Normalization of Vaping on TikTok: A Mixed-Methods Approach Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis

Sungwon Jung, Dhiraj Murthy, Bara S Bateineh, Alexandra Loukas, Anna V Wilkinson

Submitted to: Journal of Medical Internet Research
on: December 18, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Normalization of Vaping on TikTok: A Mixed-Methods Approach Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis

Sungwon Jung[1] MS; Dhiraj Murthy[1] PhD; Bara S Bateineh[2] MD, PhD; Alexandra Loukas[3] PhD; Anna V Wilkinson[2] PhD

[1]School of Journalism and Media University of Texas at Austin Austin US
[2]University of Texas Health Science Center at Houston School of Public Health Houston US
[3]Department of Kinesiology and Health Education University of Texas at Austin Austin US

Corresponding Author:
Sungwon Jung MS
School of Journalism and Media
University of Texas at Austin
300 W Dean Keeton St
Austin
US

Abstract

Background: Posts on social media that depict vaping in positive social situations serve to shape expectations about vaping, suggesting it is socially acceptable. TikTok, a social media platform popular among adolescents, has such content, despite restrictions on uploading depictions or promotions of controlled substances. There is a need to understand strategies employed in promoting vaping on TikTok, especially among susceptible youth audiences.

Objective: This study seeks to describe direct and indirect themes promoting vaping on TikTok and to understand how such themes may influence youth perceptions of vaping to inform public health communication and regulatory policies regarding vaping endorsements on TikTok.

Methods: We collected 14,002 unique posts from TikTok using 50 vape-related hashtags (i.e., #vapetok, #boxmod, etc.) and used K-means, an unsupervised machine-learning algorithm, to identify clusters. Following an in-depth examination of the posts in each cluster, we categorized posts into themes. Thereafter, we retrieved images from thematically organized video datasets. We assessed the performance of three machine-learning-based model architectures (i.e., ResNet50, VGG16, and ViT) when extracting and clustering visual features. We evaluated results and selected the best-performing model for the qualitative analysis. We examined the characteristics of 25% of the samples from each image cluster to determine the accuracy of classification. Finally, we selected 50 videos randomly from each of the five themes to compare with the theme descriptor.

Results: We identified five major themes. One theme ‘vaping marketing’ (8.3%) reflected direct marketing. The other four themes reflected: 'general vape' (19.6%), ‘TikTok influencer’ (27%), ‘vape brands’ (14.6%), and ‘vaping cessation’ (9.1%). The ResNet50 model successfully classified clusters based on image features (obtaining an F1 score of 0.97 on average, the highest of the three models). Content analyses indicated that vaping was depicted as a routine part of daily life and TikTok influencers subtly integrated vaping into popular culture (e.g., games) and social practices. Moreover, analyses revealed that the ‘vaping cessation’ theme had the lowest success of classification into the appropriate cluster, underscoring the relative low frequency of cessation-focused material on TikTok.

Conclusions: Results from both computational and qualitative methods on text and visual data indicate that vaping is normalized among youth on TikTok. The identified themes emphasize that various elements, such as everyday conversations, promotional content, and the influence from popular figures, work together to create an environment on TikTok where vaping is portrayed as a normal and accepted part of daily life. Furthermore, the high accuracy of our computational models, validated through qualitative content analysis, reinforces the reliability of our findings. This comprehensive examination provides valuable insights for regulatory policy and public health initiatives designed to address challenges posed by the normalization of vaping on social media platforms.

(JMIR Preprints 18/12/2023:55591)
DOI: https://doi.org/10.2196/preprints.55591

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Normalization of Vaping on TikTok: A Mixed-Methods Approach Using Computer Vision, Natural Language Processing, and Qualitative Thematic Analysis

Abstract

Background: Social media posts that portray vaping in positive social contexts shape people’s perceptions and serve to normalize vaping. Vaping content is particularly widespread on TikTok, a platform favored by young people. Despite restrictions on depicting or promoting controlled substances, vape-related content is easily accessible on the platform. There is a need to understand strategies employed in promoting vaping on TikTok, especially among susceptible youth audiences.

Objective: This study seeks to comprehensively describe direct (i.e., explicit promotional efforts) and indirect (i.e., subtler strategies) themes promoting vaping on TikTok using a mixture of computational and qualitative thematic analyses of social media posts. Additionally, we aim to describe how these themes might play a role in normalizing vaping behavior on TikTok for youth audiences, thereby informing public health communication and regulatory policies regarding vaping endorsements on TikTok.

Methods: We collected 14,002 unique TikTok posts from 50 vape-related hashtags (e.g., #vapetok, #boxmod, etc.). Using the K-means unsupervised machine-learning algorithm, we identified clusters and then categorized posts qualitatively based on themes. Next, we organized all videos from the posts thematically and extracted visual features of each theme using three machine-learning-based model architectures: ResNet50, VGG16, and ViT. We chose the best-performing model, ResNet50, to thoroughly analyze the image clustering output. To assess clustering accuracy, we examined 25% of the samples from each video cluster. Finally, we randomly selected 50 videos (5% of the total videos) from each theme, which were qualitatively coded and compared with the machine-derived classification for validation.

Results: We successfully identified five major themes from the TikTok posts.‘Vape product marketing’ (8.3%; n=1160), reflected direct marketing, while the other four themes reflected indirect marketing: ‘TikTok influencer’ (27%; n=3775), 'general vape’ (19.6%; n=2741), ‘vape brands’ (14.6%; n=2042), and ‘vaping cessation’ (9.1%; n=1272). The ResNet50 model successfully classified clusters based on image features, achieving an average F1 score of 0.97, the highest among the three models. Qualitative content analyses indicated that vaping was depicted as a normal, routine part of daily life, with TikTok influencers subtly incorporating vaping into popular culture (e.g., gaming, skateboarding, and tattooing) and social practices (e.g., shopping sprees, driving, and grocery shopping).

Conclusions: The results from both computational and qualitative analyses of text and visual data reveal that vaping is normalized on TikTok. Our identified themes underscore how everyday conversations, promotional content, and the influence of popular figures collectively contribute to depicting vaping as a normal and accepted aspect of daily life on TikTok. Our findings also highlight the relatively low frequency of cessation-focused material on TikTok, as indicated by the lower success rate in classifying ‘vaping cessation.’Our study provides valuable insights for regulatory policies and public health initiatives aimed at tackling the normalization of vaping on social media platforms.

Keywords: electronic cigarettes; vaping; social media; natural language processing; computer vision

Introduction

Background

TikTok is a short video-sharing platform that has grown in popularity, primarily among youth and young adults, since its launch in 2016. Though TikTok restricts uploading videos containing "the depiction, promotion, or trade of drugs or other controlled substances," such content persists on the platform. The e-cigarette or vaping industry and pro-vaping influencers are increasingly using social media platforms, including TikTok, to disseminate content that glamorizes and normalizes vaping [1,2]. This surge of promotional content is concerning as it shapes perceptions, especially those of younger people, suggesting that vaping is a desirable and socially acceptable behavior [3]. The portrayal of vaping as humorous, cool and its association with a particular lifestyle, can indirectly send misleading messages, implying that vaping is an integral and normal aspect of daily life [4]. Additionally, despite existing policies, some e-cigarette retailers continue to directly market e-cigarette products on social media that are explicitly designed to appeal to specific targeted audiences, such as youth [5]. This includes offering exclusive deals, employing attractive packaging, introducing a variety of flavors, etc. The underlying strategies of these marketing efforts are to desensitize people to the presence and use of e-cigarettes [6], making it appear that vaping is a routine part of everyday life [4]. This trend is alarming as it downplays the health risks associated with vaping, and shapes the attitudes and behaviors of young people who are vulnerable and open to suggestion [7].

Previous work based on qualitative content analysis has examined how TikTok posts depict e-cigarettes and vaping (e.g., [1,2]). These studies are based on manual, human coding, and require a large number of human labor hours, a method which does not effectively scale to the large volume of content currently produced and consumed on TikTok [8,9]. The current study employs a hybrid method to analyze large-scale TikTok posts depicting e-cigarettes and vaping. A hybrid method combines computational and qualitative thematic analysis of both textual and visual data, respectively, to ensure precision and reduce bias. This approach extends existing research by examining both linguistic (i.e., text) and visual (i.e., video) elements of e-cigarette content on TikTok. The purpose of this study is to identify direct or indirect themes promoting e-cigarettes on TikTok and to describe how these themes might play a role in normalizing vaping behavior on TikTok for youth audiences.

Normalization of Vaping

Normalization in sociology refers to the process by which an individual adjusts their behavior, thoughts, beliefs, or feelings to conform to the social norms and expectations of their culture or group [10]. This process can occur through repetition, ideology, and/or propaganda [11]. The normalization of vaping refers to the process by which vaping becomes perceived as a typical, socially acceptable behavior within society [12,13]. The shift in perception and attitude typically occurs gradually through multiple factors that interact and reinforce each other such as widespread advertising, media representation, cultural trends, peer influences, product accessibility, and perceived safety and harm reduction messages [14,15,16,17].

Marketing for vaping devices and flavor options across digital platforms and print media could play an important role in shaping public perceptions of vaping [14]. Similarly, the portrayal of vaping in movies, TV shows, and online content contributes significantly to its normalization, especially when famous characters, and models are depicted using e-cigarettes, subtly implying that vaping is acceptable and trendy [15]. Peer influence also has a strong impact on individuals' attitudes toward vaping; if vaping is common within a peer group, others in the group are more likely to vape to fit in [16]. The ease of accessing e-cigarettes, boosted by online sales, and in the absence of strict regulations, further increases their use. Lastly, the perception of vaping as a safer alternative to traditional smoking, driven by harm reduction messages, lowers barriers to experimentation and regular use [18]. All these factors collectively create an environment where vaping is increasingly normalized, especially among younger people.

The normalization of vaping carries considerable implications for public health, especially if the risks associated with vaping trends are not adequately addressed or recognized. For instance, research shows that the normalization of vaping, especially among youth, can increase dependence risk, normalize smoking behavior, and potentially lead to the initiation of traditional tobacco cigarettes, especially among those who might never have smoked otherwise [19,20].

Social Media Marketing Strategies and Normalization of Vaping

Social media platforms have become a pivotal tool in shaping public behaviors in recent years [21,22]. TikTok, for instance, offers users the opportunity to share their thoughts, experiences, and preferences to a vast audience [23]. Recent data indicate a surge in content related to vaping use on these platforms. Notably, the majority of messages are positive and favorable towards vaping [24]. Content in these messages ranges from personal testimonials about the benefits of vaping over traditional smoking to endorsements by influencers of specific e-cigarette brands and flavors [25]. Such messages might lead to gradually establishing the belief that vaping is a normalized and is socially acceptable behavior [26,27], regardless of any associated potential health risks or controversies [28,29].

Both direct (i.e., explicit promotional efforts of e-cigarettes) and indirect marketing (i.e., subtler strategies presenting the promotion in a lifestyle-oriented manner) through social media platforms, including TikTok, significantly influence the popularity of e-cigarette use [1,2]. Individuals who are exposed to promotional or user-generated e-cigarette content have a lower perception of harm, have more positive attitudes toward vaping, and have a higher risk of initiating e-cigarette use [26,27]. For example, a recent longitudinal study found that adolescents who reported seeing e-cigarette posts at least one time a week on TikTok had a higher risk of lifetime and current e-cigarette use [24]. Findings also indicated that the frequency of TikTok use was a significant predictor of initiating e-cigarette use [24]. Findings from another study indicated that when young people on social media (friends, influencers, etc.) are portrayed as happy and popular while using e-cigarettes, it creates a perception that vaping is a normative behavior within their social circles, and as such is a behavior to emulate [30].

TikTok's Community Guidelines [31] regarding "Drugs, controlled substances, alcohol, and tobacco" clearly state that users should not share content that involves "the purchase, sale, trade, or solicitation of drugs or other controlled substances, alcohol or tobacco products (including e-cigarettes)." However, e-cigarette content is prevalent on TikTok, often masked by indirect advertising methods [32]. Instead of traditional advertisements, e-cigarette brands collaborate with influencers to promote their products in their daily life posts and videos, making the promotion more organic. Influencers also link e-cigarette use with appealing lifestyle such as wellness, party scene, or luxury [2]. The repetitive e-cigarette content posted by influencers could potentially normalize the act of vaping, and make users more receptive to vaping, especially if the content is presented without highlighting potential risks. This psychological phenomenon, often termed the "mere exposure effect," suggests that people tend to develop a preference for behaviors merely because they are familiar with them [33].

Additionally, content featuring individuals from different backgrounds engaging with e-cigarettes has the potential to appeal to wider demographics. Individuals often connect more with content that mirrors their own experiences, identity, ethnicity, and age, which may prompt them to think about experimenting with e-cigarettes [34]. The dynamic interplay and use of these indirect marketing strategies on TikTok and other platforms contribute to the normalization of vaping by integrating it into popular culture, social practices, and lifestyle images among various backgrounds.

Thus, in this study, we aimed to identify and describe themes--either direct or indirect--that capture how e-cigarettes are promoted on TikTok and understand how these themes might play a role in the perception and normalization of vaping among young people.

The Need to use Computational Methods to study Vaping on Social Media

Given the rising popularity of TikTok as a social media platform, several qualitative studies have successfully examined e-cigarette-related content on TikTok. Sun et al [30] found that the themes capturing e-cigarettes depicted in TikTok videos included humor and jokes, lifestyle, marketing, vaping tricks, nicotine and addiction, creativity, and caution. Purushothaman et al [35] concluded that over two-thirds of the posts contained references to e-cigarettes. Although these studies advance our understanding of how e-cigarettes are depicted on social media, they are limited to the examination of a small number of posts because the content was analyzed manually. The volume of social media posts has rapidly risen and outpaced thematic analysis based on manual content analysis. Computational methods-both natural language processing and computer vision-can help researchers analyze e-cigarette-related content on social media at scale [36,37].

Media scholars increasingly employ computational analytic methods (e.g., developing a computational interface to facilitate the work of human coders) to manage ever-larger datasets [38]. Computational analytic methods have the potential to overcome many of the sampling and coding limitations of conventional content analysis. Early studies in computational social science on the TikTok platform have primarily concentrated on text analysis using natural language processing [e.g., 39]. In the vast domain of healthcare, natural language processing serves as a powerful tool with diverse applications, offering insights, and solutions to various challenges faced by clinicians, researchers, and healthcare providers [40]. Natural language processing techniques are instrumental in mining textual data from online health forums, social media platforms, and patient-generated content [41]. By analyzing conversations, posts, and comments, natural language processing algorithms can identify emerging health trends, public perceptions of diseases and treatments, adverse drug reactions, and even sentiments toward healthcare providers and services. For example, Sun et al [42] utilized topic modeling, a form of natural language processing and machine learning that automatically identifies topics or themes in a collection of documents [8], to examine COVID-19 vaccine-related posts on TikTok through text data. Furthermore, using natural language processing, Bharti et al [43] created a multilingual conversational bot to deliver essential healthcare education, information, and advice to individuals with chronic conditions.

More recent social media research has extended computational analytic methods scope to encompass computer vision techniques, a method designed for analyzing social media images and videos [44,45]. Computational techniques rooted in deep learning have facilitated the automated examination of vaping-related visual content shared on social media platforms such as Instagram and TikTok [45]. Moreover, machine learning models based on computer vision have been found to effectively and precisely recognize e-cigarette-related content (e.g., vaping devices and vapor clouds) on predominantly visually oriented social media platforms [44]. The transition from text-focused analysis to computer vision analysis signifies an evolution in TikTok research methodologies, offering a more comprehensive understanding of the platform's multifaceted content and societal implications. Nevertheless, there is still a dearth of studies applying computational content analysis methods to the context of vaping. Research is particularly lacking in the realm of combining computational methods with both text and images, especially to detect vaping-related themes.

