Sentiment Analysis of Twitter data based on Cannabis Legalization

Smith Tripornkanokrat[1*], Ketsarin Boonkanit[2],
Nuttachai Kulthammanit[3] and Veerachai Suwatvanich[4]
[1,2,4]WeStride Institute of Technology
[3]Department of Computer Engineering, Faculty of Engineering,
Southeast Asia University, Bangkok, Thailand
*Corresponding author: smith@we-stride.com

Received: 21 September 2024 / Revised 1st:15 November 2024 / Revised 2nd: 5 December 2024 /Accepted: 7 December 2024

Abstract- This study delves into Thai public sentiment towards cannabis legalization by analyzing Twitter data from 2019 to 2024. Despite recent legalization for medical and industrial purposes, our findings reveal a persistent negative public perception and only 0.2-5.5% support for medicine based on positive tweets. Thais express significant concerns about potential societal harms, such as increased drug use and negative impacts on youth and levels of hate speech and harassment are around 31-35%, sexual content is 21-23%, and 7-11% for dangerous content were found on tweets. While proponents highlight medical benefits and personal freedoms, the broader online conversation remains dominated by negative associations linked to cannabis, including crime and societal decay. Employing advanced natural language processing and un-supervised learning techniques like RoBERTa, LDA, Word Cloud and Clustering are used to identify the group of similar content and are the core to provide score on tweets, we identified three distinct sentiment clusters: strongly opposed, mixed, and supportive. Our results clearly show that while discussions about cannabis have grown, negative sentiment continues to prevail, especially when linked to political issues and perceived threats to social order. Most common negative words are "พรรค", "ขาย", "ยาเสพติด", and "เด็ก" while positive words are "กำลังใจ", "ฟิน" and "ดีมาก" These findings underscore the complex interplay between public opinion, policy changes, and cultural attitudes toward cannabis in Thailand.

Keywords:- policy changes, sentiment analysis, public opinion, natural language processing, sentiment clusters.

1. Introduction

The legalization of cannabis in Thailand has sparked intense debates and differing perspectives among the government and the public. While the government has presented arguments in favor of legalization, citing potential benefits in various sectors, public opinion remains divided on the issue. The timeline presented details a significant shift in Thailand's cannabis policy from 2019 to 2024. Initially, the focus was on strictly controlled medical research and utilization. However, in 2022, there was a major policy shift that decriminalized cannabis, allowing for broader usage, including in food and beverages [1,2,3,4,5,6]. Between 2019-2021, gradual liberalization, with a focus on medical research and limited personal cultivation. 2022, major policy shift decriminalizing cannabis for broader use. 2023-2024, a pushback against the initial liberalization, with discussions and legislation aimed at re-imposing stricter controls on cannabis, particularly for recreational use. The timeline highlights the dynamic and changing nature of cannabis policy in Thailand, there's a tension between the government's focus on medical applications and the public's interest in recreational use.

Public sentiment around cannabis legalization has shown notable division. Study [18] found that negative sentiment among Thai Twitter users significantly rose following legalization, while positive sentiment declined. It further highlighted public concerns over cannabis-containing food, underscoring the need for clear regulations and public education on safe cannabis use. Cannabis legalization also impacts Thailand’s tourism sector, as evolving cannabis policies shape traveler experiences and raise public health considerations. Study [19] reveals that recreational substance use among international travelers poses both direct and indirect health risks, including neuropsychiatric issues, accidents, and risky behaviors. These findings emphasize the importance of public health interventions to promote safer travel practices, especially0020in destinations like Thailand, where cannabis policies are evolving.

This leads us to the questions, How have public attitudes towards cannabis evolved over this period? The legalization of cannabis has sparked widespread discussions on social media platforms, highlighting diverse opinions and potential implications for society This study exclusively uses Twitter data due to its suitability for capturing public sentiment on controversial issues like cannabis legalization. Twitter offers a unique platform where people express their opinions openly, providing a broad spectrum of perspectives. Its real-time and continuous updates allow researchers to track shifts in sentiment as they happen, a crucial feature for analyzing evolving public opinions. Additionally, Twitter is widely used for discussing social and political issues, making it a valuable source for identifying trends in public sentiment related to policy changes [20].

