**Report**

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

**Student 1 ID:**1002073766 **Student 2 ID:**1002080027

**Student 1 Name:** Sai Sumana Adurugatla **Student 2 Name:** Perin Devi Behara

**1.Title of the paper**:

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

**2**. **Authors, University, Country:**

Alex Sebastian Schott & Juhan Lee - Yale School of Medicine - Connecticut, USA

Hassan Dashtian & Dhiraj Murthy -The University of Texas at Austin-Texas, USA

Dr Grace Kong - Yale School of Medicine - New Haven, USA

**3**. **Year of publication**- 2022

**4. Number of times the paper is cited by (referenced by) another papers-** 15.

**5. Number of research papers referenced in the paper**- 36

**6. Conference paper**- Tobacco Control 2023; 32:739-746.

**7. Submission deadline for 2023 of the conference & related link**

Submission Deadline: October 19, 2023

[Link:](https://tobaccocontrol.bmj.com/content/32/6/739) [Understanding e-cigarette content and promotion on YouTube through machine learning | Tobacco Control (bmj.com)](https://tobaccocontrol.bmj.com/content/32/6/739)

**8. Best paper name, authors, and Link to find its abstract or pdf:**

**Best paper name:** Affective Learning Objectives for Communicative Visualizations.

**Authors:** Elsie Lee-Robbins and Eytan Adar.

**Link:** [Affective Learning Objectives for Communicative Visualizations (arxiv.org)](https://arxiv.org/pdf/2208.04078.pdf)

**Justification:** By claiming that all visualizations are inherently affective, this award-winning study transforms the field of visualization design. Its qualitative research-based taxonomy encourages designers to make thoughtful judgments, increasing its influence and earning it the Best Paper Award.

