Analyzing Topics of JUUL Discussions on Social Media Using a Semantics-assisted NMF model

Hejing Liu[1,2,3] Qiudan Li[1,3] Riheng Yao[1,2,3] Daniel Dajun Zeng[1,2,3]
[1]The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Shenzhen Artificial Intelligence and Data Science Institute (Longhua)
{liuhejing2018, qiudan.li, yaoriheng2017, dajun.zeng }@ia.ac.cn

***Abstract*— JUUL has become a widely used brand of e-cigarettes which takes more than 70% of the market. Social media provides a popular platform for users to discuss the preference and perceptions of JUUL. The discussions are valuable for real-time monitoring of JUUL use. Current research on topic analysis of JUUL discussions mainly relies on human work, which takes much time and effort. This paper adopts a Semantics-assisted NMF topic analysis model to automatically discover topics from JUUL-related short posts on Reddit. By successfully merging the semantic relationships into traditional NMF, this model outperforms in discovering topics with keywords that are important but have a lower word frequency among the posts. Experimental results show the potential of this model in JUUL surveillance and control practice.**

Keywords*— *e-cigarettes, JUUL, topic analysis, short posts.

I. INTRODUCTION

In recent years, the use of electronic nicotine delivery systems (also called "e-cigarettes") has fast increased especially in youth [1]. In 2018, the percentage of e-cigarettes use reaches to 20.8% among high school students and 4.9% among middle school students [1]. JUUL is a novel product which entered the market in 2015 and now takes more than 70% share of the e-cigarettes [2]. Social media such as Twitter and Reddit has provided a valuable platform for users to communicate with each other on the opinions, preferences, and experiences of JUUL. Analyzing the JUUL-related discussions could contribute to tobacco surveillance and control, especially in youth tobacco prevention.

Existing research mainly focuses on manually analyzing topics and information diffusion to obtain users' use, recognition, and perceptions of JUUL [3][4][5]. Chu et al. [3] followed JUUL's official account on Twitter and discovered the messages are spread by adolescents. Allem et al. [4] analyzed Twitter posts related to JUUL by manually concluding topics using word frequency, and found that the discussions are around the product, purchasing and school. Brett et al. [5] analyzed the topics in JUUL-related posts on Reddit using human coder and found valuable content from 364 posts. However, the labor work on large amount of data takes much effort and time. Therefore, it is much worthy to automatically reveal the hidden semantic information among JUUL-related discussions, which may help for real-time monitoring people's action and perception on e-cigarettes use, further declining the tobacco use among adolescents.

Latent Dirichlet allocation (LDA) model was commonly adopted to automatically capture the thematic relationship for e-cigarette topics [5]. Based on word-documents co-occurrence, LDA method could generate each topic using probabilistic model. However, JUUL-related posts are a huge amount of short, informal and noisy texts. With limited words in each post, simply using document-level information may not be enough to reveal the topics precisely. In order to capture the internal information provided by the words themselves, Biterm Topic Model (BTM) was proposed [7]. BTM firstly extracts biterm set, i.e. the word-pairs occurred in each document then uses it to model on the whole document collection. With large quantities of text, BTM could overcome the problem of data sparsity in short posts. Yet both LDA model and BTM focus on calculating occurrences to find out the mathematical relationships in documents, the semantic information still needs to be further explored. Recently, Shi et al. [8] proposed a Semantics-assisted non-negative matrix factorization (SeaNMF) model. Under the help of word embedding techniques, SeaNMF model could learn the internal semantic relationships using Skip-Gram view, which better solved the challenge of lacking for document-level word co-occurrence. Applying this model to analyze JUUL-related discussion may uncover more specific and valuable topics with more meaningful keywords.

This paper adopts a SeaNMF based topic analysis method for identifying JUUL-related topics. First of all, JUUL-related short posts, words and their contexts are denoted. Second, the external semantic information from posts and internal semantic information from words are figured out. Both word-posts and word-contexts correlation are learned. Eventually, JUUL-related topics are generated using those two correlations.

II. TOPIC DISCOVERY IN JUUL-RELATED SHORT POSTS

Fig.1 shows the framework of the SeaNMF model for discovering JUUL-related topics in short posts.

A. Word-post and Word-context Correlation Calculation

In order to better understand topics of JUUL-related posts, it is critical to get the external and internal information, i.e. the post-level and word-level information. The former can be

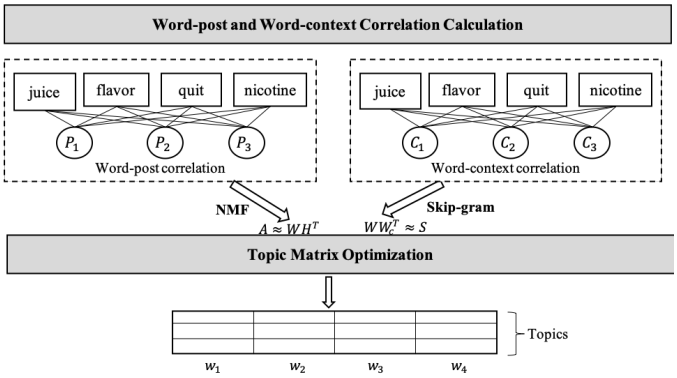

Fig. 1 Framework of Semantics-assisted Model

obtained by calculating the word-post correlation and the latter can be learned through the embedding of words and their contexts.

