JOURNAL OF
EMERGING INVESTIGATORS

Article

Using machine learning to understand social media discourse on the co-use of tobacco and cannabis

Vineeth Godavarti[1], Helen Lam[2], Hye Min Kim[3], Donghee N. Lee[4], Elise M. Stevens[5]

[1] Westford Academy, Westford, MA
[2] Brigham and Women's Hospital, Boston, MA
[3] Department of Communication, University of Massachusetts, Boston, MA
[4] Population Sciences in the Pacific Program, University of Hawaii, Honolulu, HI
[5] Department of Population and Quantitative Health Sciences, Division of Preventive and Behavioral Medicine, UMass Chan Medical School, Worcester, MA

SUMMARY

The simultaneous use of tobacco products and cannabis is increasingly common among youth aged 18 to 24 years. Co-use has a deleterious effect on physical and mental health. Social media is being used to promote and discourage co-use. However, there is no comprehensive monitoring or surveillance on how co-use is discussed in these platforms. We conducted this research to analyze the content and sentiments of social media posts about co-use and develop a surveillance system using Deep Learning. We hypothesized that the majority of posts would have a positive sentiment promoting co-use, possibly driven by the rapid increase of tobacco and cannabis use. We entered co-use–related keywords into a social media monitoring platform and analyzed the sentiment data over a specific period. In addition, we developed a codebook with pre-defined features and manually classified the sentiment of a random subset of posts. Results from the manual coding study showed that approximately 3% of the posts had a positive sentiment towards co-use, while 28% had a negative sentiment. Approximately 60% of the posts mentioned physical or mental health effects of co-use. However, despite the harmful health effects, over 90% of the posts did not mention quitting. The manually coded data was used to train and develop a surveillance system with a Long Short-Term Memory (LSTM) classification model. The model achieved an accuracy of 78% on the test data and was further optimized to attain 91% accuracy, making it a potential tool for public health officials and researchers in the future.

INTRODUCTION

Young adults are reporting an increased use of cannabis along with tobacco products (co-use), with over 20% of people between the ages of 18 to 24 years describing use of both products in the past month (1,2). Co-use is characterized as the combined use of a tobacco product (i.e., cigarettes, e-cigarettes) and a cannabis product (e.g., blunts, edibles, beverages) in the past 30 days. Other definitions of co-use include consumption of both tobacco and cannabis simultaneously (e.g., vaping devices filled with THC or hollowed out cigars with cannabis), or separately in the same time period (3). Co-use increases the risk of becoming addicted to, and making it harder to quit, both substances, and has a significant negative impact on physical and mental health (4-9). As more states legalize recreational and medicinal cannabis, it is of the utmost importance to act against co-use addiction before it gets worse (10,11).

Young adults frequently rely on social media for communicating and gathering information on various topics, including access to vital health-related content (12-14). As such, co-use may be either promoted through advertising or discouraged through discussion of harmful impacts on social media. In a recent content analysis study on tobacco marketing, it was reported that pro-tobacco messages were prevalent on Instagram, with content that appealed to young people (15). Other studies have found that social media content related to cannabis predominantly presents a positive perspective, emphasizing its health and social benefits (16). Studies indicate that pro-cannabis advocates and individuals with industry ties may be leveraging social media as a persuasive tool, particularly to influence younger audiences (16).

While previous surveillance analyses have independently examined tobacco and cannabis content, there has yet to be a comprehensive study of social media posts discussing both substances together. Surveillance studies play an important role in shaping regulatory policies, helping construct educational campaigns, and improving the effectiveness of product warnings (17). There is no comprehensive monitoring or surveillance on how co-use is being discussed in these online landscapes. This gap in the scientific and public health literature provides an opportunity to help understand the social media landscape discussing co-use.

Social media monitoring tools help brands and businesses track online conversations, manage their reputation, and analyze mentions across the web. Many platforms use artificial intelligence (AI) and machine learning (ML) to report key metrics. Mention.com is one such platform using Natural Language Processing (NLP) techniques like sentiment analysis to classify social media posts as positive, negative, or neutral (18). For example, a post such as ‘cannabis is not carcinogenic’ is classified as having a positive sentiment while a post ‘the tobacco smoke made me feel nauseous’ is classified as negative. By using specific keywords such as co-use, it is possible to analyze and monitor online conversations on this topic. However, the scope of the surveillance and monitoring on the social media monitoring platform is usually limited to the specific domains like sentiment, volume, engagement, or reach of posts. To address this limitation, we conducted a manual coding or classification process that systematically analyzed social media posts mentioning

https://doi.org/10.59720/25-117

tobacco and cannabis for specific pre-defined features. We developed a codebook following manual coding approaches used in the past, and refined it upon qualitative review of the collected posts (**Table 1**). We expanded the scope of content analyses beyond what was provided in Mention.com (15, 19). Additionally, we used the manual coding data to train and develop a novel ML model to classify the sentiment of social media posts for use in monitoring or surveillance. We hypothesized that the majority of social media posts would have a positive sentiment towards co-use, driven by the rapid increase in tobacco and cannabis use.