**9. List of references listed in the reverse order of number of times it is referenced**

|  |  |  |  |
| --- | --- | --- | --- |
| **Serial**   **No.** | **References** | **Year of Publication** | **Number**  **of Citations** |
| 20 | Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. <https://nlp.stanford.edu/projects/glove/> 2014. | 2014 | 37951 |
| 16 | Stephen A, Galak J. The effects of traditional and social earned media on sales: A study of a microlending marketplace. Soc Sci Res Network 2012. 10.2139/ssrn.1480088 | 2012 | 882 |
| 15 | Food and Drug Administration. Deeming tobacco products to be subject to the federal food, drug, and cosmetic act. Fed Regist April 25, 2014;79:No. 80. | 2016 | 621 |
| 12 | Huang J, Kornfield R, Szczypka G, et al. A cross-sectional examination of marketing of electronic cigarettes on twitter. Tob Control 2014;23:iii26–iii30. 10.1136/tobaccocontrol-2014-051551 | 2014 | 316 |
| 32 | Morean ME, Kong G, Camenga DR, et al. High school students’ use of electronic cigarettes to vaporize cannabis. Pediatrics 2015;136(4):611–616. 10.1542/peds.2015-1727. | 2015 | 289 |
| 1 | Park-Lee E, Ren C, Sawdey MD, et al. Notes from the field: E-cigarette use among middle and high school students - national youth tobacco survey, united states, 2021. MMWR Morb Mortal Wkly Rep 2021;70(39):1387–1389. 10.15585/mmwr.mm7039a4 | 2021 | 258 |
| 23 | Luo C, Zheng X, Zeng D, et al. Portrayal of electronic cigarettes on YouTube. BMC Public Health 2014;14:1028. <http://www.biomedcentral.com/1471-2458/14/1028>. | 2014 | 119 |
| 7 | Vogel EA, Ramo DE, Rubinstein ML, et al. Effects of social media on adolescents’ willingness and intention to use e-cigarettes: An experimental investigation. Nicotine Tob Res 2021;23(4):694–701.10.1093/ntr/ntaa003. | 2021 | 114 |
| 29 | Cheney M, Gowin M, Wann TF. Marketing practices of vapor store owners. Am J Public Health 2015;105(6):e16–21. 10.2105/ajph.2015.302610. | 2015 | 84 |
| 5 | Kong G, LaVallee H, Rams A, et al. Promotion of vape tricks on youtube: Content analysis. J Med Internet Res 2019;21(6):e12709. 10.2196/12709 | 2019 | 75 |
| 30 | Kong G, LaVallee H, Rams A, et al. Promotion of vape tricks on YouTube: Content analysis. J Med Internet Res 2019;21(6):e12709. 10.2196/12709. | 2019 | 75 |
| 4 | World Health Organization. Report on the global tobacco epidemic. Geneva: 2019 | 2021 | 71 |
| 31 | Vassey J, Metayer C, Kennedy CJ, et al. #vape: Measuring e-cigarette influence on Instagram with deep learning and text analysis. Front Commun 2020;4(75). 10.3389/fcomm.2019.00075. | 2020 | 46 |
| 8 | Zheng X, Li W, Wong S-W, et al. Social media and e-cigarette use among us youth: Longitudinal evidence on the role of online advertisement exposure and risk perception. Addict Behav 2021;119:106916. 10.1016/j.addbeh.2021.106916 | 2021 | 36 |
| 9 | Lee J, Tan ASL, Porter L, et al. Association between social media use and vaping among Florida adolescents, 2019. Prevent Chronic Dis 2021;18:E49–E49. 10.5888/pcd18.200550. | 2021 | 28 |
| 13 | Fu R, Kundu A, Mitsakakis N, et al. Machine learning applications in tobacco research: A scoping review. Tobacco Control 2021:tobaccocontrol-2020–056438. 10.1136/tobaccocontrol-2020-056438. | 2023 | 26 |
| 10 | Ceci L YouTube - Statistics & Facts. Statista Jul 12, 2021 | 2022 | 25 |
| 34 | YouTube. Community guidelines. <https://www.youtube.com/yt/policyandsafety/communityguidelines.html> 2022. | 2017 | 25 |
| 36 | Visweswaran S, Colditz JB, O’Halloran P, et al. Machine learning classifiers for Twitter surveillance of vaping: Comparative machine learning study. J Med Internet Res 2020;22(8):e17478. 10.2196/17478. | 2020 | 25 |
| 3 | Tarasenko Y, Ciobanu A, Fayokun R, et al. Electronic cigarette use among adolescents in 17 european study sites: Findings from the global youth tobacco survey. Eur J Public Health 2022;32(1):126–132. 10.1093/eurpub/ckab180. | 2022 | 23 |
| 27 | Amin S, Dunn AG, Laranjo L. Exposure to e-cigarette information and advertising in social media and e-cigarette use in Australia: A mixed methods study. Drug Alcohol Depend 2020;213:108112. 10.1016/j.drugalcdep.2020.108112. | 2020 | 23 |
| 26 | Kong G, Morean ME, Bold KW, et al. Dripping and vape tricks: Alternative e-cigarette use behaviors among adolescents. Addict Behave 2020;107:106394. 10.1016/j.addbeh.2020.106394 | 2020 | 20 |
| 14 | Kim K, Gibson LA, Williams S, et al. Valence of media coverage about electronic cigarettes and other tobacco products from 2014 to 2017: Evidence from automated content analysis. Nicotine Tob Res 2020;22(10):1891–1900 | 2020 | 19 |
| 24 | Guy MC, Helt J, Palafox S, et al. Orthodox and unorthodox uses of electronic cigarettes: A surveillance of YouTube video content. Nicotine Tob Res 2019;21(10):1378–1384. 10.1093/ntr/nty132 | 2019 | 18 |
| 6 | Alpert JM, Chen H, Riddell H, et al. Vaping and instagram: A content analysis of e-cigarette posts using the content appealing to youth (cay) index. Subst Use Misuse 2021;56(6):879–887. 10.1080/10826084.2021.189923 | 2021 | 17 |
| 21 | Song M, Zhao X, Liu Y, et al. Text sentiment analysis based on convolutional neural network and bidirectional lstm model. 24th International Conference on Automation and Computing Singapore: Springer Singapore; 2018:55–68. | 2018 | 17 |
| 25 | Camenga DR, Morean ME, Kong G, et al. Appeal and use of customizable e-cigarette product features in adolescents. Tob Regul Sci 2018;4:51–60. 10.18001/TRS.4.2.5. | 2018 | 16 |
| 22 | Langdetect 1.0.9. <https://pypi.org/project/langdetect/> 2021. | 2021 | 15 |
| 33 | Lee J, Kong G, Kassas B, et al. Predictors of vaping marijuana initiation among us adolescents: Results from the population assessment of tobacco and health (PATH) study wave 3 (2015–2016) and wave 4 (2016–2018). Drug Alcohol Depend 2021;226:108905. 10.1016/j.drugalcdep.2021.108905. | 2021 | 15 |
| 2 | Gottschlich A, Mus S, Monzon JC, et al. Cross-sectional study on the awareness, susceptibility, and use of heated tobacco products among adolescents in guatemala city, guatemala. BMJ Open 2020;10(12):e039792. 10.1136/bmjopen-2020-039792 | 2020 | 14 |
| 18 | Tor Project. <https://www.torproject.org/> March 11, 2022. | 2022 | 13 |
| 28 | Yang JS, Lee E. A qualitative assessment of business perspectives and tactics of tobacco and vape shop retailers in three communities in orange county, ca, 2015–2016. Arch Public Health 2018;76:57. 10.1186/s13690-018-0307-z. | 2018 | 11 |
| 17 | Dashtian H, Murthy D, Kong G. An exploration of e-cigarette–related search items on youtube: Network analysis. J Med Internet Res 2021 | 2022 | 7 |
| 35 | YouTube. Age-restricted content. <https://support.google.com/youtube/answer/2802167?hl=en> 2022. | 2022 | 4 |
| 11 | Ceci L Number of YouTube users worldwide from 2016 to 2021. Statista August 31, 2021. | 2021 | 2 |
| 19 | Natural Language Toolkit. <https://www.nltk.org/2021>. | 2021 | 1 |