By examining Twitter data, this research seeks to understand the prevailing sentiment among Thai users, identifying patterns of support, opposition, or neutrality. The findings will contribute to a broader understanding of public attitudes towards cannabis legalization in Thailand, shedding light on the potential impact on policy-making and societal norms. This study aims to analyze the sentiment of Thai Twitter users regarding cannabis legalization in Thailand, utilizing natural language processing and unsupervised learning techniques including by leveraging RoBERTa (a state-of-the-art language model) [7], Latent Dirichlet Allocation (LDA) [8], and K-means clustering [9,10], Word Cloud [11] and Gemini 1.5 Flash to gain deeper insights into public opinion on this controversial topic. Chat-GPT and LLaMA can be also used for sentimental analysis, however Gemini shows significant rating difference across languages [12] which may be more suitable for Thai language. RoBERTa is used for sentiment analysis and determining the sentiment expressed in a piece of text (e.g., positive, negative, neutral). LDA is used to identify the main topics discussed in a collection of documents, here in this research is used for group clustering to classified neutral sentiment. K-means is applied to cluster the positive, negative, or neutral given from RoBERTa into a specified number of clusters (K), Elbows and Silhouette method are used to evaluate k clusters [13]. Word Clouds are visual representations of text data where the size of each word indicates its frequency of occurrence to identify common themes and keywords to list the most common top ten words. Gemini 1.5 flash is a large language model developed by Google AI that can be used to perform this task indirectly by text classification into different categories, including sentiment categories. Due to limitations in coding, Octoparse was chosen as an efficient web scraping tool [14]. It enables the extraction of large datasets from Twitter without requiring advanced programming knowledge, facilitating data collection, organization, and analysis. Its user-friendly interface, along with support for multiple output formats like Excel and CSV, enhances accessibility and ensures compatibility with the analysis methods employed in this study.

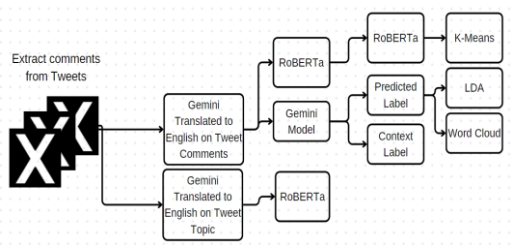

Fig.1 : The workflow methodology

2. Propose Method

This workflow demonstrates a multi-step process for conducting sentiment analysis on Thai tweets as illustrated in Fig. 1. It involves extracting comments from tweets, translating them into English, using RoBERTa for sentiment prediction, the Gemini model is applied to get the score prediction between 1 to 5 (1 means very negative, 3 means neutral and 5 means very positive) and also pull the context score including hate speech, harassment, sexual and dangerous scores between 1-5

Table 1: The top 10 common words on comments based on predicted label

Index	1	2	3	4	5	6	7	8	9	10
word_sentiment_1	กัญชา	คน	ทำ	ดี	พรรค	ขาย	เค้า	เสรี	ประเทศ	เรื่อง
word_sentiment_2	กัญชา	คน	ทำ	พรรค	ประชาชน	ขาย	ดี	ยาเสพติด	เสรี	เด็ก
word_sentiment_3	กัญชา	คน	ทำ	เค้า	ขาย	ดี	ไทย	พรรค	ปลูก	เรื่อง
word_sentiment_4	กัญชา	คน	เลือก	พรรค	ทำ	ประชาชน	ดี	เห็นด้วย	นะคะ	เหล้า
word_sentiment_5	กก	พี่	เลือก	กำลังใจ	รูป	ได้ดี	หน้าตา	ฟิน	คน	ดีย์