In this research, we collected, classified, and analyzed the content of social media posts for specific pre-defined features related to co-use and investigated their relationship to online sentiment (i.e., positive, neutral, or negative) using ML tools. Long Short-Term Memory (LSTM) is a tool specifically built to process sequential data. LSTMs are useful in creating models where word order, context, and long-term dependencies in text comments within social media post are important. For example, in sentiment analysis of the phrase "not great", the word "not" modifies the following word "great", requiring the model to remember this context over time. Upon completion of the research, our Long Short-Term Memory (LSTM) classification model achieved an accuracy of 91% when determining the sentiment of test data posts. This approach could be further refined in future AI models to enhance the sentiment classification of social media posts related to co-use, providing valuable insights for regulatory policies and health communication campaigns.

RESULTS

To classify the sentiment towards co-use in social media posts, we took two approaches. In the first approach, we input keywords related to co-use directly in a social media monitoring platform (Mention.com) and analyzed the results from the dashboard directly. In the second approach, we developed a codebook with pre-defined features and used it to manually classify a random subset of social media posts across those features (**Table 1**).Finally, we developed an LSTM model and tested its ability to classify the sentiment of posts using training data from both approaches.

Content analysis of social media monitoring platform show majority of posts on co-use had a negative or neutral sentiment

The content analysis was based on approximately 168,000 posts from the three social media sites, extracted from a social media monitoring platform (Mention.com), containing the keywords about tobacco and cannabis (18). Most of the posts were on X (91.9%), followed by Instagram (7.84%), while TikTok had the least number of posts (0.17%).

Within the 168,000 posts, 'tobacco' was the most prevalent word (66,066 posts), followed by 'twitter' (48,928 posts) and 'weed' (46,347 posts) **(Figure 1A)**. 'Smoking', 'alcohol', and 'cigarette' were the next most quoted words (approximately 28,000-41,000 posts).

In addition to publishing a word cloud in the dashboard, the sentiment of the posts was classified as positive, negative, or neutral using an ML algorithm (20). Positive posts are those that express support, praise, or enthusiasm about a brand, product, or topic (e.g., "all I smoke is weed and tobacco now") while negative posts contain criticism, complaints, or dissatisfaction (e.g., "his followers are prohibited from consuming intoxicating substances like alcohol, tobacco, or weed). Neutral posts do not express any support or criticism (e.g. "have you smoked weed in a pipe?"). 42.9% of the posts had a negative sentiment, 23.5% of the posts had a positive sentiment, and 33.6% of the posts were considered neutral **(Figure 1B)**.

Manual coding of social media posts on co-use reflect predominantly negative or neutral sentiment

Since the data provided in the social media monitoring platform was limited to sentiment and a few other features, we developed a framework (or codebook) similar to what was done in previous content analysis studies and adapted it to include specific features for this study (21-23). We manually classified each post from the various social media sites for several features such as type of tobacco products or sentiment (**Table 1**). A total of 1000 posts from the entire dataset were randomly selected and coded by two coders (500 each) for various topics about the content of the post (**Table 1)**. We considered a post as relevant for co-use only if both tobacco and cannabis were mentioned. Based on this criterion, 49.7% of the posts (497/1000) were classfied as relevant.

Among the relevant posts, 71.4% mentioned 'nicotine' or

Domain	Type of feature coded
Type of tobacco products mentioned	Cigarettes, e-cigarettes, cigars, cigarillos, hookah, pipe tobacco, smokeless, snus, dissolvable, nicotine pouch, tobacco/nicotine generally mentioned, can't tell, other
Type of cannabis products mentioned	Smoked, Vaped' Blunt, Edibles Beverages, Cannabis generally mentioned, Can't tell, Other
Valence/Sentiment of Post	Pro-tobacco/pro-cannabis use, Anti-tobacco/anti-cannabis use, Pro-tobacco/anti-cannabis use, Anti-tobacco/pro-cannabis use, Neutral tobacco/neutral-cannabis use, Neutral tobacco/pro-cannabis use, Pro-tobacco/neutral-cannabis use, Anti-tobacco/neutral-cannabis use, Neutral tobacco/anti-cannabis use, Can't tell, Not Applicable
Type of Effect Mentioned	Physical, Mental, Emotional, Social, Environmental, Financial, Occupational, Buzz, Other, None mentioned
Mention of Quitting	Quitting tobacco, quitting cannabis, Quitting tobacco, no mention of quitting cannabis, Quitting cannabis, n mention of quitting tobacco, No mention of quitting either tobacco or cannabis
Mention of Replacement	Replacing cannabis with tobacco, Replacing tobacco with cannabis, No mention
Topics that are Mentioned	Regulation/Policy, Personal, Experience/Anecdote/Narrative, Health, Product Marketing
Features from Media monitoring platform Content Analysis	Type of social media - X, Instagram, TikTok, Word cloud, Sentiment - positive, negative, neutral, Emotion - joy, anger, disgust etc, Reach metric - number of unique viewers

Table 1: Features of social media posts included in the manual coding codebook. Domain refers to the feature of the post that was coded. The second column describes the types of features of each domain that was included in the analysis.