**10. Maximum times any of its referenced papers has been cited**

Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. <https://nlp.stanford.edu/projects/glove/> 2014. - number of citations(**37951**)

The paper "GloVe: Global vectors for word representation" introduced a groundbreaking approach to representing words as vectors, capturing semantic relationships and contextual information efficiently. Its impact extends across natural language processing applications, contributing to improved word embeddings and widespread adoption in research and development.

**11. Dataset:** - YouTube video Data Set.

Data used from this study is publicly available data from YouTube and I have mailed to author regarding the data. There is no source code available in the paper and as well as in the internet.

**12. Code of Original Research Paper:**

There is no code for this research paper but related code I took from GitHub this is (click to access [code link)](https://github.com/forrabota/elektronines-cigaretes.github.io)

[GitHub - forrabota/elektronines-cigaretes.github.io](https://github.com/forrabota/elektronines-cigaretes.github.io)

**13. Domain Name:** E-Cigarette Video Content Analysis- Supervised learning

**Description:**

The use of machine learning algorithms in examination of videos pertaining to electronic cigarettes on digital platforms, particularly YouTube. These algorithms extract textual information from video titles and identify different themes, such as comparison content, product reviews, tutorials, and vaping techniques. This study is an essential monitoring tool to find e-cigarette trends and marketing tactics used on internet platforms and YouTube e-cigarette content.

**14. Problem Explanation:**

This paper addresses the prevalent issue of youths being exposed to YouTube information about electronic cigarettes, or "e-cigarettes." The research shows the need for regulatory measures to mitigate the impact of e-cigarette promotion on social media platforms accessible to young people by using machine learning to analyze and categorize various aspects, such as video themes, featured products, uploaders, and marketing strategies.

**15.BenefitsofTechniquesused:**  
**Non-survey paper:**

The research paper's use of machine learning provides notable benefits for non-review papers. Firstly, it enables efficient analysis of a large dataset of 3,830 English videos on YouTube, automating the identification of diverse e-cigarette content aspects. The high accuracy achieved, with an F1 score of up to 0.87, ensures reliable predictions. Machine learning's capability to recognize patterns and handle complexity offers a nuanced understanding of e-cigarette promotion on social media. Additionally, its scalability allows for application to larger datasets and potential extensions to similar studies, providing valuable insights for policymakers and aiding in the development of regulations addressing youth exposure to e-cigarette content.

**16. Summary**:

The paper employs machine learning to analyze e-cigarette content on YouTube, examining video themes, products, uploaders, and sales. With 3,830 videos analyzed, diverse e-cigarette products were featured, primarily by 'Vape Enthusiasts' and retailers. About 43.2% of videos included discount/sales, and ML models achieved up to 87% accuracy. Findings stress the need for regulating e-cigarette promotion on social media to protect youth. In summary, the study identifies trends and marketing strategies, emphasizing the necessity for federal regulations to safeguard youth from e-cigarette promotion on platforms like YouTube.

**17. Literature review:** Literature review and its methods.

The paper describes a research study about machine learning (ML) to analyze e-cigarette content on YouTube, particularly focusing on video themes, featured products, video uploaders, and marketing/sales strategies.

**Objective:** The study aims to understand e-cigarette content on YouTube, a popular social media platform among youth, using ML techniques.

**Methods:**

* The researchers identified e-cigarette content using 18 search terms.
* Fictitious youth viewer profiles were used to predict four ML models based on metadata.
* Models were designed to analyze video themes, featured products, channel types (uploaders), and discount/sales.
* (1) human labelling, (2) text preprocessing, (3) training the supervised ML classification algorithms and (4) evaluating the algorithm’s performance.