Table 2 : The top 11-20 common words on comments based on predicted label

index	11	12	13	14	15	16	17	18	19	20
word_sentiment_1	ประชาชน	ยาเสพติด	เลือก	เด็ก	ปลูก	ไทย	ไอ้	รู้	แบบนี้	หนู
word_sentiment_2	เรื่อง	ประเทศ	แบบนี้	เลือก	นายก	ไอ้	สังคม	ปชช	กัญชา เสรี	ไทย
word_sentiment_3	เสรี	ดู	เลือก	ยาเสพติด	ประเทศ	เด็ก	หนู	รู้	แบบนี้	ซื้อ
word_sentiment_4	สส	การแพทย์	ขาย	เค้า	ท่าน	ควบคุม	เขต	ชอบ	ไทย	สูบ
word_sentiment_5	เย	ย	Great	ทิม	ดีมาก	สะดุด	สู้	อนาคต	แน่นอน	หนู

(1 means very low and 5 means very high score), employing LDA for topic modeling when the predicted label is neutral to get deeper insight detail, and creating Word Cloud to list the most top ten common words, while K- Means help to cluster on RoBERTa score on 3 dimensions including roberta_pos, roberta_neu and roberta_neg.

2.1 Extract Comments from Tweets and ROBERTa on Tweet Topic

The data is collected from the Octoparse including 3634 comments and 85 tweet topics various times between 2019-2024 by filtering the interested year and comments on tweet topics must greater than 5 to collect data efficiency, after apply filtering, the tweet topics are selected

manually until it reaches all the available topics, considering after processing RobBERTa, we receive 5 positive tweet topics, 52 neutral tweet topics and 32 negative tweet topics or 5.6% , 58.4% , 36.0% respectively.

2.2 NLP Processing on Tweet Comments

2.2.1 Predicted and Context Label

All of the comments are translated to the English language and process separately, one goes to the gemini model to predict score and context label. This returns into json format and later to dataframe. The predicted labels are giving 50.9%, 32.0%, 11.3%, 5.5% and 0.2% on scores 3, 2 ,1 ,4 ,5 respectively. The response comment numbers on each tweet topic type are taken into account giving 55.9% on neutral tweets, 39.4% on negative tweets and 4.7% on positive tweets. In addition to visualizing how people respond to each tweet topic type, the percentage of predicted label and tweet types are plotted in Fig. 2 left, the results are 10.4 - 12.5% very negative, 31.0 - 44.7% negative, 41.2 - 52.6% neutral, 1.8 - 6.9% positive and 0-0.4% very positive on tweets. Furthermore, the sentiment trends over time are illustrated in Fig. 2 right, the Kernel Density Estimation (KDE) is used. When comparing negative, neutral, and positive tweets with levels of hate speech, harassment, sexual, and danger content it was found that the percentages were 33-35% , 31- 33% , 21- 23% , and 7- 11% , respectively.

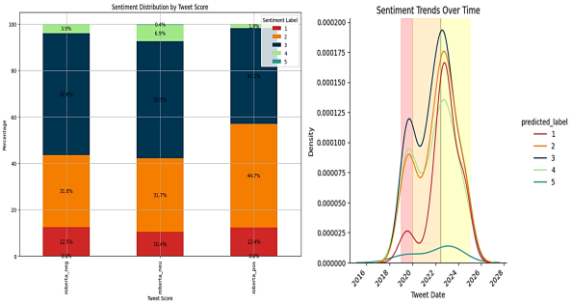

Fig.2: The visualization on how people respond to each tweet topic type (left) and the sentiment trends over time (right)

2.2.2 LDA and Word Cloud on Predicted Label

Considering that the most top 10 frequent words appear in tweet comments based on predicted labels are shown in Table 1 and Table 2 for the top 11 - 20 frequent words. Thai tweets comments are cleaned and preprocessed to remove stop words using thai_stopwords on **pythainlp** [15] (common words like “เพิ่งจะ”, “ที่ได้”, “พึ่ง”) and punctuations (e.g. comma, period, question mark and others) for all relative predicted label. While Gensim LDA model [16] method is considered to show the topic clusters, beside the relative predicted label on word cloud (word sentiment 1 - word sentiment 5) and the other one is unspecific, so totally 5 clusters or 5 number of topics are chosen to be compared beside the context label based on pre-processing on Thai tweets comments given $\lambda = 0.3$ since the lambda in the range of 0.3 to 0.6 is required [17]. The result of clusters density showed 73.3%, 23%, 1.5%, 1.2% and 1% of token words (e.g. cluster 0 is 73.3% and cluster 1 is 23%) and classify the noun, verb and emotion words using the gemini model, shown on Table 3.

