5-2018

Predictive Modelling Using Unstructured Data From Online Forums: A Case Study on E-cigarette Users

K W. Khong
The University of Nottingham Malaysia Campus, Kokwei.khong@nottingham.edu.my

S L. Tan
The University of Nottingham Malaysia Campus, ksax4tsl@exmail.nottingham.edu.my

S Teng
The University of Nottingham Malaysia Campus, Shasha.Teng@nottingham.edu.my

F S. Ong
The University of Nottingham Malaysia Campus, FonSim.Ong@nottingham.edu.my

PREDICTIVE MODELLING USING UNSTRUCTURED DATA FROM ONLINE FORUMS: A CASE STUDY ON E-CIGARETTE USERS

Khong, K.W.
The University of Nottingham Malaysia Campus
Kokwei.khong@nottingham.edu.my

Teng, S.
The University of Nottingham Malaysia Campus
Shasha.Teng@nottingham.edu.my

Tan, S.L.
The University of Nottingham Malaysia Campus
ksax4tsl@exmail.nottingham.edu.my

Ong, F.S.
The University of Nottingham Malaysia Campus
FonSim.Ong@nottingham.edu.my

Abstract:

In the age of the digital economy, social media, forums and other online platforms have played active parts in our daily activities. The amount of data digitized and recorded in these platforms have surged exponentially. Many believed that this underexplored unstructured data sources have huge potential in offering insights to policy makers and companies. This paper aims to propose a hybrid approach using inductive and deductive reasoning to identify motivational factors to use e-cigarettes for predictive modelling. A total of 790 comments and discussions relevant to e-cigarette use and motivations to use e-cigarette were scraped and stored from online forums like Reddit, Vapingunderground and e-cigarette-forum. A series of text analytics were conducted on the text corpus and the cluster analysis enabled us to build a predictive model. Using Bayesian Structural Equation Modelling, we concluded that the constructs derived by clustering, i.e. Cost and Convenience and Enjoyment, have significant associations with smokers trying to quit smoking. While health-related issues were inherent to the notion of quitting smoking, enjoyment, cost and convenience were motivational factors which will generate favourable response towards quitting smoking. The findings showed encouraging results from a methodological standpoint and offered insights to policy makers and companies on health-related issues pertaining to the use of e-cigarettes.[1]

Keywords:

Unstructured data, predictive modelling, hybrid approach, online forum, text analytics, Bayesian SEM

1. Introduction

Social media, forums and other online platforms serve as social support systems for sharing ideas, information and beliefs. Our daily activities on these platforms are digitized and recorded. To date, analysis of this data provides us with an opportunity to have a better understanding of social behaviours (Torii, Tilak, Doan, Zisook, & Fan, 2016). It offers insights for social sciences to have access to demographic groups that are heavily concerned with health-related issues. A surge of vibrant data mining research has been conducted using online data as the unique source of health-related information (Torii, Tilak, Doan, Zisook, & Fan, 2016). This underexplored data source is used to study online discussions regarding

[1] This is a project funded by the Ministry of Higher Education under the Fundamental Research Grant Scheme.

health-related information. Given the advancement of technology, increasing dialogues regarding health information emerged surrounding social media, forums and other online platforms. Collecting and processing this dataset, which used to take weeks or months, is captured and completed in hours or real time (Torii, Tilak, Doan, Zisook, & Fan, 2016).

Electronic cigarettes (or e-cigarettes) has gained momentum as an emerging tobacco product across the world in recent years. There are many research studies spanning topics such as health effects, cessation, and marketing (Cole-Lewis, et al., 2015). E-cigarettes are often perceived as alternatives to conventional cigarettes for smoking cessation. However, little is known to verify such claims from scientific research (Kavuluru & Sabbir, 2016). On one hand, Bullen at al. (2013) suggested that e-cigarettes are modestly effective at helping smokers to quit. There are other studies findings indicated that there was no association between e-cigarette use and quitting (Grana, Popova, & Ling, 2014). Given considerably ranged findings and divergent opinions on e-cigarette use, with increased and latest dialogues on social media, forums and other online platforms, it is critical to study health effects of e-cigarettes, identify conversation trends in attitudes and behaviours to develop campaign strategies for public health surveillance and relevant cessation interventions (Cole-Lewis, et al., 2015).

