HHS Public Access

Author manuscript

*Tob Control*. Author manuscript; available in PMC 2023 November 01.

Published in final edited form as:

*Tob Control*. 2023 November ; 32(6): 739–746. doi:10.1136/tobaccocontrol-2021-057243.

# Understanding e-cigarette content and promotion on YouTube through machine learning

**Grace Kong, PhD**[1], **Sebastian Alexander Schott**[1], **Juhan Lee, PhD**[1], **Hassan Dashtian, PhD**[2], **Dhiraj Murthy, PhD**[2]

[1]Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA

[2]School of Journalism and Media/Computational Media Lab, The University of Texas at Austin, Austin, TX, USA

## Abstract

**Introduction:** YouTube is a popular social media used by youth and has e-cigarette content. We used machine learning to identify the content of e-cigarette videos, featured e-cigarette products, video uploaders, and marketing and sales of e-cigarette products.

**Methods:** We identified e-cigarette content using 18 search terms (e.g., e-cig) using fictitious youth viewer profiles and predicted four models using the metadata as the input to supervised machine learning: 1) video themes, 2) featured e-cigarette products, 3) Channel type (i.e., video uploaders), and 4) discount/sales. We assessed the association between engagement data and the four models.

**Results:** 3830 English videos were included in the supervised machine learning. The most common video theme was “product review” (48.9%) followed by “instruction” (e.g., “how to” use/modify e-cigarettes; 17.3%); diverse e-cigarette products were featured; “vape enthusiasts” most frequently posted e-cigarette videos (53.1%) followed by retailers (19.0%); 43.2% of videos had discount/sales of e-cigarettes; and the most common sales strategy was external links for purchasing (31.5%). “Vape trick” was the least common theme but had the highest engagement (e.g., >2 million views). “Cannabis” (53.9%) and “instruction” (49.9%) themes were more likely to have external links for purchasing ($p<.001$). The 4 models achieved an F1 score (a measure of model accuracy) of up to .87.

**Discussion:** Our findings indicate that on YouTube videos accessible to youth, a variety of e-cigarette products are featured through diverse videos themes, with discount/sales. The findings highlight the need to regulate the promotion of e-cigarettes on social media platforms.

Corresponding Author: Grace Kong, PhD, 34 Park Street, New Haven, CT 06519. Grace.kong@yale.edu.

CONTRIBUTORSHIP STATEMENT

GK: conceptualized and designed the study; obtained funding for the study; interpreted the results; wrote the first draft;
DM: designed the study, acquired data, analyzed the data; interpreted the results; wrote sections
JL: wrote sections; interpreted the results;
AS: analyzed the data; interpreted the results; wrote sections
HD: assisted in data analysis

CONFLICT OF INTEREST

The authors do not have any conflict of interest.

## INTRODUCTION

E-cigarette use among youth is an epidemic in the United States (U.S.).[1] In 2021, past-month e-cigarette use among U.S. high school students was 11.3%, with 43.6% of these students using e-cigarettes on 20 or more days in the past month. E-cigarette use among youth is also high around the globe.[2–4] A main source of youth appeal of e-cigarettes may stem from how these products are portrayed on social media. Social media is mostly unregulated and provides a unique opportunity for the e-cigarette industry to market and promote their products. E-cigarettes are promoted on social media as part of the youth culture with themes that appeal to youth, such as “vape tricks” (i.e., blowing large amounts of clouds or shapes using exhaled aerosol)[5] and images that are visually appealing, which could elicit youth to engage with the content by liking or sharing it on their own social network.[6] Proe-cigarette content on social media has the potential to shape positive social norms surrounding use among youth. Indeed, experimental and observational studies showed that youth who use social media frequently are more willing to try, have positive attitudes toward, have less harm perceptions to, and are more likely to initiate e-cigarettes.[7–9] Thus, surveillance of e-cigarettes on social media is critical to inform tobacco control efforts.

In this study, we examined YouTube videos to understand how and which e-cigarette products are promoted to youth, who are promoting these content, and how e-cigarettes are being sold. YouTube is a social media platform that allows users to view, upload, post comments, and rate videos. YouTube use is steadily growing and is currently used by 2.1 billion users around the globe.[10] YouTube continues to be popular among youth despite other emerging social media platforms; recent data show that 77% of U.S. youth use YouTube.[11]

Most studies that examined e-cigarettes on YouTube used human coders. However, human coding is limited by the number of videos that could be analyzed, examining 22 to 350 videos. YouTube provides a plethora of data that cannot be fully examined using current methods. One study identified that YouTube had 28,000 videos on e-cigarettes, posted by 10,000 unique accounts, viewed over 100 million times, and commented on more than 280,000 times.[12] Advances in methods such as machine learning (ML) can enhance current methods to analyze more content than with human alone.

A recent scoping review found that 32 of 74 studies that used ML to understand tobacco content examined social media.[13] Most of the studies in this review examined Twitter and only one study examined YouTube, which identified that 97% of YouTube videos had pro -tobacco content.[14] However, less known is whether ML could be used to identify complex content relevant for surveillance, such as the video content, e-cigarette products, video uploaders, and presence of discount/sales. Video themes, such as instructions on “how to” modify an e-cigarette or how to conduct vape tricks could reveal trends in use, which may have implications for product regulation and health messaging. For example, the U.S. Food and Drug Administration (FDA) can prohibit open system e-cigarette products if these products are used in unsafe ways.[15] Furthermore, identifying the features (e.g., flavors, nicotine concentrations, and other additives) that are marketed in popular “product review”

videos can provide insight to product characteristics and marketing strategies that appeal to youth. Identifying video themes with high engagement could also provide insight into advertising targets and wants and needs of users who are engaging with this content.

To identify who and how e-cigarettes are promoted to youth, identifying video uploader (i.e., Channel) and presence of discount/sales on videos accessible to youth, is important. There could be a combination of promotional strategies used to increase user engagement: paid media (i.e., promotion through paid advertisements), earned media (i.e., promotion through a third party, such as an influencer), and owned media (i.e., promotion through a company's own website/social media account).[16]. Furthermore, it is widely known that social media platforms, including YouTube, custom tailors their content to user characteristics, but their algorithm is proprietary.

