Ketonen, Vili; Malik, Aqdas

Characterizing vaping posts on Instagram by using unsupervised machine learning

Published in:
International Journal of Medical Informatics

DOI:
10.1016/j.ijmedinf.2020.104223

Published: 01/09/2020

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY-NC-ND

Please cite the original version:
Characterizing vaping posts on Instagram by using unsupervised machine learning

Vili Ketonen, Aqdas Malik*
Department of Computer Science, Aalto University, Konemiehentie 2, 02150, Espoo, Finland

ARTICLE INFO

Keywords:
e-Cigarettes
Electronic cigarettes
Social media
Instagram
Photos
Machine-learning
Adolescents
Young adults

ABSTRACT

Electronic cigarettes (e-cigarettes) usage has surged substantially across the globe, particularly among adolescents and young adults. The ever-increasing prevalence of social media makes it highly convenient to access and engage with content on numerous substances, including e-cigarettes. A comprehensive dataset of 560,414 image posts with a mention of #vaping (shared from 1 June 2019 to 31 October 2019) was retrieved by using the Instagram application-programming interface. Deep neural networks were used to extract image features on which unsupervised machine-learning methods were leveraged to cluster and subsequently categorize the images. Descriptive analysis of associated metadata was further conducted to assess the influence of different entities and the use of hashtags within different categories. Seven distinct categories of vaping related images were identified. A majority of the images (40.4%) depicted e-liquids, followed by e-cigarettes (15.4%). Around one-tenth (9.9%) of the dataset consisted of photos with person(s). Considering the number of likes and comments, images portraying person(s) gained the highest engagement. In almost every category, business accounts shared more posts on average compared to the individual accounts. The findings illustrate the high degree of e-cigarettes promotion on a social platform prevalent among youth. Regulatory authorities should enforce policies to restrict product promotion in youth-targeted social media, as well as require measures to prevent underage users’ access to this content. Furthermore, a stronger presence of anti-tobacco portrayals on Instagram by public health agencies and anti-tobacco campaigners is needed.

1. Introduction

Despite a significant decline in the adoption of smoked and smokeless tobacco, the last decade has witnessed an extraordinary evolution concerning the emergence of novel tobacco products. In particular, the range and use of e-cigarettes have sharply escalated, particularly among adolescents and young adults within the USA as well as other countries. According to the annual school-based cross-sectional National Youth Tobacco Survey (NYTS), e-cigarettes has now become the most prevalent tobacco product among high school and middle school students. An e-cigarette is a battery-operated device combined with a liquid solution, known as e-liquid or e-juice, which contains nicotine and other chemicals. By pressing a button or by inhaling in an electronic device, the battery powers up an element called an atomizer, which consists of a small heating element or coil that vaporizes the e-liquid solution.

Numerous scientific studies and news reports have presented growing concerns about the increasing popularity of e-cigarettes usage (termed as vaping) among adolescents and young adults [1–3]. Although e-cigarettes are often marketed as smoking cessation products, and a safer alternative to combustible cigarettes [4,5], the effectiveness of these claims is neither endorsed by the FDA or supported by recent scholarly investigations [6–8]. Multiple studies have also associated e-cigarettes use with the succeeding use of traditional smoking, alcohol, marijuana, and other drugs [1,9,10]. Even though research on the long-term health implications of e-cigarettes compared to traditional smoking products is still infancy, the latest evidence highlights the potential risks from flavorings, nicotine, and numerous toxicants triggered during the vaping activity. Moreover, several recent inquiries on the health impacts associated with vaping indicate the use among adolescents and young adults can potentially harm their brain development, and various constituents can induce transient lung inflammation, and increase blood pressure, heart rate, as well as platelet activation [11–13].

In recent years, the content generated on various social media outlets has been increasingly used for social intelligence and health surveillance. Data gained from these platforms provide novel insights that can support and inform regulatory policies, enforcement effort, as...
well as direct future research on numerous aspects of public health, including vaping. Recent studies on social media commentary about tobacco and other substances highlight the ever-influencing role of these platforms, in particular among youth. For instance, information about vaping spreads heavily through online (as well as offline) social networks through a family member, friends, or by following celebrities [14,15]. Beyond exposing adolescents and young adults to numerous campaigns, promotions, and marketing vaping products and related accessories, social media channels act as a gateway to learn, experiment, use, and experience sharing about the vaping activity [14,16,17]. Furthermore, postings on these platforms tend to portray and evoke a highly positive image of e-cigarettes in general in the form of youth culture, luxurious lifestyle, sex appeal, and feelings of freedom and relaxation that can ultimately lead to social acceptability among the vulnerable population [2,5]. Seminal scholarly work further highlights active streams of content posted and followed by adolescents and young adults in the form of commentary, images, and videos on different social media platforms, including Twitter [17,18], Instagram [2,3,19] [2,318], and Reddit [20].