Table 3: The LDA topic model clustering and classify noun, verb and emotions

index	Nouns	Verbs	Emotions
Cluster 0	กฎหมาย,กฎ,พี่,ร้าน,หัว,เมือง,เดิม,รัฐ	อยู่,ขาย,ปลูก,ตาม,สูบ,หมาย,ไว้,รับผิดชอบ ,ปลด,ควบคุม,ใส่,ซื้อ,ล็อค,หาเสียง,ฟัง,คุม ,ตั้ง	น่ะ,แหละ,รึ
Cluster 1	ละคร,หนูผี,แมร่ง,แลนด์,เกม,ความ น่าเชื่อถือ,สัตว์,รถ,นายใหญ่,เศรษฐกิจ, คะแนนเสียง	รัก,แลก,กลับไปกลับมา,ยอมรับ,เดือดร้อน, พลิกลิ้น,เลีย,ต่อต้าน,ปา,เอาใจ,เชื่อถือ	เดือดร้อน,แมร่ง
Cluster 2	ตั๋ว,กากี่นั้ง,เด,หด,Marijuana,พิย วิทยา,มีม,กอล์ฟ	ถอนฟ้อง,ตลก,เยี่ยม,แคร์,เสียบ,คบ, เป็,ปลิ้น,กสิ	ระยำ,อุ้ย,อิอิ,5555555555,น่า สมเพช,รำคาญ,แหม่,เชื่อได้
Cluster 3	รูป,ความต้องการ,ร่าง,หอย,คำสั่ง	ขำ,เคลียร์,เห็นดีเห็นงาม,รับฟัง	โหด,เพลีย,เจ๋ง,เศรษ
Cluster 4	ความเสี่ยง,การลงทุน,สนามกอล์ฟ,ผู้ ลงทุน,เขาใหญ่,คอน,ร้านอาหาร,ปาน	ลงทุน,เสด,เตร่,ชดใช้,อ่อ,ตระบัด,เต, แทร่,say,พังพินาศ,กลับกลอก,กลบ ,วิเคราะห์	ฮึ,กร๊าก,ก่,กร๊ากก,เต็ก,หม,อิ๊

2.2.3 K-Means on RoBERTa

K-means clustering can automatically group data and its efficiency can be improved if we know the exact number of clusters. The clustering process utilizes cluster centroids by calculating the minimum distance between the cluster centroid and the data points. Subsequently, the cluster centroid is updated to the mean position of all nearby data points, and this process is repeated until the cluster centroids converge to their optimal positions. By employing the elbow and silhouette methods to determine the optimal number of clusters. The plot often resembles an arm, with a distinct "elbow" where the curve bends sharply, and also used to avoid overfitting or underfitting. It was found that three clusters yielded the best results. The interpretation of each cluster is as follows: Cluster 0 likely represents primarily negative sentiments, accounting for 53.6% of the data. Cluster 1 may indicate a mix of sentiments with a slight positive leaning, comprising 35.9% of the data. Cluster 2 is likely to represent primarily positive sentiments, making up 10.5% of the data, illustrated in Fig.3.