To address these gaps, we created fictitious viewer profiles separated by gender (male/female), age group (16 years old, 24 years old), and race/ethnicity (white, black, Hispanic) to mimic youth searching for e-cigarettes on YouTube. Additionally, to identify e-cigarette content and source of these videos, we used supervised ML to classify: 1) Video Themes, 2) Featured E-cigarette Products, 3) Channel Type, 4) Discount/Sales (See Table 1 for definitions). We also assessed whether video themes differed by featured e-cigarette products and Channel type, and whether presence of discount/sales differed by video themes, featured e-cigarette products, and Channel type.

## METHODS

### Search procedures:

We created 16 fictitious youth profiles to simulate youth searching for e-cigarettes on YouTube.[17] For each fictitious youth profile, we used Orbot, a mobile app that uses an anonymized "Tor" network, which is a set of servers around the world that relays traffic through each other to obfuscate IP address and other information such as geographic location.[18] Through this app, we searched the following words on YouTube in June 2020: "e-cigarette", "e-cig", "electronic cigarette", "e-liquid", "ENDS", "e-juice", "vape", "vaping", "vape juice", "box mods", "cigalikes", "disposable e-cigs", "disposables", "disposable vape", "pod mods", "vape mods", "vape pens," and "vape pods." Also using Orbot, we scraped 140 videos, which is equivalent to the number of videos shown on the first 7 pages for each search word. For each fictitious profile, we used a new SIM card, phone number, and performed a factory reset of the Android phone to conduct these searches to ensure that previous search did not affect the search outcomes. We obtained 4201 unique videos and extracted metadata such as the title, description of the video, view counts, number of likes/dislikes, comments, date of when the video was posted, and the Channel name from the YouTube Application Programming Interface (API; an application that allows access to data from a service/program such as YouTube).

### Supervised machine learning (ML) procedures:

The overall ML procedures involved: 1) human labeling, 2) text pre-processing, 3) training the supervised ML classification algorithms, and 4) evaluating the algorithms' performance.

### Human labeling:

We used deductive and inductive approaches to create the codebook (Table 1). The deductive approach involved creating a codebook with constructs and definitions based on the existing literature and topics relevant to inform e-cigarette regulation. The inductive approach involved refining each category through a human viewing randomly selected videos, reading the video titles and the descriptions, and following the external links to verify that these links led to e-cigarette retailers to confirm that they were retailers and the presence of sales of e-cigarette products. A second independent coder trained in the codebook used similar procedures to confirm these categories and amended them after discussing with the first coder. Then, a third independent coder trained in the modified codebook, randomly labeled 10% of the videos used for the training set to determine inter-rater reliability, but due to some videos being removed from YouTube, the coder labeled 74 videos. We obtained a Cohen's kappa of 0.93, $p < .001$. Finally, the second coder viewed and labeled 1000 videos. This labeled dataset was used as an input to supervised ML algorithms to train and test the rest of the corpus of videos.

### Text pre-processing:

We used the Python Natural Language Toolkit (NLTK)[19] to pre-process the texts of video titles and descriptions. We tokenized the data to split the raw text into tokens and removed punctuations, stop-words (e.g., "is," "the," "on"), and capitalizations, and applied stemming (i.e., removing suffixes such as "ing") and lemmatization (i.e., identifying groups of words with the same root word. For example, "e-cig" and "e-cigarette" are considered the same). We used a tokenizer to give each word a unique number with a capped vocabulary size of 800 (1200 for classifying the Channel Type because the data contained a larger vocabulary) and then computed the most common tokens.

### Supervised machine learning:

The ML model we developed is a sequential model with 2 layers. The first is an embedding layer, which uses the weights taken from the GloVe word embeddings. GloVe is a set of word embedding, which uses a word co-occurrence matrix that was built using a collection of billions of textual inputs.[20] This layer turns each word into a vector which represents its semantic meaning. The next layer is a bidirectional long short-term memory networks (BLSTM) layer, followed by the output layer. BLSTM is a deep, neural machine learning method that takes the complete sequential information of words before and after the target word to provide the context to enhance classification. This architecture has shown to be optimal for processing semantic data.[21]

In classifying Video Themes and Featured E-cigarette Product, we used video titles and descriptions as inputs to the ML model. We did not include video transcripts because they did not improve the model prediction. We excluded non-English videos (n=371), using the Python langdetect library,[22] from the training set and the prediction set. Videos which were classified as "Other irrelevant" (n=239; i.e., non-e-cigarette videos identified from the Video Theme construct) were excluded for classifying Featured E-cigarette Product, Channel Type, and Discount/Sales.

For classifying sales, we scraped the first 6 external links (excluding links to other social media platforms) provided on YouTube's video description. We developed a parser to extract the meta description (i.e., HTML element that provides a brief summary of the website) on the external website to obtain the web developer's website description as additional input for the classification. This description captured whether the website sold e-cigarette products (e.g., "buy vape now"). For classifying Channel Type, in addition to video titles and descriptions, we also added up to 50 of the most recent videos posted by each Channel to expand our training dataset for this construct, which eliminated the need for humans to label more videos.

### Performance evaluation:

The metrics used for performance evaluation were precision scores (true positive [i.e., the ML algorithm correctly classifying the categories within each construct] divided by sum of both true and false positive), recall (true positive divided by sum of true positive and false negative), F1 (harmonic average between precision and recall scores), and accuracy (sum of true positive and true negative divided by sum of true positive, true negative, false positive, and false negative).

### Data analysis:

Using the BLSTM classification model we developed, trained, and tested, we classified the videos to: Video Theme, Featured E-cigarette Product, Channel Type, Discount/Sales. We calculated the median and the interquartile range (IQR) for engagement variables (i.e., number of views, likes, dislikes, comments, upload date [2007–2018; 2019–2020]) that were non-normally distributed, and we calculated the frequencies for categorical variables (Table 2). To assess the association between each construct and engagement variable, we conducted Krusal-Wallis tests and Pearson chi-square tests (Table 2). We also conducted Pearson chi-square tests to assess whether Video Themes differed by Featured E-cigarette Product and Channel Type (Figure 1), and whether Discount/Sales differed by Video Theme, Featured E-cigarette Product, and Channel Type (Figure 2). The ML procedures and descriptive analyses were conducted with Python and STATA 16.0, respectively. The observational study of publicly available dataset was deemed exempt as human subjects research from the Yale Institutional Review Board (HIC# 2000028350).

