากคะ เมาได้แต่ทำไมต้องทำเหี้ยคะ
b. to self: personal regrets such as losing belongings, having shameful or embarrassing behaviors	- ช่วยด้วยกุอายคน กุไม่เคยเมาขนาดนี้มาก่อน กุอายเพื่อนไอเหี้ย
3) Health-related problems	
a. Physical - hangovers, vomiting, headaches, muscle pain	- ช้างอร่อย กินง่ายก็จริง แต่นักดื่มรู้ๆกันอยู่ว่า เช้ามาคือตาย บางคนปวดขาด้วยเวลานอน เลี่ยงได้เลี่ยง ถ้าดื่มขวดนึงแค่พอนอนหลับอ่ะไม่เป็นไร https://t.co/JD2ZJjwkFR - ของมึนเมาที่ไม่ชอบที่สุดคือเบียร์ แดกทีไรท้องอืดตลอด ไม่ชอบ - จะเลิกดื่มเหล้า เบียร์ของมึนเมาแล้วคือแพ้อะ เมื่อคืนก็เพิ่งแพ้มาหมดสภาพยันตอนนี้ ไม่เอาแล้วค้าบ
b. Psychological - feeling depressed, dysphoria after drinking	- เมาแล้วร้องไห้เลิกได้เลิกนะ5555555
4) Car accidents	
	- เวรเหอะ รถจอดอยู่หน้าบ้านโดนคนเมามาชนท้ายตอนตี 2

The results of the open coding are not surprising: Twitter users posting about alcohol in Thai Language expressed well-known physiological and psychological effects linked to the consumption of ethanol and potential negative consequences linked to these effects. Alcohol, or more precisely ethanol, is a depressant-type psychoactive substance affecting five main neurotransmitter actions. Ethanol is an agonist of Dopamine, Endorphin, Serotonin and Gamma-aminobutyric acid (GABA) and an antagonist of glutamate (Koob & LeMoal, 2001). In other words, ethanol through its action on dopamine and serotonin provides a sensation of euphoria and increases self-confidence (dopamine) as well as enhances social behaviors (serotonin), while reducing stress (GABA), pain (Endorphin) and inducing a feeling of relaxation (agonist action on GABA and antagonist action on glutamate). These physiological effects are described in the Positive tweets.

As an imbalance in the two main inhibitory/excitatory neurotransmitters (excess of GABA due to the agonist action and reduction of glutamate linked to the antagonist action), individuals who have consumed large quantity of alcohol tend to have reduced cognitive and motor functions, which is expressed in the tweets describing car/pedestrian accidents and the difficulty to handle inebriated individuals. In addition, when the action of ethanol start to wear off, the inverse neurophysiological actions emerge: hangover (linked to dehydration, but also to the release of large quantity of glutamate), antisocial behaviors (rudeness, brawl, violent behaviors) and a feeling of dysphoria linked to the depletion of the dopamine neuroreceptors.

Because the numbers of tweets coded as Positive or Negative were small, there are most certainly other themes that the coders were unable to uncover. The coding of an additional number of tweets could help to discover additional themes in the non-neutral PC tweets.

4. Computational Data Analysis Results

a. Supervised Machine Learning algorithm f-score.

Based on the coding of the Qualitative Content Analysis, three standard supervised Machine Learning (ML) algorithms were tested before the automated coding of the full 10,888,196 million tweets dataset: Naïve Bayesian (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM). The accuracy (f1-score) of these three MLs were: NB = 0.73; KNN = 0.69; SVM = 0.79. The same three standard ML algorithms were also tested to classify PC tweets per sentiment (f1-score NB = 0.81; KNN = 0.72; SVM = 0.83). A f1-score of 0.8 or above is considered to be good, therefore, we selected the SVM algorithms to categorized both complete datasets.

b. Category-labeling results

The full dataset was then labeled using the trained SVM algorithm: 53.4% (5,817,855/10,888,196) of tweets were coded as PC, 14.7% (1,599,041/10,888,196) as "Pornography-related" and 31.9% (3,471,300/10,888,196) as Irrelevant. These results are relatively similar to the manual categorization produced during the qualitative content analysis (see H.3).

c. Sentiment analysis results.

An SVM supervised ML algorithm was also run to label tweets per sentiment. 97.8% of these tweets were coded a Neutral, 1.4% as Negative and 0.8% as Positive. These results are not in line with the results of the manual coding (see H.3).

This can be explained by the fact that the precision scores of the sentiment SVM classifier for the Positive and Negative tweets were maximal (1.00) compared to their recall scores (0.06 and 0.08 respectively). These high precision, but low recall scores tend to indicate that the results of the sentiment classification were very accurate (very few False Positive i.e., tweets wrongly labeled) but that the quantity of tweets classified as Positive and Negative were very limited (hence the low recall score). This is a direct consequence of the unbalanced dataset that was used to train the sentiment SVM algorithm: only 4.9% (n=359) of the tweets contained in the training dataset were coded as Positive and 14.0% (n=1,030) as Negative tweets. A balanced dataset (i.e., the training dataset contains an equal number of items for all categories) would most likely provide better results and increase the number of tweets labeled correctly. In future works focusing on social media data, a larger number of posts needs to be manually coded to establish a balanced dataset for sentiment analysis.

d. Terms frequency analysis

A standard Term Frequency analysis was then run to isolate the 30 most frequent words in both Negative and Positive tweets labeled by the SVM algorithm.

The whole corpus of 94,493 tweets labeled as Negative Personal communication tweets contained 2,371,858 words in total (11,909 unique words). The most common 30 terms are presented in Table 3:

Table 3. Most common words in tweets labeled as Negative

Terms	Translation	Count	%Corpus
เมา	drunk	83,981	3.5%
คน	person	77,454	3.3%
กิน	eat	60,161	2.5%
ดี	good	59,068	2.5%
เหล้า	alcohol	54,296	2.3%
ตัวเอง	myself	50,272	2.1%
รู้	know	45,913	1.9%
ขับ	drive	44,391	1.9%
รับผิดชอบ	be responsible	39,081	1.6%
พ่อ	father	35,589	1.5%
เหี้ย	damn	32,683	1.4%
แบบนี้	like this	31,121	1.3%
ลูก	child	30,266	1.3%
ลุง	uncle	28,873	1.2%
ลูกชาย	son	28,820	1.2%
อ่าน	read	27,274	1.1%
เค้า	He	26,461	1.1%
บ้าน	house	25,848	1.1%
เจอ	meet	25,245	1.1%
อะ	Ah!	25,165	1.1%
อันตราย	dangerous	24,496	1.0%
จิ	Chi	23,808	1.0%
ให้เกียรติ	honor	23,715	1.0%

กร่าง	swagger	23,696	1.0%
เอ๊ะ	Eh!	23,693	1.0%
นักร้อง	singer	23,692	1.0%
รา	mold	23,686	1.0%
ใช่ปะ	Is that right?	23,667	1.0%
ชน	crash	23,608	1.0%
อาวุธ	weapons	23,446	1.0%

Several of the most common terms are directly related to potential harmful events such as "ขับ," "อันตราย," "ชน," or "อาวุธ."