JOURNAL OF
EMERGING INVESTIGATORS

https://doi.org/10.59720/25-117

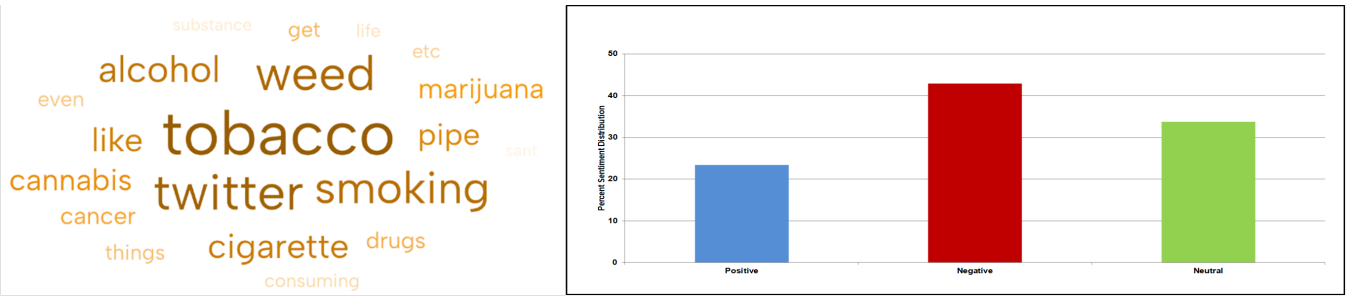

Figure 1: Data extracted from Media monitoring platform Dashboard. A) Word cloud generated from Mention.com social media monitoring platform. The word cloud represents the most frequently discussed words in comments from approximately 168,000 posts on the three social media platforms. 'Tobacco' was the most prevalent word, followed by 'twitter' and 'weed'. B) Distribution of the sentiment of posts classified in Mention.com's dashboard. Mention.com used an algorithm to classify the sentiment of the posts as positive, negative, or neutral. Majority of the 168,000 posts analyzed were classified as having a negative sentiment.

'tobacco', while 24.5% of the posts mentioned 'cigarettes' specifically. 'E-cigarettes', 'cigars', and 'hookah' were mentioned in about 12% of the relevant posts. Other tobacco products were rarely mentioned, with a range of 0-7 mentions. For types of cannabis products, 'cannabis' was generally mentioned in 78.3% of the relevant posts, while 'smoking cannabis' was coded in 20.5% of the posts. 'Vape' occurred in 2.0% of the posts, while 'blunt' appeared in 3.8% of the posts. All other types of cannabis products had very few occurrences (0-3 posts in total).

The manual classification data for sentiment of the posts showed that 51.5% of the relevant posts were neutral for both tobacco and cannabis products, while 28.4% of the posts had a negative sentiment towards both tobacco and cannabis products **(Figure 2)**. Only 3% of the posts were pro-tobacco and pro-cannabis. Since most of the posts did not have a positive sentiment towards co-use, the observations did not support our hypothesis that the majority of posts would have a positive sentiment towards co-use. Some of the posts had a positive sentiment for one of the products, with either a neutral or negative sentiment towards the other product. For example, 1.6% of the posts were pro-tobacco products and either anti- or neutral for cannabis products, while there was a slightly higher percentage of 5.4% that were pro-cannabis products and either anti- or neutral for tobacco products. The sentiment of about 7.9% of the posts was labeled 'other' since it was not possible to classify the sentiment as pro-, anti-, or neutral.

Next, the types of effects of tobacco or cannabis mentioned in the posts were classified. These data were not captured in Mention.com. Out of the relevant posts, 31.4% mentioned a physical effect (defined as affecting body functions and causing diseases) relating to the dangerous effects of tobacco and cannabis **(Figure 3)**. Additionally, 25.8% of the relevant posts discussed mental health (i.e., affecting brain function and leading to mental conditions such as anxiety and depression or, in some cases, releasing stress). Some posts described experiencing both physical and mental health effects. Furthermore, other effects, such as social effects relating to interactions and relationships, as well as environmental effects relating to impacting others through second-hand smoke, accounted for 4.8% and 7% of the relevant posts. Other classified effects included emotional, financial, or occupational impacts, which had very few mentions (fewer than 10 posts each).

To determine whether the association between co-use and the various effects described in the posts was statistically significant, we performed a statistical hypothesis test (chi-square test) (**Table 2**). The chi-square analysis showed a correlation between consumption of tobacco and cannabis products and the physical, mental, social, and environmental effects described in the posts. The chi-square values for the four different effects were 17.7, 23.9, 9.2, and 59.2, respectively. The associations between all classified effects and co-use were statistically significant ($p < 0.01$).

By analyzing posts mentioning quitting tobacco and/or cannabis products, we found that 90% of the posts did not discuss quitting either product, while just 9.1% mentioned quitting both. Very few posts mentioned quitting only one of the two products—three posts discussed quitting tobacco but not cannabis, while only one post mentioned quitting cannabis but not tobacco. There was no mention of replacing cannabis with tobacco products; 11 posts discussed switching from tobacco to cannabis products, while 97.6% of the

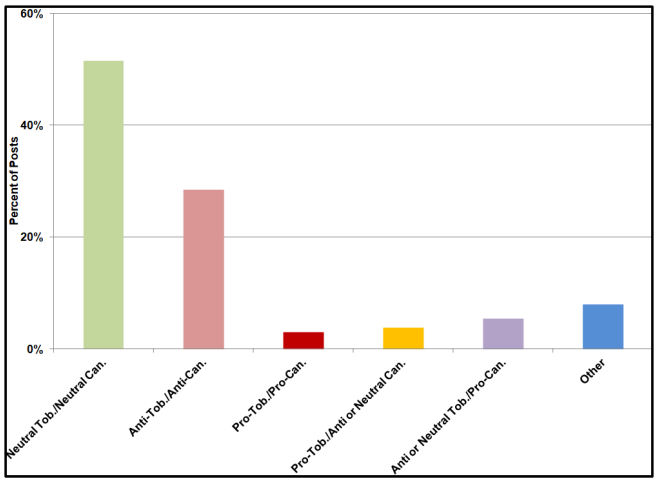

Figure 2: Distribution of sentiment data from manual coding. The distribution of the sentiment based on manual coding of 497 posts using the features described in Table 1. The sentiment of each of the posts was analyzed after reading them and classified as being pro-, anti-, or neutral towards tobacco (Tob.) products or cannabis (Can.) products, or both. Majority of the posts expressed a neutral sentiment towards co-use followed by negative sentiment, while positive sentiment was least prevalent.