Kavuluru and Sabbir (2016) identified e-cigarette proponents and their behaviours along popular themes on Twitter. One of the limitations in this work is the polarity of tweet themes is not discussed (Kavuluru & Sabbir, 2016). Despite the pervasiveness of e-cigarette topic to public, Cole-Lewis et al. (2015) agreed that there was limited evidence on public knowledge and attitudes towards e-cigarettes. Furthermore, the role of entry effects in attitude toward e-cigarettes is not addressed (Myslín, Zhu, Chapman, & Conway, 2013). These information is critical to guiding the public health decision makers for health information communication, surveillance and interventions. Lazard et al., (2016) reviewed past studies regarding sentiments of tweets by e-cigarette users. They found that e-cigarette contents on Twitter skewed favourably towards e-cigarettes (Cole-Lewis, et al., 2015; Kavuluru & Sabbir, 2016; Myslín, Zhu, Chapman, & Conway, 2013). They believed that a large-scale and inductive analysis of topics and themes surrounding social media, forums and other online platforms is needed (Lazard, et al., 2016). Similar viewpoints with Lazard et al. (2016), Torii et al. (2016) were motivated to obtain valuable information from a vast amount of online reviews, summarize different types of illness and discover patterns of online health data. Online data that appears to be underutilized will continue to uncover trends of public views on e-cigarette use (Cole-Lewis, et al., 2015).

It is noticeable that majority of previous studies related to text mining e-cigarette tweets are seeking to identify conversation trends, ideally with revealing public views on e-cigarette use. Little is known about motivational factors associated with e-cigarette use within social media, forums and other online platforms contexts. There are many studies carried out to search for reasons for e-cigarette use from different countries (Hummel, et al., 2015; Kinouani, Castéra, Laporte, Pétrègne, & Gay, 2016; Kong et al., 2016; Li, Newcombe, & Walton, 2015). Top reasons for use e-cigarettes are curiosity, influence of friends, quit smoking (Kong et al., 2015). Most of these studies have been conducted through surveys or focus groups. Analysing the amount and contents of online data related to e-cigarettes is a promising way to gain insights compared to small sample size and possible biased self-reported surveys. However, previous studies found 90% of tweets are from commercial users (Huang, Kornfield, Szczypka, & Emery, 2014), and for advertising (Kim, Hopper, Simpson, & Porter, 2015). Online forums act as other resources for online data regarding e-cigarettes (Sharma,

Wigginton, Meurk, Ford, & Gartner, 2017). This work will focus on online forum discussions regarding e-cigarette use.

2. The Aim

The study aims to propose an approach, discussed later in section 3, of using unstructured data scraped from online forums for predictive modelling. Using text mining programmes, unstructured data on motivational factors to use e-cigarette from online forums were scraped. In this study we scraped e-cigarette related information from multiple online forums by leveraging scalable analytic technologies like SAS Text Miner. The content analysis of reasons for e-cigarette use reflects perceptions on e-cigarette related discussion in online forums. This study will reveal not only conversation trends regarding e-cigarette use, but also users' perceptions and motivations on the use of e-cigarettes. Equally important to the aim is that the current study provides insights to motivational factors to e-cigarette use as an alternative approach to traditional approaches like surveys and interviews.

3. Approach to Analyse Unstructured Data

In this paper, we proposed a hybrid approach encapsulating both inductive and deductive reasoning to explore motivational factors to use e-cigarettes. Inductive approach is an approach to analyse raw data that are predominantly qualitative in nature to derive concepts, themes and dimensions of an area of study (Strauss & Corbin, 1998; Thomas, 2006). This enables us to explore and identify "frequent, dominant, or significant themes inherent in the raw data, without the restraints imposed by structured methodologies" (Thomas, 2006, p 238). For example, inductive approaches are used to look at new concepts, frameworks and theories based on a dataset. In deductive approaches, hypotheses are formulated based on existing theories and models (Fereday & Muir-Cochrane, 2006; Gallaire & Minker, 1984). These hypotheses are then put through a series of statistical tests. Often predictive models are derived based on these hypotheses. For example, deductive approaches are used to test a theory where researcher use instruments to gather data and subsequently test the theory based on the dataset. Several studies using the hybrid approach were conducted especially on social media platforms and this approach can depict complex reasoning based on huge datasets (Barbieri, et al., 2010; Zeng, et al., 2010). Based on these studies, the approach was able to manage the ever changing of knowledge drawn from stream-based content from new media channels. The use of unstructured data from the text corpus represented potentially huge datasets from new media channels and the knowledge drawn from these contents suited the use of the hybrid approach. In this paper the inductive approach was first deployed to understand the unstructured data scraped from multiple online forums. Based on the themes of these unstructured data, a model will be identified using deductive reasoning and a predictive model was derived to test a series of hypotheses. It is important to highlight that the paper uses the thematic analyses to identify themes in the text corpus. We identified the themes based on terms derived from the text corpus as suggested in the study by Chabi et al., (2011). The following depicts the data collection and data analyses of this hybrid approach.