The whole corpus of 50,592 tweets labeled as Positive contained 728,790 words in total (8,466 unique words in total). The most common 30 terms are presented in Table 4:

Table 4. Most common words in tweets labeled as Positive

Terms	**Translation**	**Count**	**%Corpus**
น่ารัก	cute	55,390	7.6%
เมา	drunk	47,208	6.5%
พี่	elder sibling	15,716	2.2%
ตอน	section	11,741	1.6%
คุย	talk	10,361	1.4%
น้อง	sibling	10,056	1.4%
เบียร์	beer	9,508	1.3%
ชอบ	like	9,368	1.3%

คน	person	7,908	1.1%
เรื่อง	subject	7,607	1.0%
ขนาด	size	7,251	1.0%
ชา	tea	6,569	0.9%
อ้อน	beg	6,412	0.9%
จอง	reserve	6,392	0.9%
ขวด	bottle	6,285	0.9%
อะ	Ah!	5,979	0.8%
ใส่	wear	5,941	0.8%
งง	confused	5,824	0.8%
ฉัน	Me	5,816	0.8%
ก.ก.	kg	5,798	0.8%
เบอร์	number	5,712	0.8%
จีบ, แทะโลม	court	5,630	0.8%
ดังลั่น	blatant	5,597	0.8%
ปาร์ตี้สละโสด	stag (party)	5,588	0.8%
อาร์	R	5,576	0.8%
สนุก	fun	5,362	0.7%
ดื่ม	drink	5,339	0.7%
มีความสุข	be happy	5,321	0.7%
วง	loop	5,248	0.7%
ดี	good	5,223	0.7%

A large number of the most frequent words are positive terms in conjunction to alcohol such as “น่ารัก,” “ชอบ,” “สนุก,” “มีความสุข,” or “ดี”.

e. Temporal heatmaps of Personal Communication tweets

Based on the results of the supervised machine learning tweets categorization, a temporal heatmap of the Personal Communication was generated using the catplot Python library (Figure 4). Importantly, we did not included tweets collected after the 23rd of May 2023 to prevent misrepresentation of the data, considering the limited number of tweets we were able to collect after that date (see F.1).

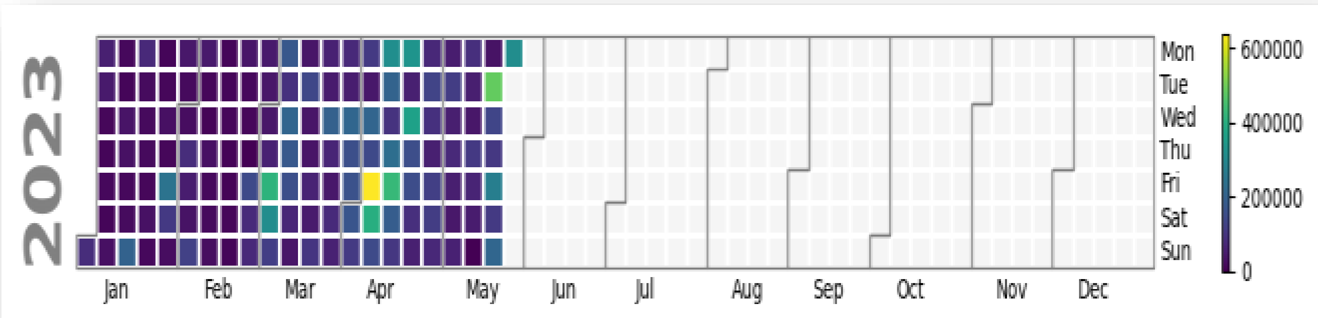

Figure 4. Temporal Heatmap of Personal Communication tweets per day and month

The 7th of April 2023 was the day where the maximum number of tweets labeled as PC were posted. This day corresponded to the first Friday before Songkran break and the celebration that unfold during the next 10 days.

f. Temporal heatmaps of Negative Personal Communication tweets

Similarly, the results from the supervised machine learning aiming at categorizing tweets per sentiment were used to design a temporal heatmap of tweets labeled as Negative (Figure 5).

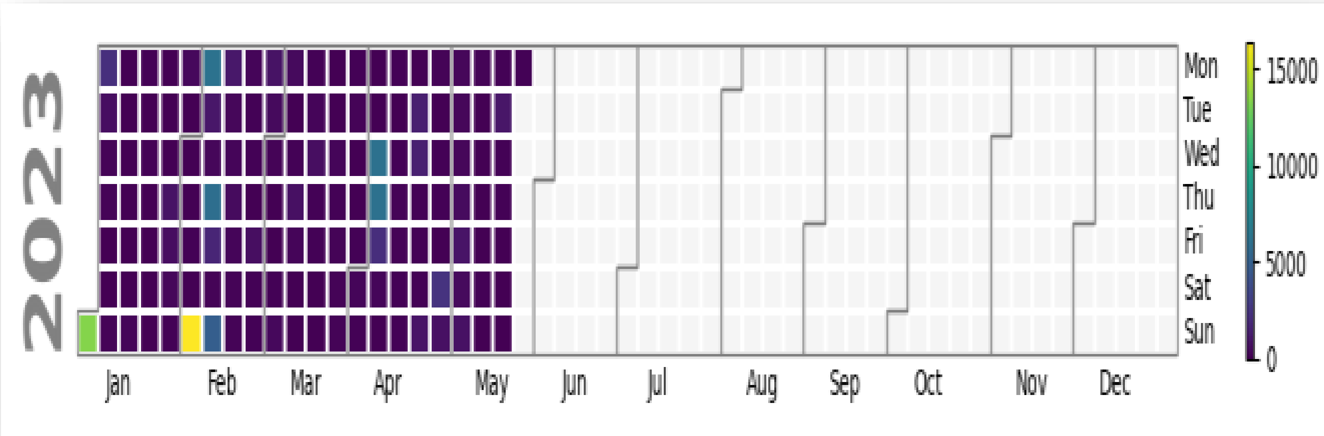

Figure 5. Temporal Heatmap of Negative Personal Communication tweets per day and month

