Original Paper

Influence of User Profile Attributes on e-Cigarette–Related Searches on YouTube: Machine Learning Clustering and Classification

Dhiraj Murthy[1], PhD; Juhan Lee[2], PhD; Hassan Dashtian[1], PhD; Grace Kong[2], PhD

[1]Computational Media Lab, School of Journalism and Media, Moody College of Communication, The University of Texas at Austin, Austin, TX, United States

[2]Department of Psychiatry, Yale School of Medicine, New Haven, CT, United States

Corresponding Author:
Dhiraj Murthy, PhD
Computational Media Lab, School of Journalism and Media
Moody College of Communication
The University of Texas at Austin
300 W Dean Keeton (A0900)
Austin, TX, 78712
United States
Phone: 1 512 471 5775
Email: dhiraj.murthy@austin.utexas.edu

Abstract

Background: The proliferation of e-cigarette content on YouTube is concerning because of its possible effect on youth use behaviors. YouTube has a personalized search and recommendation algorithm that derives attributes from a user's profile, such as age and sex. However, little is known about whether e-cigarette content is shown differently based on user characteristics.

Objective: The aim of this study was to understand the influence of age and sex attributes of user profiles on e-cigarette–related YouTube search results.

Methods: We created 16 fictitious YouTube profiles with ages of 16 and 24 years, sex (female and male), and ethnicity/race to search for 18 e-cigarette–related search terms. We used unsupervised (k-means clustering and classification) and supervised (graph convolutional network) machine learning and network analysis to characterize the variation in the search results of each profile. We further examined whether user attributes may play a role in e-cigarette–related content exposure by using networks and degree centrality.

Results: We analyzed 4201 nonduplicate videos. Our k-means clustering suggested that the videos could be clustered into 3 categories. The graph convolutional network achieved high accuracy (0.72). Videos were classified based on content into 4 categories: product review (49.3%), health information (15.1%), instruction (26.9%), and other (8.5%). Underage users were exposed mostly to instructional videos (37.5%), with some indication that more female 16-year-old profiles were exposed to this content, while young adult age groups (24 years) were exposed mostly to product review videos (39.2%).

Conclusions: Our results indicate that demographic attributes factor into YouTube's algorithmic systems in the context of e-cigarette–related queries on YouTube. Specifically, differences in the age and sex attributes of user profiles do result in variance in both the videos presented in YouTube search results as well as in the types of these videos. We find that underage profiles were exposed to e-cigarette content despite YouTube's age-restriction policy that ostensibly prohibits certain e-cigarette content. Greater enforcement of policies to restrict youth access to e-cigarette content is needed.

KEYWORDS

XSL•FO
RenderX

Introduction

Nicotine exposure through e-cigarettes, particularly during adolescence, poses negative health outcomes such as brain maldevelopment and subsequent substance use [1]. In 2022, 9.4% (representing 2,550,000 students) of US middle and high school students reported using e-cigarettes in the past 30 days [1]. e-Cigarettes are also popular among adults (5.1% used them in the past 30 days in 2020), but these are most often used by young adults (15.6%) [2,3]. e-Cigarette use among adolescents and young adults (referred to as "youth" from here onward) may be driven, in part, by its heavy presence and positive portrayal on social media [4,5]. There is accumulating literature documenting e-cigarette promotion on social media. e-Cigarettes are portrayed on social media as fashionable, acceptable, and cool [6]. There are also themes that specifically appeal to youth, such as cartoon-based images on Instagram [7] and vape tricks (ie, blowing large vapor clouds or shapes like rings) on YouTube [8]. Studies have suggested that positive perceptions such as e-cigarette use being socially acceptable is related to its use among youth [9,10]. Studies have also shown that such positive portrayals of e-cigarettes on social media platforms have contributed to youth appeal and use behaviors [11]. For example, Lee et al [12] used state-level population data and found that the daily use of social media platforms, namely, Instagram, Snapchat, Facebook, and Twitter, was associated with e-cigarette use among adolescents, suggesting that youth may be exposed to e-cigarette–related information on social media. Given the high rate of social media usage by youth [13] and the unregulated environment [14], surveillance of e-cigarette–related content on social media platforms is warranted.

Social media platforms custom-tailor content to user characteristics [15]. However, these algorithms are proprietary, and it is unclear how information regarding e-cigarettes is featured to youth users. In this study, we examined how user profile attributes (ie, age and sex) influence the e-cigarette content being shown to youth users on YouTube—an online video streaming social media platform that has more than 2 billion users and is viewed more than 1 billion hours/day [16]. Users can upload and watch videos on YouTube and interact with other users by posting comments, reacting to videos (like/dislike), sharing content, and subscribing to YouTube channels. YouTube was the most frequently used social media platform in 2021, with 81% of the social media users reporting having used the platform [17]. Moreover, YouTube is the most popular platform among youth [8].

e-Cigarette content is prolific on YouTube. For instance, Huang et al [18] analyzed 28,000 e-cigarette–related YouTube videos and found that those videos had received more than 100 million views, indicating high engagement by users [18]. Further, e-cigarettes are frequently positively portrayed on YouTube and pro–e-cigarette videos are commonly sponsored by the e-cigarette industry [19]; 85% of the e-cigarette–related videos on YouTube are sponsored by e-cigarette marketers, including e-cigarette companies or people endorsing e-cigarette companies [20]. Pro–e-cigarette videos include portrayals of e-cigarettes as safer, cleaner, and less malodorous than combustible cigarettes [21]; videos showcasing or teaching how to conduct

vape tricks (ie, using e-cigarettes to blow large, thick amounts of exhaled aerosol or shapes) [8]; modification of e-cigarette devices for unintended purposes such as increasing the temperature and using other substances in it [19,22,23]; instructions on how to use e-cigarettes (eg, how to puff) [24]; product reviews [25]; and health information or misinformation about e-cigarette use [26]. Concerningly, these e-cigarette contents are readily available on YouTube without a warning label/statement [27], and these videos are easily accessible to youth [4]. In sum, there are abundant e-cigarette–related videos on YouTube. However, less known is specifically what content youth are exposed to. All users do not receive the same results when they search for the same terms on YouTube. This is partially due to YouTube's personalized search and recommendation algorithms, which consider, to some extent, a user's age, sex, and the history of the searched items by that specific user [28,29].

YouTube's search and recommendation algorithms are responsible for creating personalized content for users from an ever-growing collection of videos. Similar to other social networks, YouTube has undergone a paradigm shift toward using deep machine learning—systems based on artificial neural networks—as a solution for scaling the systems used by YouTube's search and recommendation algorithms [30]. However, the opaque nature of the search and recommendation algorithms poses questions concerning whether algorithmic visibility can be evaluated. Search and recommendation algorithms may be developed to take viewers' demographic profiles (eg, age, sex) as inputs in determining what search results users receive. Therefore, YouTube's search and recommendation algorithms have important public health implications. For instance, males have consistently shown a higher level of e-cigarette use among adolescents and adults [31], and data suggest that e-cigarette–related videos such as vape tricks videos mostly feature young men and seem to be targeting this population [8]. A recent study identified that e-cigarette content on YouTube contained themes related to product reviews (provide reviews of e-cigarette products), instructional videos (teach viewers how to use, modify, or create e-cigarette products), health information (provide health information related to e-cigarettes), vape tricks (feature different vape tricks), cannabis (cannabis vaping–related topics), and other (a variety of other themes such as news clips related to e-cigarette use) [19]. However, less known is whether these video themes are differentially exposed to users by their demographic attributes. Such information is important to inform tobacco regulatory actions in restricting marketing that targets at-risk populations such as underage youth, and it can be used to inform how prevention strategies such as countermarketing can be targeted to these populations.