## RESULTS

Of the 4201 videos in our dataset, we excluded 371 non-English videos, so 3830 videos were used in supervised ML to classify Video Themes. In classifying the Featured E-cigarette Product, Channel Type, Discount/Sales, we also excluded "irrelevant, non-e-cigarette videos" (n=239), which were determined from Video Themes. Table 3 lists the performance evaluation for each construct.

### Video Theme:

The most common video theme was product review (48.9%), followed by instruction (17.3%) and health information (11.3%). In general, engagement (e.g., likes) was low for all video themes except for videos that featured vape tricks. Vape trick videos had

the most views (Median=2,290,086 [IQR=10,156,343]) followed by instruction videos (Median=47,728 [IQR=181,672]). Higher proportion of instruction (69.6%) and vape trick (67.4%) videos were uploaded earlier (i.e., 2007–2018) and cannabis (57.1%) videos later (i.e., 2019–2020).

### Featured E-cigarette Product:

The most featured e-cigarette product was "non-specific" device (29.8%; e.g., vaping kit, e-cigarette subscription service), followed by box mods (25.1%), e-liquid (14.7%), and disposable pods (11.9%). Videos that did not feature a specific device and videos that featured box mods, vape pens, and pod systems had more engagement than videos that featured "other" e-cigarette products. Videos featuring cigalikes, e-liquid, and vape pens were uploaded in earlier years, whereas videos featuring disposable pods and pod systems were uploaded in recent years (Table 2).

### Channel Type:

The most common Channel Type was vape enthusiasts (54.0%) followed by retailers (20.3%). Overall, engagement was the highest for videos uploaded by private users followed by vape enthusiasts. A larger proportion of retailers posted videos in earlier years (61.9%) than recent years (38.1%).

### Discount/Sales:

43.2% of the videos had discount/sales. The most common discount/sales strategies were external links for purchasing an e-cigarette product (34.1%), followed by "other" promotional strategies (7.5%; e.g., e-cigarette product giveaways, non-e-cigarette merchandise such as t-shirts with e-cigarette images), and discount codes (1.6%; e.g., "enter XYZ code to get 20% off" to purchase e-cigarette products). Videos that had "other" promotional strategies had the most engagement and were posted in recent years, whereas videos that offered discounts and external links for purchasing were posted in earlier years (Table 2).

### Video Theme by E-cigarette Product and Channel Type (Figure 1):

The most featured e-cigarette products for each video theme (Figure 1A) was box mods (34.7%) in product review and e-liquid (24.1%) and box mods (22.9%) in instruction. Health information, "other relevant," vape trick, and cannabis themed videos did not feature a specific e-cigarette product.

Video Theme by Channel Type comparisons (Figure 1B) showed that vape enthusiasts were most likely to post vape trick (75.0%), product review (69.5%), and instruction videos (44.2%). Retailers were second most likely to post instruction (39.7%) and product review videos (16.4%). The medical community most often posted health information videos (41.5%). "Other Channel" and retailers most often posted cannabis videos, 50.0%, 18.6%, respectively.

### Discount/Sales by Video Theme, E-cigarette Products, and Channel Type (Figure 2):

Discount/sales by video theme (Figure 2A) comparisons showed that the majority of vape tricks (90.9%), health information (88.6%), and "other relevant" (73.7%) video themes did not have direct discount/sales. The most common promotional strategy used in cannabis vaping videos was external links for purchasing (53.9%). Instruction and product review videos were mixed; some had no direct discount/sales and some had links for purchasing (Product review: 53.8% no discount/sales, 36.9% purchasing links; Instruction: 46.1% no discount/sales, 49.9% purchasing links). "Other" promotional methods were more common among cannabis (15.4%) and product review themes (7.1%) than other themes.

Comparisons of discount/sales by e-cigarette product type (Figure 2B) and Channel Type (Figure 2C) showed that e-liquid videos (57.5%) had more purchasing links relative to other e-cigarette products. Retailers (62.8%) had greatest purchasing links than purchasing links presented in other Channel types.

## DISCUSSION

### Principal Findings

To the best of our knowledge, our study is the first to use supervised machine learning (ML) to classify complex constructs such as video theme, e-cigarette product, video uploader (i.e., Channel type), and discount/sales of e-cigarettes on YouTube. Consistent with prior studies on YouTube, we also identified that common video themes were product reviews and instructions on how to use/modify/create e-cigarette products, some of which had external links for purchasing.[5, 23, 24] These similar themes may not be surprising because we used similar search terms and did not restrict the years of the video upload. However, it is notable that these themes persisted over time. For example, instruction videos that were posted earlier were still prominent, which suggests relative popularity of this topic. This finding is consistent with findings from survey and qualitative studies that observed that youth find the ability to customize their e-cigarettes appealing, engage in this behavior (e.g., changing flavors, PG/VG ratio, voltage)[25], and learn how to do this through viewing YouTube videos.[26, 27]

We also observed new trends; "other" promotional strategies (which also had the most engagement) were posted in recent years, suggesting that the e-cigarette industry is using novel methods to promote their product. Such "other" promotional strategies included giveaways of e-cigarette products and sales of non-e-cigarette merchandise that alludes to e-cigarettes, such as clothing with e-cigarette-related logos/images. These findings highlight the utility of conducting surveillance of e-cigarettes on YouTube to identify novel promotional trends and trends that persist over time to inform e-cigarette regulation.