Interestingly, a larger than usual number of negative tweets were posted on the 1st of January and the 5th of February 2023. The first date being the day following New Year Eve, the second date was linked to an uproar of the X community when a drunk man climbed on stage of K-pop concert: “แล้ว ถ้า มี อาวุธ นี่ ทำ ไง อะ ใคร จะ รับผิดชอบ อันตราย มาก ทำไม เค้า ต้อง มา เจอ อะไร แบบนี้ คน มอง นักร้อง เป็น อะไร กัน หมด แล้ว คำ ว่า ให้เกียรติ นี่ ไม่ มี เลย ใช่ปะ เมา แล้ว กร่าง ก็ กิน อยู่ ที่ บ้าน ค่ะ #เอ๊ะ จิ รา กร.”

A second temporal heatmap based on the negative tweets was generated to investigate the day of the week and hour of the day, negative tweets about alcohol were the most frequently posted (Figure 6).

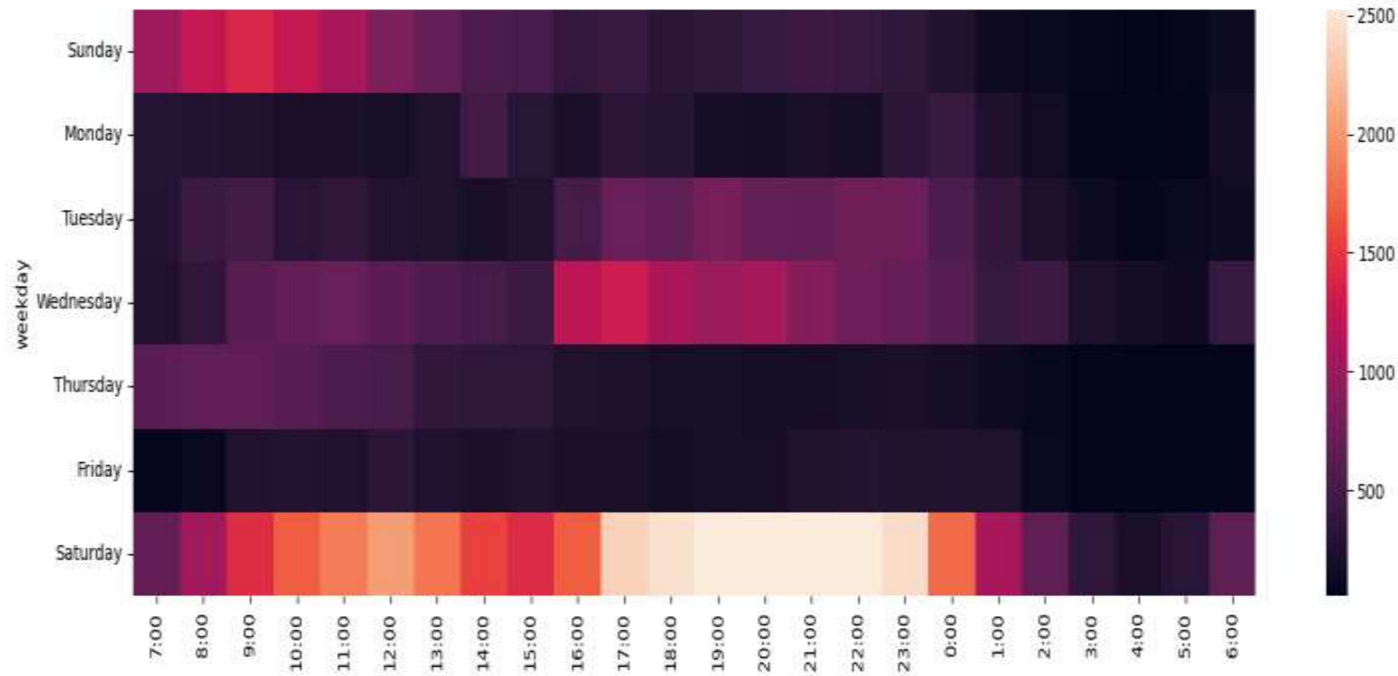

Figure 6. Temporal Heatmap of Negative tweets per weekday and hour

Most negative tweets were posted on Saturday (from 16:00 to midnight). The lowest number of negative tweets were posted during the night (01:00 to 06:00) on Monday but also on Thursday evening and Friday. It appears that Twitter/X users tweeting about negative events linked to alcohol are tweeting about events they witnessed/experienced after their Friday night or during their Saturday evening. This explanation requires, nevertheless, additional information to be validated: interviews with social media users who post about alcohol would most likely accurately explain this phenomenon.

g. Temporal heatmaps of Positive Personal Communication tweets

In comparison, Personal communication tweets labeled as Positive were more likely to be posted on Sunday, Monday and Thursday evenings (19:00 to 22:00) and more rarely on Saturday (Figure 7).