In the US, Instagram is one of the leading social media platforms among youth [21]. Moreover, with over 1 billion monthly active users, Instagram is among the most popular social networks worldwide [22]. According to a recent report on digital technologies and platforms, 36.2% of the global Instagram users worldwide are 24 years old or younger, and more than two thirds (71.2%) of the users are under 35-year-olds [23]. From the research perspective, Instagram’s image-centric platform provides a valuable data source to probe across various research topics and disciplines. Similarly, photos are widely present in many other social media platforms as well, in various forms, such as profile images, memes, selfies, and affective imaging, which makes them imperative to investigate. The use of visual content on these social spaces exceeds plain archiving or informative purposes as they fulfill many other gratifications, including status-seeking, attention-seeking, and self-presentation [24,25]. Prevalence, engagement, and perceived realism of visual content compared to other forms of modalities in computer-mediated platforms (in particular, social media) also make them extremely relevant and worth studying. Because of the mass adoption of handheld devices across the globe, advancements in camera technology, and digital ubiquity, pictures have made their way as a preferred mode of communication [26]. From the engagement perspective, studies have shown that compared to other modalities, pictures obtain far more engagement indices such as number of likes, comments, and shares [27]. Moreover, “realism heuristic” associated with visuals invokes higher credibility and trust in humans due to higher affinity with the real world. Finally, plain and apparent depictions that require minimal decoding makes visual content more appealing than text and other modalities [28]. Given the prominence and engagement associated with pictures on various social media platforms [24,29], in particular Instagram among youth [21,22,25], there is a decisive need to further explore and understand this form of content in conjunction with vaping.

By deploying a large data dataset, the current research aims to (a) identify the main categories of vaping-related images shared on Instagram by leveraging deep neural networks and unsupervised machine learning methods, and (b) explore the statistical differences between the content created by business users and individuals in these categories. Furthermore, this study seeks to determine the type of content created by business users and individuals in these categories. The chosen methodologies and subsequent findings from the current study may guide researchers in effectively identifying and characterizing critical discourse and visual content associated with vaping on other computer-mediated platforms. Likewise, the study findings may aid various public health entities and anti-tobacco campaigners in developing suitable content to counter e-cigarettes marketing, associated misconceptions, and discourage its usage among the vulnerable cohorts of the society.

2. Related Work

Most of the prior work analyzing smoking-related content on Instagram has focused on a small subset of image data or solely inspected the textual content of the posts. The results derived from a limited sample size in the prior studies may provide an insufficient representation of Instagram content specifically related to vaping, such as the relative proportion of the most common image types, which may be assessed reliably through large-scale investigations. It is also worth mentioning that most of the prior studies that used applied machine-learning techniques have focused so far on textual content within the context of Twitter.

2.1. Small-scale content analysis

Laestadius and colleagues performed a qualitative content analysis on 85 posts (43 #ecig, 42 #vape) on Instagram by manually inspecting images, text, and hashtags of each post [3]. In addition, they monitored the increasing volume of posts tagged with the eight most popular hashtags related to e-cigarettes and vaping from March 2014 to October 2015. Their content analysis discovered that the business users created over half of the posts (58.8%), and none of the posts exhibited negative representation of e-cigarettes. The results also showed that the posts did not include any leading brands, and half of the posts described or depicted mod style e-cigarettes. Another exploratory study used manual content analysis of 1800 posts related to the nine most popular search terms of e-cigarettes from Instagram and Pinterest [30]. The captured images were classified into 11 categories using the hand-coding technique. In their study, the three most popular categories for the Instagram images were e-cigarette related marketing (60%), e-cigarette customization (38%), and e-liquids/ flavors (18%).