We also observed that "vape enthusiasts" and retailers were the most common Channel who posted e-cigarette videos about product reviews reviews and instructions, which is a novel way to promote e-cigarettes without using traditional paid ads that could be tracked and regulated. This finding is consistent with previous qualitative studies that observed that vape shop retailers considered social media as a new effective marketing channel to use strategies, such as featuring new e-cigarette products.[28, 29] We used the term "vape enthusiasts" to

refer to Channels that primarily posted videos about e-cigarettes.[30] However, we cannot confirm whether they are “influencers” who get paid by the e-cigarette industry to promote their product because funding source is not disclosed. “Vape enthusiasts” may also include individuals who are not paid by the industry but who are trying to become “influencers” through collecting followers through posting e-cigarette videos. Indeed, a recent study observed that youth who use e-cigarettes are motivated to become influencers to promote vaping products to get paid.[31] It is also notable that a non-negligible number of “private users” are also posting external purchasing links, and it is unclear whether these “private users” have ties to the e-cigarette industry.

Our findings also indicate that youth who search for e-cigarettes may be inadvertently exposed to cannabis vaping content. Cannabis vaping among youth is high.[32] Moreover, recent research showed that e-cigarette use is one of the predictors for cannabis vaping among youth.[33] More research is needed on how cannabis and nicotine vaping products are promoted to prevent youth use of both substances.

## Regulation of e-cigarette promotion on YouTube

Overall, our findings underscore the presence of e-cigarette promotion on YouTube videos accessible to youth. While some of the promotions fall under “earned media,” in which third party retailers such as vape shops are marketing and selling e-cigarette products through posting videos themes such as product reviews and instruction, most of the promotion do not fall under media promotions that can be regulated. For instance, e-cigarettes are not advertised using traditional paid advertisements(i.e., “paid media”), the presence of “influencers” cannot be verified due to the lack of clear financial disclosure (i.e., “earned media”), and e-cigarette manufacturers/brands are not directly selling their products on their YouTube accounts (i.e., “owned media”). However, we did not examine content such as paid advertisement banners, so it is possible that paid advertisement exists on YouTube.

Currently, YouTube has self-imposed policies that attempt to restrict tobacco content including e-cigarettes. YouTube prohibits the sales of e-cigarettes through posting contact information including external links.[34] YouTube also discourages e-cigarette content through limiting advertising revenues that could be earned on videos with tobacco content.[34] Finally, e-cigarette content such as product review of e-cigarettes are restricted to underage youth.[35] Despite these efforts, our study findings indicate the need for stricter enforcement of these policies to protect youth from e-cigarette content.

The duty to protect youth does not fall on YouTube alone. The government entities could set restrictions on tobacco marketing, including e-cigarette marketing on venues frequented by youth, such as social media. There should be greater efforts to counteract pro-e-cigarette content on social media. For instance, we observed that medical community’s videos that showcased health risks of e-cigarette use through interviews with experts and research presentations were prominent but had the lowest engagement (e.g., < 2,600 views). Health information videos are competing with many diverse pro-e-cigarette content on YouTube, and novel methods are needed to make this content appeal to youth and increase engagement.

## Strengths/Weaknesses

Social media's algorithm, including YouTube's, tailors the search results based on a variety of factors such as one's profile information, search and view history and other conglomerate factors, which are proprietary information. Existing studies have not used any personalization when identifying e-cigarette or other tobacco content, and our study is the first to use fictitious youth viewer profiles to search for e-cigarette videos. Despite this strength, fictitious viewer profiles are only the initial step in identifying the types of content that youth may be exposed to, and future research should leverage this method to obtain more relevant and accurate information on e-cigarettes on social media. For instance, actual youth may be searching for other terms related to e-cigarettes and not just a single word. Relatedly, our findings should be interpreted with the caveat that our search words were broad and general and did not include brand-specific search words like, "JUUL" or "Puffbar." Despite these limitations, our findings showed that youth who do not have any viewing history related to e-cigarettes could be exposed to diverse e-cigarette content, including concerning content on YouTube, such as direct sales and discounts of e-cigarettes and cannabis vaping.

Another strength is that the ML performance was robust and comparable to those identified in other studies that examined complicated themes, such as user sentiment to e-cigarette content on Twitter,[36] which demonstrates that ML models can be used to identify relevant themes for e-cigarette surveillance on social media. However, we acknowledge areas for improvement; our lowest F1 score was in classifying the featured e-cigarette product. We used the title and the video description in our ML model to classify each construct, but future research may improve model performance through using visual classification. Visual classification also may be able to identify the characteristics of "models" featured on e-cigarette videos and other visual components like the presence of warning labels.

Although we analyzed many videos (N=3830), our method trades off volume for personalized results. We could not conduct automated queries to the YouTube API to obtain more videos because our method of using fictitious profiles involved having to factory reset the phone for each fictitious viewer profile and then web scraping results for each search term. Since we did not examine all videos, it is possible that emerging themes such as COVID-19 related themes were not detected at the time of our search (July 2020).

Our search was conducted using English so the content that we obtained may be relevant to the English-speaking U.S. population. Non-English speakers may be exposed to e-cigarette content in other languages so non-English content should be examined in future studies, as both YouTube and e-cigarettes have global reach.

## Summary

In summary, we identified video themes, featured e-cigarette products, who posted these videos, and discount/sales on youth-accessible YouTube videos. Our findings highlight the utility of using advanced ML methods to conduct surveillance of e-cigarette use trends and marketing/sales strategies on social media platforms such as YouTube. Our findings also

underscore the need for comprehensive federal regulations to protect youth from exposure to promotion of e-cigarette and cannabis vaping on YouTube.

## FUNDING STATEMENT

Research reported in this publication was supported by grant number R01DA049878 from the National Institute on Drug Abuse (NIDA) and U.S. Food and Drug Administration (FDA) Center for Tobacco Products (CTP). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the FDA.