Methods

Overview

The goal of this study was to understand the role of the demographic factors (ie, sex, age) of YouTube users' profiles in influencing the variations in e-cigarette–related search results presented to users. To accomplish this goal, we developed a

3-step framework, which combined computational methods and human labeling (Figure 1). First, we used an unsupervised machine learning method, the k-means method, which used the distribution of words in the video data (ie, titles and descriptions) to cluster the videos into themes. Human-labeled data sets using titles and descriptions of the YouTube videos were then used to confirm the themes identified in our k-means clustering results. Second, we used this human-labeled data set to train a supervised machine learning method, that is, the graph convolutional network (GCN), to classify all the videos in our data set based on their identified themes. Finally, we performed unsupervised network analysis to measure how YouTube video results varied by user attributes (ie, age and sex). We examined whether there were differences in the video themes between different age and sex profiles. The application of these machine learning–based methods is novel in tobacco regulatory science work using social media data. Our approach is also scalable to large volumes of data and can be extended to a variety of social media platforms.

Figure 1. Overall framework of data collection, preprocessing, and analytics.
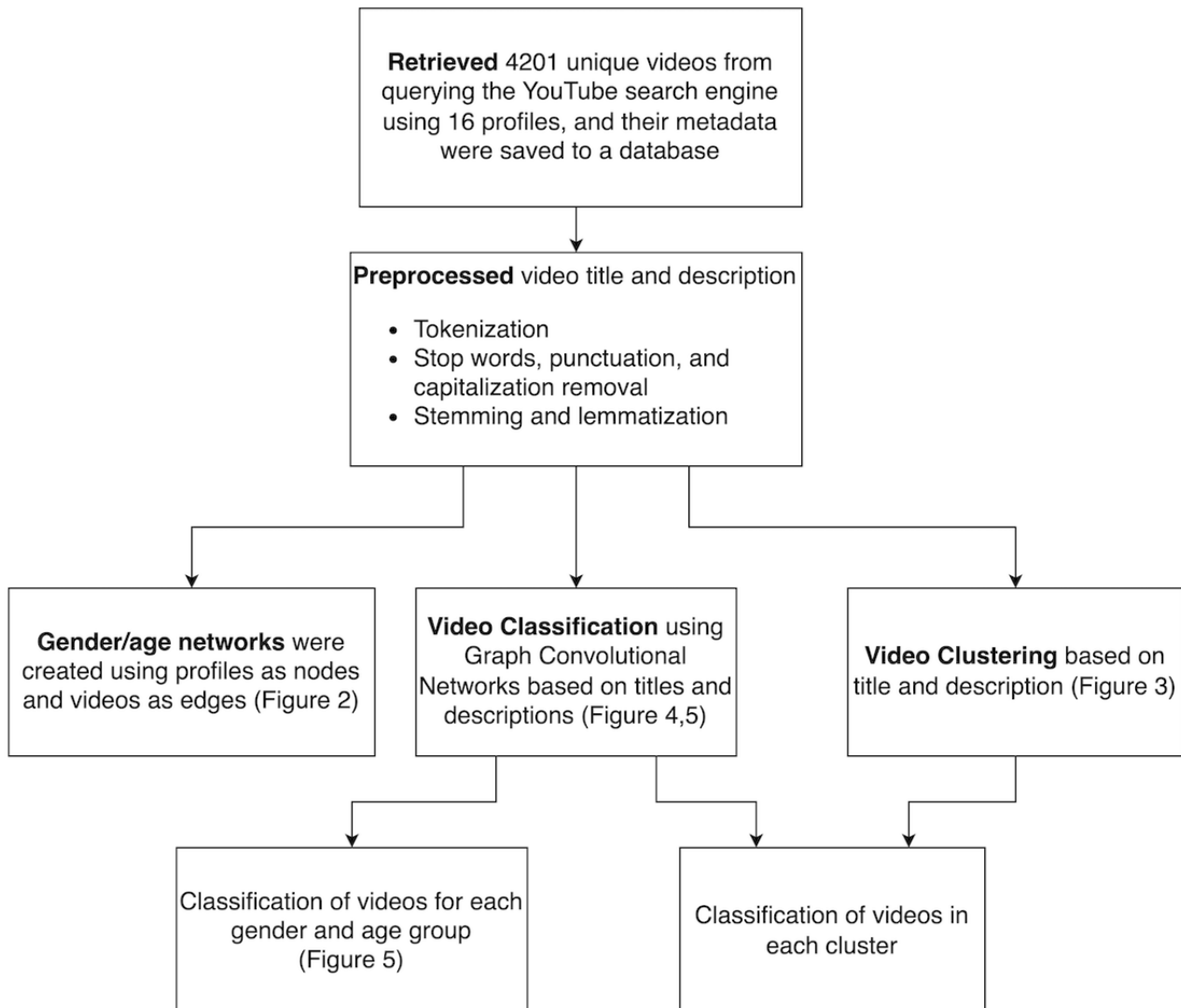
Ethical Considerations

This research is not deemed as human subjects research according to the definition provided by the Office of Human Research Protections, US Department of Health and Human Services. We examined publicly available data, and we did not report any identifying information of the content observed on social media. Additionally, this observational study was deemed exempt as human subjects research by the Yale Institutional Review Board (HIC 2000028350).

Search Methods

We created 16 fictitious profiles on YouTube that sought to vary and reflect particular demographic attributes (ie, age, sex, and race) [32]. Profile photos were not added. To attempt to reflect particular racial and ethnic attributes, we created profiles by using common African American, Hispanic, and White first and last names by using existing name data [33]. The profiles consisted of African American females and males aged 16 and 24 years (4 profiles), Hispanic females and males aged 16 and 24 years (4 profiles), and 2 sets of White females and males aged 16 and 24 years (8 profiles). We oversampled White users to be more reflective of the e-cigarette use population. To create

each fictitious profile on YouTube, we used a new SIM card and phone number and performed a factory reset of an Android phone. Sex and age were entered during each fictitious account creation. No other demographic metadata were included during account creation. We used a mobile phone rather than a web browser to conduct our searches to best replicate how youth access YouTube content [13].

During the course of a week in June 2020, we collected data for 2-3 profiles per day. Once we collected 140 videos per profile per search term, we factory reset the Android phone and moved to the next profile. For each profile, the following keywords were searched for each profile by using Orbot, a mobile app that allows one to use an anonymized Tor bridge (to avoid location or IP address personalization): box mods, cigalikes, disposable e-cigs, disposables, disposable vape, e-cig, e-cigarette, e-juice, electronic cigarette, e-liquid, ENDS, pod mods, vape, vaping, vape juice, vape mods, vape pens, vape pods. Studies typically examine the first page [18,20,34] of the search results on YouTube, which has 20 videos, or the first 2 pages, which has 40 videos. However, some users may search through multiple pages if they do not find what they are looking for in the first few pages. Thus, for each of our 16 fictitious YouTube profiles, we searched through 7 pages (140 videos) for each of our keywords (n=5875). This approach is therefore far more aggressive than previous work [35]. After removing duplicates (n=1674), we arrived at the final sample (N=4201) of unique videos. We collected video metadata such as title, description, transcript, view counts, likes/dislikes, comments, date published, channel name, and category. The methods are further explained in Dashtian et al [32].