2.2. Large-scale content analysis utilizing machine learning

Within the context of Instagram, a handful of studies have applied machine-learning techniques on large-scale image analysis. One of the recent study used deep learning image classification to classify 49,655 Instagram image posts, separating men, women, different vaping device types, and e-liquids [31]. They discovered that in the year 2019, 40% of the vaping-related images depicted e-juices, 30% of the images featured different e-cigarette types, and 13% of the images depicted humans. Another study presented a method to automatically identify hookahs (water pipes) on Instagram by leveraging feature extraction using a convolutional neural network (CNN) and classification using a support vector machine (SVM) [32]. Czaplicki et al. analyzed 14,838 Instagram posts related to Juul, one of the most novel and popular vaping product produced by Juul Labs. They used a combination of machine learning methods, keyword algorithms, and human coding to categorize the textual content of the captured posts [2]. Their analysis showed that over one-third of the posts were promotional, and half of the posts were related to youth or lifestyle.

Most of the previous studies utilizing machine learning have analyzed e-cigarettes related content on Twitter and Reddit. These platforms provide easy access to their data using public APIs (Application Programming Interface) compared to Instagram, where public content cannot be accessed as efficiently on a large scale. These investigations have primarily focused on analyzing the textual content of the posts. Prior work studying e-cigarette related content on Twitter has displayed consistent results where most of the tweets expressed positive sentiment towards e-cigarettes [17, 18, 33]. A recent study by Kavuluru et al. used
supervised machine learning to identify e-cigarette proponents on Twitter [34]. The model used various textual features, such as profile biography and recent tweets. Their results showed that e-cigarette promoters tweeted at least ten times more frequently than other Twitter users. Another study performed a qualitative content analysis on 364 public posts on Reddit related to topics around Juul and youth [20]. Results from the study showed that a large portion (41.1%) of the Juul-related posts presented a mixed polarity of positive and negative sentiment. Overall, 60% of the posts showed negative perceptions of youth use, while only 37% of posts by youth showed negative perceptions of youth use.

3. Materials and methods

3.1. Data collection

The current study collected the public image posts and their related metadata from Instagram tagged with the hashtag #vaping. We decided to use this specific hashtag since the hashtag is very generic and, therefore, often used with vaping-related images. Consequently, using this hashtag allowed us to collect a large and diverse set of vaping-related images to discover the core characteristics of vaping-related images shared on the platform. In total, 560,414 images posted by 45,232 unique users were retrieved from 1 June 2019 to 31 October 2019. We obtained the data through the public search functionality on the Instagram website, which supports searching for public posts for a given hashtag or username. To automate the large scale data collection from Instagram, we followed methods undertaken by several recent studies [35,36]. Similar to these studies, our data collection procedure leveraged Selenium1 and Instagram Scraper2 to collect the public image posts as well as related metadata from the related users’ profile pages. The metadata collected from the Instagram profile page directly provided us with the type of the account, stating whether the type is a business account or personal. In addition, when the user had specified a link to an external website address on their profile, such as their blog or a business website, we gathered this information to separate different profile types. From the business profiles, we collected the user-specified business category of the account that was appended in the Instagram profile page metadata. The profile related details were used to support the analysis by exploring the differences between posts created by these distinct entities. Out of the 45,232 unique posters within the dataset, the profile data of 44,977 (99.4%) of the unique posters was retrieved since the remaining accounts were either deleted or had regional viewing restrictions.

Although the collected profile information directly specifies whether the user has set the account type as a business account or not, further inspection of the collected data revealed that a noticeable number of e-cigarette related businesses have, nevertheless, specified their Instagram account type as a non-business account. Thus, to better separate the business users and individual users, this study makes an explicit definition between a personal account and a business account. First, we define a personal account as an account that has its type marked as a non-business account. However, if the user has also specified an external web address in their profile and it does not link to a public website, such as YouTube or WordPress, we define the account as a business account, since a non-generic URL indicates a business-related website. Additionally, we consider users with the business category set as ‘Creators & Celebrities’ as personal users, since this category is often used by the individual social influencers.

3.2. Feature extraction and K-Means clustering

We leveraged an image classification approach utilizing unsupervised machine learning techniques to cluster the images into homogenous subgroups. The core principle behind clustering algorithms is that nearby data points in the feature space are similar, which allows partitioning the data points into disjoint groups [37]. A pre-trained CNN is used to extract features from the images, and then similar images are clustered into different subgroups using the k-means clustering algorithm. As an initial step, a subset of 50,000 random images were selected from the collected dataset because a subset of the dataset is enough to discover the main underlying groups of the images. The complete dataset was later classified into different categories using the clusters found with this smaller subset. Afterward, we removed the duplicate images (2314) using a perceptual hashing algorithm called dHash.