Fundamentals of Computational Image Clustering

Computational image clustering is a popular computer vision method that groups images based on visual similarity [46]. Computer vision involves developing algorithms and models to help computers interpret and understand images and videos [47]. To mimic human vision, computer vision teaches machines to recognize objects, scenes, and patterns as well as understand visual inputs

[48]. Computer vision performs object recognition and content categorization using image features, patterns, and structures [49]. The goal is to find patterns or similarities in a large number of images without knowing their content or categories [50]. Each cluster comprises images with similar visuals [51]. This procedure enhances visual data navigation and comprehension, particularly for large image collections. There are two steps to image clustering: feature extraction and feature-based clustering.

Several machine learning computational models are capable of extracting features. In this step, visual attributes from input images are extracted. Initially, visual elements such as edges, patterns, and colors may be included [52]. The ultimate goal is to identify objects or patterns in images and capture more complex and higher-level features. Image pixel data are converted into numerical representations for computational analysis in this process [53]. Clustering algorithms group similar images into coherent categories after extracting these visual features [54]. The algorithms categorize images into meaningful groups to facilitate retrieval, analysis, and interpretation.

A convolutional neural network (CNN) can extract features and cluster images. CNNs are AI models that specialize in image classification [55]. The main steps in the CNN procedure are feature extraction and classification [56]. Vision Transformers (ViTs) is a deep-learning model architecture commonly used in computer vision tasks. Transformer is a deep learning model architecture designed for natural language processing tasks, characterized by its attention mechanism and ability to capture long-range dependencies [57]. Transformers' self-attention and parallelization make them good at sequential data processing [58]. K-means clustering is a popular unsupervised machine learning method for splitting a dataset into non-overlapping clusters. This algorithm determines data similarity [59]. Using the feature vectors from the previous step, K-means clustering groups similar images [60].

Using Computational Image Clustering on Vape-related Content on Social Media

Multimodal content clustering has not been used to study the normalization of vaping on social media. However, image clustering using computational methods (i.e., CNNs and K-means together) has been used successfully to study visual content on social media. For example, Ketonen and Malik [61] employed ResNet and VGG CNN models with K-means clustering to analyze and describe Instagram posts related to vaping. They found seven clusters: e-liquid, e-liquids, e-cigarettes, product packages, people, statements, and miscellaneous, with an accuracy of 90.76%. One limitation of this study is that it relied solely on posts under a single hashtag, which raises the possibility that not all posts are directly relevant to e-cigarettes. Furthermore, the utilization of images from the identified clusters for training a classifier for predicting a category for each image in the dataset may result in the inclusion of certain inaccuracies within the final group labels. Vassey et al [62] measured the influence of e-cigarettes on Instagram using the Inception v3 model, one of the CNN-based models developed by Google. The Vassey et al [63] model correctly classified over 90% of the images containing men, women, e-juices, and mod system devices across all samples. However, their model's training was limited to recognizing a single class per image despite specific images containing two distinct categories.

Setiawan and Purnama [64] also applied the DarkNet19 CNN architecture to extract image features from tobacco leaf images. They used k-means to cluster these images (with an accuracy ranging from 0.81 to 0.93 for all three classes). Previous research has successfully integrated feature extraction techniques using CNN architectures based on machine learning and clustering methods. However, previous research was constrained to the use of posts with a single hashtag, and their model needed to be updated for strong, robust performance attempted to overcome the limitations of previous social media studies related to vaping by collecting large amounts of data from multiple hashtags and employing a model with robust performance.

The current study aimed to employ a hybrid computational and qualitative approach to analyze a substantial volume of TikTok content related to e-cigarettes, utilizing K-Means clustering and pre-trained Vision Transformer and CNN models. By identifying direct and indirect themes in the promotion of e-cigarettes on TikTok, we sought to elucidate how these themes contribute to the normalization of vaping behavior among youth audiences. This comprehensive analysis addresses gaps in existing research methodologies and provides insights into the multifaceted factors influencing the portrayal and perception of vaping on social media platforms. Findings from this study are necessary to inform the development of interventions targeting adolescents and young adults that counteract the glamorization of vaping, as evidenced by the strategies employed by e-cigarette companies [64]. Findings may also be used to inform legislative actions, educational campaigns, and digital interventions aimed at de-normalizing harmful behaviors and safeguarding public health, a need further emphasized by recent findings on the prevalence of pro-vaping content on platforms like TikTok [65].

Methods

We used a hybrid approach, both computation and qualitative methods, to derive vaping-related themes from TikTok.

Figure 1. Project architecture; relevant figures and tables are referenced in parentheses.

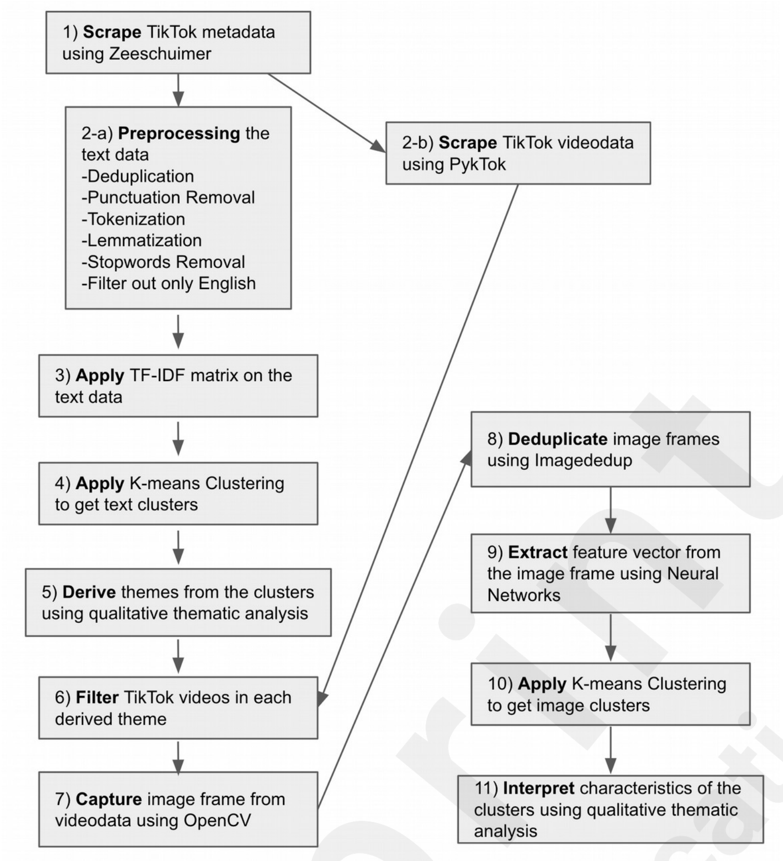
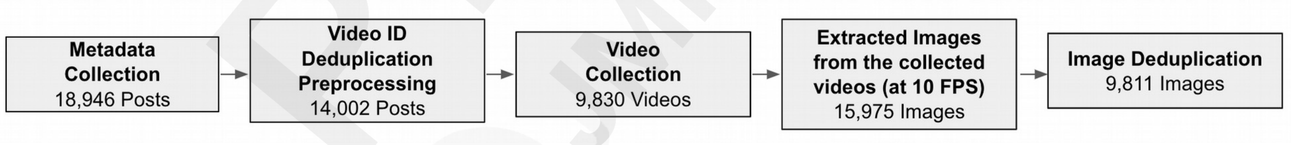

1) **Scrape** TikTok metadata using Zeeschuimer
2-a) **Preprocessing** the text data
-Deduplication
-Punctuation Removal
-Tokenization
-Lemmatization
-Stopwords Removal
-Filter out only English
2-b) **Scrape** TikTok videodata using PykTok
3) **Apply** TF-IDF matrix on the text data
4) **Apply** K-means Clustering to get text clusters
5) **Derive** themes from the clusters using qualitative thematic analysis
6) **Filter** TikTok videos in each derived theme
7) **Capture** image frame from videodata using OpenCV
8) **Deduplicate** image frames using Imagededup
9) **Extract** feature vector from the image frame using Neural Networks
10) **Apply** K-means Clustering to get image clusters
11) **Interpret** characteristics of the clusters using qualitative thematic analysis

Figure 2. Primary dataset evolution for the analysis; a flowchart of a data processing pipeline, starting with metadata collection, followed by video ID deduplication, extraction of images and finally image deduplication



This research has been assessed by the University of Texas at Austin's Institutional Review Board (IRB) concluding that it does not meet the criteria for human subject research, hence exempting it from requiring IRB review and approval (see Ethical Considerations section).

We collected 18,946 posts from TikTok from 2018-2023 using a set of vape-related hashtags (e.g., #vapetok, #boxmod, #vaper, etc.; see Multimedia Appendix for all hashtags). Though TikTok has a research application programming interface (API) for academics, it often returns inaccurate metadata and we therefore used a data scraping approach, which obtained accurate data and metadata. Data scraping was done in two steps. Using Zeeschuimer [66], a Firefox browser extension facilitates the scraping of data from social media platforms, including Twitter, Instagram, TikTok, and LinkedIn. Prior studies have demonstrated Zeeschuimer's efficacy as a data mining tool [67,68]. In accordance with previous studies, we used the Zeeschuimer tool to obtain e-cigarette-related content from TikTok's social media data.

Zeeschuimer enables the export of the collected metadata in Newline Delimited JSON (NDJSON) format. To scrape the data using Zeeschuimer, we scrolled down the TikTok hashtag URLs for each targeted hashtag. We collected 20,575 NDJSON files for each hashtag. The metadata fields included fundamental video information such as video_id, video_timestamp, video_duration, video_description, video_is_ad, video_height, video_duration, video_location, author_name, and user engagement elements such as video_sharecount, video_commentcount, and video_playcount. Then, we augmented our dataset with one additional step; we collected TikTok videos using the Pyktok [69] package. Pyktok obtains data directly from the JSON objects embedded within Tiktok web pages and from APIs that are not publicly documented. Based on the video_id collected using Zeeschuimer, we verified the valid video_id that is public and downloaded the TikTok videos. Accordingly, we obtained 18,946 data points for a total of 50 hashtags.

Ethical Considerations

This study has undergone review by the IRB at the University of Texas at Austin (ID: STUDY00006075) and has been determined not to constitute human subject research, as defined by regulations set forth by the Department of Health and Human Services (DHHS) and the Food and Drug Administration (FDA). Consequently, IRB review and approval were not required for this study.

As the data utilized in this study are publicly available, informed consent was not required.

To ensure confidentiality and privacy, all data have been anonymized. Specifically, we have taken measures to prevent the identification of individual participants, including blurring identifiable information such as user IDs or TikTok IDs in any images utilized. Additionally, due to the publicly available nature of the data, no compensation was required or provided for participants involved in this study.

Text Clustering Methods

We combined the text data, which is from the ‘video_description’ column, and then preprocessed them using Python's NLTK [70] library to perform lemmatization (finding the basic form of the word inflected in multiple forms in a sentence), punctuation removal, and capitalization removal. All the collected text was tokenized or divided into word tokens, which are grammatically indivisible language elements. After the basic text preprocessing, we used the Python package langdetect [71] to detect and filter out only English. Following this, we deduplicated the dataset, resulting in a total of 14,002 posts to define the clusters. Because the TikTok dataset lacks correct answer labels, we created a cluster using the K-means algorithm, a type of unsupervised learning. The document was first converted to a vector using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The TF-IDF weight is a statistical value used in information retrieval and text mining that indicates how important a word is in a specific document when there is a document group consisting of many documents [72]. We set the minimum document frequency to 5 as a parameter of the TF-IDF vectorizer. Words that appear less than five times were excluded from the word dictionary. Also, we set the ngram_range to (1, 2), which means that up to two groups of words were treated as if they were one word.

Then we normalized the word score of each document vector to keep it between 0 and 1 so that vector clustering could be performed effectively. We used cosine similarity between vector distances to perform clustering. We utilized k-means clustering as a computational clustering method. This is a clustering method in which the sum of the squared distances between the average vector in each cluster and the vectors in the cluster is minimized. Because the center point is chosen at random, the clustering result may be different or poor. To determine the optimal number of clusters K, we used the elbow method, which has been proven to be an excellent technique [73]. The elbow method is selected after considering the section where the inertia value, representing the sum of the distances between clusters, drops rapidly as the optimal number of k.

After identifying the optimal number of text clusters, we conducted qualitative text thematic analysis to assess the coherence and relevance of the clusters, further validating the robustness of our findings. We met twice as a group to evaluate the uniqueness of the themes derived from the elbow method and final themes were decided by consensus. We investigated the posts that belonged to each cluster in-depth, analyzed the data composition of each cluster, and grouped them into themes.

Video Clustering Methods

After conducting the thematic analysis on the text corpus, we conducted image clustering on derived themes. We examined whether the themes could be categorized according to the video's visual elements. Within the themes, we also identified sub-clusters. The average length of the videos with themes was between 22 and 31 seconds. Theme 1 encompassed 2,741 videos, Theme 2 1,272 videos, Theme 3 1,160 videos, Theme 4 had the largest collection of 3,775 videos, and Theme 5 had 2,042 videos.

We initially sampled every 25th video from the total videos. We discovered that if the video lasted longer than 10 seconds, most of the first few seconds were not devoted to vaping. Following this observation, to secure the vape-related moments in the TikTok posts, we opted for a standard of 10 Frames Per Second (FPS) and extracted frames at regular intervals of 10 FPS from each video we gathered. This approach was employed to mitigate any potential bias that may arise from the researcher's subjective selection process.

Subsequently, the Python-based Imagededup [74] package was employed to eliminate duplicate images. As mentioned above, the Imagededup has been widely engaged in computer vision research, demonstrating its efficacy in image deduplication [75,76]. The software package utilized a built-in CNN encoder for image deduplication. The CNN encoder employed the mobilenet_v3_small model to extract features from the images in the original image directory.

By computing the cosine similarities following the encoding of the images, the CNN algorithm selectively identified and removed only those image files duplicated from the original image directory. Cosine similarity is a mathematical metric used to compare the similarity of two vectors, extended to image comparison [77]. To preserve visual features, the CNN encoded each image into a vector. When these vectors were compared using the cosine similarity metric, the outcome indicated how similar the two images were regarding visual content. A cosine similarity score of 1 indicated that the images are identical, whereas a score of 0 indicated dissimilarity. The strict cosine similarity threshold was set at 0.85, which means that only image pairs with a cosine similarity score of 0.85 or higher would be recognized as duplicates. Images with a lower similarity score would be considered distinct.

After deduplication, we obtained 2,520 images for Theme 1 from 3,833 images. The same procedure was replicated on Themes 2, 3, 4 and 5. We obtained 1,242 images after removing duplicates from an initial pool of 2,172 images for Theme 2. Additionally, we obtained 1,269 images for Theme 3 from 1,821 images, and 4,276 images for Theme 4 from 7,495 images. Finally, we obtained 1,773 images for Theme 5 from 2,475 images.

CNN-based Clustering

The CNN model was trained on a large dataset to understand visual features, which is referred to as pre-training [78]. To classify images, feature extraction employs CNN layers to recognize increasingly complex patterns and objects. CNNs can classify images by generating a probability distribution over output categories [78]. For example, CNNs can recognize e-cigarettes and other similar items in images. After pre-training, CNN classification layers that classify images are removed, leaving only feature extraction layers [79]. These feature extraction layers extract high-level attributes from images. The remaining backbone, feature extraction layers, convert input images into feature vectors [80]. These feature vectors are useful for image clustering because they capture image shapes, textures, and object parts.