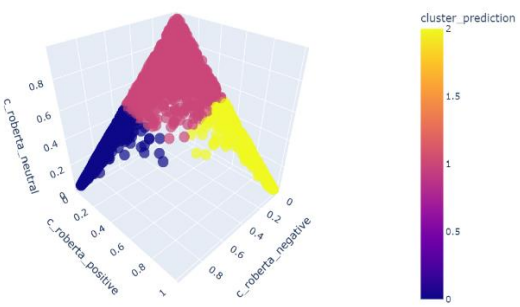

Fig.3: The visualization on cluster prediction on RoBERTa results

3. Summary and Discussion

A sentiment Analysis of public opinion on Thailand's cannabis legalization reveals a predominantly negative attitude. Many Thai citizens still view cannabis as a drug rather than a potentially beneficial substance, like for medical use. This negative perception becomes more intense when linked to political issues, particularly corruption. Fig. 2 on the right shows that as cannabis legalization gained traction, so did the intensity of opposition. Most comments, regardless of the news angle, leaned towards neutral or negative. Additionally, the majority of comments expressed hate or threats rather than concerns about danger or sexual content. This study confirms the results of Study [18], showing that negative opinions are becoming more common each year, while positive opinions are becoming less frequent. Interestingly, while the distribution of news articles wasn't perfectly balanced (with more neutral news than negative or positive) , the number of comments for each type of news was fairly similar. However, there was a significant disparity in the level of agreement. A small minority of people expressed support for legalization, while most comments were negative, often targeting political parties and their credibility, as well as concerns about societal impacts, especially on children. Those in favor often cited medical and recreational uses. We understand that cannabis can be beneficial when used appropriately, and it can also stimulate both household and medical economies. The government should strictly define the scope of use for each gender and age group, based on the results of LDA and Word Cloud analysis. These analyses reveal that the public is primarily concerned about the recreational use of cannabis and fears that uninformed children may misuse it. However, cannabis legalization policies can also be a political tool, with opponents of a particular party expressing their dissent through political comments. Therefore, to foster greater public consensus and reduce conflict, we should prioritize addressing the societal impacts of cannabis, particularly on children, before fully legalizing it. This approach would allow us to focus on the benefits of cannabis while mitigating potential negative consequences. When analyzing the language used in these comments, only a small percentage contained explicit insults or threats. Most negative comments were more subtle, like using the term 'หนู' (a slang term often used disparagingly towards those who support legalization). Neutral comments were most prevalent for both negative and neutral news, while positive news received a relatively higher proportion of negative comments, as shown in Fig. 2 on the left. Using K- Means clustering with RoBERTa, we identified three main clusters. This analysis provides strong evidence that a majority of comments were against cannabis legalization. A significant portion expressed a mix of negative and neutral sentiments, while genuine support was relatively rare. However, the K-Means clustering with RoBERTa model can only group the sentiments of individual tweets. To gain deeper insights into the social and cultural factors influencing opinions based on location, it would be beneficial to understand how negative or positive

sentiments are distributed across different regions of Thailand. This would enable us to mitigate negative impacts or foster understanding in specific areas.

In future research, given that Twitter data may not represent the viewpoints of all population segments and we cannot always accurately collect the locations of commenters, utilizing more diverse datasets would yield more robust results. By incorporating data on gender, age, residential location, occupation, and other relevant factors, we can more precisely address the underlying causes and better comprehend the relationships between positive, neutral, and negative sentiments.

References

[1] (2024, May 13). “Marijuana ในเงื้อมมือรัฐ คืน "กัญชา" กลับสู่บัญชียาเสพติด”: Thai PBS https://www.thaipbs.or.th/news/content/339943

[2] (2024, March 1). “กัญชาเสรีไทยใกล้ถึงจุดจบหรือไม่ เมื่อรัฐบาลประกาศตั้งเป้าห้ามใช้เพื่อ “สันทนาการ” ภายในสิ้นปี 2567”: BBC NEWS ไทย https://www.bbc.com/thai/articles/cn4lymm38myo

[3] (2022, Dec 05). “ร่าง พ.ร.บ.กัญชา: ส่องมาตรการควบคุมกัญชาหลังถูกปลดจากบัญชียาเสพติด“ : iLaw https://www.ilaw.or.th/articles/5534

[4] (2024, Jul 04). “อดีต ปัจจุบัน อนาคต นโยบายกัญชา“ : Thai PBS Policy Watch https://policywatch.thaipbs.or.th/article/life-43

[5] (2023, May 26). “ย้อนรอย “กัญชา” จากปลดล็อก สู่ MOU กลับเป็นยาเสพติด“ : The Active Thai PBS https://theactive.net/read/timelines-of-cannabis/

[6] (2021, Mar 10). “กัญชา - กัญชง เรียบเรียงไทม์ไลน์ อ่านง่าย สบายสมอง (2564)“ : Growcery https://theactive.net/read/timelines-of-cannabis/

[7] Adoma, A. F., Henry, N. M., & Chen, W. (2020, December). Comparative analyses of bert, roberta, distilbert, and xlnet for text- based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 117-121). IEEE.