Preprocessing Data

We converted the text into numerical form so that we can apply machine learning algorithms such as clustering and classification to the data. The preprocessing steps included tokenization, stop words removal, stemming, and lemmatization. Tokenization is the process of splitting a set of texts into words (also called tokens) and then removing certain characters such as blank sequences and punctuation. Stop words are usually frequent in English text (eg, a, an, the, that, I, be, other, with). The goal of both stemming and lemmatization is to find the base form of a word from its inflectional forms and derivatives (eg, vaped, vaping have a base of vape). We used Porter stemmer, an algorithm which has been successfully used by others for the stemming of health-related texts for machine learning purposes [36].

Video Clustering (Unsupervised Machine Learning)

K-means automatically arranges texts into clusters such that text data within clusters are relatively similar in terms of content when compared to text data in other clusters [37]. Another health-related work [38] has successfully used the k-means clustering algorithm for automated text classification. We therefore chose to use k-means to categorize video types. In our case, the input to the k-means clustering is preprocessed text (video title and the description provided by the uploader to describe the video). We used the elbow method to find the optimum value for the number of clusters (k). The elbow method provides a good indication that the underlying model and

number of (k) fits best at that point and has been successfully used in other health-related machine learning studies [39]. We examined the results visually to discern the point at which diminishing returns are observed (ie, an elbow appears). K-means seeks to cluster around optimal centroids (ie, cluster centers). The best placement of initial centroid positions is a standard method for maximizing the k-means clustering process. To avoid any bias, we randomly selected initial centroids and iterated the algorithm several times for each k to confirm that the initial centroids do not change our optimized clustering results. We measured cosine similarity to generate a measure of similarity between each video and the other videos in the search results. Cosine similarity is a measure mostly used for k-means clustering of text documents. The distance matrix was then converted into a 2D array by using multidimensional scaling.

Video Classification (Human Labeling)

Members of the research team with expertise in e-cigarettes randomly selected videos from the full corpus of the collected videos (n=1000) [19] and labeled the videos by the following classes: (1) product review (ie, an individual(s) providing a review of an e-cigarette product), (2) health information (ie, health information related to e-cigarette use), (3) instructional (ie, a tutorial on how to use an e-cigarette or how to modify an e-cigarette), and (4) other (which consists of a variety of other themes, including cannabis, television/news clips, vape tricks). Interrater reliability (Cohen κ) was 0.93, indicating "almost perfect" agreement between coders. These categories were used in previous research [32]. Please refer to Kong et al [19] for more information on how these themes were determined and labeled.

Text Classification Using GCNs (Supervised Machine Learning)

We used GCN, which is a supervised machine learning method, to classify data (ie, titles and descriptions) by theme to better understand the unique clusters identified through k-means clustering. In GCN, word frequency and word co-occurrence information are used to build the word-to-word and word-to-video edges (ie, as common videos between pairs), respectively. We also classified the nodes (ie, entities in the network) instead of the actual videos. The entities in the network represented just the nodes in the graph. These do not refer to the themes. GCN has shown strong performance for classification with a small portion of labeled data similar to the data used in our study [40].

To model the global word co-occurrence, we built a large 2-mode graph (ie, 2 types of nodes). Our graph contains word nodes (which represent single words) and document nodes (which represent whole documents with many words). See Multimedia Appendix 1 for a visual rendering of the relationship between the document nodes and word nodes. Specifically, the first mode of nodes consists of words and the second mode of nodes consists of documents with titles and descriptions (ie, with many words). One document represents 1 video (title and description together). Document nodes and word nodes are interconnected and intraconnected. The number of nodes in the text graph |V| is the number of documents (document nodes)

plus the number of unique words in the documents (word nodes). We set feature matrix X = I as an identity matrix, which means every word or document is represented as a 1-hot vector as input to text GCN. One-hot encoding converts categorical data into binary values suitable for machine learning algorithms. We build edges (ie, connections) between nodes based on word occurrence in documents (document-word edges) and word co-occurrence in the whole corpus (word-word edges). The weight of the edge between a document node and a word node is the term frequency-inverse document frequency of the word in the document. Term frequency is the number of times the word appears in the document, and inverse document frequency is the logarithmically scaled inverse fraction of the number of documents that contain the word. After performing clustering and classification on preprocessed data, we calculated the percentage of each video type (derived from classification) in each category (derived from clustering).

Profiles Network

The frequency of common videos between different ages and sexes can be used as a measure to quantify the strength of the relationships between these variables. For example, the overlap of videos among the same sex and age profiles can be used to discern whether users with these attributes (eg, both female and male, adolescents or young adults) receive similar information from YouTube's search engine. Furthermore, the connections between nodes in a network provide information about the structure of the network. We can also use the number of connections of a node in each demographic group to identify the most influential nodes in the network. Specifically, the network of 4 demographic groups can be represented as nodes with their edges representing common videos between pairs of groups. To show the connections, we plotted a line between two groups and calculated the number of common videos between them. Lines with a larger value represent more common videos between a pair than lines with smaller values. We assessed 2 separate networks: one with common videos between age and sex and another that assessed a combination of the two.

Results

Video Clustering (Unsupervised Machine Learning)

To better understand which content shows up for different demographic profiles, we identified the types of videos in our data set by using k-means to cluster videos. Figures 2A and B illustrate the video clusters as 3 clusters and 4 clusters, respectively. The former had 3 distinct topical clusters, whereas the latter had 3 distinct topical clusters and 1 diffuse cluster (that likely represents the "other" content cluster). The elbow method indicated that the plateau (ie, the first stable k value in the sum of squared distances) is at k=3 (Figure 2C). In some cases, the elbow method has ambiguity [41]. However, in our case, we had a clear result that videos can be automatically clustered into 3 main clusters.