## REFERENCES

1. Park-Lee E, Ren C, Sawdey MD, et al. Notes from the field: E-cigarette use among middle and high school students - national youth tobacco survey, united states, 2021. MMWR Morb Mortal Wkly Rep 2021;70(39):1387–1389. 10.15585/mmwr.mm7039a4. [PubMed: 34591834]
2. Gottschlich A, Mus S, Monzon JC, et al. Cross-sectional study on the awareness, susceptibility and use of heated tobacco products among adolescents in guatemala city, guatemala. BMJ Open 2020;10(12):e039792. 10.1136/bmjopen-2020-039792.
3. Tarasenko Y, Ciobanu A, Fayokun R, et al. Electronic cigarette use among adolescents in 17 european study sites: Findings from the global youth tobacco survey. Eur J Public Health 2022;32(1):126–132. 10.1093/eurpub/ckab180. [PubMed: 34694383]
4. World Health Organization. Report on the global tobacco epidemic. Geneva 2019.
5. Kong G, LaVallee H, Rams A, et al. Promotion of vape tricks on youtube: Content analysis. J Med Internet Res 2019;21(6):e12709. 10.2196/12709. [PubMed: 31215510]
6. Alpert JM, Chen H, Riddell H, et al. Vaping and instagram: A content analysis of e-cigarette posts using the content appealing to youth (cay) index. Subst Use Misuse 2021;56(6):879–887. 10.1080/10826084.2021.1899233. [PubMed: 33749515]
7. Vogel EA, Ramo DE, Rubinstein ML, et al. Effects of social media on adolescents' willingness and intention to use e-cigarettes: An experimental investigation. Nicotine Tob Res 2021;23(4):694–701. 10.1093/ntr/ntaa003. [PubMed: 31912147]
8. Zheng X, Li W, Wong S-W, et al. Social media and e-cigarette use among us youth: Longitudinal evidence on the role of online advertisement exposure and risk perception. Addict Behav 2021;119:106916. 10.1016/j.addbeh.2021.106916. [PubMed: 33798917]
9. Lee J, Tan ASL, Porter L, et al. Association between social media use and vaping among florida adolescents, 2019. Prevent Chronic Dis 2021;18:E49–E49. 10.5888/pcd18.200550.
10. Ceci L YouTube - Statistics & Facts. Statista Jul 12, 2021.
11. Ceci L Number of YouTube users worldwide from 2016 to 2021. Statista August 31, 2021.
12. Huang J, Kornfield R, Szczypka G, et al. A cross-sectional examination of marketing of electronic cigarettes on twitter. Tob Control 2014;23:iii26–iii30. 10.1136/tobaccocontrol-2014-051551. [PubMed: 24935894]
13. Fu R, Kundu A, Mitsakakis N, et al. Machine learning applications in tobacco research: A scoping review. Tob Control 2021:tobaccocontrol-2020–056438. 10.1136/tobaccocontrol-2020-056438.
14. Kim K, Gibson LA, Williams S, et al. Valence of media coverage about electronic cigarettes and other tobacco products from 2014 to 2017: Evidence from automated content analysis. Nicotine Tob Res 2020;22(10):1891–1900. [PubMed: 32428214]
15. Food and Drug Administration. Deeming tobacco products to be subject to the federal food, drug, and cosmetic act. Fed Regist April 25, 2014;79:No. 80.
16. Stephen A, Galak J. The effects of traditional and social earned media on sales: A study of a microlending marketplace. Soc Sci Res Network 2012. 10.2139/ssrn.1480088.
17. Dashtian H, Murthy D, Kong G. An exploration of e-cigarette–related search items on youtube: Network analysis. J Med Internet Res 2021.
18. Tor Project. https://www.torproject.org/ March 11, 2022.
19. Natural Language Toolkit. https://www.nltk.org/2021.

20. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. https://nlp.stanford.edu/projects/glove/ 2014.
21. Song M, Zhao X, Liu Y, et al. Text sentiment analysis based on convolutional neural network and bidirectional lstm model. 24th International Conference on Automation and Computing Singapore: Springer Singapore 2018:55–68.
22. Langdetect 1.0.9. https://pypi.org/project/langdetect/ 2021.
23. Luo C, Zheng X, Zeng D, et al. Portrayal of electronic cigarettes on youtube. BMC Public Health 2014;14:1028. http://www.biomedcentral.com/1471-2458/14/1028. [PubMed: 25277872]
24. Guy MC, Helt J, Palafox S, et al. Orthodox and unorthodox uses of electronic cigarettes: A surveillance of youtube video content. Nicotine Tob Res 2019;21(10):1378–1384.10.1093/ntr/nty132. [PubMed: 29961828]
25. Camenga DR, Morean ME, Kong G, et al. Appeal and use of customizable e-cigarette product features in adolescents. Tob Regul Sci 2018;4:51–60. 10.18001/TRS.4.2.5. [PubMed: 33163582]
26. Kong G, Morean ME, Bold KW, et al. Dripping and vape tricks: Alternative e-cigarette use behaviors among adolescents. Addict Behav 2020;107:106394. 10.1016/j.addbeh.2020.106394. [PubMed: 32222561]
27. Amin S, Dunn AG, Laranjo L. Exposure to e-cigarette information and advertising in social media and e-cigarette use in australia: A mixed methods study. Drug Alcohol Depend 2020;213:108112. 10.1016/j.drugalcdep.2020.108112. [PubMed: 32574981]
28. Yang JS, Lee E. A qualitative assessment of business perspectives and tactics of tobacco and vape shop retailers in three communities in orange county, ca, 2015–2016. Arch Public Health 2018;76:57. 10.1186/s13690-018-0307-z. [PubMed: 30349691]
29. Cheney M, Gowin M, Wann TF. Marketing practices of vapor store owners. Am J Public Health 2015;105(6):e16–21. 10.2105/ajph.2015.302610.
30. Kong G, LaVallee H, Rams A, et al. Promotion of vape tricks on youtube: Content analysis. J Med Internet Res 2019;21(6):e12709. 10.2196/12709. [PubMed: 31215510]
31. Vassey J, Metayer C, Kennedy CJ, et al. #vape: Measuring e-cigarette influence on instagram with deep learning and text analysis. Front Commun 2020;4(75). 10.3389/fcomm.2019.00075.
32. Morean ME, Kong G, Camenga DR, et al. High school students' use of electronic cigarettes to vaporize cannabis. Pediatrics 2015;136(4):611–616. 10.1542/peds.2015-1727. [PubMed: 26347431]
33. Lee J, Kong G, Kassas B, et al. Predictors of vaping marijuana initiation among us adolescents: Results from the population assessment of tobacco and health (PATH) study wave 3 (2015–2016) and wave 4 (2016–2018). Drug Alcohol Depend 2021;226:108905. 10.1016/j.drugalcdep.2021.108905. [PubMed: 34304122]
34. YouTube. Community guidelines. https://www.youtube.com/yt/policyandsafety/communityguidelines.html 2022.
35. YouTube. Age-restricted content. https://support.google.com/youtube/answer/2802167?hl=en 2022.
36. Visweswaran S, Colditz JB, O'Halloran P, et al. Machine learning classifiers for Twitter surveillance of vaping: Comparative machine learning study. J Med Internet Res 2020;22(8):e17478. 10.2196/17478. [PubMed: 32784184]