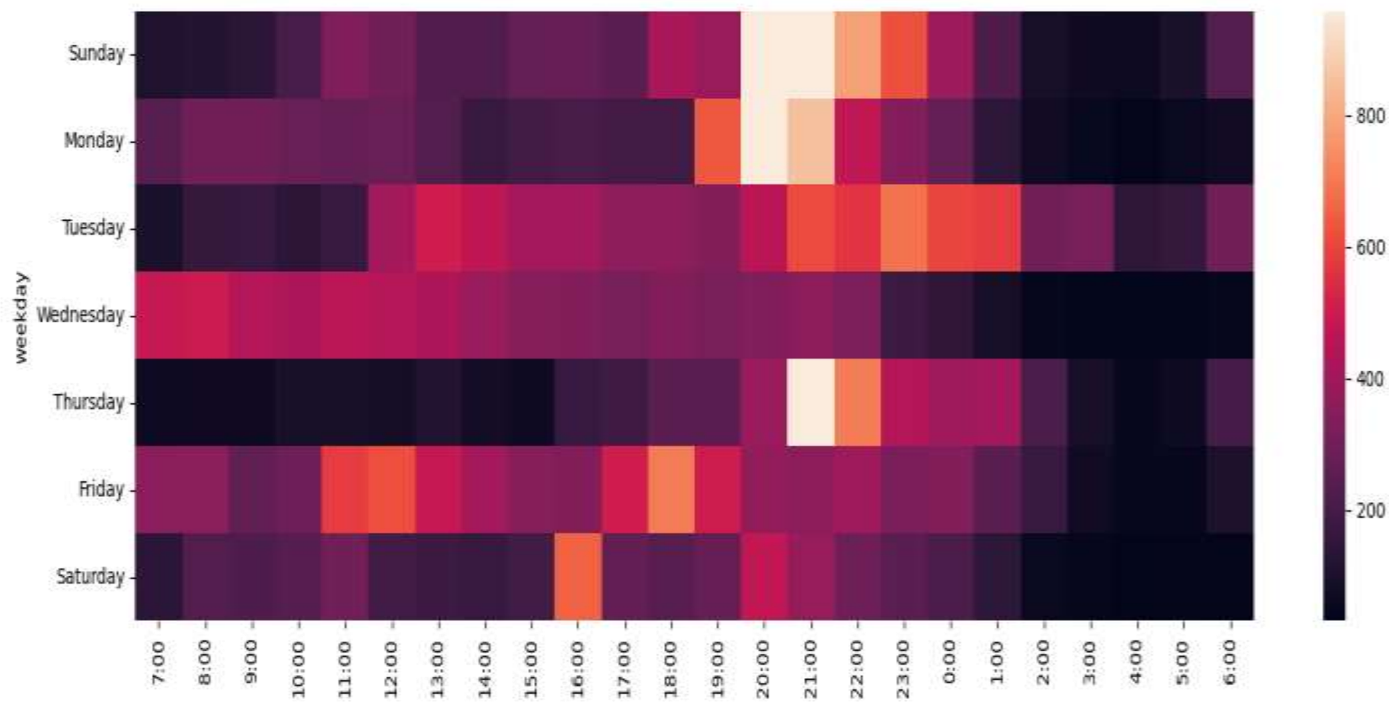

Figure 7. Temporal Heatmap of Positive tweets per weekday and hour

I. Discussion

The main objective of this study was to describe the behaviors, sentiments, attitude, and knowledge on alcohol consumption extracted from Thai tweets collected from the 01/01/2023 to the 12/07/2023. Qualitative content analysis and computational techniques were combined to analyze the 10,888,196 unique tweets collected until Twitter/X revoked the ability of researchers to use their API. The main results of this research are:

1) Only one alcohol-related prevention tweets was found in the Qualitative Content Analysis of 15,000 tweets;
2) Personal communication tweets that were not coded as Neutral were more likely to be Negative suggesting a rather negative attitude of Twitter/X users who posted about alcohol over the study period;

3) The perceptions, alcohol-associated functions, negative consequences and representation concerning alcohol found in the Personal Communication tweets coded as positive and negative were in line with the literature;
4) Collected tweets do contain mentions of negative events and consequences of HED with "drive" one of the most frequent terms associated to tweets labeled as Negative;
5) Negative tweets were more likely to be posted on Saturday evening (16:00-22:00) rather than any other days and hours of the week;
6) Social media data analysis could provide additional data (e.g., opinions, attitudes, and new trends) and should be used as complementary sources of information to traditional data collection tools such as annual household surveys.

Unfortunately, the type of information collected through Twitter/X (and most, if not all social media) cannot be compared to data collected through more traditional methods such as a national household survey (for example the 2021 Health Behavior of Population survey) due to the following reasons.

First, surveys such as the Health Behavior of Population survey focused on a representative sample of the Thai general population, while the present study collected information from a very specific subset of the Thai population, i.e., Thais who use and have access to Internet, more specifically who use social media, more specifically Twitter/X, who are active Twitter/X users, who made their tweets public, and who tweeted about alcohol. In other words, the present study provides results extracted from tweets that were produced by a fraction of the general population and cannot be considered as representative of the general population, greatly limiting the possibility for comparison with the Health Behavior of Population survey results.

Second, social media data, especially Twitter/X, do rarely display sociodemographic variables, limiting the possibility to compare social media data with epidemiologic survey results. It is possible to evaluate the gender and age range of social media users through a series of algorithms trained to analyze and categorize textual patterns. However, such results are, to the best, approximations and would require to collect a very large number of posts that can be linked to a profile with gender and age information in order to reliably train machine learning models.

Third, regarding the consumption of alcohol, there is no indication regarding the frequency nor the quantity of alcohol consumed by the Twitter/X users who have tweeted about alcohol during the data collection period.

Fourth, concerning sentiment, while the results of this research suggest that a majority of non-neutral Personal communication tweets were negative, the Health Behavior of Population survey does not contain reference to sentiment regarding alcohol.

Fifth, and in line with the previous point, Twitter/X users who have tweeted negatively about alcohol, might also enjoy drinking on a more or less regular basis: reacting negatively to an event or a social news, does not represent an overall indicator of negative attitude toward an object or idea.

Sixth, while the present research was able to collect tweets that contains expression(s) and word(s) related to HED (e.g., drunk, too much drinks), it is impossible to verify that the X users tweeting about being drunk were indeed inebriated and that they drank 6 or more standard alcoholic beverages in a single occasion over the last 30 days. Nevertheless, this type of limitations also applies to more standard data collection tools.

J. Suggestions/Future Research/Policy Proposals

The main aim of this exploratory research was to describe the behaviors, sentiments, attitude, and knowledge related to alcohol consumption in Thai tweets. Three main suggestions can be drawn from the current findings.

First, social media are widely utilized by the younger section of the Thai population and represent an important channel to disseminate prevention messages regarding a variety of health risky behaviors. Among the 15,000 tweets that were manually coded only one tweet was classified as bearing a prevention message, which represents a missed opportunity. Our study has also mapped the temporal alcohol-related tweets activities in terms of moment of the year, day of the week and hours of the day users are more likely to communicate about alcohol through Twitter/X. This type of results can be further used to disseminate adequate health prevention messages at the right time to social media users using X, Facebook or TikTok through advertisement campaigns.

Second, due to the impossibility to obtain data for free from Twitter/X (plus the fact that Twitter/X is rapidly losing audience), we recommend to utilize other social media for future research. After discussion with the panel, TikTok seems to be a more viable option: a qualitative content analysis of videos related to alcohol combined with an analysis of the metadata of watched videos would provide substantial understanding of the messages shared by influencers and the beliefs about alcohol expressed online. At least one study conducted by Russell and colleagues (2021) in the United States has shown promising results (Russell et al., 2021).