https://doi.org/10.59720/25-117

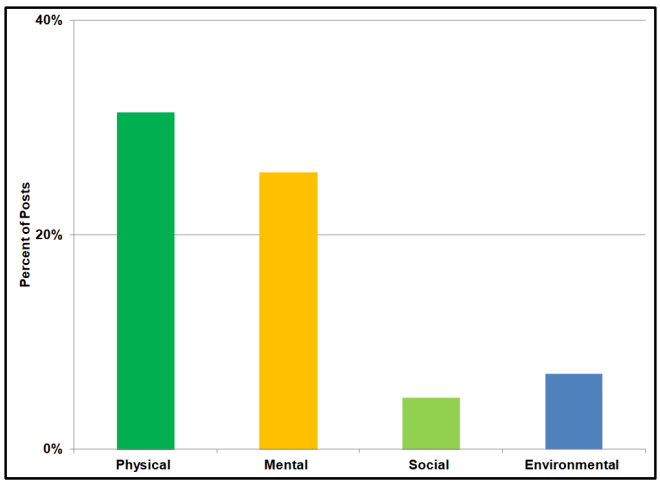

Figure 3: Discussion of the various types of effects from the manually coded data of relevant posts. The distribution of posts discussing the physical, mental, social, or environmental effects of co-use is shown. Examples of physical effects include health impacts mentioned such as lung disease, cancer. Mental effects include discussion on conditions such as depression, anxiety depression and stress. Social effects include any comments on relationship behaviors, and environmental effects include references to smell, second-hand smoke.

posts did not mention replacing one product with the other.

The final domain of interest was related to the nature of the posts. About 54.5% of the posts were users sharing anecdotes or narrating personal experiences. 32% of the posts were related to health, while 12.3% of the posts discussed regulations or public policies. Just 2.2% of the posts were related to product marketing.

An LSTM ML model was developed to classify the sentiment of social media posts with 91% accuracy

Next, we used the sentiment classification of posts (positive, negative, or neutral as described previously) to train and develop an ML model. To augment the overall data required for training the model, we included all 1000 posts (not just the 497 relevant posts. The 503 posts that were not relevant in the manual coding were classified as neutral for this study.

For initial training we used half of the initial 1000 posts, with 70% used to train the ML model and 30% used as test data. In addition, we also used the sentiment classification for the same 500 posts from the social media monitoring platform to train (350 posts) and test (150 posts) the model. The primary goal was to compare the sentiment classification between the two approaches.

The precision of the model can be described by a confusion matrix. A confusion matrix is a table that is used to describe the performance of a model by analyzing the differences between the actual values and the predicted values (25). In the 3x3 matrix, the actual sentiments are presented in rows, and the predicted sentiments (model output) are shown in columns. Each cell in the matrix provides the count of instances that are classified correctly or incorrectly. When we trained the LSTM model using sentiment classifications from Mention.com's dashboard, the model accuracy of sentiment classification was 63% **(Figure 4A)**. The model prediction accuracy for negative sentiment was the highest at approximately 80% (53/66), while the lowest accuracy was 28% for classifying positive sentiment (9/32). The neutral sentiment was coded at an accuracy of approximately 63% (33/52). The LSTM model, based on sentiment classification by manual coding, had an overall accuracy of 78%. The model prediction for neutral sentiment was the highest at approximately 95% (91/96), while the lowest accuracy was 18% for coding for positive sentiment (3/17) **(Figure 4B)**. The negative sentiment was coded at an accuracy of approximately 62% (23/37).

When the model was trained with data derived from manual coding, the accuracy of the test data was highest for neutral sentiment, while training the model with sentiment classification from Mention.com's dashboard resulted in the negative sentiment being coded with the highest accuracy. To understand the reason for this bias, we compared the distribution of sentiment classification in the training data between the manual coding approach and the social media monitoring platform to look for any class imbalance i.e. bias in model predictions towards the highest represented class in the training data **(Figure 5)**. The distribution was not balanced across all three sentiments and differed between the two coding approaches. The majority of posts in the data from the manual coding were coded neutral (approximately 60%). Conversely, the distribution of the data from the social media monitoring platform showed a higher number of posts being coded as negative (approximately 42%).

To address the issue of class imbalance described above, we took a resampling approach with just the manual coding dataset, which focused on increasing the representation of minority classes in the data. The number of samples in each sentiment increased to 579, for a total of 1737 samples in the dataset. Once again, we used 75% of the posts (1302 posts) to train the LSTM model, and the remaining 25% (435 posts) to test the model. This approach led to a significant improvement in the overall model accuracy to 91% **(Figure 6)**. Furthermore, the accuracy of coding the individual sentiments was also significantly improved to 96% for positive sentiment, 90% for negative, and 87% for neutral. Thus, having a balanced training dataset resulted in more consistent accuracy for coding each sentiment.