To extract the features from the images, we experimented with two different popular CNN model architectures, pre-trained on the large ImageNet dataset consisting of over 14 million images and 1000 labels [38]. The underlying idea of using a pre-trained CNN for feature extraction lies under the assumption that the first few layers of a CNN learn to recognize primitive image elements, such as borders, corners, basic shapes, and colors. These elementary elements are present in any image, enabling us to use the pre-trained neural network to extract distinctive features from any image.

The models that this study evaluated were ResNet [39] (50-layer and 152-layer variations), and VGG [40] (16-layer and 19-layer model). The classifier part, i.e., the last fully connected layers, of each CNN was removed to extract image features instead of performing image classification. Additionally, we also inspected the clustering quality by experimenting using earlier or later layers of the models as the last layer of the feature extractors to see which yielded the best features for the clustering. Before performing the feature extraction, the images were first resized to 224 × 224 since the CNNs were pre-trained on ImageNet with this image size. Then, each image was given as an input to the selected CNNs to extract the image features. After the feature extraction phase, the study continued by clustering the images by applying the k-means algorithm on the extracted features. To evaluate the quality of the clusters resulting from the k-means algorithm, we performed a manual inspection on the ten images closest to each cluster center. This subjective evaluation considered the clustering quality good when the samples within each cluster appeared to be perceivably related to each other, and the samples between different clusters seemed as distinctive as possible from the other clusters.

Two of the models chosen for feature extraction, VGG-16 and VGG-19, produce a particularly high number (20,588) of the extracted features compared to ResNet, which outputs a smaller number of 2048 features for a given input image. This high dimensionality of the extracted features may adversely affect the clustering algorithms by blurring the clusters [41]. In particular, k-means clustering is prone to the “curse of dimensionality” since the algorithm is based on the minimization of the squared Euclidean distance in the feature space of the data points. Therefore, we performed dimensionality reduction with principal component analysis (PCA) on the feature vectors produced by VGG-16 and VGG-19. Our experiments confirmed that reducing the feature space of the feature vectors generated by VGG-models resulted in visually more consistent clusters.

After performing multiple experiments using the subjective visual evaluation criteria specified earlier for each feature extractor, and testing a varying number of clusters from 5 to 12 as an input to the k-means algorithm, the clustering results showed that the image features extracted by VGG models yielded much more perceivably coherent clusters compared to the ResNet model variations. Furthermore, the 1https://www.selenium.dev/ 2https://pypi.org/project/instagram-scraper/ 3https://pypi.org/project/dhash/
experiments showed that using PCA to reduce the dimensionality of the extracted features made the clusters perceivably more consistent. Our experiments achieved the best image clusters by using the VGG-19 model for feature extraction, yielding 20,588 features, followed by PCA reducing the number of features to 160 (explaining approximately 35% of the variance of the data), followed by k-means clustering using \( k = 7 \). Thus, our experiments indicate that the images in the dataset are separable into seven distinct categories (See Fig. 1). The identified categories are (1) e-liquid and (2) e-liquids (single e-liquid bottle or multiple e-liquids), (3) e-cigarette, (4) product package (e-cigarette or e-liquid product packages), (5) persons (one or more persons posing in a photo), (6) statement (images with textual statements, promotions, or discounts), and (7) miscellaneous.

3.3. Classification of the complete dataset

After characterizing the seven main categories in the dataset using image clustering of the randomly selected subset, the complete dataset was classified by training a VGG-19 model using the labeled data from the obtained clusters. During the next step, the class label for each image in the dataset was predicted.

To predict the labels on the complete dataset of over 500,000 images, we selected 500 images closest to the cluster center of each cluster as the labeled training data set, yielding a total of 3500 training samples (i.e. 500 samples from each cluster). At this point, we manually inspected each image to verify that they belong to the correct category to improve the classification accuracy, and removed any images not relating to the reviewed category. After manually verifying and cleaning the training data, we split this labeled data into training (80%) and validation (20%) sets. We initialized the VGG-19 model with the weights from a VGG-19 pre-trained with the ImageNet dataset to achieve better classification performance since the size of our training dataset is limited. The last layer of the model was replaced with a new output layer that predicted the seven different output classes. Only the weights of the output layer were trained with our data samples, while the weights of other layers remained the same.