Third, in order to accurately understand the sociodemographic differences between the social media user subpopulation and the general population, an online survey targeting social media users should be conducted. Such a survey would not only help to grasp the differences between the social media user subpopulation and the general population, but it would also offer the possibility to assess how social media usage and exposure impact alcohol usage, beliefs, knowledge and practices.

K. Challenges and Limitations

As just mentioned, Twitter/X drastic changes in data collection policy has negated our ability to mine tweets for the whole duration of the project. Twitter/X does not appear as a viable data source for social sciences and epidemiological research anymore.

This exploratory study aimed at testing if social media could provide additional information regarding alcohol use and HED among the youth. Despite the fact that social media are most likely to be use by the younger portion of the population, this study cannot ascertain the age range of any of the Twitter/X users who posted about alcohol.

In addition, this study suffers from limitations mostly inherent to the field of Social Media Data analysis. First, social media users might only post about the positive aspects of their daily routine, underreport negative behaviors or refrain to express opinions that are considered socially undesirable attitudes (Althubaiti, 2016). Second, social media research solely collects data from digital media. This implies that data are collected only from the subpopulation who has access to Internet and is using social media, introducing a selection bias (Olteanu et al., 2019). Third, data collection was limited to tweets written in the Thai Language, which does not guarantee that all collected tweets were posted from within the Kingdom of Thailand. Fourth, our exploratory study did not cover a full year period and is subject to seasonality linked for example to Songkran festival.

L. References

Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: asymmetry in quantity and its impact. *EPJ Data Science*, *10*(1), 32.

Alessa, A., & Faezipour, M. (2019). Preliminary Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression With Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study [Original Paper]. *JMIR Public Health Surveill*, *5*(2), e12383. https://doi.org/10.2196/12383

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, 211-217.

Assanangkornchai, S., Li, J., McNeil, E., & Saingam, D. (2018). Clusters of alcohol and drug use and other health-risk behaviors among Thai secondary school students: a latent class analysis. *BMC public health*, *18*(1), 1-10.

Assanangkornchai, S., Mukthong, A., & Intanont, T. (2009). Prevalence and patterns of alcohol consumption and health‐risk behaviors among high school students in Thailand. *Alcoholism: Clinical and Experimental Research*, *33*(12), 2037-2046.

Assanangkornchai, S., Sam-Angsri, N., Rerngpongpan, S., & Lertnakorn, A. (2010). Patterns of alcohol consumption in the Thai population: results of the National Household Survey of 2007. *Alcohol & Alcoholism*, *45*(3), 278-285.

Baliunas, D. O., Taylor, B. J., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., & Rehm, J. (2009). Alcohol as a risk factor for type 2 diabetes: a systematic review and meta-analysis. *Diabetes care*, *32*(11), 2123-2132.

Beaunoyer, E., Arsenault, M., Lomanowska, A. M., & Guitton, M. J. (2017). Understanding online health information: evaluation, tools, and strategies. *Patient education and counseling*, *100*(2), 183-189.

Bègue, L., Pérez-Diaz, C., Subra, B., Ceaux, E., Arvers, P., Bricout, V. A., Roché, S., Swendsen, J., & Zorman, M. (2012). The role of alcohol consumption in female victimization: Findings from a French representative sample. *Substance use & misuse*, *47*(1), 1-11.

Beguerisse-Díaz, M., McLennan, A. K., Garduno-Hernández, G., Barahona, M., & Ulijaszek, S. J. (2017). The 'who'and 'what'of# diabetes on Twitter. *Digital health*, *3*, 2055207616688841.

Boonchooduang, N., Louthrenoo, O., Charnsil, C., & Narkpongphun, A. (2017). Alcohol use and associated risk behaviors among adolescents in northern Thailand. *ASEAN Journal of Psychiatry*, *18*(2), 1-7.

Callegaro, M., & Yang, Y. (2018). The role of surveys in the era of "big data". In *The Palgrave handbook of survey research* (pp. 175-192). Springer.

Cavazos-Rehg, P. A., Krauss, M. J., Costello, S. J., Kaiser, N., Cahn, E. S., Fitzsimmons-Craft, E. E., & Wilfley, D. E. (2019). "I just want to be skinny.": A content analysis of tweets expressing eating disorder symptoms. *PloS one*, *14*(1), e0207506.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S. J., & Bierut, L. J. (2015). "Hey Everyone, I'm Drunk." An Evaluation of Drinking-Related Twitter Chatter. *Journal of studies on alcohol and drugs*, *76*(4), 635-643.

Chaveepojnkamjorn, W. (2012). Alcohol consumption patterns among vocational school students in central Thailand. *Southeast Asian Journal of Tropical Medicineand Public Health*, *43*(6), 1560.

Chaveepojnkamjorn, W., Chaveepojnkamjorn, W., Pichainarong, N., & Pichainarong, N. (2009). Drinking behavior among female high school students in central Thailand.

Chaveepojnkamjorn, W., & Pichainarong, N. (2010). Factors associated with alcohol consumption among male high school students in central Thailand.

Chaveepojnkamjorn, W., & Pichainarong, N. (2011). Current drinking and health-risk behaviors among male high school students in central Thailand. *BMC public health*, *11*(1), 1-8.

Chersich, M., & Rees, H. (2010). Causal links between binge drinking patterns, unsafe sex and HIV in South Africa: its time to intervene. *International journal of STD & AIDS*, *21*(1), 2-7.

Curtis, A., Vandenberg, B., Mayshak, R., Coomber, K., Hyder, S., Walker, A., Liknaitzky, P., & Miller, P. G. (2019). Alcohol use in family, domestic and other violence: Findings from a cross-sectional survey of the Australian population. *Drug and Alcohol Review*, *38*(4), 349-358.

Daniulaityte, R., Carlson, R., Brigham, G., Cameron, D., & Sheth, A. (2015). "Sub is a weird drug:" A web-based study of lay attitudes about use of buprenorphine to self-treat opioid withdrawal symptoms. *The American Journal on Addictions*, *24*(5), 403-409. https://doi.org/10.1111/ajad.12213

Dingwall, G. (2013). *Alcohol and crime*. Willan.

Eck, A., Cazar, A. L. C., Callegaro, M., & Biemer, P. (2021). Big data meets survey science. In (Vol. 39, pp. 484-488): SAGE Publications Sage CA: Los Angeles, CA.