We used the PyTorch [81] library for the image clustering. We resized the images to a uniform size of 224x224 pixels. We converted the images to a tensor format, which is a multi-dimensional array or data structure that generalizes scalars, vectors, and matrices, suitable for deep learning. We used a ‘DataLoader’ function provided by PyTorch, which creates batches of images from the dataset to be used during model training and evaluation from the dataset. The function shuffles the data to ensure randomness during training. Then, we loaded a pre-trained CNN model (i.e., VGG-16, ResNet50), which is a built-in deep neural network model that has been trained on a large dataset for image classification tasks. We specifically opted for the VGG-16 [82] and ResNet50 [83] models due to their widely recognized reputation for robustness and accuracy in tasks related to image recognition [84,85]. We removed the final fully connected layers from the model, leaving only the convolutional layers. These convolutional layers were used for feature extraction. The images were passed through the VGG-16 network, and the output features were collected. These features capture high-level visual information about each image and were used for clustering.

Features extracted from the images were stored in the ‘features’ list. We turned on the evaluation mode of the model and then iterated a loop through the batches of images in our dataset. For each batch, the model processes the images without gradient computation (no gradients are backpropagated) using torch.no_grad(), and the output features are appended to the features list. The extracted features were concatenated into a single tensor and reshaped for subsequent clustering. We then determined the optimal number of clusters for the features we obtained using the silhouette score. The silhouette score measures how similar each data point in one cluster is to other data points in the same cluster compared to data points in other clusters. Our process involved iterating through different numbers of clusters, ranging from 2 to 10, to explore various clustering configurations. For each iteration, we computed the Silhouette score for the corresponding clustering number. The Silhouette score ranges from -1 to 1, where a higher score indicates better-defined clusters.

By systematically evaluating the Silhouette scores across different cluster numbers, we identified the number of clusters associated with the highest Silhouette score, indicating the optimal clustering number. After determining the optimal number of clusters, we perform K-means clustering using this number of clusters. The K-means algorithm groups the image features into clusters based on their similarity. We later saved images belonging to each cluster in separate subdirectories. For each cluster, we also created a Comma-Separated Values (CSV) file to store information about the images in that cluster for qualitative analysis.

Vision Transformers-based Clustering

The ViT extended the Transformer architecture to process images for computer vision. ViT considers them non-overlapping patches or tiles, whereas CNNS considers images as pixels. The ViT model treats each patch as a "word," employing self-attention mechanisms such as natural language processing (NLP) Transformers [86]. This allows the model to capture image patch dependencies and relationships over time. In ViT, positional embeddings aid the model in understanding the relative positions of patches in an image [87]. Maurício et al [88] discovered that ViTs perform similarly to traditional CNNs in image classification benchmarks. The ViT model was used in recent health studies and performed well [89,90]. The process of extracting ViT features is similar to that of CNN models. We extracted features from our dataset images using pre-trained ViT models (e.g., ViT-base, ViT-large). When trained on large image datasets, these models can encode images into rich feature vectors. The ViT model is used to extract features from images. The model transforms each image into a high-dimensional feature vector (representation). This vector represents the content and appearance of the image. We utilized ‘Transformers [91],’ a Pytorch-based Python library created by

HuggingFace [92] to employ the Vision Transformer model. We specifically chose ViT due to its proven effectiveness as a transformer-based model in comprehending visual content through tokenization [93].

After data transformation (resize and convert into tensors explained above), we initialized the ViT model using ViTFeatureExtractor and ViTForImageClassification functions. Due to computing capabilities, we employed the ‘vit-base-patch16-224-in21k’ for the pre-trained model. The model's name suggests that it is moderately sized (base), processes images as 16x16 pixel patches, accepts input images of size 224x224 pixels, and has been trained on a dataset containing 21,000 distinct image classification categories. The code iterates through the dataset to extract image features using the ViT model, and the extracted features are stored in the features list and concatenated into a single tensor for further analysis. We replicated the same step from CNN-based image clustering for the feature-based clustering task. We computed the silhouette score to figure out the optimal number of clusters.

Video Classification

We switched from clustering to classification to evaluate the quality of the clusters produced by the unsupervised K-means algorithm in each model. This approach was used to determine whether the features learned during the clustering phase are informative and can be used to predict cluster assignments accurately. We divided all the features and cluster labels to a training (80%) and testing (20%) datasets for classification. We used a simple neural network classification model with two linear layers and ReLU activation in between. ReLU stands for Rectified Linear Unit, and it is an activation function commonly used in artificial neural networks, particularly in deep learning models. The model underwent training with cross-entropy loss and the Adam optimizer. Cross-entropy loss or log loss, is a measure of how well a probabilistic classification model predicts the probability distribution of the true labels given the input data. The Adam optimizer efficiently adjusts the model's parameters during training by adapting learning rates and incorporating momentum that accelerates the learning in the relevant directions [94]. We trained the model for ten epochs on each theme. We then calculated the F1 score, which is a classification metric that combines precision and recall into a single value and provides a balanced measure of a model's performance.

Qualitative Thematic Analysis

After obtaining the output clusters from three individual models (i.e., ViT, VGG16, and ResNet50), we checked the samples from each cluster for each model to verify the clustering result. We determined the best model based on the performance, then we performed qualitative thematic analysis on each theme after looking at the final clustered images for each cluster on the themes. In validating the image clusters, we encountered challenges associated with interpreting visual similarities and patterns. To overcome this hurdle, we conducted a thorough examination of the clustered images, comparing them both within and across clusters to verify the consistency and accuracy of our clustering results. We initially picked 25% of the samples from each cluster, comprehensively analyzed each cluster's images to determine if the images were correctly clustered, and examined the characteristics of the cluster. For the second phase, we randomly selected 50 videos from each of the five themes to compare the content against the theme descriptor.

Results

Text Clustering Results

The elbow method determined the optimal number of clusters in the text dataset to be 15 (Figure 3). Our final sample was 14,002 TikTok posts after deduplication, but 3,012 posts were unrelated to our research scope and removed from our sample. From the total number of 10,990 e-cigarette-related posts, we successfully identified five themes (summarized in Table 1).

Figure 3. Elbow Method Result from K-means Text Clustering indicating that the optimal number of clusters is 15.

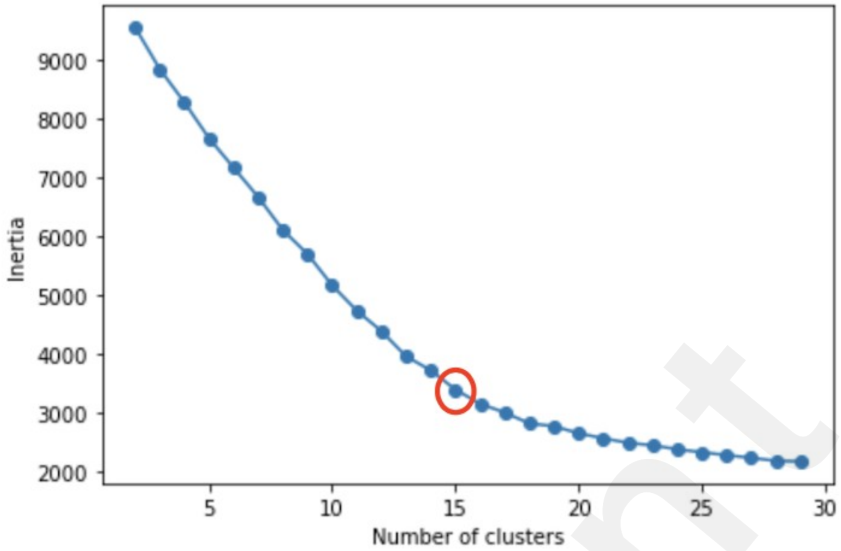


Table 1. Result of Textual Thematic Analysis; Thematic analysis revealed five themes, with 'Clusters' column indicating the aggregated clusters, 'Post Frequency' column denoting the frequency of posts per theme, and 'Indicator' column showcasing example words from each theme.

Theme	Clusters	Post Frequency	Indicator (word or hashtag)
1. General vape (including components (e.g., coil, cartridges, and pod) and use (e.g., blinker))	4	2741 (19.6%)	Vapetricks, cigarroeletronico, pod, vapelife, smoke, vapeo, vapor, vapemod, vapers, coil, vapepen, vapersontiktok, nicotine, know, funny, smoke, y'all, relatable, stop,Vapeporn, vapefamily, wonderwaterdrip, girlvaper, vapersontiktok, Carts, blinkersonlyfoo, fakecarts, blinkereyes
2. Vaping cessation	1	1272 (9.1%)	Vapelife, vapingislife, quit, Addiction, quitvaping, health, vaping is bad, stop, stopvaping
3. Vape product marketing (including flavors)	1	1160 (8.3%)	ultimatejuice, flavour, liquid, ultimatepuff, premiumliquid, eliquidshop
4. TikTok Influencer (includes skate, gamer, vape, and other influencers)	3	3775 (27%)	Skateboarding, skater, nintendo, vapetiktoker, Game, nintendo, retro, xbox, retrogrames, videogames, sega, console, games, gaming, thevideogamecollector, kingtinotazo (specific influencer)
5. Vape brands	3	2042	Elf, geekbar, elfbars, pods, fresh

		(14.6%)	mary, whereismymary, spookymary, scarymary, lostmarybm600 Juulgang, juullife, juulchallenge, juuling, juulsquad, mint, mango

The computational textual analysis identified 5 distinct clusters and our qualitative analysis indicated that these 5 clusters were organized by the following themes: general vape, vaping cessation, vape product marketing, TikTok influencer and vape brands. These themes provided insights into the online discourse surrounding vaping, from general discussions to specific aspects such as cessation, influencers, and brands.

Theme 1: General Vape

This theme encompassed general discussions and content regarding vaping, such as its components and usage. General vape consisted of a significant number of posts (19.6% of the total) and a variety of indicators, including "vapetricks," "coil," and "vapelife." These posts appeared to focus on various aspects of vaping, from the technical, such as coils and pods, to the experiential, such as user experiences. Additionally, terms such as "blinker", which is a light that flashes on and off to prevent overheating on the vaping device, suggested discussions of vaping techniques and tricks. This theme reflected a diverse and active vaping community on social media, engaging in discussions about their vaping experiences and preferences, and possibly promoting or sharing vaping-related knowledge.

Theme 2: Vaping Cessation

The vaping cessation theme had the second lowest frequency of posts (9.1% of the total). This theme discussed quitting vaping, addressing addiction, and vaping's negative health effects. Indicators such as "quitvaping," "stopvaping," and "vaping is bad" suggested that a large number of individuals utilized social media platforms to seek support, share their experiences, and declare their intentions to quit vaping. This theme emphasized the awareness and concerns regarding the addictive nature of vaping, as well as the efforts of some vapers to overcome their addiction.

Theme 3: Vape Product Marketing

This theme pertained to the marketing of vape products, with a particular focus on flavors. This theme was characterized by a substantial post frequency, accounting for 8.3% of the total content. The posts associated with this theme often included indicators such as "ultimatejuice," "flavour," "liquid," "ultimatepuff," "premiumliquid," and "eliquidshop." These indicators were indicative of the prominence of discussions and content related to the marketing and promotion of various vape products, with an emphasis on the diverse range of flavors available in the vaping industry. This theme reflected the significance of flavor offerings in the marketing strategies of vape product manufacturers and highlighted the considerable attention and discourse surrounding this aspect within the vaping community or industry.

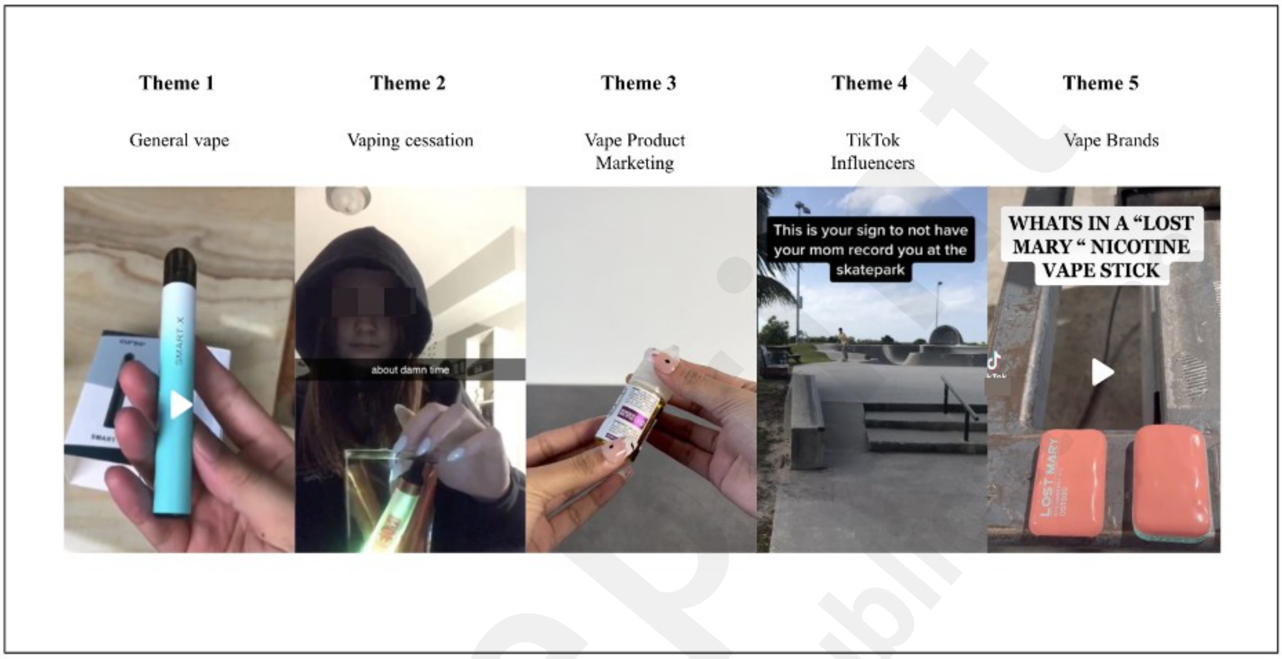
Theme 4: TikTok Influencers

This theme comprised a significant portion of the conversation (27% of the total), entailing discussions regarding TikTok influencers, and focusing on skateboarding, gaming, and vaping categories. The use of indicators such as "skateboarding," "gamer," and "vapetiktoker" emphasized the diversity of discussed influencers. This theme indicated that influencers may have shared their content or discussed their impact on the platform. This theme reflected the prominent role of influencers in shaping online conversations, especially on platforms such as TikTok.

Theme 5: Vape Brands

This theme focused on specific vape brands and their products, accounting for 14.6% of all posts. The vape brand’s theme included terms such as "Juulgang," "elfbars," and "mint." Users engaged in conversations regarding various e-cigarette brands, potentially discussing product preferences, flavors, and experiences. Specific brand names such as "Juul" indicated many users discussing this brand. This theme focused on the impact of brand loyalty and product diversity on the vaping community.

Figure 4. An example of themes derived from text clustering; selected representative images from actual posts corresponding to each theme.