Figure 2. Each dot in (A)/(B) indicates a video and each color represents a cluster. (A) Videos categorized into 3 clusters. (B) Videos categorized into 4 clusters. (C) Elbow method results, which show the sum of the squared distances as a function of the number of clusters (k).
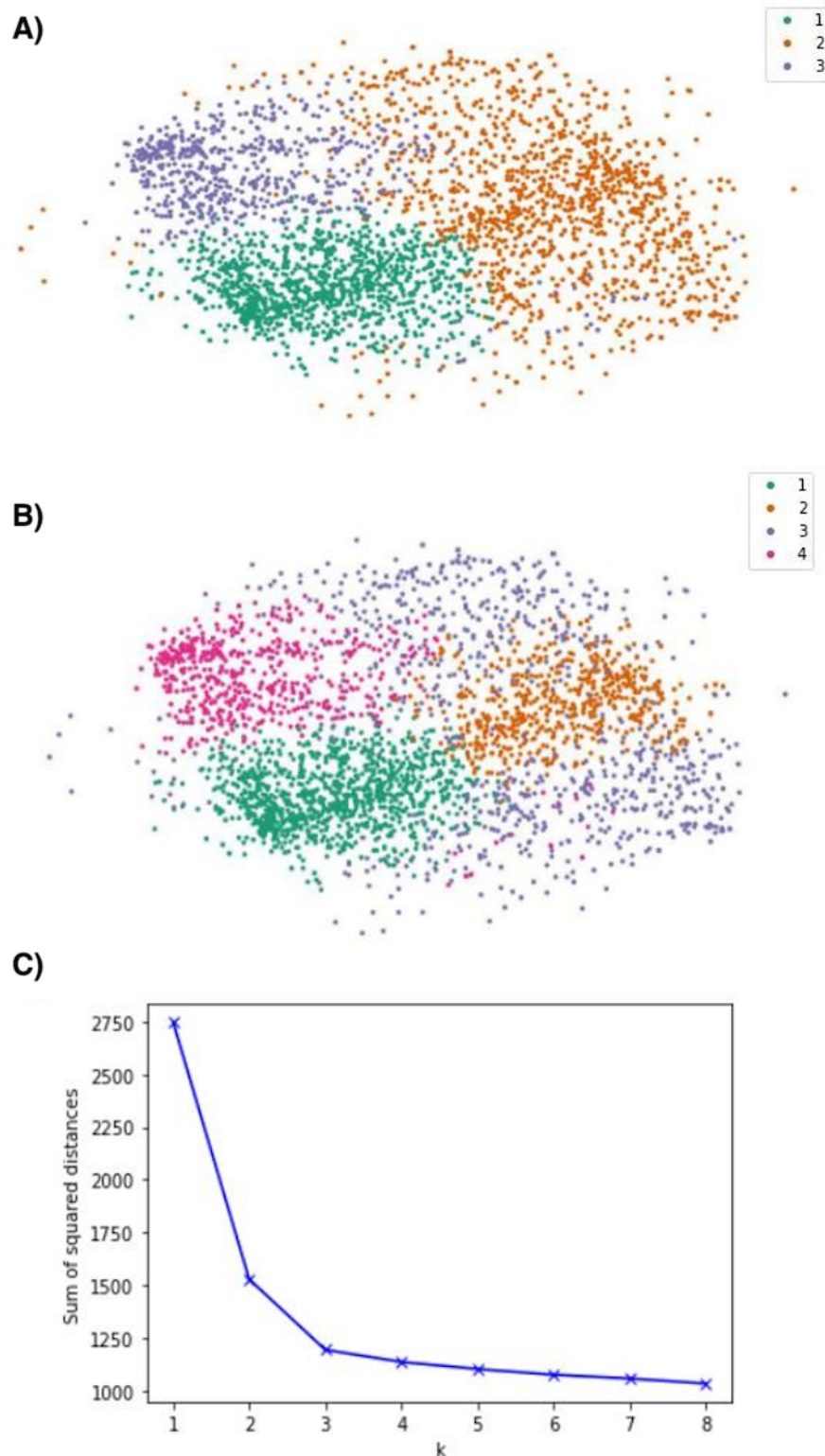
Video Classification (Human Labeling)

Human labeling identified 3 distinct classes: (1) product reviews, (2) instructional, and (3) health information. We also included a fourth catchall class of "others" for any videos that did not fit into the other 3 distinct classes. Product reviews are videos that provide reviews of e-cigarette products, instructional videos provide instructions on how to use/modify/create e-cigarette products, health information videos provide information on the health risk of e-cigarettes, and other videos are topics that do not fall into these 3 classes and include a range of topics such as cannabis and vape tricks. We found that GCN was able to successfully classify videos based on these 3 distinct classes as well as a separate "other" class. Overall, product review was the most common type of videos identified (49.3%), followed by instructional (26.9%), health information (15.1%), and other

(eg, cannabis, television/news clip) (8.5%). We further estimated the prevalence of each video type exposure by demographic attributes (Figure 3). For all demographic groups, except the 16-year-old group, product review videos showed the highest percentage in the search results, followed by instructional videos. Instructional videos showed the highest percentage in the search results of 16-year-old students. We estimated the prevalence of video themes separated by age and sex (Figure

4). The product review label was the dominant class for 24-year-old male (39.4%) and 24-year-old female (38%) profiles. Instructional videos showed the highest percentage in the search results of 16-year-old female (42.5%) and 16-year-old male (30.9%) profiles; notably, the 16-year-old female profile had the highest percentage of search results for this label. All profiles were least exposed to health information videos.

Figure 3. Prevalence of video type shown, split by demographic variables. The percentage of each label (class) is shown based on the results from graph convolutional networks. TV: television.
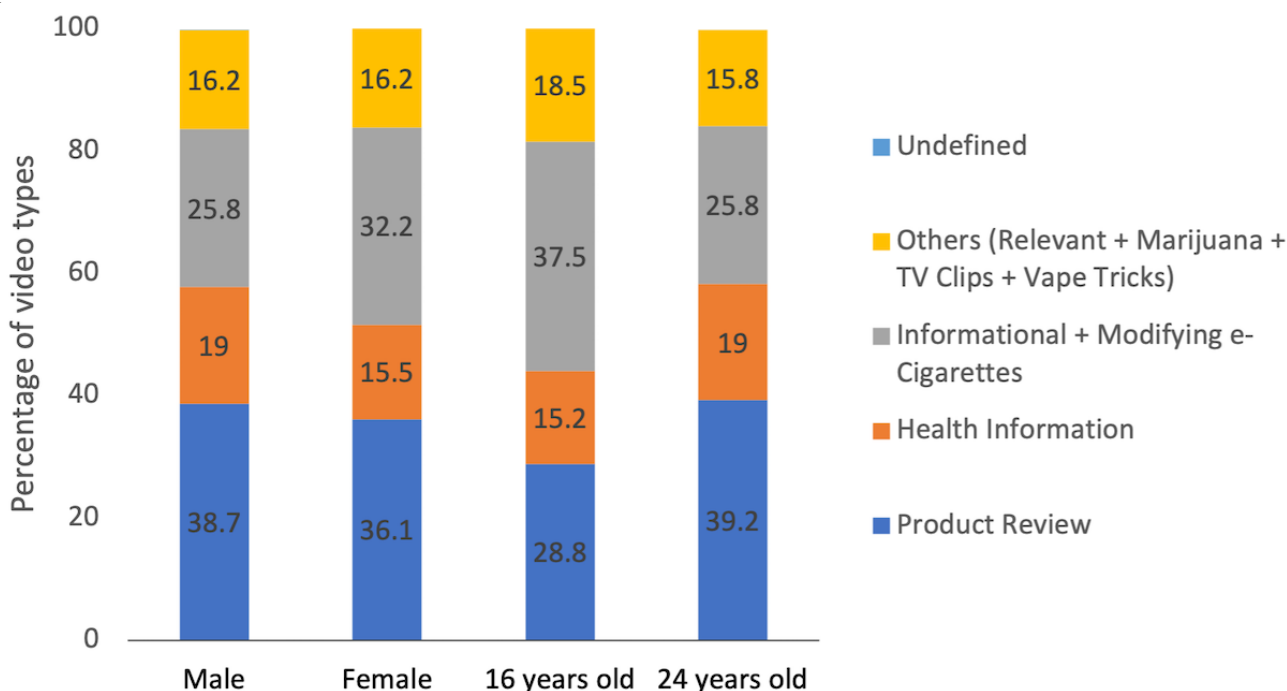
Figure 4. Results of the classification of videos in each demographic group. We grouped YouTube profiles based on age (24 or 16 years old) and sex (male and female). TV: television.
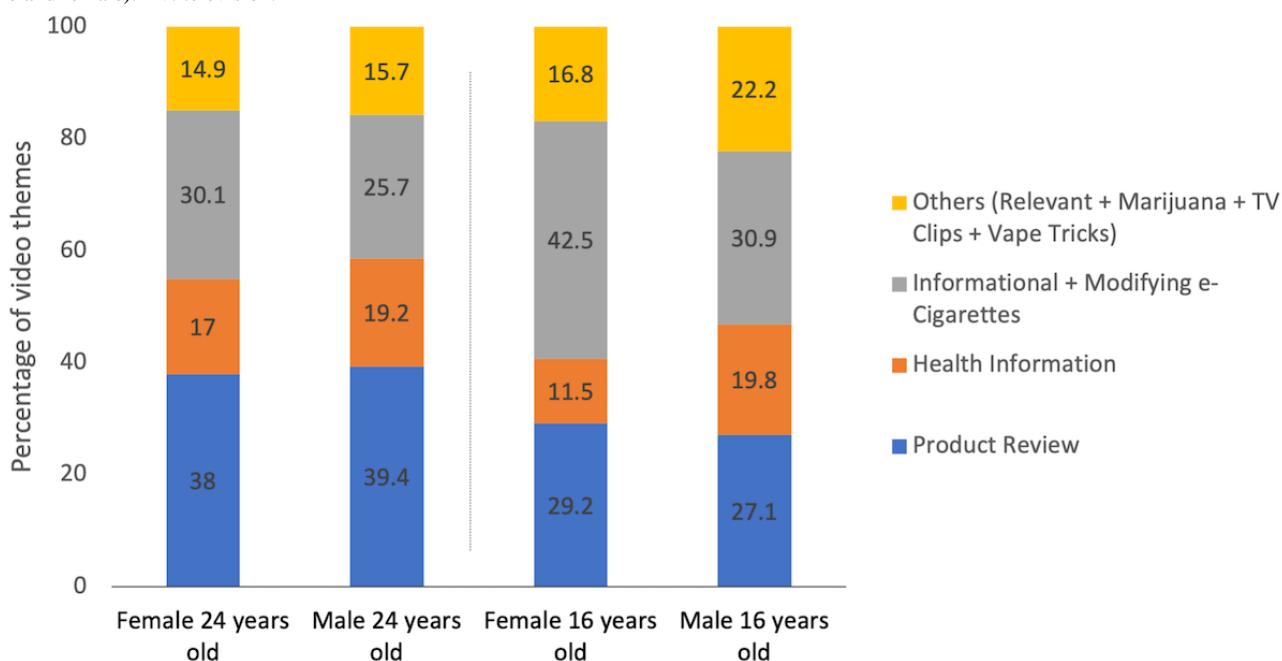
Text Classification Using GCNs (Supervised Machine Learning)