3.1 Text Mining and Data Collection

This study used text mining methods to uncover motivational factors associated with e-cigarette use in online forums; resulting in unstructured text data collected for analysis. In other words, software tools actively engaged in the process of extracting information important to audience from different text sources (Lazard, et al., 2016). Lazard et al. (2016) acknowledged that the central challenge of text mining is to study unstructured data and discover meaningful associations and patterns from a large quantity of written messages.

Inspired by previous research (Chen, Zhu, & Conway, 2015; Zhan, Liu, Li, Leischow, & Zeng, 2017), the data of this study was acquired from three online forums including Reddit, Vapingunderground and e-cigarette-forum. Reddit is popular among young people while other two forums are dedicated to e-cigarette. These popular forums were selected to acquire a general sense of what the nature of discussion is like regarding reasons for e-cigarette use as well as users' attitude towards e-cigarette use. This study used open source web scraper to collect data after searching "why do you vape" in these online forums. According to Kim et al. (2016), data from recent time frame enhances the quality of the study as perceptions and attitude towards specific topics are up to date. Consequently, contents between January 2014 and September 2017 were then downloaded and stored in a SQL database. A total of 790 comments and discussions relevant to e-cigarette use and motivations to use e-cigarette were scraped and stored. The pages of discussion contents included basic data such as dates, user IDs and member levels. Due to ethical concerns, this study will not present user IDs.

3.2 Data Analysis

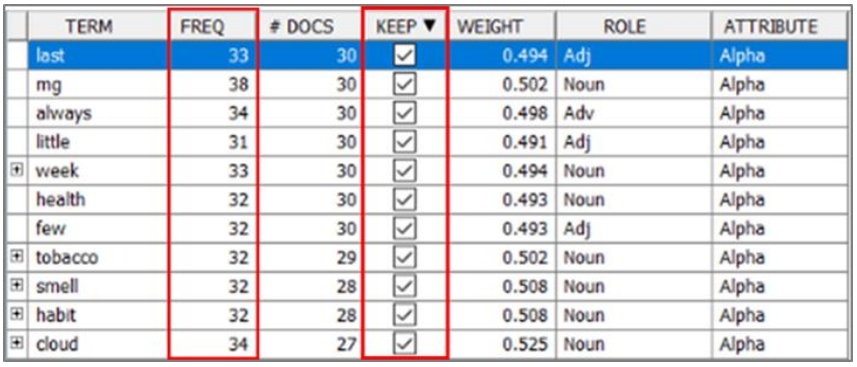
This study conducted textual content analysis using SAS Text Miner. We performed several computationally expansive steps such as text parsing, filtering, transformation and mining supported with manual coding and analysis implementation. These steps enabled us to uncover terms related to reasons for e-cigarette use from a large amount of discussion contents now called the text corpus. During text parsing, we extracted, stemmed and filtered words using a natural language processor in the SAS Text Miner programme. The tokenization process took place on the text corpus and terms were derived together with the frequencies (number of occurrence the terms appear in the text corpus). The terms frequency result is shown in Figure 1.