Video Clustering Results

Table 2. Performance Metrics for Video Classification (F1 Score) for each model; the ResNet model performed consistently well across all five themes compared to the other two models

	General Vape (Theme 1)	Vaping Cessation (Theme 2)	Vape Product Marketing (Theme 3)	TikTok Influencers (Theme 4)	Vape Brands (Theme 5)
ResNet50	**0.974**	**0.972**	**0.960**	**0.975**	**0.975**
VGG16	0.932	0.986	0.953	0.955	0.980
ViT	0.943	0.941	0.847	0.986	0.778

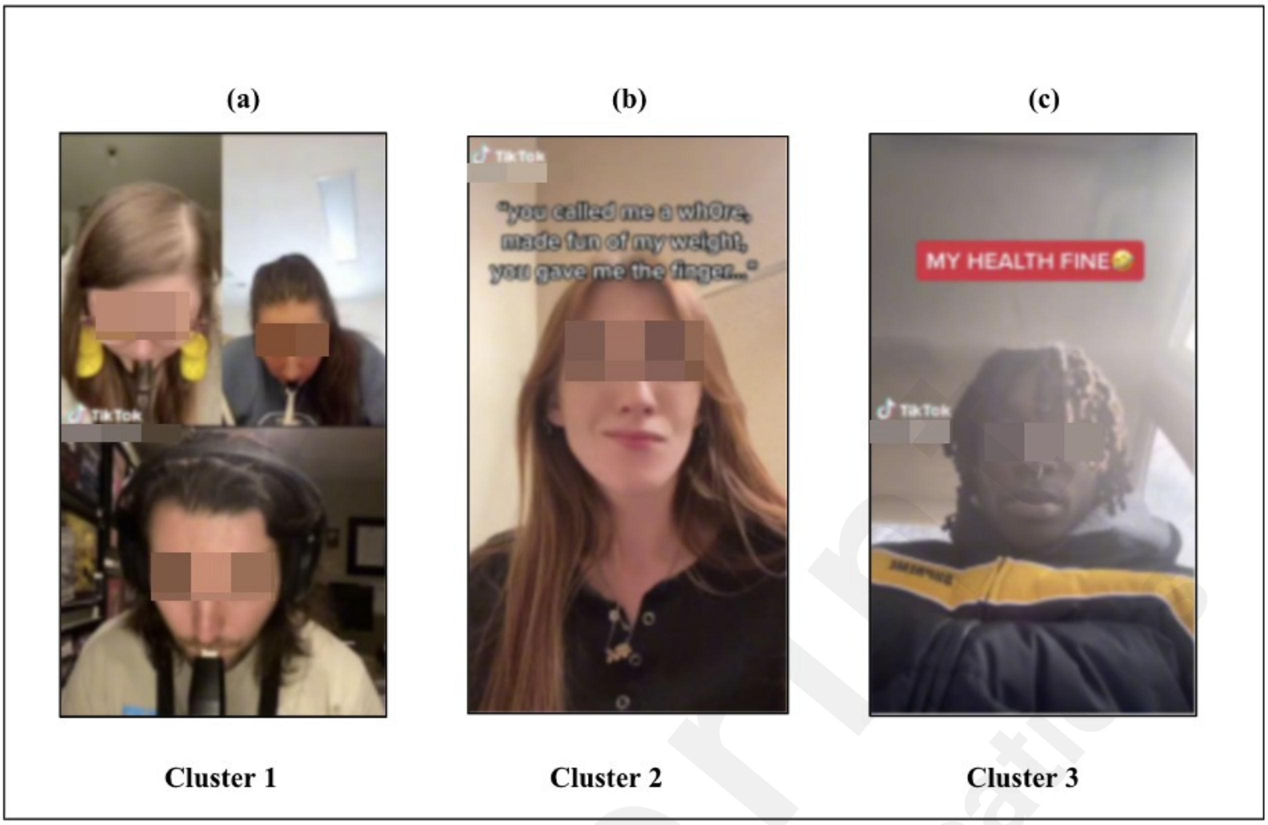
We chose to use the ResNet50 model for the qualitative thematic analysis based on the F1 score. Table 2 displays the ResNet50 model's consistently high performance on each theme (0.97 on average), which is likely due to the improved performance of neural networks with more layers.

The image clustering results on each theme from the preceding text thematic analysis revealed the normalization of vaping on TikTok. For example, TikTok influencers in domains such as skateboarding and gaming, and TikTok vloggers uploading their daily shopping lists depict vaping as an ordinary component of daily life. Furthermore, individuals featured in the video vape, regardless of location (i.e., house, car, etc.). These findings suggest a concerning trend of normalizing vaping as an everyday practice on the TikTok platform.

Theme 1 (General Vape)

The findings of our image clustering analysis indicated that the images in the 'general vape' theme formed three distinct sub-clusters. We characterized each cluster individually after carefully examining the images of each cluster. Three clusters can be identified as follows: 1) videos portraying individuals actively participating in the act of vaping by using vaping devices (see Figure 5(a)); 2) videos featuring individuals who choose not to utilize vaping devices, opting instead to partake in conversations that are unrelated to the act of vaping (see Figure 5(b)); and 3) videos showcasing individuals participating in conversations that cover a range of topics, including vaping-related subjects like the demonstration of vaping devices as well as broader subjects like personal health, in which the creators seek to satirize anti-vaping messages (see Figure 5(c)).

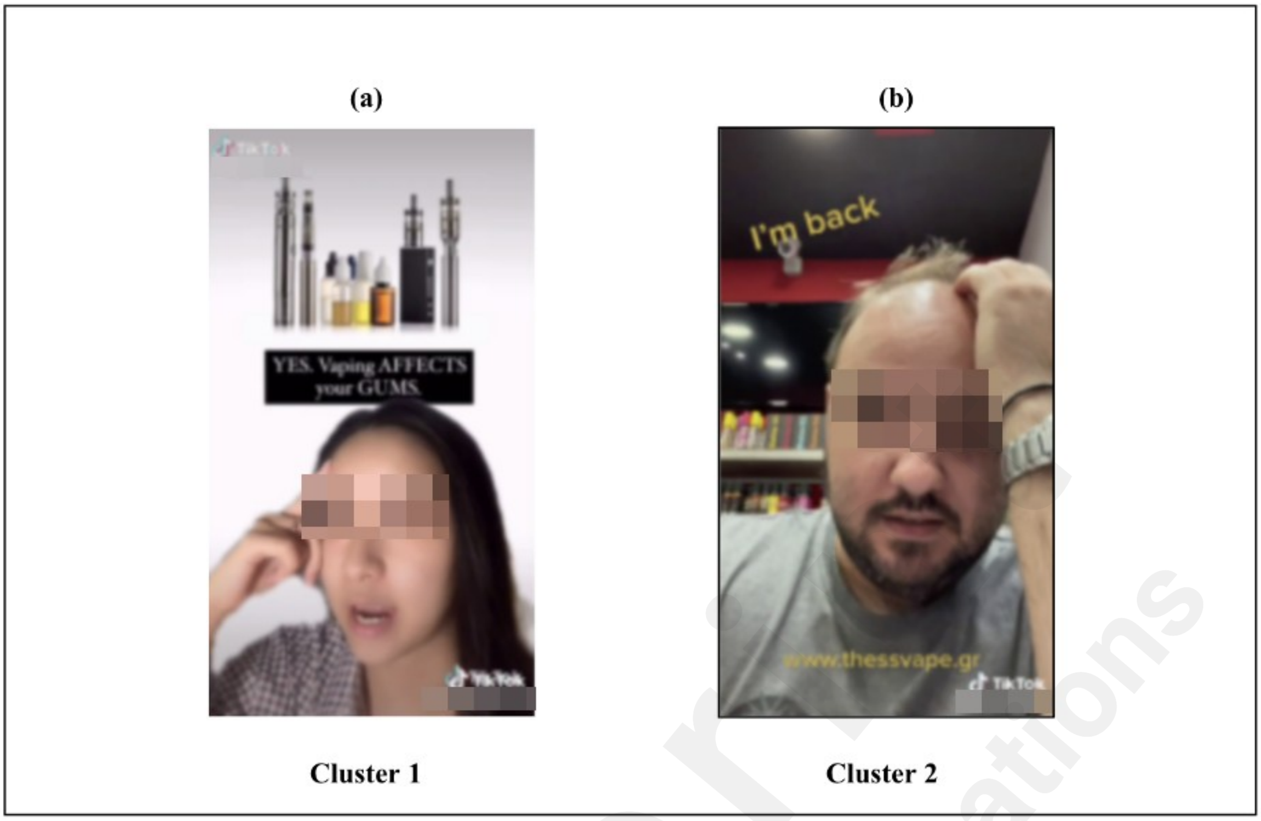
Figure 5. Sample image clustering results from Theme 1 (General Vape); three distinct sub-clusters were formed.


(a)
(b)
(c)
TikTok
"you called me a wh0re, made fun of my weight, you gave me the finger..."
MY HEALTH FINE
Cluster 1
Cluster 2
Cluster 3

Theme 2 (Vape Cessation)

The findings of our image clustering analysis indicated that the images in the 'vape cessation' theme formed two distinct sub-clusters: 1) experts or users with knowledge about vaping appear to inform video viewers about the link between vaping and potential diseases (e.g., lung-related diseases) and urge them to stop it (see Figure 6(a)); 2) users who are actually quitting vaping or have quit vaping share their own experiences (see Figure 6(b)).

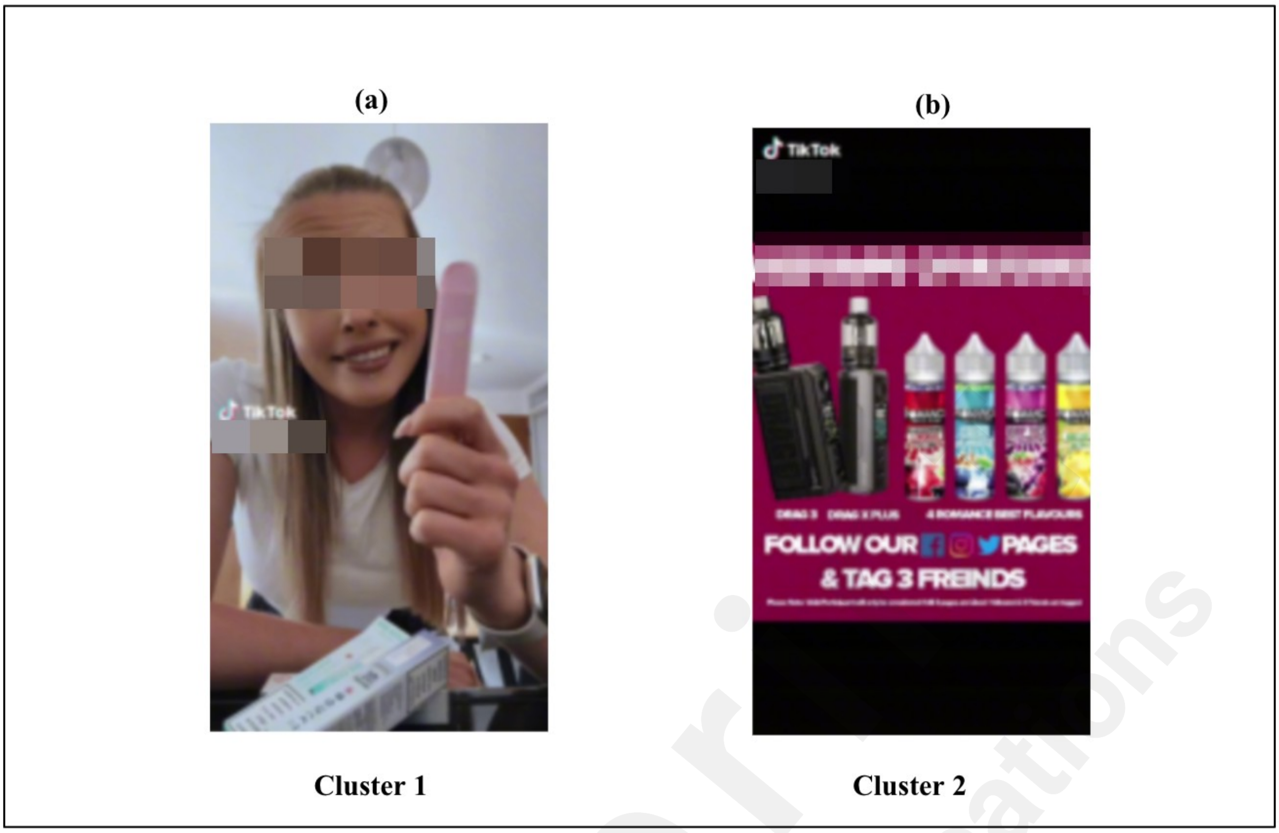
Figure 6. Sample image clustering results from Theme 2 (Vape Cessation); two sub-clusters were formed.



Theme 3 (Vaping Marketing)

We identified two distinct clusters from the Vaping Marketing theme: 1) videos that actively showcased and introduced vaping devices, illustrating the various aspects and functionalities associated with these products (see Figure 7(a)); and 2) videos that featured local shops actively selling and advertising vaping devices, indicating a localized and direct promotional approach (see Figure 7(b)). Despite TikTok's explicit guidelines opposing the direct promotion of vapes, these clustering results underscore the effectiveness of direct advertising or marketing of vaping devices at the micro level, encompassing both individual users and local shops. This observation sheds light on the nuanced ways in which TikTok's platform is utilized for promotional activities that may, at times, circumvent platform guidelines to reach a targeted audience.

Figure 7. Sample image clustering results from Theme 3 (Vaping Marketing); two sub-clusters were formed.

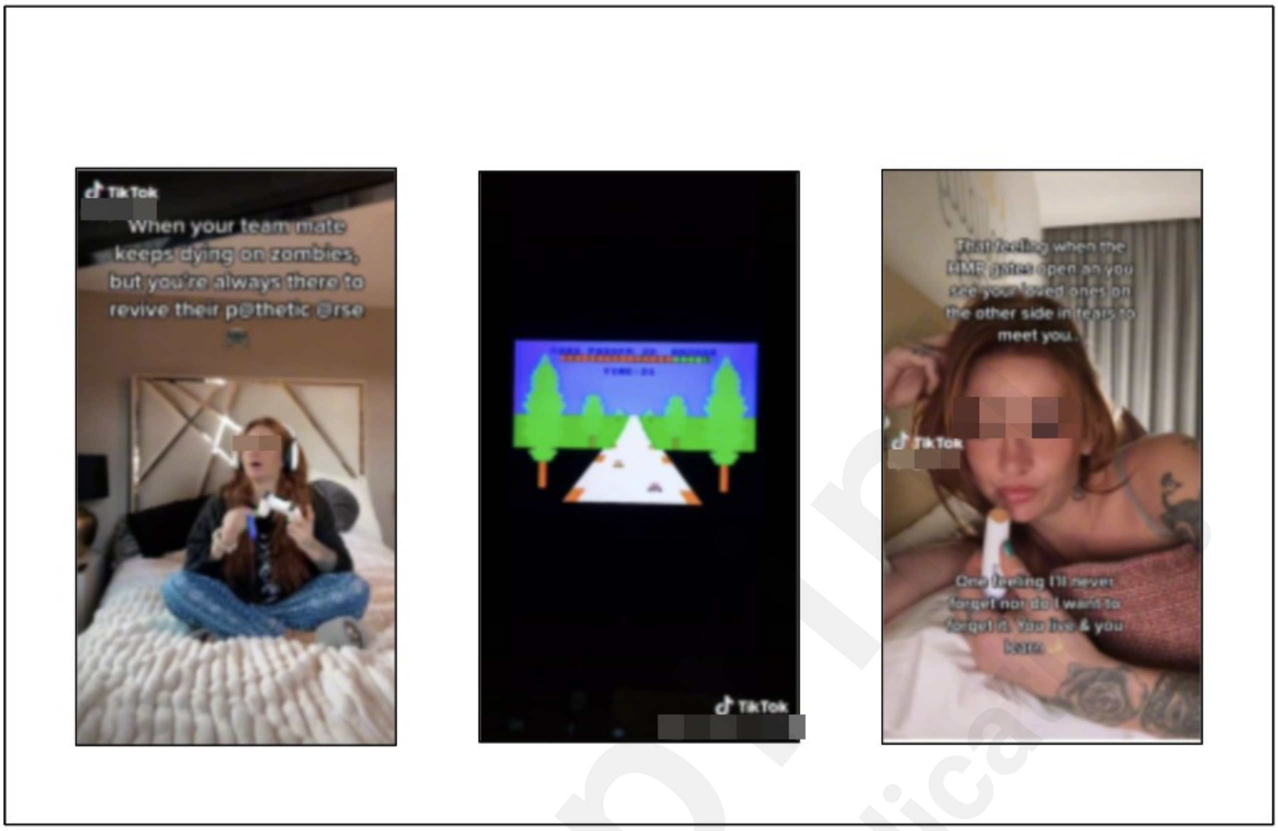


Theme 4 (TikTok Influencers)

We could not find differences between images in each cluster in the ‘TikTok Influencers’ theme. These findings demonstrate that the output of a black-box computer algorithm cannot be wholly trusted. Instead, a human-in-the-loop process, in which researchers interpret the results and determine their significance is required. Although a vape appeared in the video, it was not the focal point but a supplementary element. In other words, TikTok videos in the TikTok influences cluster? were depicting vaping alongside daily or quotidian activities such as playing video games, driving, grocery shopping, and reading books (see Figure 8). There were also vape influencers, but instead of actively promoting vape products, they flaunted adding new decorations to vapes or customizing cartridges. In other words, most images in the 'TikTok Influencers' theme depicted vaping as a natural part of daily life.