**Results:**

* The study analyzed 3,830 English videos.
* Diverse e-cigarette products were featured.
* 'Vape Enthusiasts' (54.0%) and retailers (20.3%) were the top video uploaders.
* 43.2% of videos featured discount/sales

**ML Model Accuracy:**

* The four ML models achieved an F1 score (a measure of accuracy) of up to 0.87.

**Gaps and Future Directions:**

* Enhance fictitious profiles, broaden search terms, incorporate visual classification, explore alternatives for volume-personalization trade-off, include non-English content, and implement automated data collection for temporal relevance.

**Summary:**

* The research successfully identified video themes, featured products, uploaders, and sales strategies using advanced ML methods on youth-accessible YouTube content.
* The findings underscore the importance of using ML for surveillance of e-cigarette trends and marketing strategies on social media, emphasizing the need for comprehensive federal regulations to safeguard youth from e-cigarette promotion on platforms like YouTube.

**18. Pros and Cons:**

**Pros:**

Innovative Methodology: Using fictitious youth profiles for e-cigarette content identification on YouTube.

Algorithms used: Recognition of social media algorithms, emphasizing the importance of personalization factors in content identification.

Robust ML Performance: Demonstrated robust performance of machine learning models in identifying e-cigarette themes on social media.

Broad Content Coverage: Analysis of a substantial number of videos (3830), offering a comprehensive overview of e-cigarette content on YouTube.

**Cons:**

Fictitious Profiles Limitation: Acknowledgment that fictitious profiles are an initial step, and future research should enhance this method for better accuracy.

Broad Search Term Issues: Limitation in using broad terms, missing brand-specific content.

e.g., 'JUUL' or 'Puff bar.'

Visual Classification Needed: Recognition of the need for improvement in classifying featured e-cigarette products, suggesting visual classification for enhancement.

Volume vs. Personalization Trade-off: Acknowledgment of a trade-off between analyzing volume and obtaining personalized results due to the methodological constraints.

Language Limitation: Acknowledges that the English-based search may limit relevance to non-English speakers globally, recommending future studies in other languages.

Temporal and Emerging Theme Limitations: Temporal limitation of the study conducted in July 2020, potentially missing emerging themes like COVID-19 related content.

**19. Address deficiencies:**

Addressing some deficiencies in this paper

Limitation on Fictitious Profiles: Improve profiles to resemble actual user behavior more closely or utilize actual profiles to verify results.

Issues with Broad Search Terms: Extend search phrases (such as brand-specific ones) and use natural language processing (NLP) to grasp context.

Requires Visual Classification: Using pre-trained models for features, improve the machine learning model using visual categorization.

Trade-off between Volume and Personalization: For a balanced approach, investigate other approaches and consider automated queries for larger samples.

Language Restrictions: For worldwide representation, incorporate non-English content and work with multilingual specialists.

Limitations of Temporal and Emerging Themes: Implement automatic data collecting to track developing topics and conduct follow-up investigations.

**20. Contributions:**

Grace Kong (GK): GK is the guarantor of the study. As a guarantor, accepts responsibility for the finished work and the study's conduct, has access to the data, and controls the decision to publish. She conceptualized and designed the study, obtained funding for the study, interpreted the results, and wrote the first draft.

Dhiraj Murthy (DM): Designed the study, acquired the data, analyzed the data, interpreted the results, wrote sections.

Juhan Lee (JL): Wrote sections, interpreted the results.

Alex Sebastian Schott (ASS): Analyzed the data, interpreted the results, wrote sections.

Hassan Dashtian (HD): Assisted in data analysis.

**21. Future recommendations:**

Authors suggest using fictitious viewer profiles to identify types of content exposed to, future research leverage this method to obtain more relevant and accurate information on e-cigs. Future research may improve model performance by using visual classification. Non-English speakers may be exposed to e-cigarette content in other languages, so non-English content should be examined in future studies.

Some recommendations which I suggest are Collaborate with social media platforms to integrate machine learning models for automatic detection and removal of e-cigarette promotion. Utilize natural language processing (NLP) for text-based content and deep learning for image and video analysis. Implement real-time content moderation to proactively identify and restrict e-cigarette promotion before it reaches a wider audience.

**22. Machine learning algorithms:**

Supervised Machine Learning Algorithms & Classification Algorithms

Deductive and Inductive Approaches

Python Natural Language Toolkit

Bidirectional Long Short-Term Memory Networks (BLSTM) layer.

BLSTM is a Deep, Neural Machine Learning method.