DISCUSSION

Combined use of cannabis with tobacco products is widespread among youth between the ages of 18 to 24 years (1,2). Co-use has harmful effects on both physical and mental well-being and poses a risk to public health (4-9). Social media is a powerful medium for communication among young

Type of Effect	Chi-square Statistic (degrees of freedom)	p-value
Physical	17.7 (2)	< 0.001
Mental	23.9 (2)	< 0.001
Social	9.2 (2)	0.01
Environmental	59.2 (2)	< 0.001

Table 2: Chi-Square Analysis of co-use and physical, mental, social, and environmental effects. The association between co-use and physical, mental, social, and environmental effects mentioned in posts is statistically significant, as seen by the high chi-square values relative to the critical factor (5.9) and low *p*-values.

https://doi.org/10.59720/25-117

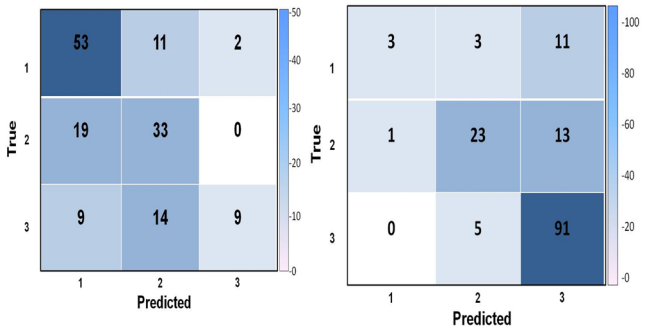

Figure 4: Accuracy of LSTM model in classifying sentiment of social media posts. The x-axis represents the true sentiment (manual coding or from the media monitoring platform) while the y-axis represents the sentiment coded by the neural network model. The values across the diagonal from left to right represent true positives. A) Model results from training (350 posts) on published values in Mention.com's dashboard. The 1, 2, and 3 axis labels represent the negative, neutral, and positive sentiment, respectively. Overall model accuracy was calculated to be 63%. B) Model results based on training (350 posts) on the sentiment data from the manual coding effort. The 1, 2, and 3 axis labels represent the positive, negative, and neutral sentiment, respectively. Overall model accuracy was calculated to be 78%.

adults for a variety of subjects, including health information (12-14). The goal of this research was to analyze the content of social media posts on the co-use of tobacco and cannabis products from three social media platforms, examine the sentiments in the posts, and build a surveillance system using deep learning. We used two approaches in this study. The first approach involved analyzing the data from over 168,000 social media posts using the data analytics features from Mention.com, a social media monitoring platform, directly. The second approach involved developing a codebook with specific pre-defined features and manually coding a random set of 1000 social media posts using keywords of tobacco and cannabis. These manually coded data then served as training data for developing an LSTM model that could be used to classify the sentiment of future posts.

Results from the manual coding study suggested widespread discourse on tobacco and cannabis products, which was consistent with the 'word cloud' data from Mention.com. Additionally, the manual coding approach revealed that over 50% of the posts were neutral towards co-use, while only 28% of the posts had a negative sentiment. However, when analyzing the larger dataset from Mention.com, the website's data analytics reported that approximately 43% of posts had a negative sentiment, while 33.6% had a neutral sentiment. The differences observed in the distribution of sentiment data are likely due to the differences in the sizes of the data sets (1000 posts for manual coding versus >168,000 posts for the platform dashboard). Another reason could be that the ML algorithm used by the social media monitoring platform interpreted the sentiment of posts differently from the manual coding approach.

The manual coding data revealed that approximately 60% of the posts mentioned a physical or mental health impact due to co-use. Importantly, the association between co-use products (cigarettes, e-cigarettes) and mention of these health effects was statistically significant. However, despite the negative sentiments expressed and the potentially harmful health effects, more than 90% of the posts had no mention of quitting. This observation is concerning as it suggests that although the negative effects of co-use on health are discussed online, there is little evidence of reducing co-use rates through quitting. ML tools such as recurrent neural network models (RNN), including LSTM models, are designed to process sequential data, such as text, speech, where the order of elements is important. They are particularly useful in natural language processing (NLP) and sentiment analysis (26). The LSTM ML model developed in this study coded the sentiment of posts with an overall accuracy of 78%. The model was further tuned and optimized to improve the overall accuracy to 91% by including a larger training dataset and incorporating resampling techniques to balance the number of samples with positive, negative, and neutral sentiments.

A few limitations were encountered in this research. One limitation was that the model tended to predict neutral sentiments more accurately than positive or negative sentiments. This is due to the limited number of positive and negative sentiment posts pulled from the platform, limiting the model's training dataset. While resampling techniques improved the prediction