We used text classification using GCN, a supervised machine learning technique, to classify the text of video titles and description into human-labeled classes (ie, product review, health information, instructional, other). We found that the accuracy of the GCN model for the classification of e-cigarette–related YouTube videos is 0.72 for the parameters that we set. The precision, recall, and F1-score values were 0.70, 0.78, and 0.74, respectively.
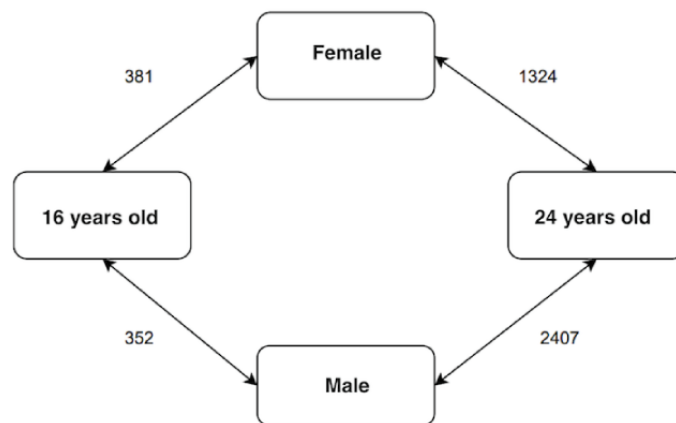
Profiles Network

The connections between the profile groups based on the common videos that were retrieved from the YouTube search are shown in Figure 5. The number of common edges between 16-year-old and 24-year-old pairs was the lowest among the other pairs. As shown in Figure 5A, the connection between the nodes of 24 years old and male is very strong, as indicated by

the edge weight of n=2407 (ie, the number of common videos). We also constructed another network by using a combination of age and sex. The videos of all the profiles were grouped into 4 subsets: 24-year-old male, 16-year-old male, 24-year-old female, and 16-year-old female. Similar to that in the previous network, each node in the network represents one of these groups, and common videos between pairs of groups are shown as an edge. Compared to the previous network (Figure 5A), the network of combined age and sex (Figure 5B) had fewer edges (connections). When we examined the network of age and sex together, we imposed further restrictions on the videos that belonged to a specific node. Thus, the number of videos and therefore, the number of connections between nodes in the network of age and sex was smaller than that of age or sex alone. Figure 5B shows that 24-year-old male and 24-year-old female profiles have the highest number of common edges, while 16-year-old male and 16-year-old female profiles have the lowest number of common edges.

Figure 5. Network of demographic attributes and videos. Edge weights are provided next to the edge line between 2 pairs, and these edge weight values indicate the number of common videos between 2 corresponding nodes (ie, between the demographic attributes of sex and age).
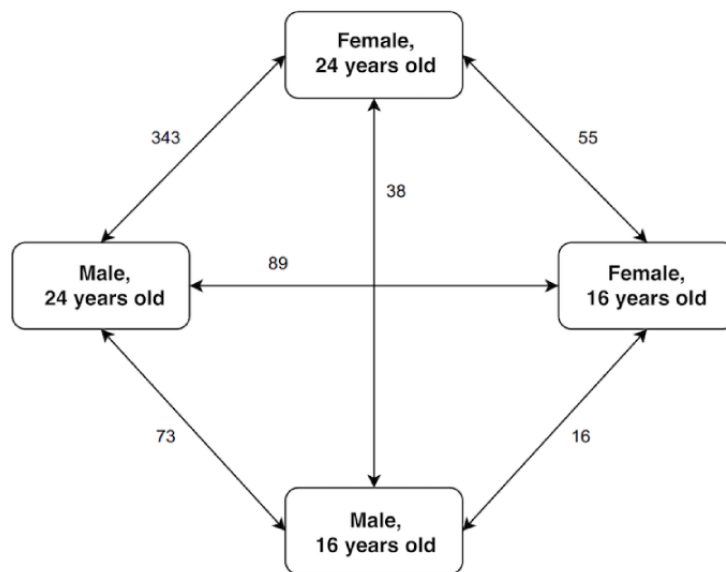
Discussion

Principal Findings

In this study, we examined how YouTube profile attributes, specifically age and sex, affected e-cigarette–related YouTube search results. Our profile network analysis indicated that there were more common videos between male and female 24-year-old profiles relative to other demographic groupings. Using our own human-labeled data, we developed a GCN machine learning model that was able to classify the videos into 4 main classes. We found that the highest proportion of younger age groups (16 years old) was exposed to instructional videos (37.5%), while the highest proportion of young adult age groups (24 years old) were exposed to product review videos (39.2%). Additionally, the group with the highest proportion of exposure to instructional videos was 16-year-old females relative to other age/sex pairs. Our findings are consistent with prior studies that observed that common video themes related to e-cigarettes on YouTube were product reviews and instructional videos on how to use/modify/create videos [8,19,22,25]. However, our results uniquely contribute to the literature by demonstrating that demographic attributes factor into YouTube's algorithmic systems in how video themes are differentially shown to profiles with different age/gender attributes.