Figure 8. Sample image clustering results from Theme 4 (TikTok Influencers); no differences were

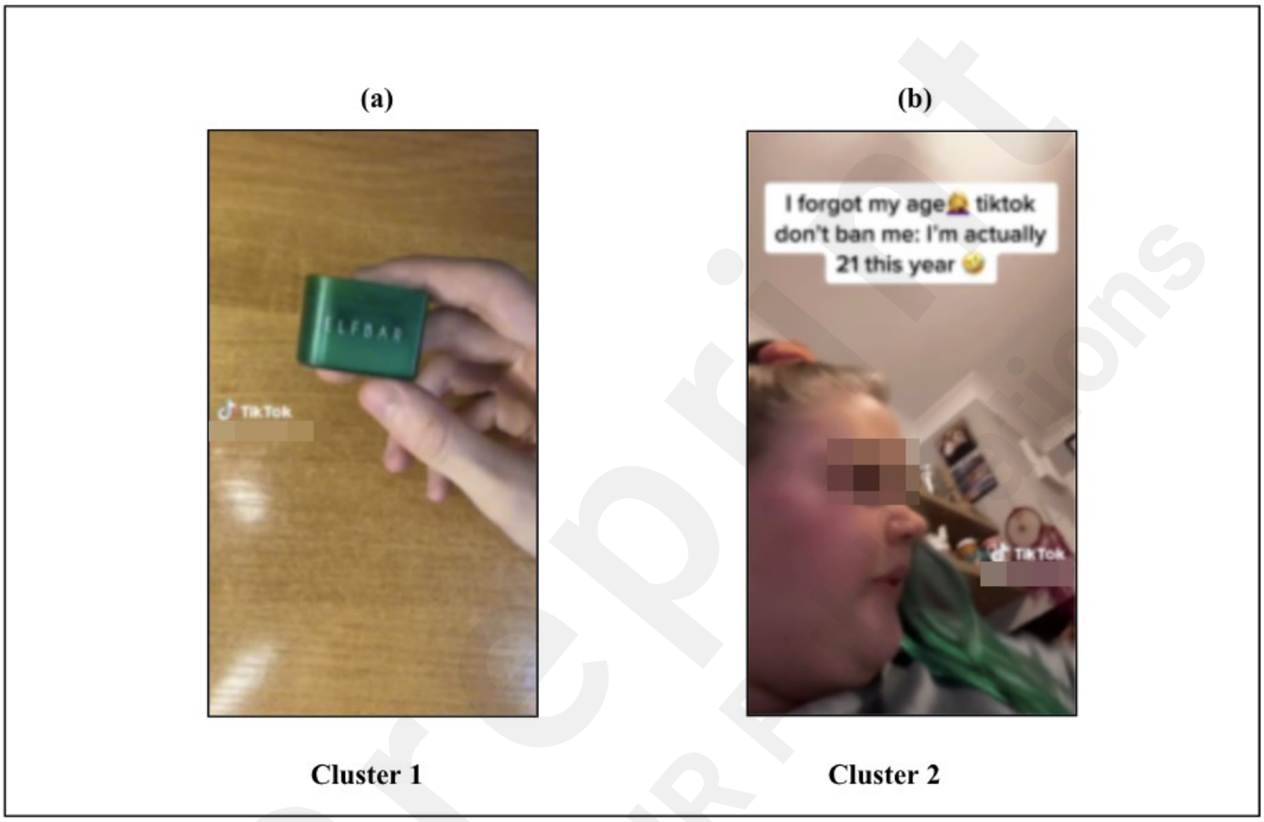
found between images within each cluster.


TikTok
When your team mate keeps dying on zombies, but you're always there to revive their p@thetic @rse
TikTok
TikTok
That feeling when the HMP gates open an you see your loved ones on the other side in tears to meet you..
One feeling I'll never forget nor do I want to forget it. You live & you learn

Theme 5. Vape Brands

We identified two distinct clusters from Theme 5: 1) videos demonstrating how to use a vape device after opening the packaging, reviewing the product, and vaping (see Figure 9(a)); and 2) vaping-related context and vape use in daily settings (see Figure 9(b)). A TikTok user depicted in Figure 9(b) uploaded a shopping haul-related video. In the video, she displays a new vape (Lost Mary) and declares, "TikTok, please don't ban me; I'm actually 21 this year." In other videos, point of view (POV) is frequently used as a text caption. POV is frequently used in captions and on-video captions to indicate that the viewer should view the scene from their point of view. One user, for instance, uploaded a video with a text caption such as "POV: you discovered elf bar made a calendar" and sang along with the background music, which contains the lyrics "Oh, I want you so badly. It's my greatest wish." These findings demonstrate that TikTok users are actively participating in normalizing vaping as an everyday practice. This is exemplified by the platform's how-to videos and everyday vaping scenarios, which reflect a concerning trend in the platform's content.

Figure 9. Sample image clustering results from Theme 5 (Vape Brands); two sub-clusters were formed.


(a)
(b)
ELFBAR
TikTok
I forgot my age🤦 tiktok don't ban me: I'm actually 21 this year 😉
TikTok
Cluster 1
Cluster 2

Qualitative Thematic Analysis Results

In our qualitative thematic analysis, 50 videos were randomly selected from the five themes. These videos were viewed and the theme of individual videos was compared with the theme descriptor to validate the themes. Of the 50 videos selected at random from Theme 1 (general vape), 40 were vape-related. The 10 non-vape-related videos included a concert in Kuala Lumpur (which featured a smoky environment), a driving Public Service Announcement (PSA) (which had a blue blinking light), three car-related videos featuring blinkers, a festive scrunchy, a disposable camera, jlcpcb labs (blinking circuits), and a calligraphy pen.

For Theme 2 (vaping cessation), of the 50 videos selected at random, only 2 were not vape-related. One of the non-vaping-related videos promoted clothing, and the other promoted instant coffee. However, a total of 15 videos covered content directly related to vaping cessation. In addition, two more were related to quitting cigarettes, one of which suggested that vaping was the best way to quit cigarettes. Three of the videos promoting cessation used dentists, several videos provided tips on how to quit vaping, and others shared information on the health and mental health benefits of quitting, such as feeling good after quitting vaping.

Of the 50 videos selected at random from Theme 3 (vaping marketing), 12 were not vape-related. The TikTok videos that were not vape-related included an episode of a story, three TikTok videos of live music concerts that were filled with smoke, a girl bemoaning the loss of her boyfriend, four TikTok videos discussing video games, one highlighting WWI Prussian rifle cartridges, a Wing Stop taste tester (mentions a HOT BOX meal), one tattoo pen and one calligraphy pen.

Only one of the 50 videos selected at random from Theme 4 (TikTok influencers) was not vape-related. Six videos provided an anti-vaping message, and one showed the aftermath of a vape exploding in a bedroom. The majority of videos, over 40, shared images of vape pens, vape juice, people in vape shops, and presented positive images of young people vaping.

Finally, of the 50 videos selected from Theme 5 (vape brands), 7 were not vape-related. These included a video by ‘coolkidclaire,’ three boys dancing or playing football (which could be mistaken for blinking), a video of a clothing store, one about food, and another about a car (another possible blinker-related error).

Of the 250 videos reviewed, just under 13% were not vape-related. However, most of these non-vape-related videos presented a smoke-filled environment, mentioned blinking in some form, or presented the image of a pen. As such most videos had words or aspects of images that were similar in nature to vaping-related content, and content that could easily be misclassified by a machine, highlighting the need for human interaction in this process. Moreover, only Theme 2 included a variety of vaping-related images that were not related to the theme’s descriptor of vaping cessation.

Although the influencer theme included some videos providing tips about how to quit and encouraging quitting, Theme 2 included 15 videos with this message. Again, in comparison to the other four themes in which most videos fit the general descriptor of the theme, this was less the case with quit-vaping-related TikTok videos. Thus, the only theme in which there was any noteworthy misclassification was in Theme 2, vaping cessation, wherein 15 of the 50 videos were specifically cessation-related.

However, misclassification was not a failure of the model because a) only 2 videos in Theme 2 were not directly vape-related and b) classifying cessation-related content was challenging due to the complexity of material along with the use of humor and double entendres by TikTok video creators. Many TikTok videos were characterized by a light and fun tone through the juxtapositioning of ideas, words, and images, which was also present in many of these vaping-related videos. Moreover, although relatively few in number, these vaping cessation videos are important because of the tone used by the creator, a tone that might resonate well with other TikTok users.

Finally, the nature of the misclassifications with Theme 2 underscores the possibility that when humor and/or double entendres are an integral aspect of the message, a computer-driven classification approach will benefit from maintaining a human-in-the-loop. While we can rely on machine learning algorithms to sort through thousands of images and text entries, improving the accuracy of the process requires humans to interpret images and text that are context-specific and rely on socially inferred meaning.

Discussion

Principal Findings

This study is one of the first to utilize a combination of computational methods to analyze both text and image data from large vaping-related content posted on TikTok from 2018-2023 (n=14,002). Our analysis identified five main themes: general vaping, vaping cessation, marketing, influencers, and vape brands. Notably, the majority of these themes/posts (three out of the five) indirectly promoted vaping and may contribute to creating an online environment that normalizes e-cigarette use. Relatively few videos provided anti-vaping messages; the theme identified as Vaping Cessation included more posts that indirectly promoted vaping and contributed to creating an online environment that normalizes e-cigarette rather than supporting vaping cessation.

We found a relatively low proportion of posts explicitly engaged in direct marketing. It is noteworthy that when direct marketing was observed, it primarily focused on promoting various vaping brands and flavors. Our findings of indirect promotion of vaping products on TikTok, particularly promotion of vaping brands and flavors align with research by Vassey et al [45], which revealed a significant increase in the prevalence of e-juice flavor names and e-cigarette brand names on Instagram and TikTok. Of note, existing studies indicate that the variety of available flavors, ranging from fruity to sweet and other appealing tastes, is one of the top reasons cited by young people to experiment with e-cigarettes [95,96].

Further findings indicated that popular TikTok users or influencers primarily promoted vaping products indirectly. Influencers, often associated with youthfulness on social media are skilled at subtly endorsing products, thereby making them more relatable than less traditional advertisements. Findings showed that influencers make vaping look like a natural part of their daily lives as they engage in activities like gaming, partying, traveling, etc., which may lead to indirectly normalizing the behavior. Partnering with influencers can be effective in shaping attitudes and behaviors because it does not trigger the same resistance as overt direct advertising [97]. Nonetheless, these findings align with and expand prior research documenting a notable presence of influencers' role in promoting vaping social media platforms on behalf of tobacco brands [1]. For example, a recent study shows that 20% of the influencers exclusively posted about e-cigarette use, while 80% of them posted about e-cigarettes and other topics on Instagram [3]. Moreover, when influencers create content related to vaping, TikTok algorithms actively promote this content to a larger audience, significantly expanding its reach and impact [97].

Our study identified a substantial amount of content that is against TikTok's current policies, underscoring the need for stricter regulations in online advertising, particularly when main companies employ indirect methods through smaller sellers or influencers to promote their vaping products. Furthermore, normalization of vaping through influencers or other strategies poses a significant concern for public health, especially among younger people who are more susceptible to the influences of such advertising [98].

Implications for Policy and Research

Given the findings in this study, social media platforms should consider implementing more robust measures to ensure videos target their intended audiences. For example, social media platforms should; a) require age verification for users accessing any vaping content or engaging with vaping-related content, b) require enhanced posts that highlight the potential health risks associated with vaping, and c) prevent content that contains misleading information about the safety of vaping. Policymakers could develop regulations to enforce more strict advertising policies explicitly prohibiting the promotion of vaping products, including restrictions on content that indirectly promotes vaping [99,100].

Policymakers may also consider implementing advanced algorithms to detect and flag content that indirectly promotes vaping. These algorithms can identify visual patterns and keyword text associated with vaping-related content. Furthermore, policies should develop and enforce specific guidelines for influencers, including transparency requirements regarding affiliations with vaping brands, and restrictions on promoting vaping to underage populations. In addition, investing in development prevention programs aimed at de-normalizing vaping use among youth could be crucial in mitigating the spread of vaping culture and its associated risks.

Future work and limitations

We successfully implemented a novel approach that combines two machine learning methods

—natural language processing and computer vision—alongside qualitative thematic analysis. Specifically, our approach leveraged computational methods, specifically K-means clustering for text data and state-of-the-art Vision Transformer and CNN models for image data. A computational, ML-based methodology allowed us to analyze a large volume of textual and visual content efficiently. We uncovered key insights into the prevalence and acceptance of vaping on TikTok using these computation techniques. However, we acknowledge a few limitations. Incorporating video transcription (i.e., which automatically transcribes spoken language into written text) could provide additional textual data for analysis in future work. The video transcription could significantly improve the text corpus for thematic analysis. Additionally, due to our hashtag-based data scraping approach, we encountered limitations in extracting comprehensive vaping-related information. For instance, our objective was to scrape content from posts containing the hashtag #blinker, which denotes a visual indicator akin to the LED lights found on a vaping device, including an e-cigarette or vape pen. Our scraped results, on the other hand, resembled irrelevant content, such as a drive-by PSA car with a blue blinking light or a calligraphy pen. The algorithmic flow of TikTok necessitates a human-in-the-loop approach, which we implemented successfully in this study.

Future work could benefit from incorporating video transcription, further enriching textual data for even more comprehensive thematic analysis. Though our classification of vaping-related messages was successful (F1=0.97), our classification of vaping cessation messages leaves room for future work. Specifically, cessation-related messages are difficult for machines to understand as they contain subtle cues (e.g., humor). These findings indicate that humans are vital to the classification process. Future work should evaluate methods to improve a machine’s ability to interpret context-specific images and text and extract socially inferred meaning in complex areas such as humor and cessation. Furthermore, exploring the integration of more advanced natural language processing models, such as Bidirectional Encoder Representations from Transformers (BERT) [101], could further enhance the depth and accuracy of our text analysis. These models offer contextual understanding, capturing intricate relationships within text that may provide deeper insights into the themes and sentiments expressed in vaping-related content on TikTok.

Conclusions

This study explored the thematic analysis of vaping-related content on TikTok, employing a hybrid computational and qualitative approach to both textual and visual data, respectively, to ensure precision and reduce bias. We aimed to discern the patterns and themes within vaping-related content to understand the extent to which vaping is normalized in the context of social media discourse. We conducted a multifaceted investigation into user-generated content on social media platforms related to e-cigarette (vape) discussions, focusing specifically on TikTok. We found that vaping-related content on TikTok is marked by distinct themes, including those portraying vaping as a common activity in daily life, discussions on vaping cessation, TikTok influencers incorporating vaping into their content, and specific discussions related to vape brands and products. These themes provided valuable insights into the normalization of vaping on the platform. Images within each clustered theme highlighted the various facets of vaping, such as active vaping practices, conversations unrelated to vaping but with a vape visible in the video, and vaping as an integral part of daily life.

Consistent with these findings, the qualitative analysis revealed that our machine learning approach correctly recognized vape-related content by successfully parsing these data into smaller meaningful units (i.e., the five themes). The cessation-related TikTok videos merit additional comment and research. While few in number, this is unsurprising as they are created by TikTok users. However, their impact may be far greater than officially created content as the user-generated approach of cessation messaging might be far more appealing to other TikTok users than cessation messages created by professionals.

