Towards More Meaningful Social Media Analysis: Case Study of Using an Age Prediction Algorithm to Identify and Code Reddit Posts about E-cigarettes by Youth vs. Adults

Mario A. Navarro, PhD[1]; Robert Chew[2]; Caroline Kery[2]; Laura Baum[2]; Thomas Bukowski[2]; Annice Kim[2]
[1]Office of Health Communication and Education, Center for Tobacco Products, FDA; [2]RTI International Research Triangle Park

Abstract

BACKGROUND: Many social media investigations have been limited to manual qualitative coding or investigating the utility of machine learning classification algorithms, which either limit the scope of the investigation or fail to provide context and utility. The current study combines the two methods to 1) predict Reddit users' age into two categories (13-20, 21-54) and 2) qualitatively code Electronic Nicotine Delivery System [ENDS] related posts within the two age groups. METHODS. An algorithm using Reddit metadata was developed to classify Reddit posts as being created by 13-20 or 21-54 year old users. Three separate ENDS related search queries were conducted to pull Reddit posts related to the following topics: general vaping, Tobacco 21 minimum age laws, and flavor restriction policies. The age algorithm was then used to predict Reddit users' ages. The 25 posts with the highest karma score (number of upvotes – number of downvotes) for each query and each predicted age group were qualitatively coded (N = 150). RESULTS: Across the three queries, there were nine prominently coded themes: Tobacco 21 Policies, Flavor Restriction Policies, Harm Perceptions, Use, Products, Memes/Jokes, COVID-19, Motivations, and Access. Tobacco 21 Policy and Flavor Restriction posts were evenly distributed across the 13-20 and 21-54 groups. Opposition to flavor restriction policies was a prominent sub-theme for both groups, but more common in the 21-54 group. The 13-20 group was more likely to discuss access in light of flavor restriction policies. Harm Perception and COVID-19 discussions were more prominent among the 21-54 group than in the 13-20 group, but no dominant sub-theme emerged. The 13-20 group was more likely to post images without text (often memes), post on non-tobacco subreddits, and have higher karma scores. Adults were more likely to mention different brand names, while youth posts in this sample only mentioned JUUL by name. CONCLUSIONS: Users who were predicted to be in the 13-20 age group and the 21-54 age group posted and discussed different topics on Reddit, allowing for more nuanced insight. Future studies could utilize machine learning classification algorithms alongside qualitative coding to gain richer insights from target audiences using social media data.

Introduction

- Previous social media studies have relied heavily on thematic coding and content analysis for posted material. Few have integrated automized data science methodologies.
- Reddit, a popular social media platform, provides a source of readily available free data to perform analyses.
- Wang et al. (2015) coded posts on the Reddit to investigate Electronic Nicotine Delivery Systems [ENDS] flavor mentions.[1]
- Reddit lacks publicly available demographic information on users and thus there is a limitation in extrapolating results for specific populations.[2]
- The current exploratory study, using the Chew et al. algorithm, investigates ENDS conversations, with a focus on Flavor Restriction and Tobacco 21 Policy discussions for posts originating from predicted Underage (UA) and Of Legal Age (OLA) posters.

Methods

- There were four steps in the data cleaning process.
- 1) Reddit posts about general vaping, flavor restriction policies, and Tobacco 21 policies were identified and downloaded from pushshift.io, a data collection and archiving platform that has collected Reddit data since 2015 and made it publicly available to researchers.[5] Multiple search keywords were used to identify relevant posts about general vaping (e.g. vape, vaping, e-cigarette), flavor restriction policies (e.g. flavor policy) and Tobacco 21 policies (e.g. "minimum age laws," and tobacco specific words such as "cigarettes," "vapes," and "cigars.")
- 2) A previously developed age prediction algorithm was used to predict the age, for each author of the downloaded Reddit posts as either Underage (UA), Of Legal Age (OLA), or uncertain.[3]
- 3) The top 25 posts in each predicted age group, based on karma scores (number of upvotes – number of downvotes), were identified across all queries (150 total posts).
- 4) Two coders were trained using a standardized codebook and, after achieving high inter-rater reliability (percent agreement of at least 70%), independently coded the study sample. Posts were excluded if they mentioned marijuana/THC/CBD, were not in the English language, or not relevant to e-cigarettes.