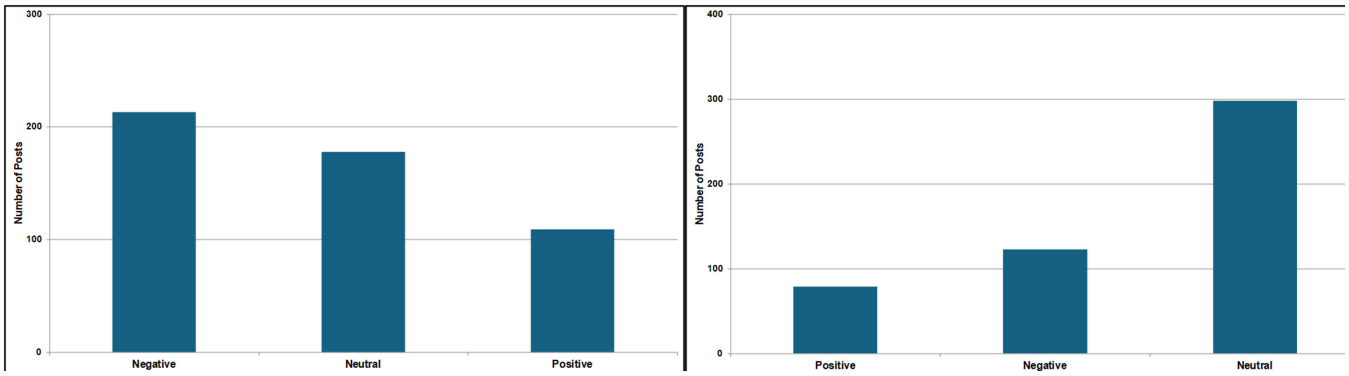

Figure 5: Distribution of sentiment data within manual coding and social media monitoring platform. A) The distribution of sentiments using manual coding data. B) The distribution of sentiment using data from the social media monitoring platform. The distribution was not balanced across all three sentiments and differed between the two coding approaches. The majority of posts in the data from the manual coding were classified as expressing a neutral sentiment. Conversely, the distribution of the sentiment data from Mention.com showed a higher number of posts being classified as negative.

https://doi.org/10.59720/25-117

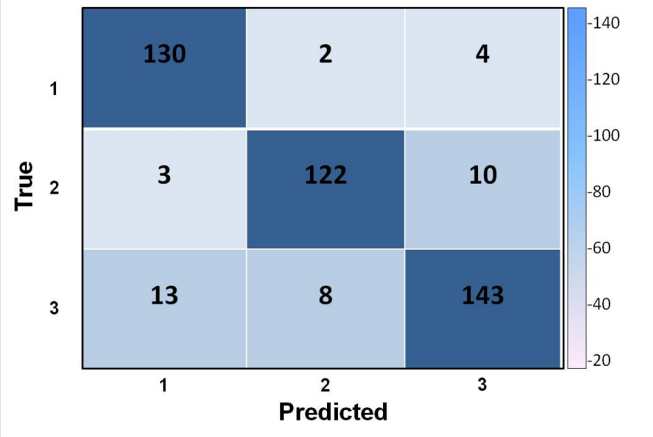

Figure 6: Impact of resampling on LSTM model accuracy in classifying sentiment. The x-axis represents the true sentiment (manual coding or from the Media monitoring platform) while the y-axis represents the sentiment coded by the neural network model. The training dataset was enhanced to increase the representation of the minority classes. The values across the diagonal from left to right represent true positives. Overall model accuracy was calculated to be 91%.

accuracy of all three sentiments, a more comprehensive and balanced dataset for initial training could have further improved the model's accuracy. Additionally, we also encountered platform bias since the majority of the posts were from X. A potential reason for this could be that our search focused on textual content while social media sites such as TikTok have the majority of their content in the form of short videos. This focus on text-based platforms may also have introduced an age-related sampling bias, as younger users, who are more active on video-centric platforms such as TikTok, may be underrepresented in the dataset. Another limitation was the subjectivity involved in manual coding of sentiment. Some posts were difficult to classify as positive, negative, or neutral towards co-use due to the ambiguous nature of the wording. To overcome this, the coding was performed by two people and applied uniformly and consistently across the dataset.

While the surveillance of social media did not reveal that most of the posts had a positive sentiment towards co-use as we originally hypothesized, it confirmed the widespread discussion of co-use topics and the extensive reach of these topics. This research demonstrates that ML models can predict the sentiment of social media posts on the use of tobacco with cannabis products. These algorithms can be used to implement automated surveillance of discussions related to co-use on various social media platforms. Automated approaches enable monitoring and surveillance over a broad range of platforms, larger time frames, and more efficiently. These data can be extremely useful for health policy makers in their efforts to reduce co-use by deploying targeted messages in social media to reduce the incidence of co-use.

MATERIALS AND METHODS

Data Collection from social media posts for content analysis

Data from social media posts were analyzed using Mention.com, a social media monitoring platform. Briefly, posts that contained keywords relating to both tobacco (e.g., vape, vaping, e-cigarettes, cigarettes) and cannabis (e.g., cannabis, marijuana, weed, pot, edibles) were used to search and collect posts from May 1, 2024 to October 9, 2024 (22-25). A total of 168,000 social media posts were analyzed. The media monitoring platform generated a 'word cloud' of words that appeared the greatest number of times in the posts. The website uses AI-based language learning models to identify the sentiment of the posts (i.e., positive, negative, or neutral) as well as the emotion (joy, anger, etc). Additionally, the website also captures the number of users who have viewed the content with its 'reach' metric.

Manual Coding: Codebook development

The manual coding approach involved a systematic, content analysis where social media posts that specifically mentioned tobacco and cannabis were examined for specific pre-defined features (16-18). Each post was tagged by the social media platform (i.e., Instagram, X, TikTok) and the date it was posted. A codebook was developed that captured features across the following domains: (1) types of tobacco products mentioned, (2) types of cannabis products mentioned, (3) sentiment of post, (4) type of effect discussed, (5) quitting, (6) replacing one product with other, and (7) theme/topic (**Table 1**). A form was created using the Qualtrics software platform to capture the name of the person coding, post ID, and whether it was relevant (i.e., had a mention of both tobacco and cannabis products). All the domains and the types of features were included in the Qualtrics form **(Table 1).** The coding process involved manually reading every single post and capturing the feature across the seven domains. For example, if the post mentioned the word cigarettes, that would be captured in the Qualtrics form under the 'type of tobacco products mentioned' domain.