Overall, our computational methods with a qualitative human-in-the-loop approach allowed us to gain a deeper understanding of how vaping is represented and normalized on TikTok. Our study provided a broader perspective, emphasizing the normalization of vaping practices rather than focusing solely on the textual or visual content of prior research. Moving forward, our findings underscore the importance of ongoing surveillance of vaping-related content on social media platforms, highlighting the need for targeted interventions to mitigate the normalization of harmful behaviors among youth populations.

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health under award number R01MD017280 to AL and AVW. We haven’t used any generative AI tool (e.g., ChatGPT by OpenAI) to generate the text manuscript.

Data Availability

The data and code generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of interest

None declared

Abbreviations

API: Application Programming Interface

CNN: Convolutional Neural Network

CSV: Comma-Separated Values

FPS: Frames Per Second

NDJSON: Newline Delimited JSON

POV: Point of View

PSA: Public Service Announcement

ViTs: Vision Transformers

BERT: Bidirectional Encoder Representations from Transformers

IRB: Institutional Review Board

DHHS: Department of Health and Human Services

FDA: Food and Drug Administration

Multimedia Appendix 1

List of TikTok hashtags. We scraped the data with 50 hashtags.

Multimedia Appendix 2

Text clustering result. We identified 15 clusters from K-Means clustering method.

References

1. **Sun T, Lim CCW, Chung J, et al. Vaping on TikTok: a systematic thematic analysis. *Tob Control*. 2023;32(2):251-254. doi:10.1136/tobaccocontrol-2021-056619**
2. Morales M, Fahrion A, Watkins SL. # NicotineAddictionCheck: puff bar culture, addiction apathy, and promotion of e-cigarettes on TikTok. *International Journal of Environmental Research and Public Health*. 2022;19(3).
3. 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. doi:10.1093/ntr/ntaa003
4. Jeffares Y. *From Smoking to Vaping: Vapers' Perspectives of Policy Related Facilitators and Barriers: A Thesis Submitted in Fulfilment of the Requirements for the Degree of Master of Science in the Department of Psychology*. Massey University; 2020.
5. Vogel EA, Hashemi R, Ramo DE, Darrow SM, Costello C, Prochaska JJ. Adolescents' perceptions of nicotine vaping-related social media content. *Psychology of Popular Media*. Published online 2023.
6. Collins L, Glasser AM, Abudayyeh H, Pearson JL, Villanti AC. E-cigarette marketing and communication: How E-cigarette companies market E-cigarettes and the public engages with

E-cigarette information. *Nicotine Tob Res*. 2019;21(1):14-24. doi:10.1093/ntr/ntx284

7. Verhaegen A, Van Gaal L. Vaping and cardiovascular health: The case for health policy action. *Curr Cardiovasc Risk Rep*. 2019;13(12). doi:10.1007/s12170-019-0634-9
8. Lewis SC, Zamith R, Hermida A. Content analysis in an era of big data: A hybrid approach to computational and manual methods. *J Broadcast Electron Media*. 2013;57(1):34-52. doi:10.1080/08838151.2012.761702
9. Su LYF, Cacciatore MA, Liang X, Brossard D, Scheufele DA, Xenos MA. Analyzing public sentiments online: Combining human-and computer-based content analysis. *Information. Communication & Society*. 2017;20(3):406-427.
10. Richards C, Glasgow L, Bidaisee S, et al. White paper on vaping: Electronic cigarettes use--A case for restrictive policies in Grenada, West Indies. *West Indies International Public Health Journal*. 2020;(1).
11. May C, Finch T. Implementing, embedding, and integrating practices: An outline of normalization process theory. *Sociology*. 2009;43(3):535-554. doi:10.1177/0038038509103208
12. McEvoy R, Ballini L, Maltoni S, O’Donnell CA, Mair FS, Macfarlane A. A qualitative systematic review of studies using the normalization process theory to research implementation processes. *Implement Sci*. 2014;9(1):2. doi:10.1186/1748-5908-9-2
13. Linkenbach JW, Lubbers DT, Brandon JM, Ooms JD, Langenberg AJ, Kilmer JR. Assessing adolescent vaping norms and perceptions in a statewide multi-community project. *Subst Use Misuse*. 2023;58(3):428-433. doi:10.1080/10826084.2023.2165413
14. Lozano P, Arillo-Santillán E, Barrientos-Gutíerrez I, Reynales Shigematsu LM, Thrasher JF. E-cigarette social norms and risk perceptions among susceptible adolescents in a country that bans E-cigarettes. *Health Educ Behav*. 2019;46(2):275-285. doi:10.1177/1090198118818239
15. Carpenter CM, Wayne GF, Pauly JL, Koh HK, Connolly GN. New cigarette brands with

flavors that appeal to youth: Tobacco marketing strategies. *Health Aff (Millwood)*. 2005;24(6):1601-1610. doi:10.1377/hlthaff.24.6.1601

16. Payne JD, Orellana-Barrios M, Medrano-Juarez R, Buscemi D, Nugent K. Electronic cigarettes in the media. *Proc (Bayl Univ Med Cent)*. 2016;29(3):280-283. doi:10.1080/08998280.2016.11929436
17. Groom AL, Vu THT, Landry RL, et al. The influence of friends on teen vaping: A mixed-methods approach. *Int J Environ Res Public Health*. 2021;18(13):6784. doi:10.3390/ijerph18136784
18. Golan R, Muthigi A, Ghomeshi A, et al. Misconceptions of vaping among young adults. *Cureus*. Published online 2023. doi:10.7759/cureus.38202
19. Soneji S, Barrington-Trimis JL, Wills TA, et al. Association between initial use of e-cigarettes and subsequent cigarette smoking among adolescents and young adults: A systematic review and meta-analysis. *JAMA Pediatr*. 2017;171(8):788. doi:10.1001/jamapediatrics.2017.1488
20. Watkins SL, Glantz SA, Chaffee BW. Association of noncigarette tobacco product use with future cigarette smoking among youth in the population assessment of tobacco and health (PATH) study, 2013-2015. *JAMA Pediatr*. 2018;172(2):181. doi:10.1001/jamapediatrics.2017.4173
21. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review: Social media and internet-based data for public health surveillance. *Milbank Q*. 2014;92(1):7-33. doi:10.1111/1468-0009.12038
22. Giustini D, Ali SM, Fraser M, Kamel Boulos MN. Effective uses of social media in public health and medicine: a systematic review of systematic reviews. *Online J Public Health Inform*. 2018;10(2):e215. doi:10.5210/ojphi.v10i2.8270
23. MacKinnon KR, Kia H, Lacombe-Duncan A. Examining TikTok’s potential for community-

engaged digital knowledge mobilization with equity-seeking groups. *J Med Internet Res*. 2021;23(12):e30315. doi:10.2196/30315

24. Vassey J, Galimov A, Kennedy CJ, Vogel EA, Unger JB. Frequency of social media use and exposure to tobacco or nicotine-related content in association with E-cigarette use among youth: A cross-sectional and longitudinal survey analysis. *Prev Med Rep*. 2022;30(102055):102055. doi:10.1016/j.pmedr.2022.102055
25. McCausland K, Maycock B, Leaver T, Jancey J. The messages presented in electronic cigarette-related social media promotions and discussion: Scoping review. J Med Internet Res. 2019;21(2):e11953. doi:10.2196/11953
26. Pokhrel P, Herzog TA, Fagan P, Unger JB, Stacy AW. E-cigarette advertising exposure, explicit and implicit harm perceptions, and E-cigarette use susceptibility among nonsmoking young adults. *Nicotine Tob Res*. 2019;21(1):127-131. doi:10.1093/ntr/nty030
27. Wang L, Chen J, Ho SY, Leung LT, Wang MP, Lam TH. Exposure to e-cigarette advertising, attitudes, and use susceptibility in adolescents who had never used e-cigarettes or cigarettes. *BMC Public Health*. 2020;20(1):1349. doi:10.1186/s12889-020-09422-w
28. Cao DJ, Aldy K, Hsu S, et al. Review of health consequences of electronic cigarettes and the outbreak of electronic cigarette, or vaping, product use-associated lung injury. *J Med Toxicol*. 2020;16(3):295-310. doi:10.1007/s13181-020-00772-w
29. Bataineh BS, Wilkinson AV, Sumbe A, et al. The association between tobacco and cannabis use and the age of onset of depression and anxiety symptoms: Among adolescents and young adults. *Nicotine Tob Res*. 2023;25(8):1455-1464. doi:10.1093/ntr/ntad058
30. Sun T, Lim CCW, Stjepanović D, et al. Has increased youth e-cigarette use in the USA, between 2014 and 2020, changed conventional smoking behaviors, future intentions to smoke and perceived smoking harms? *Addict Behav*. 2021;123(107073):107073. doi:10.1016/j.addbeh.2021.107073

31. Community guidelines. TikTok. Accessed March 20, 2024. https://www.tiktok.com/community-guidelines?lang=en

32. Rutherford BN, Sun T, Lim CCW, et al. Changes in viewer engagement and accessibility of popular vaping videos on TikTok: A 12-month prospective study. *Int J Environ Res Public Health*. 2022;19(3):1141. doi:10.3390/ijerph19031141

33. Bornstein RF. Exposure and affect: Overview and meta-analysis of research, 1968-1987. *Psychol Bull*. 1989;106(2):265-289. doi:10.1037//0033-2909.106.2.265

34. Kim H, Davis AH, Dohack JL, Clark PI. E-cigarettes use behavior and experience of adults: Qualitative research findings to inform E-cigarette use measure development. *Nicotine Tob Res*. 2017;19(2):190-196. doi:10.1093/ntr/ntw175

35. Purushothaman V, McMann T, Nali M, Li Z, Cuomo R, Mackey TK. Content analysis of nicotine poisoning (Nic sick) videos on TikTok: Retrospective observational infodemiology study. *J Med Internet Res*. 2022;24(3):e34050. doi:10.2196/34050

36. Cole-Lewis H, Pugatch J, Sanders A, et al. Social listening: A content analysis of E-cigarette discussions on Twitter. *J Med Internet Res*. 2015;17(10):e243. doi:10.2196/jmir.4969

37. Kim Y, Huang J, Emery S. Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *J Med Internet Res*. 2016;18(2):e41. doi:10.2196/jmir.4738

38. Zamith R, Lewis SC. Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*. 2015;659(1):307-318.

39. Conway M, Hu M, Chapman WW. Recent advances in using Natural Language Processing to address public health research questions using Social Media and ConsumerGenerated data. *Yearb Med Inform*. 2019;28(1):208-217. doi:10.1055/s-0039-1677918

40. Elbattah M, Arnaud É, Gignon M, Dequen G. The Role of Text Analytics in Healthcare: A

Review of Recent Developments and Applications. *Healthinf*. Published online 2021:825-832.

41. Glaz L, Haralambous A, Kim-Dufor Y, et al. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*. 2021;23(5).
42. Sun S, Liu Z, Zhai Y, Wang F. COVID-19 vaccines on TikTok: A big-data analysis of entangled discourses. *Int J Environ Res Public Health*. 2022;19(20):13287. doi:10.3390/ijerph192013287
43. Bharti U, Bajaj D, Batra H, Lalit S, Lalit S, Gangwani A. Medbot: Conversational artificial intelligence powered chatbot for delivering Tele-health after COVID-19. In: *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE; 2020.
44. Murthy D, Ouellette RR, Anand T, et al. Using computer vision to detect e-cigarette content in TikTok videos. *Nicotine and Tobacco Research*. 2024;26(Supplement_1):S36-S42.
45. Vassey J, Kennedy CJ, Chang HCH, Smith AS, Unger JB. Scalable surveillance of e-cigarette products on Instagram and TikTok using computer vision. *Nicotine Tob Res*. Published online 2023. doi:10.1093/ntr/ntad224
46. Holm EA, Cohn R, Gao N, et al. Overview: Computer vision and machine learning for microstructural characterization and analysis. *Metall Mater Trans A*. 2020;51(12):5985-5999. doi:10.1007/s11661-020-06008-4
47. Szeliski R. *Algorithms and Applications. Computer Vision*. Springer-Verlag New York, Inc; 2010.
48. Xu S, Wang J, Shou W, Ngo T, Sadick AM, Wang X. Computer vision techniques in construction: A critical review. *Arch Comput Methods Eng*. 2021;28(5):3383-3397. doi:10.1007/s11831-020-09504-3
49. Forsyth, D. A., & Ponce, J. *Computer Vision: A Modern Approach*. prentice hall professional

technical reference.; 2002.

50. Xu S, Wang J, Shou W, Ngo T, Sadick AM, Wang X. Computer vision techniques in construction: A critical review. *Arch Comput Methods Eng*. 2021;28(5):3383-3397. doi:10.1007/s11831-020-09504-3

51. Forsyth, D. A., & Ponce, J. *Computer Vision: A Modern Approach*. prentice hall professional technical reference.; 2002.

52. Holm EA, Cohn R, Gao N, et al. Overview: Computer vision and machine learning for microstructural characterization and analysis. *Metall Mater Trans A*. 2020;51(12):5985-5999. doi:10.1007/s11661-020-06008-4

53. Han K, Vedaldi A, Zisserman A. Learning to discover novel visual categories via deep transfer clustering. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2019.

54. Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*. 2008;40(2):1-60.

55. Blinn CE. *Increasing the Precision of Forest Area Estimates through Improved Sampling for Nearest Neighbor Satellite Image Classification (Doctoral Dissertation)*. 2005.

56. Liu C, Li M, Zhang Y, Han S, Zhu Y. An enhanced rock mineral recognition method integrating a deep learning model and clustering algorithm. *Minerals (Basel)*. 2019;9(9):516. doi:10.3390/min9090516

57. Galassi, A., Lippi, M., Torroni, P. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291-4308. 2020.

58. Cheon M, Yoon SJ, Kang B, Lee J. Perceptual image quality assessment with transformers. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE; 2021.

59. Murthy VSVS, Vamsidhar E, Kumar JS, Rao PS. Content based image retrieval using

Hierarchical and K-means clustering techniques. *International Journal of Engineering Science and Technology*. 2010;2(3):209-212.