It is unclear what drives the differences in exposure to e-cigarette content and the volume of this content among different demographic profiles. Previous studies have shown that age and sex affect the results delivered in search engines (eg, Google) [15]. Our findings are consistent with other research that indicate that YouTube also may use demographic information to provide the most relevant information to users [29]. Specifically, Hussein et al [29] found that once a user develops a watch history in the YouTube search engine, the demographic attributes do affect the extent of content recommended to them. However, in this study, we used the same search words between each profile and used a mobile phone that was factory reset after each profile's searches were conducted to prevent tailoring of search results. It is therefore unlikely that these factors account for differences in exposure to e-cigarette content. It appears that YouTube's search engines and recommendation algorithms are driven by the demographic factors of its users. Personalization of search engines, where individual users receive distinct results for the same search query, has also led to public concerns about the so-called "filter bubble" effects [42], where users are unable to access diverse information that a search engine's algorithm decides is irrelevant to a user [43]. Our results indicate that there might be differences in the type of exposure specific to e-cigarettes that are provided to different demographic groups. We further break down these differences in terms of age and sex attributes.

Our network of search results, which shows the influence of age and sex on search results, indicates a noteworthy difference between the number of edges (common videos) for various pairs of nodes (common videos between 2 groups) in the network, including male/female and 16-year-old and 24-year-old profiles. For example, the videos common to both 16-year-old and 24-year-old groups are the lowest. However, the second network

analysis showed that 24-year-old male group and 24-year-old female group pairs have the highest number of common videos. There is a greater number of edges between the male group and 24-year-old group than between the female and the 24-year-old group, indicating that males and 24-year-old groups have more common videos than females and 24-year-old groups. These results indicate that 24-year-old profiles are most exposed to e-cigarette content, and this exposure is greater among 24-year-old male groups compared to their female counterparts.

Our finding that e-cigarette content is mostly available to male young adult groups is consistent with research findings that show that e-cigarette–related videos on YouTube feature more males. For instance, an examination of vape tricks on YouTube showed that 80% of the vape tricks videos featured young adult males [8]. There is also research showing that males are more engaged with YouTube content than females. Khan [44] found that male users are more likely to read comments on YouTube; Molyneaux et al [45] found that there was a greater number of comments posted by male users. Perhaps, the high engagement of males on social media platforms such as YouTube can explain the higher e-cigarette use rates among males. A review on e-cigarette use behaviors among adolescents showed that e-cigarettes are used more by male adolescents than by female adolescents [46], and national data also show that e-cigarette use is higher among male adolescents and young adults [47]. However, it is important to also highlight that e-cigarette use among females is also high: up to 20% of females in middle and high school surveyed in a study in 2020 were found to use e-cigarettes [47]. It is possible that females are engaging with e-cigarette–related social media content but doing so differently from males. For instance, there was no difference between males and females in viewing YouTube videos [44] or in the rating of YouTube videos [45], suggesting that females are engaging with YouTube content similarly as males.

The lower number of e-cigarette–related videos shown to 16-year-old profiles than 24-year-old profiles may be due, in part, to the age-restriction process of e-cigarette–related content by YouTube. YouTube's current policy prohibits tobacco-related advertisements. YouTube considers content that "promotes a product that contains drugs, nicotine …" as age-restricted content [16]. They exemplified "a video reviewing brands of nicotine e-liquid" as an example of age-restricted content. This rule may explain why we observed more product review videos in the 24-year-old group (39.2%) compared to those in the 16-year-old group (28.8%). This finding also suggests that despite these self-imposed limits on e-cigarette promotional content on YouTube, there is evidence that these restrictions may be loosely implemented and content that are restricted may be shown to underage minors on this and other social media platforms [14,19,48]. It is noteworthy that in our study, the 16-year-old profiles were exposed to e-cigarette content despite YouTube's age-restriction policy that prohibits certain e-cigarette content such as product reviews. This finding is consistent with that in other studies that found that e-cigarette content such as vape tricks were readily available using non–age-verified accounts [8]. This study highlights the importance of strong policies and the enforcement of these policies to prohibit the exposure of e-cigarette–related videos

to youth on YouTube. This finding also suggests that young adults are the highest consumers of e-cigarettes among adults [49]; they may search for more information about e-cigarette products to purchase through product reviews.

Concerningly, the instruction label was observed in the highest percentage (37.5%) of search results of the 16-year-old group, and exposure to instructional videos among 16-year-old female profiles was particularly high (42.5%), suggesting that underage youth are more exposed to instructional videos, which may provide tutorials on e-cigarette use. Further, instructional videos include other content such as how to hack or modify the device to use for unintended purpose as well as to use cannabis [22,23]. The high prevalence of modification of e-cigarette content on YouTube has been shown in other studies. For instance, Massey et al [23] analyzed 168 e-cigarette–related YouTube videos and found that 20.2% of the videos were modifications of e-liquids to using cannabis. Future studies should identify whether youth modify/hack e-cigarettes and the health implications of engaging in such behaviors.

Future Work and Limitations

Several limitations in this study are noteworthy. First, we might have missed potential search terms related to e-cigarettes. For example, these may include brand-specific terms (eg, Juuling) and e-cigarette use–related slang (eg, stick). Thus, our collected videos may not represent an exhaustive list of e-cigarette–related videos. However, our study uses 18 e-cigarette–related search terms that were successfully tested and used to collect a broad range of e-cigarette–related YouTube videos [32]. Second, due to a limited number of fictitious profiles, our findings do have limits in terms of generalizability. Third, we included race/ethnicity as an element when creating profiles (ie, White, African American, and Hispanic) to be inclusive of diverse racial backgrounds. The first and last names of each profile were randomly selected by choosing names from existing data sets that were shown to be most commonly associated with a specific race/ethnicity [33,50]. However, as we created a limited number of fictitious profiles, we did not have enough data points for each race/ethnicity to incorporate machine learning to determine whether search results differed by race/ethnicity. Fourth, we only used 2 age groups (ie, 16 and 24 years), and it is possible that the search results may be different if younger or older age groups were used. Future research should therefore place an emphasis on assessing whether race/ethnicity as well as other factors (eg, viewing history, age) has an effect on search results related to e-cigarettes on YouTube. Fifth, anonymous Tor-based IP addresses may have influenced our search results; therefore, results may differ if searches were to be conducted using nonanonymized IP addresses. There may be other factors that may drive results, such as the date/time of searches as well as what content is popular on YouTube at a given time. Sixth, we cannot confirm how, whether, or to what extent YouTube's

personalized search parameters read the demographic attributes (ie, age, sex, and race) that we populated our fictitious profiles with because the algorithm is proprietary. However, we used a factory-reset Android device without any search history or cookies to avoid any implicit bias in the results. The searches were conducted using the same terms to ensure that the differences between profiles, from our vantage point, are only the demographic characteristics. Nevertheless, as we only used search results collected from a mobile device, future work can explore whether web-based results are different. Seventh, we applied our methods, that is, natural language processing, video classification, and network modeling to only a single platform, that is, YouTube. Future studies would therefore benefit from extending our methodological framework to other social media platforms. Eighth, given that after we collected 140 videos per profile per search term, we factory reset our phone and moved to the next profile; our approach does not emulate or reflect the high levels of personalization that a user who uses YouTube everyday might experience. Future studies would therefore benefit by comparing our results from collecting data from YouTube in 1 setting with fictitious profile data collection done over a longer period and with some levels of variation. Ninth, we did not undertake statistical tests comparing the proportions of content classification by profile demographics nor were we able to determine how each theme was manifested by demographic attributes (eg, was health information present more for male profiles than female profiles?). Future work could make these comparisons based on the classes identified by the GCN analysis and determine how and why content themes vary by different profile attributes. Lastly, as we did not have a control group in our data collection methods, future work would benefit from the use of a control group and the examination of some of these variables.