Eysenbach, G. (2002). Infodemiology: the epidemiology of (mis)information. *The American Journal of Medicine*, *113*(9), 763-765. https://doi.org/10.1016/S0002-9343(02)01473-0

Eysenbach, G. (2011). Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health. *American Journal of Preventive Medicine*, *40*(5), S154-S158.

Hanson, C. L., Burton, S. H., Giraud-Carrier, C., West, J. H., Barnes, M. D., & Hansen, B. (2013). Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of medical Internet research*, *15*(4), e62.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11.

Hongthong, D., Areesantichai, C., Kaunkaew, W., Chinnawattanad, T., & Nuddakul, A. (2012). Drinking risk level and alcohol consumption situation among senior high school students in a rural area of Thailand. *Journal of Health Research*, *26*(4), 187-191.

Insider Intelligence. (2022). *Twitter in 2022: Global user statistics, demographics and marketing trends to know*. Retrieved 08/30/2022 from https://www.insiderintelligence.com/insights/twitter-user-statistics-trends/#:~:text=US%20Twitter%20user%20stats&text=Nearly%2060%25%20of%20Twitter%20users,make%20up%20for%20the%20losses.

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, *35*(1), 51-59. https://doi.org/10.1027/0227-5910/a000234; 10.1027/0227-5910/a000234

Kaid, L. L. (1989). Content analysis. In P. E. L. L. Barker (Ed.), *Measurement of communication behavior* (pp. 197-217). Longman.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., & Shaw Jr, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management*, *38*(1), 1-6.

Kemp, S. (2022). *Essential Twitter statistics and trends for 2022*. Retrieved 07/01/2022 from https://datareportal.com/essential-twitter-stats?utm_source= DataReportal&utm_medium=Country_Article_Hyperlink&utm_campaign=Digital_2022&utm_term=Thailand&utm_content=Facebook_Stats_Link

Khondok, M., Homsin, P., & Srisuriyawet, R. (2012). Factors related to the early stage of alcohol drinking among male students at lower secondary schools, Bangnumprieo district, Chachoengsao province. วารสาร สาธารณสุข มหาวิทยาลัย บูรพา, *7*(1), 19-32.

Kitchua, R., Tenitsara, K., Thirasilawet, T., & Kanasi, S. (2012). *Alcohol consumption behaviors drinking impacts and quality of life in Mahasarakham University students* Thesis].[Mahasarakham]: Mahasarakham University].

Koob, G. F., & LeMoal, M. (2001). *Neurobiology of Addiction*. Elsevier.

Krauss, M. J., Grucza, R. A., Bierut, L. J., & Cavazos-Rehg, P. A. (2017). "Get drunk. Smoke weed. Have fun.": a content analysis of tweets about marijuana and alcohol. *American Journal of Health Promotion*, *31*(3), 200-208.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, *30*(3), 411-433.

Kursuncu, U., Gaur, M., Lokala, U., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2019). Predictive Analysis on Twitter: Techniques and Applications. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining* (pp. 67-104). Springer.

Lapinski, M. K. (2006). StarvingforPerfect.com: a theoretically based content analysis of pro-eating disorder Web sites. *Health communication*, *20*(3), 243-253. https://doi.org/10.1207/s15327027hc2003_4

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., & Sanz, F. (2019). Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis [Original Paper]. *J Med Internet Res*, *21*(6), e14199. https://doi.org/10.2196/14199

Luecha, T., Peremans, L., Dilles, T., Poontawee, P., & Van Rompaey, B. (2020). The prevalence of and factors related to alcohol consumption among young people in Thailand: a systematic review of observational studies. *Drugs: education, prevention and policy*, *27*(5), 337-358.

McNeil, E., Inthawong, Y., & Assanangkornchai, S. (2016). Access to alcohol and binge drinking among vocational college students: A multilevel study in a tourist destination province of Thailand. *International Journal of Alcohol and Drug Research*, *5*(2), 47-55.

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., & Brownstein, J. S. (2014). A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of medical Internet research*, *16*(10), e236. https://doi.org/10.2196/jmir.3416 [doi]

Neuendorf, K. A. (2009). Reliability for Content Analysis. In A. B. Jordan, D. Kunkel, J. Manganello, & M. Fishbein (Eds.), *Media messages and public health : a decisions approach to content analysis* (pp. 67-87). Routledge.

Nontarak, J., Geater, A. F., Assanangkornchai, S., & Aekplakorn, W. (2022). Expected Years Of Life Lost Due To Alcohol Consumption In Thai Adults: A 16-Year Follow-Up Cohort Of National Health Examination Survey 2004–2019. *Alcohol and Alcoholism*, *57*(4), 490-499.

Office, N. S. (2017). The smoking and drinking behaviour survey 2017. In: National Statistical Office Bangkok.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data 2 (2019). In.

Öström, M., & Eriksson, A. (2001). Pedestrian fatalities and alcohol. *Accident Analysis & Prevention*, *33*(2), 173-180.

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., & Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. Biocomputing 2016: Proceedings of the Pacific symposium,

Pawelek, K. A., Oeldorf-Hirsch, A., & Rong, L. (2014). Modeling the impact of twitter on influenza epidemics. *Mathematical biosciences and engineering : MBE*, *11*(6), 1337-1356. https://doi.org/10.3934/mbe.2014.11.1337 [doi]

Pengpid, S., & Peltzer, K. (2012). Alcohol use and associated factors among adolescent students in Thailand. *West Indian Medical Journal*, *61*(9).

PewResearch. (2021). *Social Media Fact Sheet*. https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=ad42e188-04e8-4a3c-87fb-e101714f1651

Pichainarong, N., & Chaveepojnkamjorn, W. (2010). Youth and alcoholic beverages: drinking patterns among high school students in central Thailand.

Pinyopornpanish, K., Jiraporncharoen, W., Thaikla, K., Yoonut, K., & Angkurawaranon, C. (2018). Sedative and analgesic drugs online: a content analysis of the supply and demand information available in Thailand. *Substance use & misuse*, *53*(4), 641-647.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Rassool, G. H., & Villar-Luis, M. (2006). Reproductive risks of alcohol and illicit drugs: an overview. *Journal of Addictions Nursing*, *17*(4), 211-213.