Manual Coding: Inter-coder reliability

Two individuals (coders) were trained to systematically examine and classify the qualitative textual data from social media posts based on the predefined set of features across the domains (**Table 1)**. To ensure both coders were interpreting and classifying the content consistently, small random samples were independently evaluated and scored for inter-coder reliability using statistical methods to determine the Cohen's kappa value (24). A random sample of 30 posts was evaluated for the features across the seven domains and captured in the Qualtrics forms by both coders. The data were then compared to determine the Cohen's kappa score, aiming for a score of over 0.7 (24). After multiple rounds of training to ensure high levels of inter-coder reliability for each variable, good inter-coder reliability was achieved with the Cohen's kappa values ranging from 0.718 to 0.946 across the variables (24). This demonstrated robust agreement between the two coders. Following the training, the two coders classified the social media posts randomly selected from three social media platforms for the features across the seven domains as described above (Instagram, X, and TikTok; *N*=1000).

Developing the ML model using Python

An ML model was developed to classify the sentiment of social media posts. Model development and training were conducted using Keras and TensorFlow libraries in Python,

https://doi.org/10.59720/25-117

commonly used for building deep learning neural networks (27). The code was uploaded to Github **(Appendix).**

A Long Short-Term Memory (LSTM) network with an embedding layer was developed for sentiment classification. The model architecture consists of an embedding layer with an input dimension of 3200 and an output dimension of 256, followed by a single LSTM layer with 256 units. A dropout rate of 0.7 was chosen to avoid overfitting of training data. The model concludes with a dense output layer containing 3 units, corresponding to the three sentiment categories, using the Softmax activation function for classification.

Briefly, the textual dataset was coded for sentiment labels (Valence): positive (1), negative (2), and neutral (3). Following that, the data was preprocessed by removing Twitter handles (e.g., @user) and special characters, numbers, and punctuation. Next, short words (words with fewer than three characters) were filtered out. The data was then broken to smaller units (tokenized) and vectorized to convert the string into numerals to train the LSTM model (28).

The training data set included 350 posts, followed by model evaluation using the remaining 150 posts as test data to code for sentiment. The model was checked for the accuracy of predictions. To improve the model accuracy, the training data were augmented by including an additional 500 posts and removing any duplicate posts. To address class imbalance, oversampling was applied by replicating minority-class instances so that positive, neutral, and negative posts were equally represented in the training dataset, reducing prediction bias. This resampling approach, commonly used in machine learning (including LSTM models), helps improve overall accuracy and sentiment-specific performance. Model performance could be further enhanced by using a larger, balanced training dataset with similar sample sizes across all three sentiments.

ACKNOWLEDGMENTS

I would like to thank Dr. Piyush Agarwal for his mentorship and guidance in developing the machine learning model.

Received: May 22, 2025
Accepted: December 9, 2025
Published: April 1, 2026

REFERENCES

1. Schulenberg, John E., *et al*. "Monitoring the Future National Survey Results on Drug Use, 1975-2017. Volume II, College Students & Adults Ages 19-55." *Institute for Social Research*, Institute for Social Research. University of Michigan, 2018. ERIC, https://eric.ed.gov/?id=ED589764. Accessed 11 May 2025.
2. Cohn, Amy M., *et al*. "Patterns and Correlates of the Co-Use of Marijuana with Any Tobacco and Individual Tobacco Products in Young Adults from Wave 2 of the Path Study." *Addictive Behaviors*, vol. 92, May 2019, pp. 122–127, https://doi.org/10.1016/j.addbeh.2018.12.025.
3. Rabin, Rachel Allison, and Tony Peter George. "A Review of Co-Morbid Tobacco and Cannabis Use Disorders: Possible Mechanisms to Explain High Rates of Co-use." *The American Journal on Addictions*, vol. 24, no. 2, 6 Feb. 2015, pp. 105–116, https://doi.org/10.1111/ajad.12186
4. Rogers, Andrew H., *et al*. "Current Cannabis Use and Smoking Cessation among Treatment Seeking Combustible Smokers." *Drug and Alcohol Dependence*, vol. 209, Apr. 2020, p. 107928, https://doi.org/10.1016/j.drugalcdep.2020.107928.
5. Liu, Jessica, *et al*. "Motivations for Tobacco, Cannabis, and Their Co-Use among U.S. Young Adults Who Engage in Same-Day Co-Use." *Substance Use & Misuse*, 29 Nov. 2024, pp. 1–7, https://doi.org/10.1080/10826084.2024.2434682.
6. Hindocha, Chandni, *et al*. "Cannabis Use and Co-use in Tobacco Smokers and Non-smokers: Prevalence and Associations with Mental Health in a Cross-sectional, Nationally Representative Sample of Adults in Great Britain, 2020." *Addiction*, vol. 116, no. 8, 22 Jan. 2021, pp. 2209–2219, https://doi.org/10.1111/add.15381.
7. Reboussin, Beth A., *et al*. "Tobacco and Marijuana Co-Use in a Cohort of Young Adults: Patterns, Correlates and Reasons for Co-Use." *Drug and Alcohol Dependence*, vol. 227, Oct. 2021, p. 109000, https://doi.org/10.1016/j.drugalcdep.2021.109000.
8. Tucker, Joan S., *et al*. "Types of Cannabis and Tobacco/Nicotine Co-Use and Associated Outcomes in Young Adulthood." *Psychology of Addictive Behaviors*, vol. 33, no. 4, June 2019, pp. 401–411, https://doi.org/10.1037/adb0000464.
9. Meier, Ellen, and Dorothy K. Hatsukami. "A Review of the Additive Health Risk of Cannabis and Tobacco Co-Use." *Drug and Alcohol Dependence*, vol. 166, Sept. 2016, pp. 6–12, https://doi.org/10.1016/j.drugalcdep.2016.07.013.
10. Tucker, Joan S., Anthony Rodriguez, *et al*. "Cannabis and Tobacco Use and Co-Use: Trajectories and Correlates from Early Adolescence to Emerging Adulthood." *Drug and Alcohol Dependence*, vol. 204, Nov. 2019, p. 107499, https://doi.org/10.1016/j.drugalcdep.2019.06.004.
11. Watkins, Shannon Lea, *et al*. "A Mixed-Methods Study to Inform the Clarity and Accuracy of Cannabis-Use and Cannabis-Tobacco Co-Use Survey Measures." *Drug and Alcohol Dependence*, vol. 224, July 2021, p. 108697, https://doi.org/10.1016/j.drugalcdep.2021.108697.
12. Auxier, Brooke. "Social Media Use in 2021." *Pew Research Center*, 7 Apr. 2021, www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. Accessed 11 May 2025.
13. "Social Media and News Fact Sheet." *Pew Research Center*, 17 Sept. 2024, www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/. Accessed 11 May 2025.
14. Lim, Megan S, *et al*. "Young Adults' Use of Different Social Media Platforms for Health Information: Insights from Web-Based Conversations." *Journal of Medical Internet Research*, vol. 24, no. 1, 18 Jan. 2022, https://doi.org/10.2196/23656.
15. Liu, Jessica, Coralia Vázquez-Otero, *et al*. "Youth-Appealing Features in Popular e-Cigarette Brand Advertising in the USA after Heightened Scrutiny in 2018." *Tobacco Control*, vol. 32, no. 4, 21 Oct. 2021, pp. 497–500, https://doi.org/10.1136/tobaccocontrol-2021-056720.
16. Park, Sung-Yeon, and KYLE J. Holody. "Content, Exposure, and Effects of Public Discourses about Marijuana: A Systematic Review." *Journal of Health Communication*, vol. 23, no. 12, 5 Nov. 2018, pp. 1036–1043, https://doi.