Conclusion

Our findings underscore the value of machine learning methods in studying how profile attributes on YouTube may influence e-cigarette–related content and move the field forward by highlighting the critical need to take into consideration how social media algorithms work in practice. We used unsupervised (k-means clustering) and supervised (GCN classification) machine learning models in combination with network models to study the variation of e-cigarette–related videos on YouTube. Our methods were designed to specifically identify the similarities and differences in the videos by using selected demographic attributes, that is, age and sex. Collectively, our results suggest that advanced computational methods can be used to help understand how YouTube's current search and recommendation algorithm customizes e-cigarette–related content based on demographic attributes such as sex and age. This suggests an urgent need for surveillance and prohibition of e-cigarette–related content on social media such as YouTube to prevent e-cigarette use among youth.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Document node and word node composition and relationship.
[DOCX File , 483 KB-Multimedia Appendix 1]

References

1. Yuan M, Cross SJ, Loughlin SE, Leslie FM. Nicotine and the adolescent brain. J Physiol 2015 Jun 23;593(16):3397-3412. [doi: 10.1113/jp270492]

2. Gentzke AS, Wang TW, Cornelius M, Park-Lee E, Ren C, Sawdey MD, et al. Tobacco Product Use and Associated Factors Among Middle and High School Students - National Youth Tobacco Survey, United States, 2021. MMWR Surveill Summ 2022 Mar 11;71(5):1-29 [FREE Full text] [doi: 10.15585/mmwr.ss7105a1] [Medline: 35271557]

3. Boakye E, Osuji N, Erhabor J, Obisesan O, Osei AD, Mirbolouk M, et al. Assessment of Patterns in e-Cigarette Use Among Adults in the US, 2017-2020. JAMA Netw Open 2022 Jul 01;5(7):e2223266 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.23266] [Medline: 35867060]

4. O'Brien EK, Hoffman L, Navarro MA, Ganz O. Social media use by leading US e-cigarette, cigarette, smokeless tobacco, cigar and hookah brands. Tob Control 2020 Dec;29(e1):e87-e97. [doi: 10.1136/tobaccocontrol-2019-055406] [Medline: 32217772]

5. Amin S, Dunn AG, Laranjo L. Social Influence in the Uptake and Use of Electronic Cigarettes: A Systematic Review. Am J Prev Med 2020 Jan;58(1):129-141. [doi: 10.1016/j.amepre.2019.08.023] [Medline: 31761515]

6. Sapru S, Vardhan M, Li Q, Guo Y, Li X, Saxena D. E-cigarettes use in the United States: reasons for use, perceptions, and effects on health. BMC Public Health 2020 Oct 09;20(1):1518 [FREE Full text] [doi: 10.1186/s12889-020-09572-x] [Medline: 33032554]

7. Dormanesh A, Kirkpatrick MG, Allem J. Content Analysis of Instagram Posts From 2019 With Cartoon-Based Marketing of e-Cigarette-Associated Products. JAMA Pediatr 2020 Nov 01;174(11):1110-1112 [FREE Full text] [doi: 10.1001/jamapediatrics.2020.1987] [Medline: 32687566]

8. Kong G, LaVallee H, Rams A, Ramamurthi D, Krishnan-Sarin S. Promotion of Vape Tricks on YouTube: Content Analysis. J Med Internet Res 2019 Jun 18;21(6):e12709 [FREE Full text] [doi: 10.2196/12709] [Medline: 31215510]

9. Romijnders KAGJ, van Osch L, de Vries H, Talhout R. Perceptions and Reasons Regarding E-Cigarette Use among Users and Non-Users: A Narrative Literature Review. Int J Environ Res Public Health 2018 Jun 06;15(6):1190 [FREE Full text] [doi: 10.3390/ijerph15061190] [Medline: 29882828]

10. Kong G, Morean ME, Cavallo DA, Camenga DR, Krishnan-Sarin S. Reasons for Electronic Cigarette Experimentation and Discontinuation Among Adolescents and Young Adults. Nicotine Tob Res 2015 Jul;17(7):847-854 [FREE Full text] [doi: 10.1093/ntr/ntu257] [Medline: 25481917]

11. Donaldson SI, Dormanesh A, Perez C, Majmundar A, Allem J. Association Between Exposure to Tobacco Content on Social Media and Tobacco Use: A Systematic Review and Meta-analysis. JAMA Pediatr 2022 Sep 01;176(9):878-885 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.2223] [Medline: 35816331]

12. Lee J, Tan AS, Porter L, Young-Wolff KC, Carter-Harris L, Salloum RG. Association Between Social Media Use and Vaping Among Florida Adolescents, 2019. Prev Chronic Dis 2021 May 13;18:E49 [FREE Full text] [doi: 10.5888/pcd18.200550] [Medline: 33988495]

13. Vogels EA, Risa GW, Navid M. Teens, social media and technology 2022. Pew Research Center. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2022/08/PI_2022.08.10_Teens-and-Tech_FINAL.pdf [accessed 2022-09-05]

14. Jackler RK, Li VY, Cardiff RAL, Ramamurthi D. Promotion of tobacco products on Facebook: policy versus practice. Tob Control 2019 Jan;28(1):67-73. [doi: 10.1136/tobaccocontrol-2017-054175] [Medline: 29622602]

15. Lorigo L, Pan B, Hembrooke H, Joachims T, Granka L, Gay G. The influence of task and gender on search and evaluation behavior using Google. Information Processing & Management 2006 Jul;42(4):1123-1131. [doi: 10.1016/j.ipm.2005.10.001]

16. YouTube. Official Blog and Support Pages. URL: https://www.youtube.com/ [accessed 2022-10-05]

17. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/04/PI_2021.04.07_Social-Media-Use_FINAL.pdf [accessed 2023-09-04]

18. Huang J, Kornfield R, Emery SL. 100 Million Views of Electronic Cigarette YouTube Videos and Counting: Quantification, Content Evaluation, and Engagement Levels of Videos. J Med Internet Res 2016 Mar 18;18(3):e67 [FREE Full text] [doi: 10.2196/jmir.4265] [Medline: 26993213]

19. Kong G, Schott AS, Lee J, Dashtian H, Murthy D. Understanding e-cigarette content and promotion on YouTube through machine learning. Tob Control. Online ahead of print . [doi: 10.1136/tobaccocontrol-2021-057243] [Medline: 35504690]

20. Paek H, Kim S, Hove T, Huh JY. Reduced harm or another gateway to smoking? source, message, and information characteristics of E-cigarette videos on YouTube. J Health Commun 2014;19(5):545-560. [doi: 10.1080/10810730.2013.821560] [Medline: 24117370]

21. Sears C, Walker K, Hart J, Lee A, Siu A, Smith C. Clean, cheap, convenient: promotion of Electronic cigarettes on YouTube. Tob Prev Cessat 2017 Apr;3:10 [FREE Full text] [doi: 10.18332/tpc/69393] [Medline: 28725876]

22. Guy M, Helt J, Palafox S, Green K, Soule E, Maloney S, et al. Orthodox and Unorthodox Uses of Electronic Cigarettes: A Surveillance of YouTube Video Content. Nicotine Tob Res 2019 Sep 19;21(10):1378-1384 [FREE Full text] [doi: 10.1093/ntr/nty132] [Medline: 29961828]