## WHAT THIS PAPER ADDS

**What is already known on this topic**

Pro-e-cigarette content is prolific on social media platforms such as YouTube. It is unknown whether machine learning can be used to classify complicated themes that can inform tobacco control, such as video themes, featured e-cigarette products, uploader type, and presence of discount/sales.

**What this study adds**

We identified YouTube videos related to e-cigarettes using fictitious youth viewer profiles. Our supervised machine learning identified video themes (e.g., product review, instruction), e-cigarette products, uploader type (e.g., retailers, “vape enthusiasts”), and presence of discount/sales.

**How this study might affect research, practice, or policy**

Despite YouTube’s policies to restrict tobacco content to youth, we found that promotion of e-cigarettes was prevalent, indicating the need for comprehensive marketing restrictions and enforcement. Machine learning can be used to monitor tobacco promotion on YouTube.

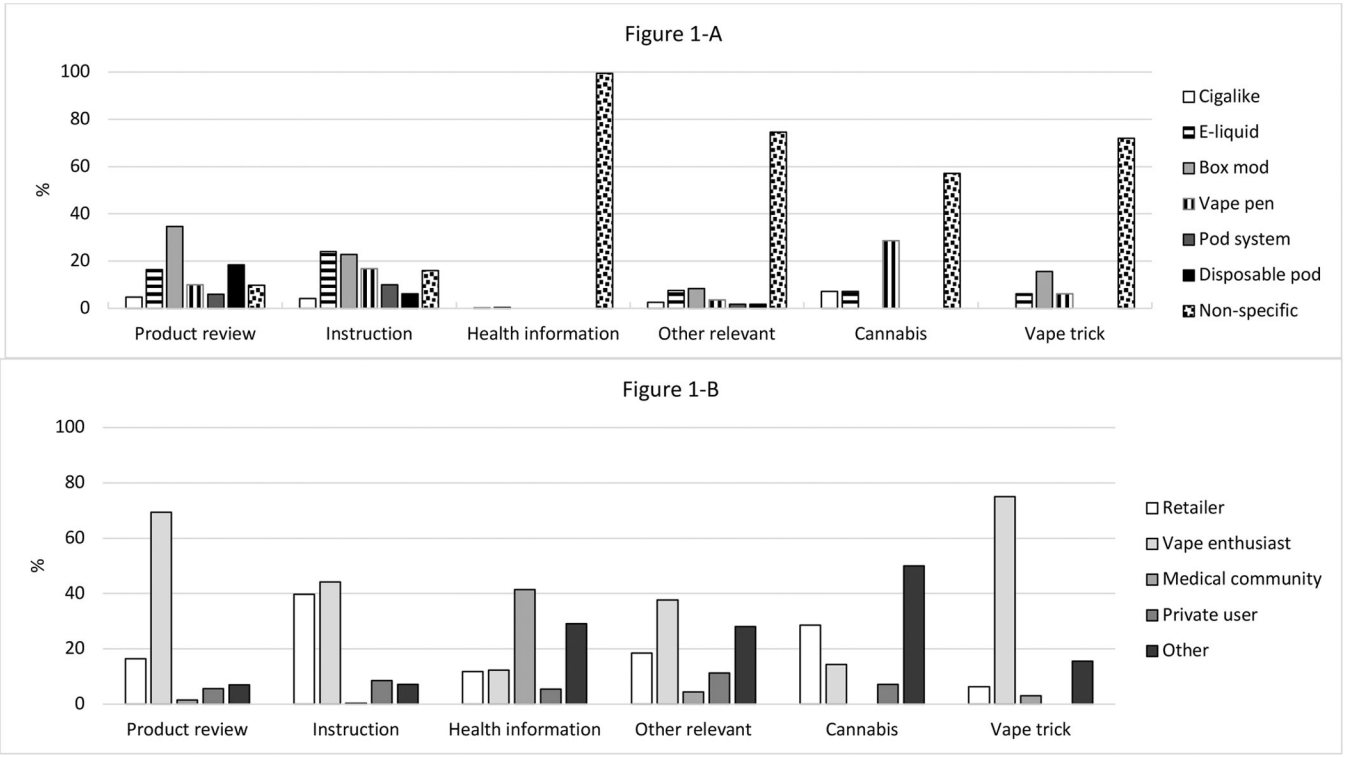


**Figure 1.**
Percentages of video theme by featured e-cigarette product (A) and Channel type (B)

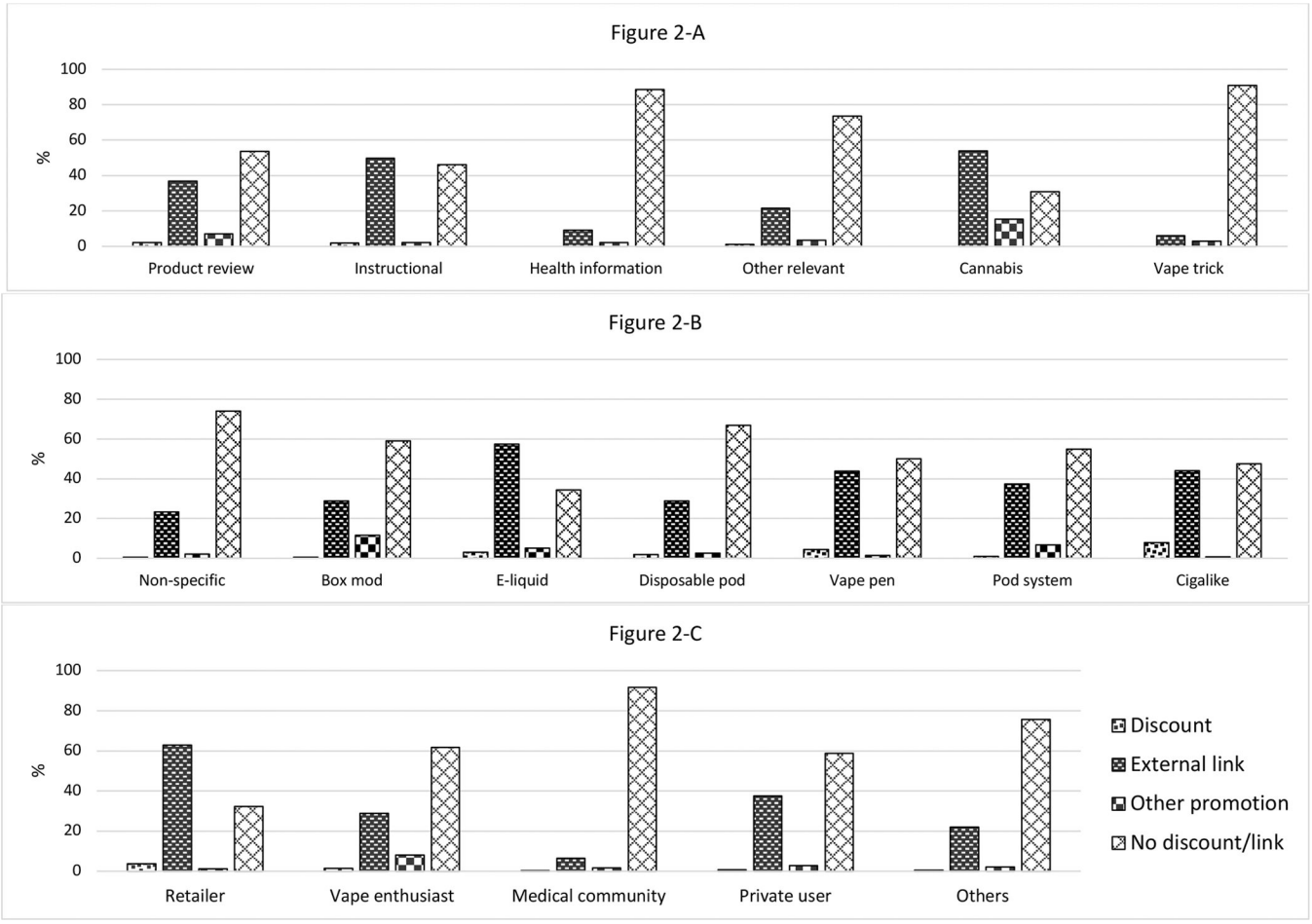


**Figure 2.**

Percentages of presence of sales/marketing by video theme (A), e-cigarette product (B), and Channel type (C)

**Table 1.**

Definition of constructs classified by supervised machine learning

| Construct | Definition |
|---|---|
| Video Theme | |
| Product review | Opinion about an e-cigarette product (e.g., top 10 e-liquid juice; “why I love my JUUL”) |
| Health information | Health information related to e-cigarette use (e.g., research presentations or interviews with a medical professional on the health risks of e-cigarettes) |
| Instruction | “How to” or instructional videos on using or modifying an e-cigarette product (e.g., mixing/making e-liquids; how to hack a device; “beginner’s guide” on how to vape) |
| Vape trick | Using an e-cigarette to blow a large volume of vapor or shapes (e.g., “selfie” video of someone doing a vape trick; compilation of various vape tricks) |