Reavley, N. J., & Pilkington, P. D. (2014). Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ*, *2*, e647. https://doi.org/10.7717/peerj.647 [doi]

Rehm, J., Taylor, B., Mohapatra, S., Irving, H., Baliunas, D., Patra, J., & Roerecke, M. (2010). Alcohol as a risk factor for liver cirrhosis: a systematic review and meta-analysis. *Drug and Alcohol Review*, *29*(4), 437-445.

Reveilhac, M., Steinmetz, S., & Morselli, D. (2022). A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion. *Multimedia Tools and Applications*, 1-36.

Richardson, J., Fendrich, M., & Johnson, T. P. (2003). Neighborhood effects on drug reporting. *Addiction*, *98*(12), 1705-1711.

Rosenbaum, C. D., Carreiro, S. P., & Babu, K. M. (2012). Here today, gone tomorrow… and back again? A review of herbal marijuana alternatives (K2, Spice), synthetic cathinones (bath salts), kratom, Salvia divinorum, methoxetamine, and piperazines. *Journal of Medical Toxicology*, *8*(1), 15-32.

Russell, A. M., Davis, R. E., Ortega, J. M., Colditz, J. B., Primack, B., & Barry, A. E. (2021). # Alcohol: portrayals of alcohol in top videos on TikTok. *Journal of studies on alcohol and drugs*, *82*(5), 615-622.

Saingam, D., Assanangkornchai, S., & Geater, A. F. (2012). Drinking-smoking status and health risk behaviors among high school students in Thailand. *Journal of Drug Education*, *42*(2), 177-193.

Schmidt, M. M., Sharma, A., Schifano, F., & Feinmann, C. (2011). "Legal highs" on the net—Evaluation of UK-based Websites, products and product information. *Forensic science international*, *206*(1-3), 92-97.

Shutler, L., Nelson, L. S., Portelli, I., Blachford, C., & Perrone, J. (2015). Drug use in the Twittersphere: a qualitative contextual analysis of tweets about prescription drugs. *Journal of addictive diseases*, *34*(4), 303-310.

Siviroj, P., Peltzer, K., Pengpid, S., Yungyen, Y., & Chaichana, A. (2012). Drinking motives, sensation seeking, and alcohol use among Thai high school students. *Social Behavior and Personality: an international journal*, *40*(8), 1255-1262.

Staines, G. L., Magura, S., Foote, J., Deluca, A., & Kosanke, N. (2001). Polysubstance use among alcoholics. *Journal of addictive diseases*, *20*(4), 57-73.

Straus, A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. In: Newbury Park, CA: Sage.

Sueki, H. (2014). The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of affective disorders*, *170C*, 155-160. https://doi.org/S0165-0327(14)00536-9 [pii]

Tantirangsee, N., Assanangkornchai, S., & Geater, A. F. (2014). Trends and associated factors of alcohol consumption among Southern Thai adolescents, 2003–2009. *International Journal of Alcohol and Drug Research*, *3*(3), 219–225-219–225.

Tonkuriman, A., Sethabouppha, H., Thungjaroenkul, P., & Kittirattanapaiboon, P. (2019). A causal model of binge drinking among university students in northern Thailand. *Journal of Addictions Nursing*, *30*(1), 14-23.

Tonpornkrang, R., Homsin, P., & Srisuriyawate, R. (2015). Factors related to binge drinking among male vocational students in Surin province. *Journal of public health nursing*, *29*(1), 29-42.

Tulloh, B., & Collopy, B. (1994). Positive correlation between blood alcohol level and ISS in road trauma. *Injury*, *25*(8), 539-543.

Turati, F., Galeone, C., Rota, M., Pelucchi, C., Negri, E., Bagnardi, V., Corrao, G., Boffetta, P., & La Vecchia, C. (2014). Alcohol and liver cancer: a systematic review and meta-analysis of prospective studies. *Annals of oncology*, *25*(8), 1526-1535.

Vantamay, N. (2012). The impacts of alcohol marketing communications on cognitive, affective, and behavioral responses among Thai youth in Bangkok. *International Journal of Business and Social Science*, *3*(18).

Vantamay, S. (2009). Alcohol consumption among university students: applying a social ecological approach for multi-level preventions. *Southeast Asian Journal of Tropical Medicine and Public Health*, *40*(2), 354.

Vichitkunakorn, P., Intusoma, U., & Assanangkornchai, S. (2021). Binge Drinking and Drunk Driving Among Current Drinkers in Thailand: Analysis of Cigarette Smoking and Alcohol Drinking Behavior Surveys From 2007 to 2017. *Asia Pacific Journal of Public Health*, *33*(2-3), 234-241. https://doi.org/10.1177/1010539520971178

Wang, Z., Jin, Y., Liu, Y., Li, D., & Zhang, B. (2018). Comparing social media data and survey data in assessing the attractiveness of Beijing Olympic Forest Park. *Sustainability*, *10*(2), 382.

WHO. (2018). *Global status report on alcohol and health 2018*.

WHO. (2019). *Young People (aged 15-24) and Road Safety in Thailand*. https://cdn.who.int/media/docs/default-source/thailand/roadsafety/yp-en-final-25-7-19.pdf?sfvrsn=64a3c846_2

WHO. (2022). *Alcohol, heavy episodic drinking (population) past 30 days*. Retrieved 01/09/2022 from https://www.who.int/data/gho/indicator-metadata-registry/imr-details/459#:~:text=Definition%3A,to%206%20standard%20alcoholic%20drinks.

Wichaidit, W., & Assanangkornchai, S. (2020). Binge-drinking and household role's associations with prevalence of domestic violence: findings from the Thailand smoking and drinking behaviour survey 2017. *Substance abuse treatment, prevention, and policy*, *15*(1), 1-10.

Zhan, Y., Liu, R., Li, Q., Leischow, S. J., & Zeng, D. D. (2017). Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms. *Journal of medical Internet research*, *19*(1), e24.