We used cross-entropy loss (log loss) as the loss criterion during the training. To learn the weights of the output layer, we used Adam optimizer [42] with a learning rate of 0.001 and 16 training epochs. To reduce overfitting, we augmented the training data by performing simple operations on the images, such as rotation and random flipping. Additionally, our model evaluation criteria considered the best performing model as the one achieving the lowest validation loss. The final trained VGG-19 model achieved the validation loss of 0.2575 and validation accuracy of 90.76%. After obtaining the best performing model from the model training phase using the previously stated model evaluation criteria, we used the trained model to predict a class label for each of the 560,414 images in the complete dataset. To predict the class for each image, we applied a softmax function on the non-normalized output of the VGG-19 network to yield a probability distribution over the seven possible labels. Then, the class with the highest probability was chosen as the final predicted label for the given input image.

4. Results

Of the classified image posts (see Table 1), e-liquid related images covered the largest portion (40.4%), E-cigarettes had the second-largest share after the e-liquid posts (15.4%), while different e-cigarette related product packages having the third-largest proportion (10.0%). Approximately every tenth image (9.9%) depicted a person or persons, mostly posing or performing e-cigarette related activities, such as vaping.

Image posts portraying persons gained the highest number of likes (158.8) on average, followed by posts with images of e-cigarettes, with 113 likes on average. Posts with images of product packages received the lowest number of average likes (51.6). Similar to the likes distribution, posts displaying persons also gained the highest number of comments on average (9.3%). Posts with product packages, statements, and promotions attracted the lowest number of comments from the users.

Examination of the creators behind the collected image posts showed that 52% of the users were personal accounts, while 47% were business-related accounts. However, at the category level, each category had significantly more posts made by business users than individuals, except for the category of images portraying people. In almost every category, the image posts by individuals gained more comments than the posts created by business users. One of the plausible reasons behind this novel finding may be that the individuals’ Instagram posts are more reflective than business user’s posts and that most users can personally relate themselves to them. On the contrary, business accounts tend to post business-related content, such as product advertisements and promotions, which may be of lesser interest to the users, consequently gaining low engagement. Due to the strong implications of this finding, this may be an important topic for further research.

Within each category, the most frequent Instagram business category of the business users, by a large margin, was "Personal Goods & General Merchandise Stores", implying high presence of e-cigarette product manufacturers and retailers marketing their products. The second most common business category was "Creators & Celebrities"

![Fig. 1. Three sample images from each identified category.](image-url)
used generally by famous individuals. The hashtags analysis (see Table 2) identified the ten most popular hashtags within each category. The presented data in this analysis omits non-English hashtags as well as the most common hashtags (e.g. #vaping) that were among the top 200 hashtags in at least five out of the seven identified categories. In categories linked to e-cigarette products, several popular hashtags refer directly to well-known e-cigarettes related businesses in the US, such as the hashtags #juul, #vaporesso, and #vandyvape. Many of the common hashtags in the “Statement” category illustrated positive sentiment towards e-cigarette use, such as #ivapeivote, #wevapewevote, and #smokefree. Popular hashtags among e-cigarettes included terms related to different types of devices, such as #squonk, #atomizer, and #vapepod. Within “Person(s)” category, several popular hashtags, such as #model and #vapebabe, refer to appealing women. In the category of “miscellaneous” images, most common hashtags describe coil-based art, where the coil refers to the heating coil present in specific e-cigarettes. These hashtags portray the broad range of e-cigarettes related acts practiced by the vaping enthusiasts.

5. Discussion

5.1. Principal findings

As a highly engaging and prevalent form of content, using social media photos for surveillance can meritoriously supplement and broaden our understanding of various health-related behaviors such as vaping. Our study suggests that a very large portion of vaping-related images shared on Instagram depict e-liquids or e-juices. Likewise, a highly active stream of images on the platform exhibit vaporizers or e-cigarettes. Numerous forms of highly appealing product packaging of vaping products and related accessories were also identified within the analytical dataset. These findings are similar to the previously reported results from both smaller-scale [2,30] and larger-scale [31] studies on Instagram that observed a large portion of the images to be covered by

---

Table 1
Category distribution of the dataset based on the predicted labels by the trained VGG-19.