Terms

TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
find	59	54	☑	0.407	Verb	Alpha
more	59	53	☐	0.0	Adv	Alpha
no	58	52	☐	0.0	Adv	Alpha
smoker	58	50	☑	0.419	Noun	Alpha
one	58	49	☐	0.0	Num	Alpha
say	57	54	☐	0.0	Verb	Alpha
too	57	50	☐	0.0	Adv	Alpha
feel	57	49	☑	0.424	Verb	Alpha
still	57	49	☐	0.0	Adv	Alpha
give	53	50	☐	0.0	Verb	Alpha

Figure 1: Terms generated from text parsing process

The text parsing process also involves the exclusion of stopwords such as prepositions, pronouns and auxiliary verbs. By eliminating extraneous terms, the text filter process kept relevant and meaningful terms that were shown by the frequency of occurrence in the dataset. The increased signal-to-noise ratio enhanced the quality measure of dataset (Kim, Huang, & Emery, 2016). Then in the filtering process, we manually and inductively discarded terms that were irrelevant to reasons for e-cigarette use. As a result, the term-by-document matrix was created (see Figure 2).



TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
last	33	30	☑	0.494	Adj	Alpha
mg	38	30	☑	0.502	Noun	Alpha
always	34	30	☑	0.498	Adv	Alpha
little	31	30	☑	0.491	Adj	Alpha
⊞ week	33	30	☑	0.494	Noun	Alpha
health	32	30	☑	0.493	Noun	Alpha
few	32	30	☑	0.493	Adj	Alpha
⊞ tobacco	32	29	☑	0.502	Noun	Alpha
⊞ smell	32	28	☑	0.508	Noun	Alpha
⊞ habit	32	28	☑	0.508	Noun	Alpha
⊞ cloud	34	27	☑	0.525	Noun	Alpha

Figure 2: Term-by-document matrix

Text clustering was then performed to partition terms into mutually exclusive groups. These groups revealed relevant themes related to reasons for e-cigarette use. The following depicts how the themes were derived as mentioned in section 3. The cluster frequency root mean square (RMS) standard deviation (Std.) test was used to test the accuracy of the clusters with a threshold value of zero (see Table 1). The RMS Std. test shows that 11 meaningful clusters were derived, and these clusters were manifested into specific themes based on the terms depicted in Table 2. Since the terms were derived based on the frequency of occurrence, the thematic analyses procedure was robust; justified by low RMS Std. values respectively in Table 1.

Cluster ID	Cluster Description	Frequency	Percent	RMS Std
3	+year +start vaping +cig +quit +old +day last +cigarette +stop +smoke +month +late back +week	132	17%	0.1202
2	+juice money +spend +buy +people +pack +flavour +long +hobby +guy +bad +life down +story people	104	13%	0.1309
1	cool +feel +look +cost +point +good +taste +hand +alternative +habit +end +switch pretty +reason +enter	88	11%	0.1281
4	nicotine fun mg +friend weed nic +pen 'a bit' +relax +help +keep vape +anxiety +buy +reason	83	11%	0.129
7	+quit +want girlfriend +smoke 'cold turkey' cold turkey +ashtray +enter +pick +kill +decide +smell +thing +die	79	10%	0.1115
5	+expand +know +easy +cigar +small +pipe +big +side home +alternative vapor +month +face pretty +hit	76	10%	0.1258
8	+high +concentrate +bong +flower access +oil herb +hit +prefer combustion +smooth +rip cleaner vape +session	63	8%	0.117
6	+enjoy 'fun hobby' pax nicotine +love +anxiety +hobby +addict +help fun +relax +trick down +drive +flavour	48	6%	0.1192
11	+hookah hand +cancer +switch +bring father 'lung cancer' +smoker +lung second +die family +tire +tobacco mg	46	6%	0.1252
10	+blow +cloud +'cigarette smoke' +trick +'smoke trick' +smoke +face +drive +people +find vapor +life fun +story +happy	44	6%	0.1169
9	+read +lung +chemical +inhale completely 'next day' +throw +gain +shit people health +bad +happy +guy	27	3%	0.1193

Table 1: Summary cluster frequency and root mean square standard deviation

From the clusters derived, using the Cartesian coordinate system we plotted these clusters to depict the distance between them. The Cartesian coordinate diagram is a two-dimensional space with X and Y axis as a result of the clustering process. Each cluster has coordinates that represent its position in the Cartesian diagram. The closer the clusters are to each other, the more related they are in terms of their semantic concepts. Using this concept, we recorded a

series of coordinates of each cluster using the Singular Value Decomposition (SVD) algorithm in the clustering process. These coordinates became the data points of the clusters for the Cartesian coordinate system. The resolution of the SVD was set at 50 resulting in 50 coordinates for further analysis. A Confirmatory Factor Analysis (CFA) was conducted to understand how the clusters manifested into respectively constructs and 3 constructs were confirmed. To further elaborate on this process, description of each theme was depicted in Table 2. Based on the description of these themes, the authors manifested them into the 3 constructs which were *Intention to Quit Smoking, Cost and Convenience* and *Enjoyment.*