| Cannabis | Any theme featuring cannabis vaping (e.g., product review of a CBD e-liquid) |
| Other relevant | Content is related to e-cigarettes but does not belong to any of the above categories (e.g., news/TV clips; e-cigarette commercials; comedy sketches) |
| Other irrelevant | Content is not related to e-cigarettes |
| E-cigarette Product Type | |
| Cigalike | Shaped liked a cigarette, could be disposable or rechargeable |
| E-liquid | Liquid that is used in an e-cigarette device |
| Vape pen | Shaped like a pen, has a tank and a battery |
| Pod system | Rechargeable, has a pod that can be inserted (e.g., JUUL) |
| Disposable pod | Resembles closed pod system, disposable (e.g., Puffbar) |
| Mod | Customizable, has battery and tank, rechargeable |
| Other | Any other e-cigarette product not listed above (e.g., starter kit, vape box subscription service) |
| Non-specific product | E-cigarette product is shown briefly but is not prominently featured (e.g., a newsclip briefly shows someone using an e-cigarette product when reporting on an e-cigarette related topic) |
| Channel Type | |
| Retailer | Online and/or brick and mortar shop that sells e-cigarette products |
| Vape enthusiast | A user who posts predominantly about e-cigarettes (>50% of videos uploaded) who may/may not be associated with an e-cigarette business/organization |
| Medical community | Medical doctors, researchers, or representatives from the health field |
| Private user | Private user who is not clearly associated with any e-cigarette or tobacco industry |
| Other Channel | Other Channel not listed above |
| Discount/Sales | |
| Discount | Discount for e-cigarette products (e.g., discount code) |
| External purchasing link | External links to purchase e-cigarette products |
| Other promotion | Other discount/sales practices (e.g., giveaways, non-e cigarette merchandise that alludes to e-cigarettes, such as clothing with e-cigarette related logos/images) |
| No discount/sales | No discount/purchasing option |