60. Murthy D, Lee J, Dashtian H, Kong G. Influence of user profile attributes on e-cigarette-related searches on YouTube: Machine learning clustering and classification. *JMIR Infodemiology*. 2023;3:e42218. doi:10.2196/42218
61. Ketonen V, Malik A. Characterizing vaping posts on Instagram by using unsupervised machine learning. *Int J Med Inform*. 2020;141(104223):104223. doi:10.1016/j.ijmedinf.2020.104223
62. Vassey J, Metayer C, Kennedy CJ, Whitehead TP. # Vape: Measuring e-cigarette influence on Instagram with deep learning and text analysis. *Frontiers in communication*. 2020;4.
63. Setiawan W, Purnama A. Tobacco leaf images clustering using DarkNet19 and K-means. In: *2020 6th Information Technology International Seminar (ITIS)*. IEEE; 2020.
64. Lavack AM. De-normalization of tobacco in Canada. *Soc Mar Q*. 1999;5(3):82-85. doi:10.1080/15245004.1999.9961068
65. Jancey J, Leaver T, Wolf K, et al. Promotion of E-cigarettes on TikTok and regulatory considerations. *Int J Environ Res Public Health*. 2023;20(10). doi:10.3390/ijerph20105761
66. zeeschuimer: A browser extension to collect social media data. Accessed June 30, 2023. https://github.com/digitalmethodsinitiative/zeeschuimer
67. Navarro C, Peres-Neto L. Hair for Freedom" Movement in Iran: Interreligious Dialogue in Social Media Activism? *Religions*. 2023;14.
68. Rozaki E. *Reading Between the Likes: The Influence of BookTok on Reading Culture*. Master's thesis. Utrecht University; 2023.
69. Freelon D. *Pyktok: A Simple Module to Collect Video, Text, and Metadata from Tiktok*. Accessed June 30, 2023. https://github.com/dfreelon/pyktok
70. NLTK :: Natural Language Toolkit. Nltk.org. Accessed July 1, 2023. https://www.nltk.org/

71. Langdetect. PyPI. Accessed July 1, 2023. https://pypi.org/project/langdetect/
72. Jones KS. A statistical interpretation of term specificity and its application in retrieval (1972). In: *Ideas That Created the Future*. The MIT Press; 2021:339-348.
73. Et-taleby A, Boussetta M, Benslimane M. Faults detection for photovoltaic field based on K-means, elbow, and average silhouette techniques through the segmentation of a thermal image. *Int J Photoenergy*. 2020;2020:1-7. doi:10.1155/2020/6617597
74. imagededup: 😎 Finding duplicate images made easy! Accessed October 1, 2023. https://github.com/idealo/imagededup
75. Carlini N, Jagielski M, Zhang C, Papernot N, Terzis A, Tramer F. The Privacy Onion Effect: Memorization is Relative. *Advances in Neural Information Processing Systems, 35, 13263-13276*. 2022.
76. Nagaraj Rao V, Korolova A. Discrimination through Image Selection by Job Advertisers on Facebook. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ; 2023:1772-1788.
77. Tao Z, Liu H, Fu H, Fu Y. Image cosegmentation via saliency-Guided Constrained Clustering with cosine similarity. *Proc Conf AAAI Artif Intell*. 2017;31(1). doi:10.1609/aaai.v31i1.11203
78. Alelyani S, Tang J, Liu H. Feature selection for clustering: A review. *Data clustering*. Published online 2018:29-60.
79. Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput*. 2017;29(9):2352-2449. doi:10.1162/NECO_a_00990
80. Pandya B, Cosma G, Alani AA, Taherkhani A, Bharadi V, McGinnity TM. Fingerprint classification using a deep convolutional neural network. In: *2018 4th International Conference on Information Management (ICIM)*. IEEE; 2018.
81. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems, 32*. Published online

2019.

82. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv [csCV]*. Published online 2014. http://arxiv.org/abs/1409.1556
83. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016.
84. Shaha M, Pawar M. Transfer Learning for Image Classification. In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE; 2018.
85. Yuan J, Fan Y, Lv X, et al. Research on the practical classification and privacy protection of CT images of parotid tumors based on ResNet50 model. *J Phys Conf Ser*. 2020;1576(1):012040. doi:10.1088/1742-6596/1576/1/012040
86. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems, 30*. Published online 2017.
87. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv [csCV]*. Published online 2020. http://arxiv.org/abs/2010.11929
88. Maurício J, Domingues I, Bernardino J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*. 2023;13(9).
89. Park S, Kim G, Oh Y, et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *Med Image Anal*. 2022;75(102299):102299. doi:10.1016/j.media.2021.102299
90. Tsai PC, Lee TH, Kuo KC, et al. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat Commun*. 2023;14(1):2102. doi:10.1038/s41467-023-37179-4

91. 🤗 Transformers. Huggingface.co. Accessed October 5, 2023. https://huggingface.co/docs/transformers/index

92. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv [csCL]*. Published online 2019. http://arxiv.org/abs/1910.03771

93. Yuan J, Barmpoutis P, Stathaki T. Effectiveness of Vision Transformer for Fast and Accurate Single-Stage Pedestrian Detection. *Advances in Neural Information Processing Systems*. 2022;35:27427-27440.

94. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv [csLG]*. Published online 2014. http://arxiv.org/abs/1412.6980

95. 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;17(7):847-854. doi:10.1093/ntr/ntu257

96. Zare S, Nemati M, Zheng Y. A systematic review of consumer preference for e-cigarette attributes: Flavor, nicotine strength, and type. *PLoS One*. 2018;13(3):e0194145. doi:10.1371/journal.pone.0194145

97. Glasman LR, Albarracín D. Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation. *Psychol Bull*. 2006;132(5):778-822. doi:10.1037/0033-2909.132.5.778

98. Vassey J, Valente T, Barker J, et al. E-cigarette brands and social media influencers on Instagram: a social network analysis. *Tob Control*. 2023;32(e2):e184-e191. doi:10.1136/tobaccocontrol-2021-057053

99. Snell LM, Nicksic N, Panteli D, et al. Emerging electronic cigarette policies in European member states, Canada, and the United States. *Health Policy*. 2021;125(4):425-435. doi:10.1016/j.healthpol.2021.02.003

100. Oconnell M, Kephart L. Local and state policy action taken in the United States to

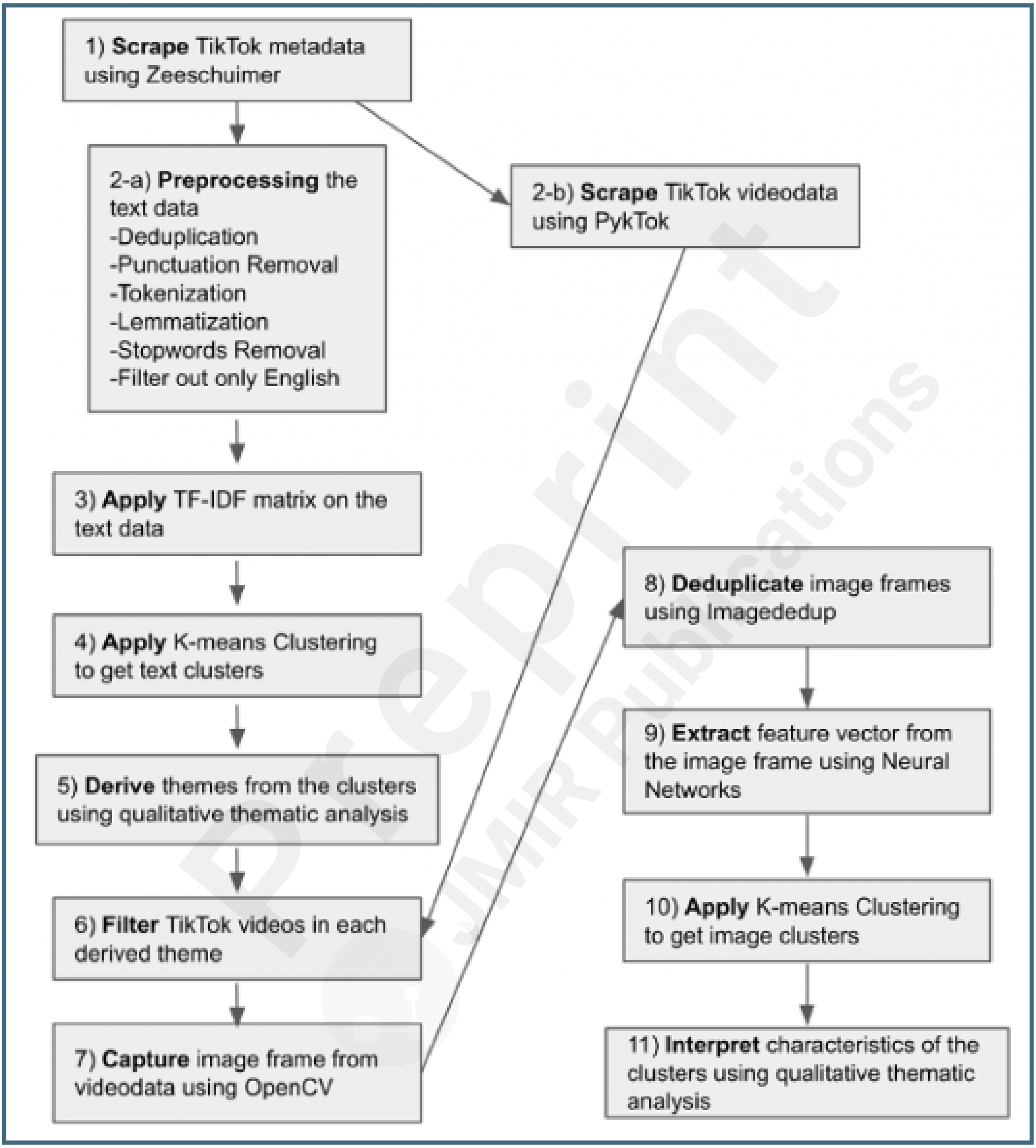
address the emergence of e-cigarettes and Vaping: a scoping review of literature. *Health Promotion Practice*. 2022;23(1):51-63.

101. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv [csCL]*. Published online 2018. http://arxiv.org/abs/1810.04805

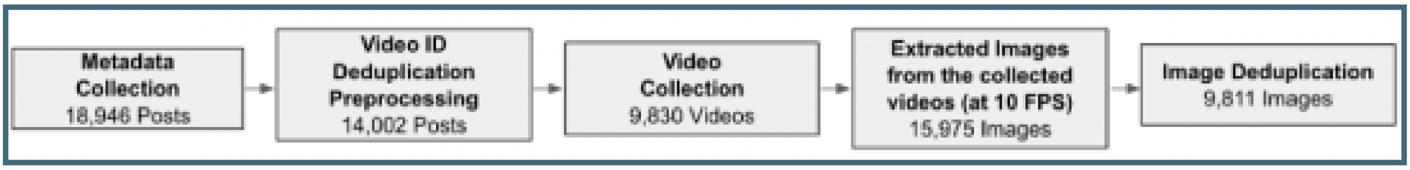
Supplementary Files

Figures

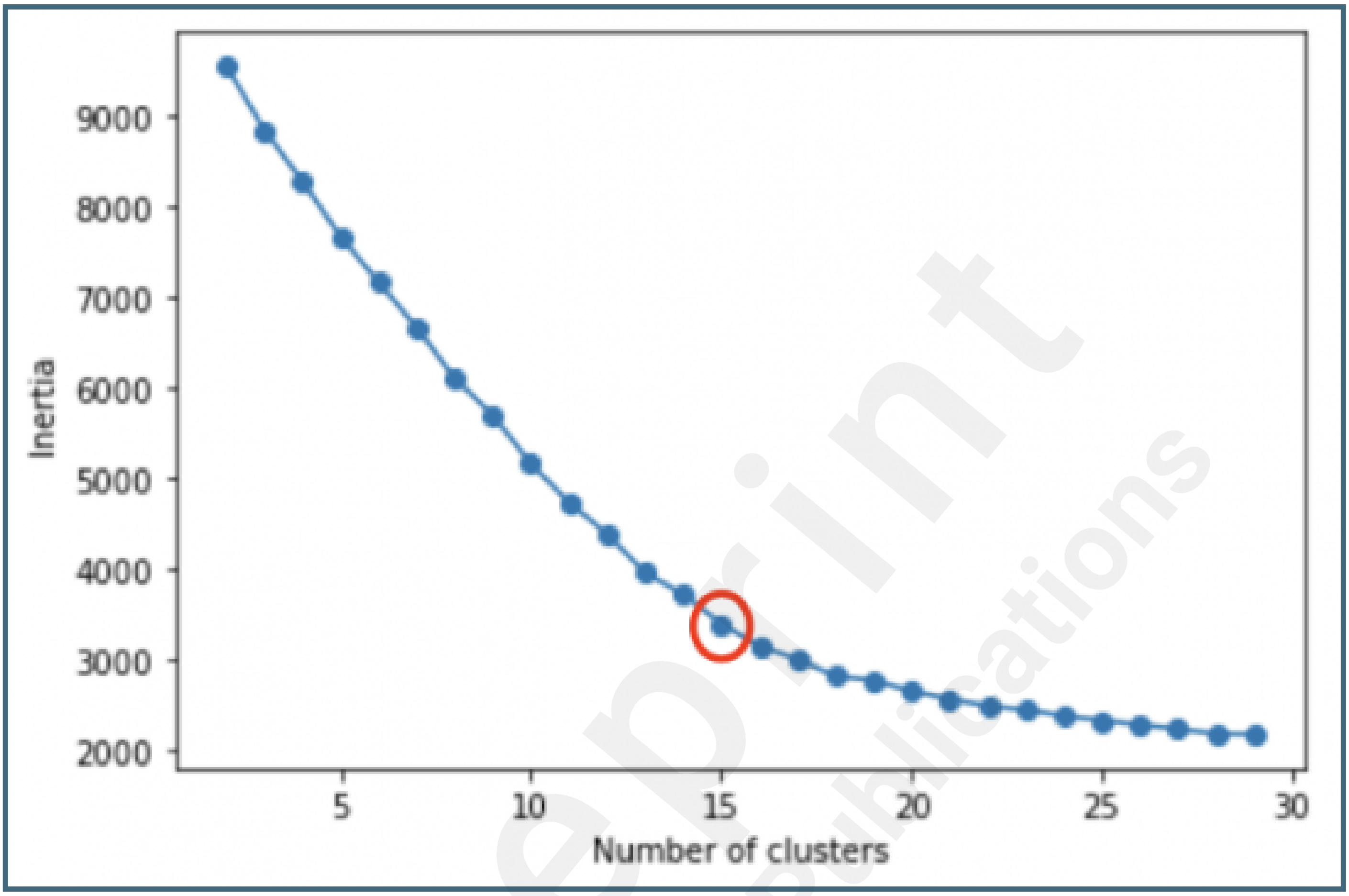
Project architecture; relevant figures and tables are referenced in parentheses.


1) Scrape TikTok metadata using Zeeschuimer
2-a) Preprocessing the text data
-Deduplication
-Punctuation Removal
-Tokenization
-Lemmatization
-Stopwords Removal
-Filter out only English
2-b) Scrape TikTok videodata using PykTok
3) Apply TF-IDF matrix on the text data
4) Apply K-means Clustering to get text clusters
5) Derive themes from the clusters using qualitative thematic analysis
6) Filter TikTok videos in each derived theme
7) Capture image frame from videodata using OpenCV
8) Deduplicate image frames using Imagededup
9) Extract feature vector from the image frame using Neural Networks
10) Apply K-means Clustering to get image clusters
11) Interpret characteristics of the clusters using qualitative thematic analysis

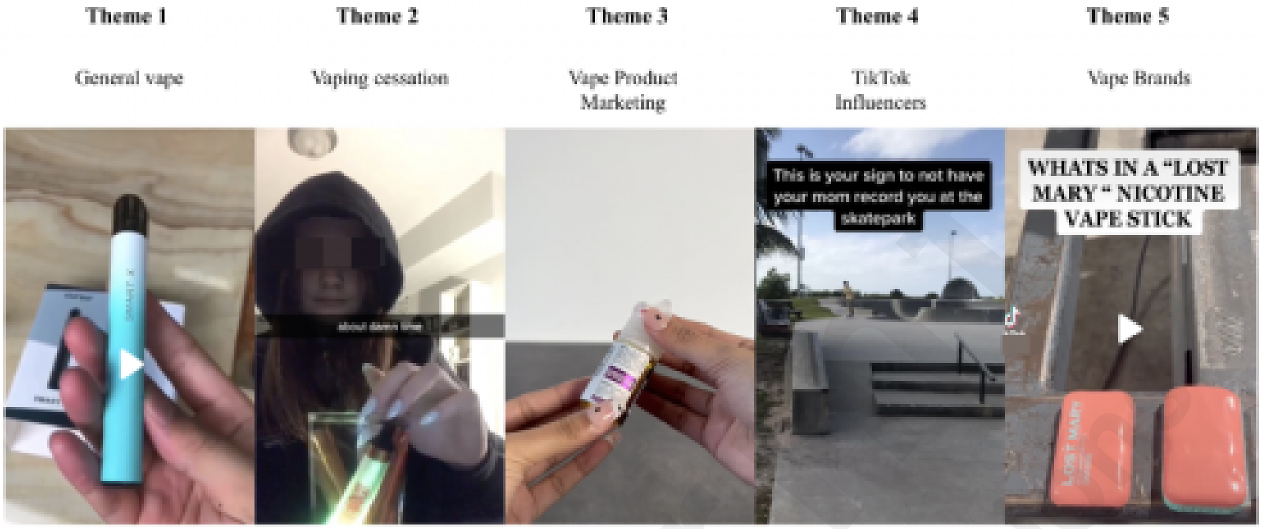
Primary dataset evolution for the analysis; a flowchart of a data processing pipeline, starting with metadata collection, followed by video ID deduplication, extraction of images and finally image deduplication.