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[9] Zul, M. I., Yulia, F., & Nurmalasari, D. (2018, October). Social media sentiment analysis using K-means and naïve bayes algorithm. In 2018 2nd International conference on electrical engineering and informatics (ICon EEI) (pp. 24-29). IEEE.

[10] Iparraguirre-Villanueva, O., Guevara-Ponce, V., Sierra-Liñan, F., Beltozar-Clemente, S., & Cabanillas-Carbonel, M. (2022). Sentiment analysis of tweets using unsupervised learning techniques and the k-means algorithm.

[11] Kabir, A. I., Ahmed, K., & Karim, R. (2020). Word cloud and sentiment analysis of Amazon earphones reviews with R programming language. Informatica Economica, 24(4), 55-71.

[12] Buscemi, A., & Proverbio, D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis. arXiv preprint arXiv:2402.01715.

[13] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP journal on wireless communications and networking, 2021, 1-16.

[14] Almaqbali, I. S. H., Al Khufairi, F. M. A., Khan, M. S., Bhat, A. Z., & Ahmed, I. (2019). Web scrapping: Data extraction from websites. Journal of Student Research.

[15 Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., Chormai, P., … & Udomcharoenchaikit, C. (2023). Pythainlp: Thai natural language processing in python. arXiv preprint arXiv:2312.04649.

[16] Tijare, P., & Rani, P. J. (2020, December). Exploring popular topic models. In Journal of Physics: Conference Series (Vol. 1706, No. 1, p. 012171). IOP Publishing.

[17] Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

[18] Lerksuthirat, T., Srisuma, S., Ongphiphadhanakul, B., & Kueanjinda, P. (2023). Sentiment and topic Modeling analysis on twitter reveals concerns over cannabis- containing food after cannabis legalization in Thailand. Healthcare Informatics Research, 29(3), 269-279.

[19] Charoensakulchai, S., Onwan, M., Kanchanasurakit, S., Flaherty, G., & Matsee, W. (2024). Recreational substance use among international travellers. Journal of travel medicine, 31(4).

[20] Lamy, F. R., & Meemon, N. (2024). Exploring Twitter chatter to assess the type and availability of cannabis-related products in Thailand. Journal of ethnicity in substance abuse, 1-21.

Authors' Biography

Smith Tripornkanokrat, bachelor's degree in Computer Engineering from the Institute of Music, Science, and Engineering at King Mongkut's Institute of Technology Ladkrabang. Currently, works at WeStride as a data mentor, focusing on data science, data engineering, and data analytics. My research interests include sound, acoustics, and sentiment analysis applied to machine learning and deep learning techniques.

Ketsarin Boonkanit holds a bachelor's degree and a master's degree in Business Administration from Kasem Bundit University and Ramkhamhaeng University, respectively. She is currently employed as an Assistant Marketing Manager at Dyno Paint Co., Ltd and is also affiliated with WeStride Institute of Technology. Prior to this, Ketsarin worked as a Marketing Key Account at Animal Food Bearing (Thailand) Co., Ltd., and as a Senior Sales Professional at Diethelm Co., Ltd.

Nuttachai Kulthammanit, bachelor's and master's degree in Computer Engineering from Chulalongkorn University. He currently works in the Department of Computer Engineering, at Southeast Asia University. His interests include microservice architecture and API design.

Veerachai Suwatvanich, bachelor's degree in Accounting- Finance from Rajamangala University of Technology Phra Nakhon. He currently works as a Financial Controller at Indian Ocean Tuna (IOT) and is also affiliated with WeStride Institute of Technology. His interests include deep learning, prompt engineering, RPA, and Fintech and Insurtech.