**Table 2.**

Engagement statistics by video theme, e-cigarette product type, Channel type, and presence of discount/sales (N=3830) [a]

| Construct | Overall | Views [b] | Likes [b] | Dislikes [b] | Comments [b] | Upload years [c] | |
|---|---|---|---|---|---|---|---|
| | | | | | | 2007–2018 | 2019–2020 |
| **Video Theme** [d] | **N(%)** [e] | Median (IRQ) | Median (IRQ) | Median (IRQ) | Median (IRQ) | **N(%)** | **N(%)** |
| Product review | 2086 (48.9) | 20827.5 (78315) | 253 (980) | 14 (50) | 60 (171) | 1030 (49.4) | 1056 (50.6) |
| Instruction | 737 (17.3) | 47728 (181672) | 379 (1392) | 31 (108) | 69 (219) | 513 (69.6) | 224 (30.4) |
| Health information | 484 (11.3) | 5302 (73685.5) | 54 (628) | 11 (113) | 19 (266) | 247 (51.0) | 237 (49.0) |
| Other relevant [f] | 401 (9.4) | 29771 (168391) | 264.5 (2672) | 31.5 (244) | 76 (517) | 230 (57.4) | 171 (42.6) |
| Cannabis | 268 (6.3) | 14602 (48588.5) | 126 (428) | 11 (41) | 41 (82) | 115 (42.9) | 153 (57.1) |
| Other irrelevant [g] | 239 (5.6) | 37189 (186131) | 843 (4497.5) | 23.5 (81) | 94 (405) | 115 (48.1) | 124 (51.9) |
| Vape trick | 49 (1.2) | 2290086 (10156343) | 23963.5 (61276.5) | 1179 (4957) | 1412.5 (2136) | 33 (67.4) | 16 (32.6) |
| **E-cigarette product type** [h] | | | | | | | |
| Non-specific product | 1079 (29.8) | 19736 (192318) | 236 (2333) | 27 (198) | 68 (453) | 601 (55.7) | 478 (44.3) |
| Box mod | 910 (25.1) | 31738.5 (115350) | 441 (1296) | 22 (72) | 94.5 (221) | 457 (50.2) | 453 (49.8) |
| E-liquid | 532 (14.7) | 25374 (84757) | 352 (1399) | 18 (58) | 72 (197.5) | 335 (63.0) | 197 (37.0) |
| Disposable pod | 431 (11.9) | 4500 (19708) | 45.5 (203.5) | 4 (15) | 19 (51) | 164 (38.0) | 267 (62.0) |
| Vape pen | 337 (9.3) | 46143 (165785) | 335.5 (1044) | 30 (105) | 68.5 (174.5) | 207 (61.4) | 130 (38.6) |
| Pod system | 195 (5.4) | 42418 (85496) | 345 (1144) | 26 (81) | 77 (180) | 84 (43.1) | 111 (56.9) |
| Cigalike | 140 (3.9) | 13331 (45767) | 45 (233) | 5 (18) | 12 (53) | 130 (92.9) | 10 (7.1) |
| **Channel Type** [h] | | | | | | | |
| Vape enthusiast | 1958 (54.0) | 28329 (121487) | 404 (1534) | 21 (82) | 82 (241) | 1034 (52.8) | 924 (47.2) |
| Retailer | 734 (20.3) | 22641 (83642) | 162.5 (668) | 12 (54) | 38 (115) | 454 (61.9) | 280 (38.1) |
| Other Channel | 442 (12.2) | 14462.5 (92015) | 115 (648) | 16 (104) | 45 (211) | 222 (50.2) | 220 (49.8) |
| Medical community | 248 (6.8) | 2559.5 (17791) | 25 (198) | 3 (31) | 5 (48) | 142 (57.3) | 106 (42.7) |
| Private user | 242 (6.7) | 57184 (264550) | 804 (4225) | 35 (175) | 139 (446) | 126 (52.1) | 116 (47.9) |
| **Discount/sales** [d, h] | | | | | | | |
| No discount/sales | 2150 (56.8) | 22024.5 (114849) | 262 (1275) | 19 (103) | 70 (254) | 1137 (52.9) | 1013 (47.1) |

*Tob Control*. Author manuscript; available in PMC 2023 November 01.

| Construct | Overall | Views [b] | Likes [b] | Dislikes [b] | Comments [b] | Upload years [c] | |
|---|---|---|---|---|---|---|---|
| | | | | | | 2007–2018 | 2019–2020 |
| External link for sales | 1289 (34.1) | 23832 (97977) | 211 (1043) | 14 (55) | 51 (159) | 766 (59.4) | 523 (40.6) |
| Other promotion | 283 (7.5) | 56523 (201345) | 960 (3165) | 32 (127) | 140 (354) | 118 (41.7) | 165 (58.3) |
| Discount | 62 (1.6) | 17300 (61071) | 116 (775) | 7.5 (47) | 22.5 (119.5) | 44 (71.0) | 18 (29.0) |

Note: The categories within each construct are ordered from highest percentage to lowest.

[a] Total videos examined: N=3830. We excluded 371 non-English videos.

[b] Kruskal-Wallis tests were conducted between each engagement variable (i.e., views, likes, dislikes, comments) and each construct (i.e., video theme, e-cigarette product type, Channel type, presence of discount/sales). All comparisons were statistically significant ($p < .001$).

[c] Chi-square tests were conducted between video upload years and each construct. All comparisons were statistically significant ($p < .001$). Row percentages are presented. We used median split to dichotomize the years of upload.

[d] The categories within the construct are coded more than once.

[e] Column percentages are presented.

[f] Refers to videos that are e-cigarette related but that do not fall into the predefined categories.

[g] Refers to videos that are non-e-cigarette related.

[h] Total videos examined: N=3624. We excluded “other irrelevant” videos.

**Table 3.**

Evaluation metrics of supervised machine learning models for each construct

| | F1 | Precision | Recall | Accuracy |
|---|---|---|---|---|
| **Video Theme** | 0.83 | 0.79 | 0.88 | 0.94 |
| **E-cigarette Product** | 0.79 | 0.79 | 0.79 | 0.93 |
| **Channel Type** | 0.86 | 0.86 | 0.86 | 0.94 |
| **Discount/Sales** | 0.87 | 0.85 | 0.89 | 0.90 |