Cluster ID	**Theme**	**Description to the Themes**	**Construct**
3	Quit Smoking 1	Intention to quit smoking due to personal desire	Intention to Quit Smoking
2	Cost, Flavour, Hobby	Attitude towards e-cigarettes related to cost, flavour and habitual notions	Cost and Convenience
1	Cost	Attitude towards e-cigarettes related to cost	Cost and Convenience
4	Enjoyment – Relax1	Motivation to use e-cigarettes based on enjoyment and relaxation	Enjoyment
7	Quit Smoking 2	Intention to quit smoking due to family and peer pressure	Intention to Quit Smoking
5	Convenience	Attitude towards e-cigarettes related to the variety of e-cigarettes products and their accessibility	Cost and Convenience
8	Enjoyment – Taste	Motivation to use e-cigarettes based on enjoyment and taste	Enjoyment
6	Enjoyment – Relax2	Motivation to use e-cigarettes based on enjoyment and relaxation	Enjoyment
11	Health - Diseases	Intention to quit smoking due to health-related reasons	Intention to Quit Smoking
10	Smoke / Tricks	Motivation to use e-cigarettes based on fun and enjoyment when doing tricks using e-cigarettes	Enjoyment
9	Health - Toxic	Intention to quit smoking due to health-related reasons	Intention to Quit Smoking

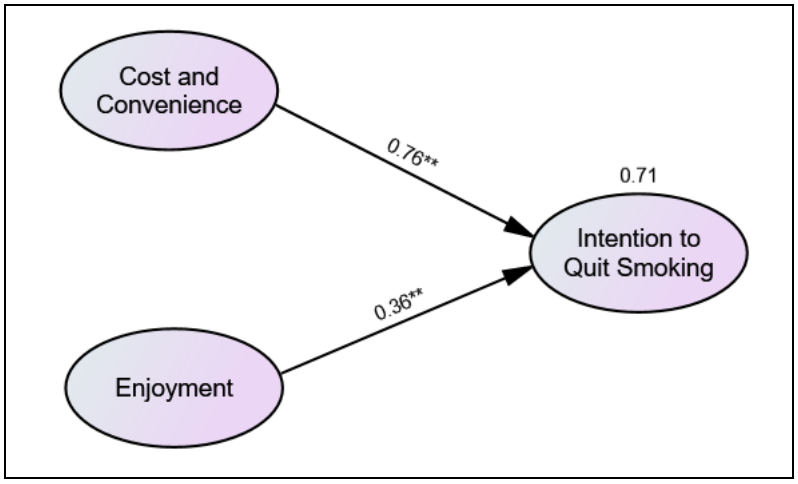
Table 2: Cluster description and construct identification

From these constructs, we used Structural Equation Modelling (SEM) to construct a predictive model. The following shows the results in Figure 3 and Table 3.

Constructs	Cluster IDs / Variables	Regression Weights
Cost and Convenience	1	0.709**
	2	0.723**
	5	0.709**
Enjoyment	4	0.845**
	6	0.627**
	8	0.385**
	10	0.587**
Intention to Quit Smoking	3	0.831**
	7	0.727**
	9	0.529**
	11	0.587**