Results & Discussion

Descriptive Statistics

- Eighteen posts were excluded from the predicted UA group and 24 posts were excluded from the predicted OLA group leaving 57 UA (General Vaping: 18, Flavor Restriction Policies: 18, Tobacco 21 Policies: 21) and 51 OLA (General Vaping: 13, Flavor Restriction Policies: 17, Tobacco 21 Policies: 21) posts.

Qualitative Results

- For both UA and OLA, the categories of Flavor Restriction Policies and Tobacco 21 Policies were the most prominent (> 40%). Between the two groups.
- For Flavor Restriction policies, opposition was a primary sub-category for both predicted age groups, but many flavor restriction posts fell into the Other sub-category for the UA group and Skepticism for the OLA group.
- For the Tobacco 21 Policy category, a similar pattern emerged for the UA and OLA with Opposition, Skepticism, and the Other sub-categories dominating the conversation. For UA, a sub-category code emerged that detailed the desire to allow 18-20 ENDS users, who were previously able to use ENDS products, to continue having the ability to purchase ENDS products (Legacy Clause, sometimes referred to by posters as "grandfather clause"). For other category and sub-category results, please see Table 1.

Discussion

- This study provides a case study of combining both machine learning methods with qualitative coding to get a better picture of what is occurring in the social tobacco landscape.
- By streamlining research processes using unique methodologies, quick turnaround studies are possible to better surveil the tobacco landscape.
- Future studies should utilize other social media platforms to compare the types of conversations held per platform.

References

[1]Wang L, Zhan Y, Li Q, Zeng DD, Leischow SJ, Okamoto J. An examination of electronic cigarette content on social media: Analysis of e-cigarette flavor content on Reddit. *Int J Environ Res Pu* 2015;12:14916-14935. doi:10.3390/ijerph121114916.
[2]Sharma R, Wigginton B, Meurk C, Ford P, Gartner CE. Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of Reddit discussions. *Int J Environ Res Pu* 2017;14: 7-21. doi:10.3390/ijerph1401007.
[3]Chew R, Kery C, Baum L, Bukowski T, Kim A, Navarro M. Predicting Age Groups of Reddit Users based on Posting Behavior and Metadata: Comparative Study of Classification Models. JMIR Public Health and Surveillance, in press.

Table 1

Post Category or Sub-Category	Underage n(%)	Of Legal Age n(%)
Flavor Restriction Policies	26 (45.61)	37 (72.54)
Support	0 (0)	1 (2.70)
Oppose	9 (34.62)	17 (45.95)
Skepticism	1 (3.85)	8 (21.62)
Access	4 (15.38)	2 (5.41)
Switching	3 (11.53)	5 (13.51)
Quitting	1 (3.85)	0 (0)
Other	8 (30.77)	4 (10.81)
Tobacco 21 Policies	27 (47.37)	21 (41.18)
Support	1 (3.71)	0 (0)
Oppose	8 (29.63)	4 (19.04)
Skepticism	4 (14.81)	1 (4.76)
Legacy Clause	4 (14.81)	0 (0)
Access	2 (7.41)	2 (9.52)
Switching	1 (3.70)	1 (4.77)
Quitting	0 (0)	0 (0)
Other	7 (25.93)	13 (61.91)
Use	11 (19.30)	22 (43.14)
Dual	1 (9.10)	0 (0)
Switching	0 (0)	2 (9.09)
Quitting	2 (18.18)	1 (4.55)
Vape Terms	5 (45.45)	11 (50.00)
Other	3 (27.27)	8 (36.36)
Motivations for Vaping	9 (15.79)	9 (17.65)
Harm Perceptions	2 (3.51)	13 (25.49)
Products	18 (31.58)	8 (15.69)
Memes/Jokes	17 (29.82)	5 (9.80)
COVID-19	1 (1.75)	6 (11.76)
Other	1 (1.75)	3 (5.88)

Table 1. Table 1 reports the frequency and percentages of each post code category and sub-category.