<table>
<thead>
<tr>
<th>Category</th>
<th>Posts</th>
<th># likes (mean)</th>
<th># comments (mean)</th>
<th># posts per user (mean)</th>
<th>Unique users</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-liquid (multiple)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>29.0 %</td>
<td>75.9</td>
<td>8.4</td>
<td>4.9</td>
<td>44.0%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>71.0 %</td>
<td>55.1</td>
<td>4.1</td>
<td>9.5</td>
<td>56.0%</td>
</tr>
<tr>
<td>Total</td>
<td>140,070 (25.0 %)</td>
<td>61.1</td>
<td>5.3</td>
<td>7.5</td>
<td>18,688</td>
</tr>
<tr>
<td>E-cigarette</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>41.6 %</td>
<td>91.1</td>
<td>8.4</td>
<td>5.5</td>
<td>49.5%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>58.4 %</td>
<td>129.5</td>
<td>7.2</td>
<td>7.6</td>
<td>50.5%</td>
</tr>
<tr>
<td>Total</td>
<td>128,541 (22.9 %)</td>
<td>113.0</td>
<td>7.7</td>
<td>6.6</td>
<td>19,451</td>
</tr>
<tr>
<td>E-liquid (single)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>41.7 %</td>
<td>112.7</td>
<td>6.4</td>
<td>6.5</td>
<td>44.0%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>58.3 %</td>
<td>63.3</td>
<td>3.7</td>
<td>7.2</td>
<td>56.0%</td>
</tr>
<tr>
<td>Total</td>
<td>86,477 (15.4 %)</td>
<td>83.5</td>
<td>8.9</td>
<td>6.9</td>
<td>12,526</td>
</tr>
<tr>
<td>Product package</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>24.3 %</td>
<td>67.7</td>
<td>6.4</td>
<td>3.0</td>
<td>37.7%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>75.7 %</td>
<td>46.6</td>
<td>3.7</td>
<td>5.6</td>
<td>62.3%</td>
</tr>
<tr>
<td>Total</td>
<td>56,076 (10.0 %)</td>
<td>51.6</td>
<td>4.3</td>
<td>4.6</td>
<td>12,078</td>
</tr>
<tr>
<td>Person(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>59.0 %</td>
<td>160.2</td>
<td>10.6</td>
<td>3.6</td>
<td>58.2%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>41.0 %</td>
<td>157.1</td>
<td>7.4</td>
<td>3.5</td>
<td>41.8%</td>
</tr>
<tr>
<td>Total</td>
<td>55,423 (9.9 %)</td>
<td>158.8</td>
<td>9.3</td>
<td>3.6</td>
<td>15,560</td>
</tr>
<tr>
<td>Statement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>28.7 %</td>
<td>64.8</td>
<td>2.8</td>
<td>2.8</td>
<td>36.9%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>71.3 %</td>
<td>59.5</td>
<td>3.8</td>
<td>4.0</td>
<td>63.1%</td>
</tr>
<tr>
<td>Total</td>
<td>43,653 (7.8 %)</td>
<td>60.7</td>
<td>3.5</td>
<td>3.8</td>
<td>12,323</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>43.1 %</td>
<td>81.3</td>
<td>6.9</td>
<td>3.6</td>
<td>46.6%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>56.9 %</td>
<td>96.0</td>
<td>3.9</td>
<td>4.1</td>
<td>53.4%</td>
</tr>
<tr>
<td>Total</td>
<td>50,173 (9.0 %)</td>
<td>89.4</td>
<td>5.2</td>
<td>3.8</td>
<td>13,047</td>
</tr>
<tr>
<td>All categories combined</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posts by personal users</td>
<td>37.6 %</td>
<td>98.5</td>
<td>9.5</td>
<td>8.9</td>
<td>52.3%</td>
</tr>
<tr>
<td>Posts by business users</td>
<td>62.4 %</td>
<td>81.6</td>
<td>4.8</td>
<td>16.2</td>
<td>47.7%</td>
</tr>
<tr>
<td>Total</td>
<td>560,414</td>
<td>87.7</td>
<td>6.6</td>
<td>12.4</td>
<td>45,232</td>
</tr>
</tbody>
</table>

Table 2
Ten most popular# hashtags in each category.