Note: ** denotes regression weights are significant at 0.05

Table 3: SEM results in a glance



Note: ** denotes regression weights are significant at 0.05

Figure 3: SEM results in a path diagram

Referring to Table 3, the construct *Intention to Intention to Quit Smoking* was manifested by clusters that concerned with health-related issues like the toxicity of smoking, diseases that came with smoking, trying to quit smoking for loved ones, taking up e-cigarettes as a substitute for smoking, etc. *Cost and Convenience* was manifested by clusters that were related with the ease of use of e-cigarettes, convenience to switch to e-cigarettes, feeling and looking cool, spending less on cigarettes, etc. *Enjoyment* was manifested by clusters that were related with the fun and love in using e-cigarettes, enjoyable experience of doing neat tricks with smoke and vapour, relaxation when using e-cigarettes, reducing anxiety and destress with e-cigarettes, etc. Based on Figure 3, the construct *Intention to Quit Smoking* was explained by *Cost and Convenience* and *Enjoyment.* The total variance explained was 71%. *Cost and Convenience* and *Enjoyment* were significantly associated with *Intention to Quit Smoking.* We further performed Bayesian SEM to train the model using Markov Chain Monte Carlo (MCMC); a machine learning algorithm. The model converged at approximately 82,000 iterations but we allowed the programme (SPSS AMOS) to carry on to approximately 154,000 iterations. At this point, the convergence statistics was at 1.0026 and the acceptance rate was optimal at 0.24 implying that MCMC was generating new parameter values 24% of the time while repeating previous parameter values 76% of the time. The MCMC standard error was 0.025 and below indicating that the posterior distribution reflected the population distribution.

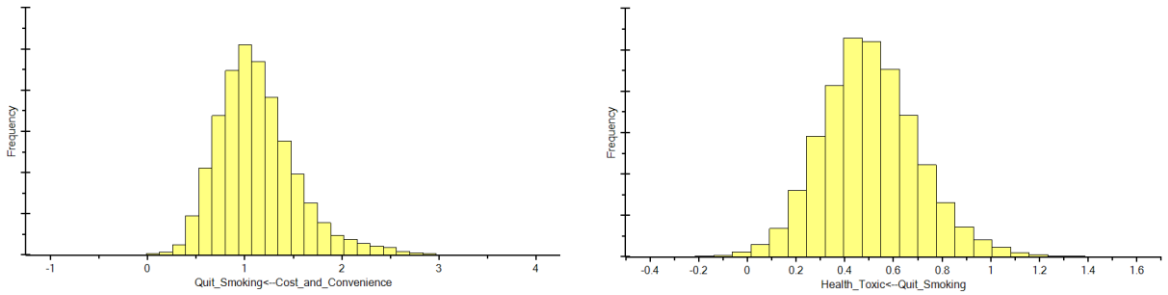


Figure 4: Posterior distribution of the associations based on the path diagram

4. Findings

Of the 790 discussion contents captured from three online forums, 5 distinctive themes were sorted using text clustering. The themes featured the prevalent terms that were produced by SAS text miner. A description of clusters was made up by relevant words. The frequency of occurrence of each theme was represented by numbers. The percentage of each theme

indicated how popular that topic was among online forum participants. Of these 5 themes, the most popular topic in the analysis of reasons for e-cigarette use was quit smoking. More than 36% of word occurrences were related to quit smoking. Comments showed that people used e-cigarettes as cessation devices to help them quit smoking (e.g. “I really don't have a reason other than quitting smoking.” “To quit smoking, but at the moment I've been vaping and slowly smoking cigarettes less”). The second topic depicted that the enjoyment of vaping motivated people to use e-cigarette (e.g. “vaping is just an enjoyable hobby for me.” “I find it enjoyable. Doing tricks, blowing clouds, hanging out with other vapers.”). This topic made up of 25% of word of occurrences. Hobby and e-cigarette flavours were included as well. The third popular topic for initiating e-cigarette use was cost which contained alternative compared to traditional tobacco products (e.g. “The fact that it's much cheaper once you get over the initial investment is definitely a plus as well.”). This reason was represented by 24% of word frequency. The fourth topic on why people started using e-cigarette was convenience (e.g. “It’s easy and fast to set up and put away, and cleaning is really infrequent. You simply plug it in and wait for it to heat up, perfect time to prepare the bowl, and then you are good to go...whenever you are done just unplug and put away.”), represented by 10% of word frequency. Smoke/tricks was the fifth topic regarding reasons for e-cigarette use (e.g. “Nice sensation from blowing clouds, different flavours and that slight nicotine hit.”). The remaining themes were similarly the repeat of top five themes. A list of themes within the context is in Table 4.