Given the JUUL corpus, suppose there are N posts and M unique words. To calculate word-post correlation, NMF [11] method is used to decompose the word-post matrix $A \in \mathbb{R}_+^{M\times N}$ into two low-rank matrices $W \in \mathbb{R}_+^{M\times K}$ and $H \in \mathbb{R}_+^{N\times K}$, where K is the number of topics. W is a word matrix in which each column represents a topic and the values are weights of the M words. H is a topic matrix in which each row represents a post and the values are weights of the K topics.

The word-context correlation matrix S can be approximated by $W \cdot W_c^T$. The matrix W is the same word matrix in NMF. The context matrix W_c is formulated as $W_c(j,:) = \vec{c_j}$, where vector $\vec{c_j} \in \mathbb{R}_+^K$ and c_j represents the context word of word w_i in vocabulary set. Base on Skip-Gram algorithm [9], each element S_{ij} is obtained as follows [8]:

$$S_{ij} = \left[\log\left(\frac{\#(w_i,c_j)}{\#(w_i)\cdot p(c_j)}\right) - \log k\right]_+ \quad (1)$$

where p is a unigram distribution, $\#(w_i,c_j)$ is the number of the appearance of word pair (w_i,c_j) in the corpus and $\#(w_i)$ is the word frequency of w_i. Each sliding window is set to the length of the corresponding short text.

B. Topic Matrix Optimization

To merge the semantic relationships of words and their contexts into the conventional NMF model, the objective function [8] is as follows:

$$\min_{W,W_c,H\geq 0} \left\| \left(\frac{A^T}{\sqrt{\alpha}S^T}\right) - \left(\frac{H}{\sqrt{\alpha}W_c}\right) W^T \right\|_F^2 + \psi(W,W_c,H) \quad (2)$$

where $\alpha \in \mathbb{R}_+$ is a scale parameter and ψ is a penalty function which can be set to for different circumstances.

Block coordinate descent (BCD) algorithm is adopted to optimize the objective function.

III. Experimental Results and Discussions

A. Dataset

We collected posts on JUUL subreddit[1] on Reddit from Dec. 27, 2018 to Mar. 20, 2019. The collection concluded 5,943 main posts and 41,111 replies. We selected posts with length less than 140 characters as our short text dataset. Finally, 1,132 main posts and 39,707 replies (40,839 short texts in total) were picked after data preprocess such as lowercase and removing symbols and stopwords.

B. Baseline Methods

- LDA [10]: A generative probabilistic model which learns the topics by modeling the word-document co-occurrence.
- BTM [7]: A generative topic model suitable for short texts which directly models the word-pair appears in each document.

C. Experimental Settings and Evaluation

The topic number is set to 7 for all of the three models.

As evaluation, we use Pointwise Mutual Information (PMI) to calculate the topic coherence score C, which is formulated as follows [8]:

$$C_k = \frac{2}{N(N-1)} \log \frac{p(w_i,w_j)}{p(w_i)p(w_j)} \quad (3)$$

where N is the number of top words in k-th topic, $p(w_i,w_j)$ is the possibility of word w_i and w_j co-occurs in the same document, $p(w)$ is the possibility of word w occurs in a random document.

D. Analysis and Results

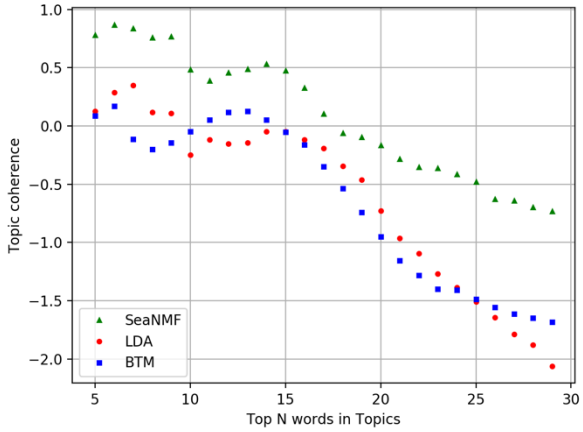

Fig. 2 Average Topic Coherence of Three Models

Fig. 2 shows the average of topic coherence of each topic discovered by all models. We calculated the score of top N words in each topic, where N ranges from 5 to 30. It can be seen that SeaNMF shows better performance in every word bucket, which proves the effectiveness of SeaNMF in learning JUUL-related topics from short social media posts.