<table>
<thead>
<tr>
<th>Rank</th>
<th>E-liquid (single)</th>
<th>E-liquid (multiple)</th>
<th>E-cigarette</th>
<th>Product package</th>
<th>Person(s)</th>
<th>Statement</th>
<th>Miscellaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>juice</td>
<td>giveaway</td>
<td>vaporesso**</td>
<td>juul**</td>
<td>model</td>
<td>juul**</td>
<td>coiibuilding</td>
</tr>
<tr>
<td>2</td>
<td>ejuices</td>
<td>juice</td>
<td>squonk**</td>
<td>liquidvape</td>
<td>vapebabe</td>
<td>ivapevite</td>
<td>coiibuilder</td>
</tr>
<tr>
<td>3</td>
<td>vapenol</td>
<td>vaporlife</td>
<td>geekvape**</td>
<td>cvv**</td>
<td>tattoo</td>
<td>wevapewvete</td>
<td>coiibuilder</td>
</tr>
<tr>
<td>4</td>
<td>vapeliquid</td>
<td>atomizer</td>
<td>vapegear</td>
<td>stopsmokingartvaping</td>
<td>vapababe</td>
<td>adultlikeflavors</td>
<td>art</td>
</tr>
<tr>
<td>5</td>
<td>vaporlife</td>
<td>vapevape</td>
<td>vapevape</td>
<td>vapevapeshotvaping</td>
<td>selfie</td>
<td>smokefree</td>
<td>cleanbuilds</td>
</tr>
<tr>
<td>6</td>
<td>guydwahvape</td>
<td>vandyvape**</td>
<td>atomizer</td>
<td>vapegiveaway</td>
<td>followme</td>
<td>vapadvocacy</td>
<td>vapesoils</td>
</tr>
<tr>
<td>7</td>
<td>vapingsnitchaine</td>
<td>flavourunjkie</td>
<td>mtl</td>
<td>vapormax</td>
<td>lifestyle</td>
<td>memes</td>
<td>weed</td>
</tr>
<tr>
<td>8</td>
<td>productphotography</td>
<td>vapegiveway</td>
<td>vandyvape**</td>
<td>pods</td>
<td>cloudcheck</td>
<td>giveaway</td>
<td>coilsmith</td>
</tr>
<tr>
<td>9</td>
<td>vapeshop</td>
<td>ejucices</td>
<td>vapepod</td>
<td>vapepod</td>
<td>art</td>
<td>health</td>
<td>westcoastvapers</td>
</tr>
<tr>
<td>10</td>
<td>vapahappy</td>
<td>vaporesso**</td>
<td>giveaway</td>
<td>startvaping</td>
<td>vapahappy</td>
<td>vapesoils</td>
<td>squnk</td>
</tr>
</tbody>
</table>

* The table omits non-English hashtags and the most common hashtags that appeared within the top 200 hashtags of at least five out of the seven identified categories.

** Related to an e-cigarette company.
different types of e-cigarettes and e-liquids. Given the age limit of 13 years to create an account on Instagram can potentially expose the vulnerable cohort to an array of products meant for adult use only. The dominant presence of images featuring sleek design features of vaporizers, a variety of alluring flavors, and attractive packages point to programmed advertisement endeavors that have also been used historically for combustible cigarettes by the tobacco industry [43,44].

Within the studied dataset, one out of every ten images consisted of selfies or depicted persons(s) using a vape product. Similar findings on social media content depicting individuals with different substances such as cigars, cigarrillos, marijuana, and hookah have also been reported earlier [31,45,46]. In line with the prior work [16,19], photos containing personal images often illustrated a person blowing vapor, smoke playing, or performing different tricks with the e-cigarettes. Furthermore, an overwhelming majority of these images portrayed and evoked a positive image of vaping general in the form of youth culture & lifestyle, sex appeal, and feeling of freedom and relaxation. The positive sentiments sketched through these images could grab the attention of adolescents and young adults and entice them towards the initiation and contribute to the normalization and social acceptability of vaping among the highly vulnerable population.

Images within the “statements” category often contained text together with company logo and various vaping products. These images revealed that most of them were about promotions of vaporizers, e-juices, and related accessories. As observed by the prior literature [5,17,47,48], statements associated with discounts, sales, giveaways, buy more - save more, and new stock arrivals were highlighted frequently in the text. These promotional images may be profoundly tempting for underage cohorts as well as the non-smoker population. Given the potential exposure to harmful effects associated with vaping among youth [1,11-13] and increased likelihood of subsequent use of combustible cigarettes and other illicit substances [9,13] extent of promotional content on Instagram is extremely alarming.