M. Annexes

Annex 1. Qualitative Content Analysis Coding scheme

Theme	Category	Label	Explanation	คำอธิบายเพิ่มเติม
A. Personal Communication				
	A1. Sentiment	Tweets showing attitudes towards alcohol drinking.		ข้อความที่แสดงความคิดเห็นหรือความรู้สึกต่อเครื่องดื่มหรือการดื่มแอลกอฮอล์
	A1.1 Neutral	Tweets contains terms implying alcohol drinking but cannot be judged whether it is positive or negative, show no emotion towards alcohol drinking	If you are not sure about the feeling/emotions (ambiguous) code as neutral	
	A1.2 Positive	Tweets implying good effects or advantages of alcohol drinking		

	A1.3 Negative	Tweets implying bad effects of alcohol drinking or showing bad feelings about alcohol drinking		
	A2. Social meaning	Tweets showing the roles of alcohol drinking		ข้อความที่มีนัยยะถึงบทบาทของแอลกอฮอล์ในทางสังคม
	A2.1 Use of alcohol for socialization or cope with problems	Tweets contains the reason to drink alcohol and/or drinking peers/occasion	It can be either solitary drinking or social drinking, showing alc. is used for a purpose	การดื่มเพื่อการสังสรรค์ ดื่มร่วมกับเพื่อนร่วมงาน/ผู้อื่น ที่มีการบอกวัตถุประสงค์ของการดื่ม
	A2.2 Self-identification	Tweets showing self-identification of the poster/ characteristics of drinkers	Implies social identity/ social meaning of alcohol (e.g., beer or badass, wine for high-so)	การใช้คำเพื่อแสดงตัวตนของตัวเอง (อาจจะเป็นในแง่โอ้อวด หรือแค่นิยามตัวเอง) หรือนิสัยการดื่มของตัวเอง

	A2.3 Group membership/ identity	Tweets mentioning exposure to or imitation of alcohol consumption by idols, k-pop, actors, etc.	Those expressing support or affection of and idol that result in alcohol drinking	ทวีตพูดถึงดารานักร้อง หรือคนที่ตัวเองชื่นชม กับการดื่ม (เป็นตัวเองดื่มตามไอดอล)
	A2.4 Social stigma	Tweets showing attitudes towards persons who drink alcohol	Contain blaming or labelling drinking person/behavior	การใช้คำเพื่อแสดงการตำหนิ หรือตีตราคนอื่น (ต่างจาก self-identification ตรงที่จะเป็นการใช้คำนั้นๆ ในแง่ลบ)
B. Advertisement		Tweets from sellers/promoters of alcohol or alcohol-related places/events		
	B1. Commercial ads	Advertisement of products contains alcohol content	Include brand or producer name	
	B2. Place or events	Advertisement of places or events that potentially serve alcoholic drinks	Also include tweets with positive comments about the places	

C. News				
	C1. News	News related to alcohol (could be adverse events (accidents, crimes) or investments, etc.	Messages that looks news format or headlines	
	C2. Knowledge sharing	Tweets aiming to share information and knowledge about alcohol	Documentary or short knowledge messages/tips	
D. Health-related content				
	D1. Observed/ known experiences from others	Effects of alcohol on physical and mental health as observed from the others	Sign, symptoms, effects, impacts occurs as a (direct/indirect) result of alcohol drinking	
	D2. Self-experiences	Effects of alcohol on physical and mental health as experienced by the poster him/herself	Sign, symptoms, effects, impacts occurs as a (direct/indirect) result of alcohol drinking	
E. Alcohol and sex				

<table>
<tr><td></td><td>E1. Use of alcohol with sexual activities</td><td>Tweets mention sex experience with the use of alcohol</td><td>Self-experience or observed experience, prostitution, risky behaviors</td><td></td></tr>
<tr><td></td><td>E2. Advertisement of porn videos/clips</td><td>Introduction or persuasion of buying sexual-related contents that involve alcohol drinking/intoxication</td><td>Tweets sharing links to clips/video/movies/novels (fictions) with sexual scene involving alcohol</td><td>เป็นข้อความโฆษณาพวกสื่อที่มีเนื้อหาเกี่ยวกับเพศและมีการดื่มแอลกอฮอล์ในเนื้อหานั้น</td></tr>
<tr><td colspan="5">F. Irrelevant</td></tr>
<tr><td></td><td>1) The message does not contain the terms related to alcohol/drinking/state of drunkenness at all</td><td></td><td>If the message contains the word "drunk", we assume that it is related to alcohol drinking UNLESS it mentions clearly that it is due to something other than alcohol.</td><td rowspan="2">ฉะนั้น เงื่อนไขการเลือกแต่ละขั้นตอนน่าจะเป็น
1) ตัดข้อความที่ไม่มีคำที่เกี่ยวข้องออกไป
2) ดูบริบทว่าคำที่เกี่ยวข้องนั้นมีนัยยะถึงการดื่มแอลกอฮอล์ใช่หรือไม่
3) ถ้าเกี่ยวข้อง ค่อยไปแยกว่า จะอยู่ในหมวดไหน
4) ส่วนที่เป็นประเภทของ</td></tr>
<tr><td></td><td>2) The message contains relevant terms but implies something other than alcohol drinking</td><td>Could be other types of drugs, behaviors, or just a metaphor</td><td></td></tr>
</table>

				เนื้อหา Personal communication vs. Ads vs. News จะเป็นการเลือกอันใดอันหนึ่ง 5) A1 กับ A2 อาจเลือกพร้อมกันได้ แต่ข้อย่อยข้างในต้องเลือกอันใดอันหนึ่ง

Annex 2. IRB certificate.

Certificate of Exemption from Ethical Review

The Committee for Research Ethics (Social Sciences)

★★★★★★★★★★★★★★

Certificate of Exemption No.	2022/019.1912
MUSSIRB No.	2023/001 (B1)
Title of Project:	MONITORING TWITTER CHATTER TO ASSESS THE BELIEFS, ATTITUDES, AND KNOWLEDGE OF THAI YOUTH AND DETECT TEMPORAL PATTERNS OF ALCOHOL USE AND ALCOHOL-RELATED RISKY BEHAVIORS
Principal Investigator:	ASST. PROF. DR. FRANCOIS RENE LAMY
Co-Investigator:	-
Name of Institution:	FACULTY OF SOCIAL SCIENCES AND HUMANITIES

The Committee for Research Ethics (Social Sciences) is in full compliance with International Guidelines of Human Research Protection such as Declaration of Helsinki, The Belmont Report, and CIOMS Guidelines.

Date of Determination: 19 December 2022

(Assoc. Prof. Pichet Kalamkasait)
Chairman

(Pol. Capt. Dr. Sutham Cheurprakobkit)
Deputy Dean for Research and Academic Services,
Faculty of Social Sciences and Humanities

Office of The Committee for Research Ethics (Social Sciences), Faculty of Social Sciences and Humanities, Mahidol University
Phuttamonthon 4 Rd., Salaya, Phuttamonthon District, Nakhon Pathom 73170. Tel. 0 2441 9180
Website: www.mussirb.com ; E-mail: mussirb310@gmail.com