https://doi.org/10.59720/25-117

org/10.1080/10810730.2018.1541369.
17. Moran, Meghan Bridgid, *et al*. "Selling Tobacco: A Comprehensive Analysis of the U.S. Tobacco Advertising Landscape." *Addictive Behaviors*, vol. 96, Sept. 2019, pp. 100–109, https://doi.org/10.1016/j.addbeh.2019.04.024.
18. Carval, Lucas. "AI: An Ongoing Journey at Mention ." *Mention*, 13 Nov. 2024, mention.com/en/blog/ai-at-mention/. Accessed 11 May 2025.
19. Brett, Emma I., *et al*. "A Content Analysis of Juul Discussions on Social Media: Using Reddit to Understand Patterns and Perceptions of Juul Use." *Drug and Alcohol Dependence*, vol. 194, Jan. 2019, pp. 358–362, https://doi.org/10.1016/j.drugalcdep.2018.10.014.
20. Araujo, Rodrigo. "Sentiment Analysis." *Mention Help Center*, 2024, en.support.mention.com/en/articles/2041310-sentiment-analysis. Accessed 11 May 2025.
21. Potter, Robert F., and Paul David Bolls. *Psychophysiological Measurement and Meaning: Cognitive and Emotional Processing of Media*. Routledge, 2012. www.perlego.com/book/1618206/.
22. Phan, Lilianna, *et al*. "Development and Pretesting of Hookah Tobacco Public Education Messages for Young Adults." *International Journal of Environmental Research and Public Health*, vol. 17, no. 23, 25 Nov. 2020, p. 8752, https://doi.org/10.3390/ijerph17238752.
23. Stevens, Elise M., *et al*. "People in E-Cigarette Ads Attract More Attention: An Eye-Tracking Study." *Tobacco Regulatory Science*, vol. 6, no. 2, 1 Mar. 2020, pp. 105–117, https://doi.org/10.18001/trs.6.2.3.
24. Cohen, J. "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*., vol. 20, no. 1, 1960, pp. 37–46, https://doi.org/10.1177/001316446002000104
25. Zhu, Ling. "Measuring Domain Shift Effect for Deep Learning In ..." *Universitat de Barcelona*, diposit.ub.edu/dspace/bitstream/2445/186091/3/tfg_zhu_ling.pdf. Accessed 01 June 2025.
26. Mienye, Ibomoiye Domor, *et al*. "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications." *Information*, vol. 15, no. 9, 25 Aug. 2024, p. 517, https://doi.org/10.3390/info15090517.
27. "Keras: The High-Level API for Tensorflow : Tensorflow Core." *TensorFlow*, www.tensorflow.org/guide/keras. Accessed 11 May 2025.
28. Friedman, Robert. "Tokenization in the Theory of Knowledge." *Encyclopedia*, vol. 3, no. 1, 20 Mar. 2023, pp. 380–386, https://doi.org/10.3390/encyclopedia3010024.

Appendix

GitHub Link to code:

https://github.com/vineethgodavarti/sentiment-analysis.git