In addition to identifying the main categories of the e-cigarettes imagery shared on Instagram, the current study also examined the statistical differences between posts created by individual and business users in different classes. We found a broad utilization of Instagram by e-cigarette companies to advertise their products specifically targeted to the young cohort. Although the ratio of business users and individuals in the overall dataset was almost even, business accounts created more posts that predominantly relate to various forms of product promotions. These results are similar to the previous work analyzing the e-cigarette related content on Instagram, which found that a large portion of the content consists of promotional posts created by businesses [2,3]. Furthermore, we also observed a strong presence of celebrities and references to women and models. Previous research has shown that e-cigarette brands often utilize celebrity endorsers and females for their product promotions [16,49]. These promotional tactics go beyond glamorizing vaping by influencing consumer engagement with content and gaining social acceptability among the youth.

Finally, in line with several previous studies [17,18,33], our study found that an overwhelming amount of content related to vaping on Instagram depicted positive sentiments and attitudes towards vaping, while the negative facet was mostly absent. Consequently, adolescents and young adults frequently interact and engage with content that actively promotes vaping while staying uninformed about its negative aspects. Given the findings presented by the current study, there is a pressing need to produce Instagram-based awareness campaigns, cessation tips, as well as portray the negative sides of vaping. Active participation of health agencies, educational institutes, and anti-tobacco activists on Instagram is also vital, which has been lagging so far. At a policy level, laws requiring social media platforms to build more robust measures to ensure prevention of vaping related content targeted to underage users are needed.

5.2. Key contributions

To the best of our knowledge, this is one of the few studies that apply large-scale unsupervised machine learning techniques to systematically classify and further examine vaping-related images on Instagram, a leading social media among youth. With our unsupervised machine learning approach, this study managed to discover more refined vaping-related image categories. In contrast to the methodological approaches adopted by large scale studies [31], our study makes no initial assumption on the categories but utilizes unsupervised machine learning methods to discover the most relevant categories from the collected dataset directly.

Given the recent outbreak, EVALI [50] in the US, with over 2800 reported deaths or hospitalized, it is significantly critical to understand how adolescents and young adults interact with vaping content on different social media. The applied methodological choices and findings from the current study can support several directions for future research, including the identification and classification of prevailing and popular content on other social media outlets, such as Facebook, Reddit, and Twitter. Likewise, the study findings can help develop suitable interventions by various public health entities as well as tobacco regulations and policy to curb social media marketing and discourage the use of vaping among the vulnerable. From the methodological perspective, the utilized unsupervised machine learning approach supported in classifying a large volume of images and associated metadata with relatively high details and precision. The state-of-the-art methodological choices endeavored by the current study can also supplement automated infodemiology applications for systematic surveillance of cross-platform data. Furthermore, the study demonstrates the utility of social media data (in particular images) as a highly useful source for timely and cost-effective risk assessments and interventions delivery that can be equally pertinent for investigating critical public health subjects including obesity, infectious diseases, food safety, substance use, and mental health.

5.3. Limitations

As the current study relied solely on a single hashtag, #vaping may not represent all the Instagram posts related to e-cigarettes. Since the study leveraged unsupervised machine learning techniques to identify the main categories within the images through visual inspection, the resulting image categories may not represent the most suitable grouping of the images. Furthermore, since we used the images from the identified clusters to train a classifier to predict a category for each image in the dataset, the final group labels may contain some noise. Future studies could evaluate alternative methods to identify the core image classes in the #vaping data on Instagram and compare their results to the findings of this study. Additionally, further work should leverage the extensive set of data obtained in this study to analyze it from new aspects, such as inspecting differences in hashtags used by individuals and businesses in the identified categories. Finally, trend analysis may be conducted to examine trends evolving overtime on the proportion of posts posted by different user groups in the identified categories.

6. Conclusion

The widespread use of social media platforms for promoting e-cigarette products raises increasing concerns that the marketing attracts new young users to become nicotine addicts. The results of this study portray the large portion of promotional content shared on Instagram under popular e-cigarette related hashtags. Since the adolescents cover a large share of Instagram’s user base, effective regulatory actions are required to prevent the e-cigarettes content from reaching the youth.
Role of funding source

None.

Contributors

Mr. Ketonen lead the data collection, analysis, and writing of the manuscript. Both authors approved the final version of the manuscript. Dr. Malik conceptualized the study, data collection, and analysis, and contributed in writing the introduction and discussion sections.

Declaration of Competing Interest

None.

References