Elbow Method Result from K-means Text Clustering indicating that the optimal number of clusters is 15.



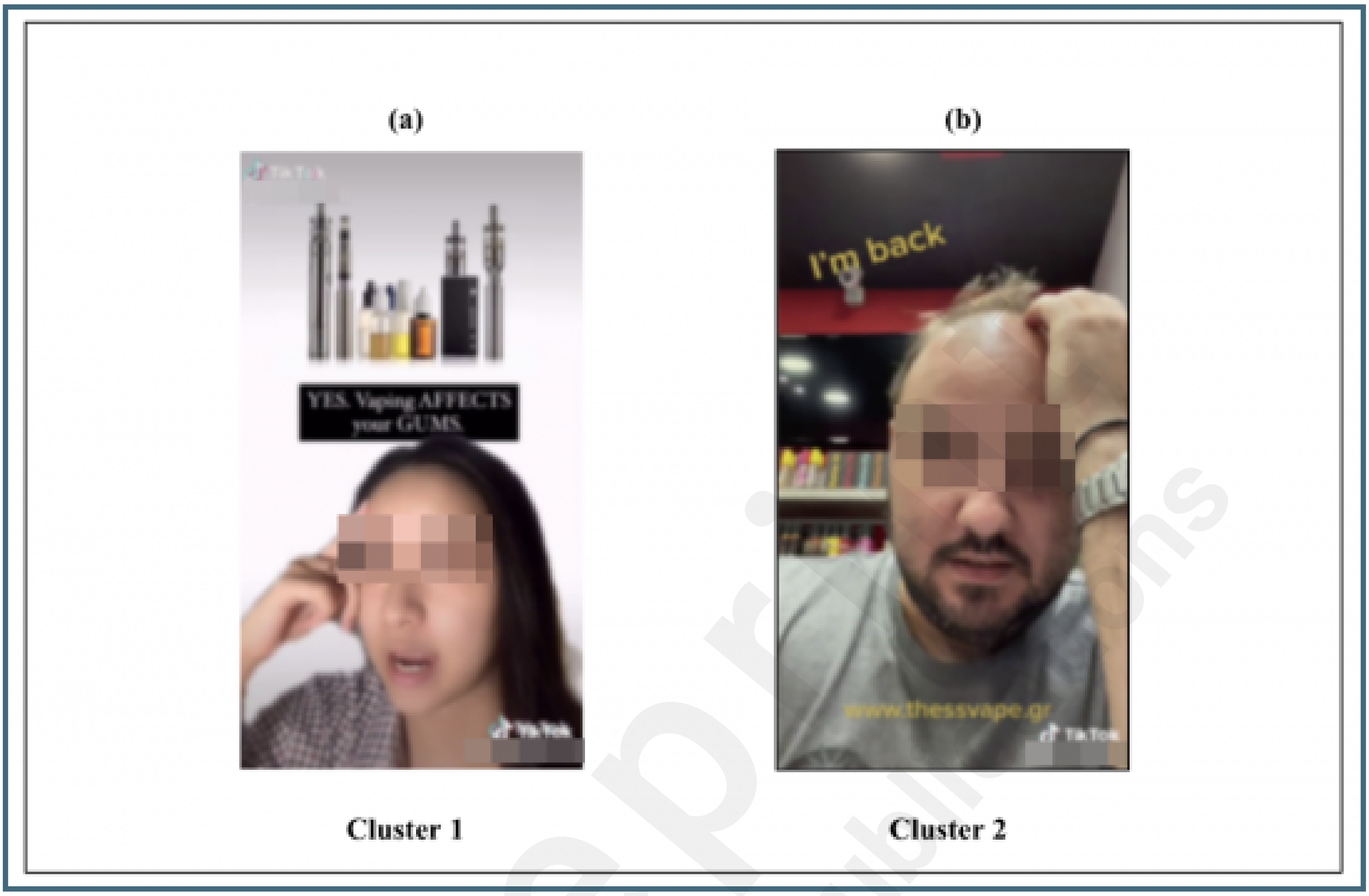
An example of themes derived from text clustering; selected representative images from actual posts corresponding to each theme.



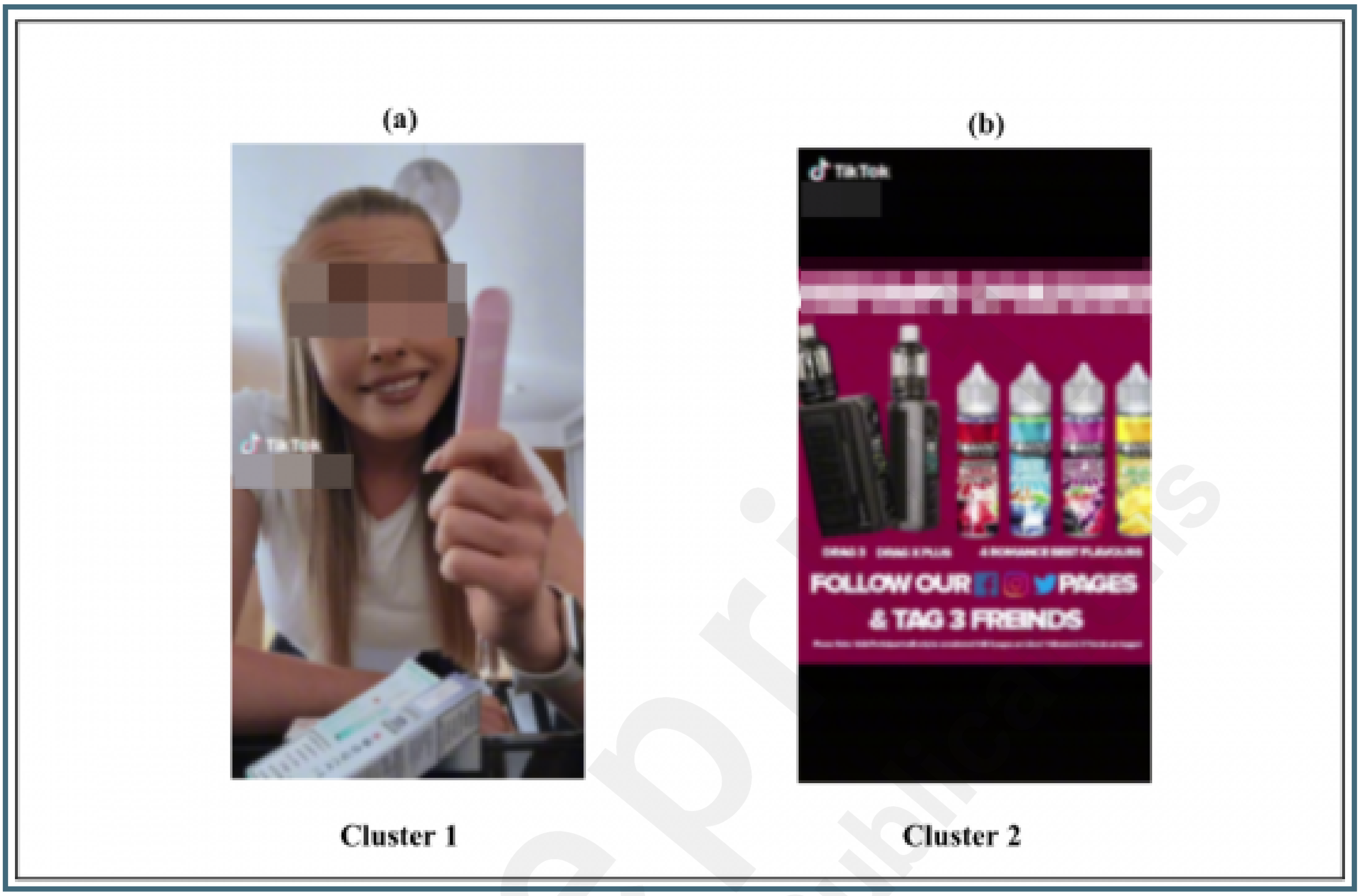
Sample image clustering results from Theme 1 (General Vape); three distinct sub-clusters were formed.

(a)
(b)
(c)
TikTok
"you called me a wh0re, made fun of my weight, you gave me the finger..."
MY HEALTH FINE
Cluster 1
Cluster 2
Cluster 3

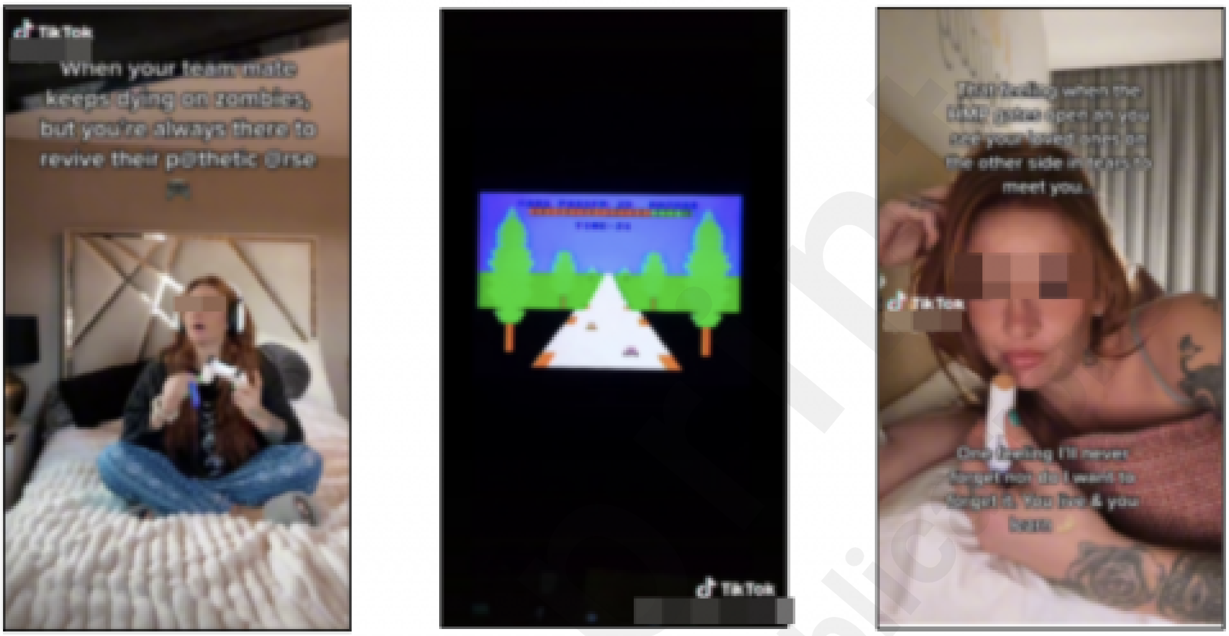
Sample image clustering results from Theme 2 (Vape Cessation); two sub-clusters were formed.



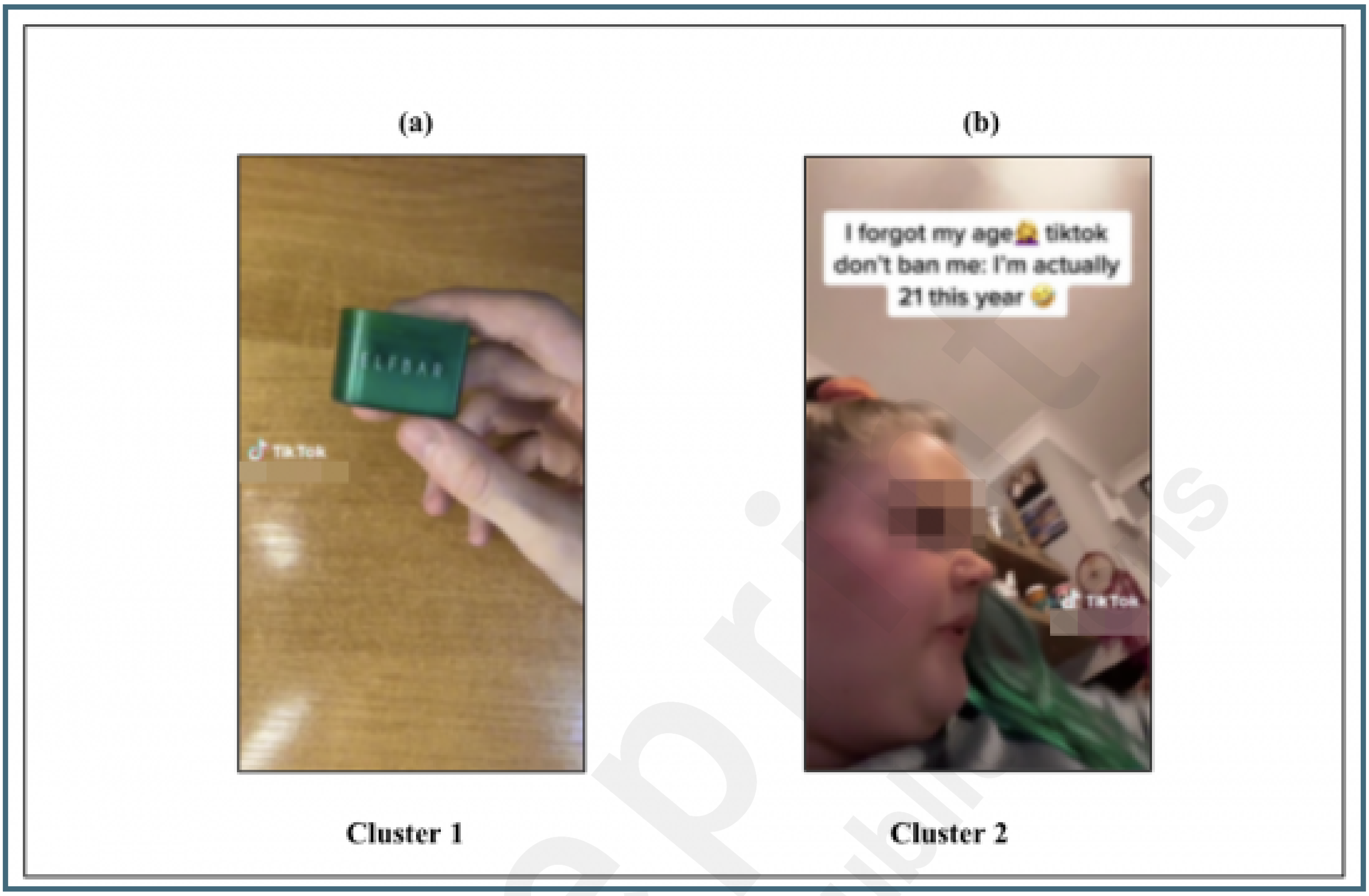
Sample image clustering results from Theme 3 (Vaping Marketing); two sub-clusters were formed.



Sample image clustering results from Theme 4 (TikTok Influencers); no differences were found between images within each cluster.


TikTok
[redacted]
When your team mate keeps dying on zombies, but you're always there to revive their p@thetic @rse
TikTok
[redacted]
That feeling when the [illegible] gates open an you see your loved ones on the other side in tears to meet you.
TikTok
[redacted]
One feeling I'll never forget nor do I want to forget it. You live & you learn

Sample image clustering results from Theme 5 (Vape Brands); two sub-clusters were formed.



I forgot my age🤦‍♀️ tiktok don't ban me: I'm actually 21 this year 🤪

Multimedia Appendixes

List of TikTok hashtags. We scraped the data with 50 hashtags.
URL: http://asset.jmir.pub/assets/e68c42900df816d79bce7b3ff1eb1f50.docx

Text clustering result. We identified 15 clusters from K-Means clustering method.
URL: http://asset.jmir.pub/assets/9a5b77cf8c0eb04dc2b2b64af78705ba.docx