Table 1 shows the Top 10 keywords and their word frequency in each topic discovered by SeaNMF. Words that

[1] https://www.reddit.com/r/juul

TABLE 1 TOP 10 KEYWORDS AND WORD FREQUENCY IN TOPICS DISCOVERED BY SEANMF

	Top 10 Keywords in Topics Discovered by SeaNMF	Topic Description
1	pod(7311) pack(2080) buy(1361) refill(687) device(703) juice(1456) **leak**(434) put(670) fake(657) **brand**(522)	Product-related
2	hit(1985) juice(1456) nicotine(1344) good(2138) nic(908) salt(726) vape(1027) **throat**(414) cigarette(839) refill(687)	Experience-related
3	flavor(1914) mango(1744) pack(2080) mint(1392) taste(1326) tobacco(952) **menthol**(429) **fruit**(365) **cucumber**(437) bought(859)	Flavor-related
4	post(1992) question(1645) comment(1679) free(1850) contact(1556) bot(1480) message(1468) subreddit(1504) user(1554) concern(1444)	Community-related
5	website(445) order(478) **site**(301) **ship**(296) state(275) sell(574) **price**(483) **code**(202) online(505) **live**(365)	Purchasing-related
6	smoking(761) kid(439) vaping(487) cigarette(839) quit(608) **health**(100) **addiction**(206) **body**(117) cigs(432) **juuling**(342)	Addiction-related
7	**bottom**(260) **side**(309) battery(573) **rubber**(100) **clean**(261) **metal**(149) **paper**(109) **tip**(226) **remove**(92) **top**(257)	Instruction-related

not detected by all baseline models are painted green.

In Table 1, topic 1 represents the overview of JUUL product, such as product descriptions and brand comparisons. Topic 2 is about throat hit, which is one of the most important experience when vaping. In this topic, the keywords "juice" "nicotine (nic)" and "salt" show the vital aspects that JUUL users will pay attention to when choosing e-cigarettes. Topic 3 refers to the flavor of JUUL where users share the taste, price and preference. Compared with baseline models, more detailed words like "menthol" "fruit" and "cucumber" are found in the result. In this topic, some of the users express the extreme favor towards certain flavor while some mention the bad experience. A user reported headache as: "*Mint and menthol give me headaches*". Finding flavors automatically is important for catching up the transformation of flavor on JUUL and monitoring the addiction and adverse reaction. Notably, the flavor keywords and the rank of their discussions are consistent with previous research [5]. Topic 4 is related to the discussion community, as it is a characteristic of Reddit. Topic 5 represents the discussion on purchasing method of JUUL, where people tend to share their experience and opinions on auto-shipping, online order and purchase limitations. Furthermore, our model discovers the other two topics which are almost missed by baseline methods. In topic 6, both keywords "health" and "addiction" indicate people concern about what health effects may cause by JUUL. Topic 7 is related to the use instruction of JUUL, such as how to clean the device and change the battery.

IV. CONCLUSION

In this paper, we used SeaNMF to automatically learn JUUL-related topics from short posts in social media. The model uncovered the detailed topics hidden among common discussions with keywords which provided important information but might have a lower frequency in the corpus. The results proved the effectiveness of SeaNMF in discovering JUUL-related topics from social media texts, which could contribute to the surveillance and control of JUUL use.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China under Grant No. 2016QY02D0305, the National Natural Science Foundation of China under Grant No. 61671450, 71621002, the Key Research Program of the Chinese Academy of Sciences under Grant No. ZDRW-XH-2017-3.

REFERENCES

[1] Cullen, K. A., Ambrose, B. K., Gentzke, A. S., Apelberg, B. J., Jamal, A., & King, B. A. (2018). Notes from the field: Use of electronic cigarettes and any tobacco product among middle and high school students—United States, 2011–2018. Morbidity and Mortality Weekly Report, 67(45), 1276.

[2] King, B. A., Gammon, D. G., Marynak, K. L., & Rogers, T. (2018). Electronic cigarette sales in the United States, 2013-2017. Jama, 320(13), 1379-1380.

[3] Chu, Kar Hai, et al. (2018) JUUL: Spreading Online and Offline. Journal of Adolescent Health 63(5):582-586.

[4] Allem, J. P., Dharmapuri, L., Unger, J. B., & Cruz, T. B. (2018). Characterizing JUUL-related posts on Twitter. Drug and alcohol dependence, 190, 1-5.

[5] Brett, E. I., Stevens, E. M., Wagener, T. L., Leavens, E. L., Morgan, T. L., Cotton, W. D., & Hébert, E. T. (2019). A content analysis of JUUL discussions on social media: Using Reddit to understand patterns and perceptions of JUUL use. Drug and alcohol dependence, 194, 358-362.

[6] Zhan, Y., Liu, R., Li, Q., Leischow, S. J., & Zeng, D. D. (2017). Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms. Journal of medical Internet research, 19(1), e24.

[7] Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web (pp. 1445-1456). ACM.

[8] Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018, April). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In Proceedings of the 2018 World Wide Web Conference on World Wide Web (pp. 1105-1114). International World Wide Web Conferences Steering Committee.

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[10] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[11] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788.