23. Massey ZB, Li Y, Holli J, Churchill V, Yang B, Henderson K, et al. Modifications to Electronic Nicotine Delivery Systems: Content Analysis of YouTube Videos. J Med Internet Res 2020 Jun 02;22(6):e17104 [FREE Full text] [doi: 10.2196/17104] [Medline: 32348288]

24. Hua M, Alfi M, Talbot P. Health-related effects reported by electronic cigarette users in online forums. J Med Internet Res 2013 Apr 08;15(4):e59 [FREE Full text] [doi: 10.2196/jmir.2324] [Medline: 23567935]

25. Luo C, Zheng X, Zeng DD, Leischow S. Portrayal of electronic cigarettes on YouTube. BMC Public Health 2014 Oct 03;14:1028 [FREE Full text] [doi: 10.1186/1471-2458-14-1028] [Medline: 25277872]

26. Tan AS, Bigman CA. Misinformation About Commercial Tobacco Products on Social Media—Implications and Research Opportunities for Reducing Tobacco-Related Health Disparities. Am J Public Health 2020 Oct;110(S3):S281-S283. [doi: 10.2105/ajph.2020.305910]

27. Jones DM, Guy MC, Soule E, Sakuma KK, Pokhrel P, Orloff M, et al. Characterization of Electronic Cigarette Warning Statements Portrayed in YouTube Videos. Nicotine Tob Res 2021 Aug 04;23(8):1358-1366 [FREE Full text] [doi: 10.1093/ntr/ntaa272] [Medline: 33400781]

28. Davidson J, Liebald B, Liu J, Nandy P, Van Vleet T, Gargi U, et al. The YouTube video recommendation system. 2010 Presented at: RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems; September 26; Barcelona, Spain p. 293-296. [doi: 10.1145/1864708.1864770]

29. Hussein E, Juneja P, Mitra T. Measuring misinformation in video search platforms: an audit study on YouTube. 2020 Presented at: Proceedings of the ACM on Human-Computer Interaction; May 28; New York, USA p. 1-27. [doi: 10.1145/3392854]

30. Covington P, Adams J, Sargin E. Deep neural networks for YouTube recommendations. 2016 Presented at: RecSys - Proceedings of the 10th ACM Conference on Recommender Systems; September 7; Boston, MA p. 191-198. [doi: 10.1145/2959100.2959190]

31. 2016 Surgeon General's report: e-Cigarette use among youth and young adults. Centers for Disease Control. URL: https://www.cdc.gov/tobacco/sgr/e-cigarettes/index.htm [accessed 2023-03-17]

32. Dashtian H, Murthy D, Kong G. An Exploration of e-Cigarette-Related Search Items on YouTube: Network Analysis. J Med Internet Res 2022 Jan 27;24(1):e30679 [FREE Full text] [doi: 10.2196/30679] [Medline: 35084353]

33. Tzioumis K. Demographic aspects of first names. Sci Data 2018 Mar 06;5:180025 [FREE Full text] [doi: 10.1038/sdata.2018.25] [Medline: 29509189]

34. Romito LM, Hurwich RA, Eckert GJ. A Snapshot of the Depiction of Electronic Cigarettes in YouTube Videos. Am J Hlth Behav 2015 Nov 01;39(6):823-831. [doi: 10.5993/ajhb.39.6.10]

35. Stellefson M, Chaney B, Ochipa K, Chaney D, Haider Z, Hanik B, et al. YouTube as a source of chronic obstructive pulmonary disease patient education: a social media content analysis. Chron Respir Dis 2014 May;11(2):61-71 [FREE Full text] [doi: 10.1177/1479972314525058] [Medline: 24659212]

36. Kloehn N, Leroy G, Kauchak D, Gu Y, Colina S, Yuan NP, et al. Improving Consumer Understanding of Medical Text: Development and Validation of a New SubSimplify Algorithm to Automatically Generate Term Explanations in English and Spanish. J Med Internet Res 2018 Aug 02;20(8):e10779 [FREE Full text] [doi: 10.2196/10779] [Medline: 30072361]

37. Curiskis SA, Drake B, Osborn TR, Kennedy PJ. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management 2020 Mar;57(2):102034. [doi: 10.1016/j.ipm.2019.04.002]

38. Benis A, Barak Barkan R, Sela T, Harel N. Communication Behavior Changes Between Patients With Diabetes and Healthcare Providers Over 9 Years: Retrospective Cohort Study. J Med Internet Res 2020 Aug 11;22(8):e17186 [FREE Full text] [doi: 10.2196/17186] [Medline: 32648555]

39. Haque MM, Rabbani M, Dipal DD, Zarif MII, Iqbal A, Schwichtenberg A, et al. Informing Developmental Milestone Achievement for Children With Autism: Machine Learning Approach. JMIR Med Inform 2021 Jun 08;9(6):e29242 [FREE Full text] [doi: 10.2196/29242] [Medline: 33984830]

40. Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: a comprehensive review. Comput Soc Netw 2019 Nov 10;6(1):1. [doi: 10.1186/s40649-019-0069-y]

41. Horne E, Tibble H, Sheikh A, Tsanas A. Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping. JMIR Med Inform 2020 May 28;8(5):e16452 [FREE Full text] [doi: 10.2196/16452] [Medline: 32463370]

42. Pariser E. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. New York, NY: Penguin Books; Apr 24, 2014.

43. Hannak A, Sapiezynski P, Molavi Kakhki A, Krishnamurthy B, Lazer D, Wilson C. Measuring personalization of web search. 2013 Presented at: Proceedings of the 22nd International Conference on World Wide Web; May 13; Rio de Janeiro, Brazil. [doi: 10.1145/2488388.2488435]

44. Khan ML. Social media engagement: What motivates user participation and consumption on YouTube? Computers in Human Behavior 2017 Jan;66:236-247. [doi: 10.1016/j.chb.2016.09.024]

45. Molyneaux H, O'Donnell S, Gibson K, Singer J. Exploring the gender divide on YouTube: an analysis of the creation and reception of vlogs. American Communication Journal. 2008. URL: https://ac-journal.org/journal/2008/Spring/3GenderandYoutube.pdf?origin=publication_detail [accessed 2023-03-17]

46. Kong G, Kuguru KE, Krishnan-Sarin S. Gender Differences in U.S. Adolescent E-Cigarette Use. Curr Addict Rep 2017 Dec;4(4):422-430 [FREE Full text] [doi: 10.1007/s40429-017-0176-5] [Medline: 29545987]

47. Wang TW, Gentzke AS, Neff LJ, Glidden EV, Jamal A, Park-Lee E, et al. Characteristics of e-Cigarette Use Behaviors Among US Youth, 2020. JAMA Netw Open 2021 Jun 01;4(6):e2111336 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11336] [Medline: 34097049]

48. Kong G, Laestadius L, Vassey J, Majmundar A, Stroup AM, Meissner HI, et al. Tobacco promotion restriction policies on social media. Tob Control. Online ahead of print [FREE Full text] [doi: 10.1136/tc-2022-057348] [Medline: 36328589]

49. Olfson M, Wall MM, Liu S, Sultan RS, Blanco C. E-cigarette Use Among Young Adults in the U.S. Am J Prev Med 2019 May;56(5):655-663 [FREE Full text] [doi: 10.1016/j.amepre.2018.12.004] [Medline: 30885517]

50. Comenetz J. Frequently occurring surnames from the 2010 census. US Census Bureau. URL: https://www.census.gov/topics/population/genealogy/data/2010_surnames.html [accessed 2021-05-05]

Abbreviations

GCN: graph convolutional network