Theme from Text Clustering	**Percentage**
Quit Smoking / Health	36%
Enjoyment	25%
Cost	24%
Convenience	10%
Smoke / tricks	6%

Table 4 Top five themes of reasons for e-cigarette use

The Bayesian SEM results showed that *Cost and Convenience* and *Enjoyment* are constructs that affect *Intention to Quit Smoking. Enjoyment* seemed to have a more robust effect on users who would like to quit smoking compared to *Cost and Convenience*. This is shown in Figure 4 where the posterior distribution depicted that the impact of *Enjoyment* on *Intention to Quit Smoking* was always positive. Additionally, the regression weights (0.76) of *Enjoyment* on *Intention to Quit Smoking* was stronger than the regression weights (0.36) of *Cost and Convenience* on *Intention to Quit Smoking.*

5. Discussion and Conclusion

This study was conducted to explore motivational factors for e-cigarette use by applying text mining and predictive modelling methods. The results uncovered top 5 reasons for e-cigarette use including quite smoking and health related reasons, enjoyment, cost, convenience and Smoke/tricks. Majority of forum participants mentioned that they started using e-cigarette to quit smoking traditional cigarettes. According to the online discussion contents, e-cigarettes helped people quit smoking and they implied that their health had improved since then. In other words, participants perceived e-cigarette as harm reduction alternative compared to traditional cigarettes. Some reported that they had successfully cut down the nicotine levels in their body after vaping. Others described promising health benefits such as recovering from bulimia. Although e-cigarette maybe regarded as replacement to traditional cigarettes, it may not be an appropriate way to promote e-cigarettes in the public (Kavuluru & Sabbir, 2016). Given that the evidence is still being tested by clinical studies, negative aspects of using e-

cigarette including side effects and heathier alternative remained unknown. Next to the main reason for e-cigarette use was enjoyment of vaping. Majority reported that their satisfaction in using e-cigarettes was due to their ability to control e-cigarettes consumption to suit their needs and preferences (Simmons, et al., 2016). Some participants found that the appealing factor of e-cigarette as a "toy" or a form of entertainment and hobby. This study findings also suggested that cost was another appealing factor and the reason for e-cigarette use by online forum users. They believed that e-cigarettes were economically feasible as vaping was cheaper than traditional smoking. This is consistent with previous research (Hummel, et al., 2015; Sumner, McQueen, Scott, & Sumner, 2014). Besides the above often cited reasons for e-cigarette use, convenience was found to be another factor that motivated forum participants to initiate e-cigarette use. The results also suggested that participants considered e-cigarette use as convenient because they can use it in places where traditional smoking was banned. The freedom to use e-cigarettes in these places became appealing to many users (Chen, Zhu, & Conway, 2015). Smoke/tricks was identified as final reason to use e-cigarette in this study. Participants mentioned that they liked to do smoke tricks. It was a distinctive feature of e-cigarette that resembled waterpipe smoking (Grant & O'Mahoney, 2016).

Based on the literature review regarding reasons for e-cigarette use, majority of the survey and interview results suggested that curiosity, quit smoking, enjoyment and convenience were the top four factors motivating people to use e-cigarettes (Cooper, Harrell, & Perry, 2016; Etter, 2016; Hummel, et al., 2015; Li, Newcombe, & Walton, 2015; Surís, Berchtold, & Akre, 2015). Compared to the text mining and CFA results generated from this study (i.e. quit smoking, enjoyment, cost and convenience), it is notable that curiosity did not appear in the findings. Such minor inconsistencies are common when different methodologies are deployed Aside from curiosity, reasons for e-cigarette use such as quit smoking, enjoyment, cost and convenience were in line with previous research (Etter & Bullen, 2011; Hummel, et al., 2015; Kong et al., 2015). Consistent with the literature, the predictive model via Bayesian SEM also exhibited encouraging results showing significant associations between enjoyment, cost and convenience on quit smoking.

5.1 Contributions

The major strength of this study is the ability to analyse a compilation of unstructured data from multiple online sources rather than a random sampling from traditional data collection process like surveys or focus groups. In addition, predictive analysis was conducted by expanding on previous established methods in relation to thematic analysis of social data. The use of MCMC as a probabilistic machine learning for deep learning can enable a large amount of data to be analysed with less manpower over a greater period of time (Cole-Lewis, et al., 2015). More exploratory studies on health-related topics can use this method to identify patterns and trends from social media, forums and other online platforms. Using a series of processes as discussed in section 3.2, we used the unstructured data in the text corpus to construct a predictive model on users' perceptions on the notion of quitting smoking enjoyment, cost and convenience of using e-cigarettes. The results in this paper showed that the hybrid approach in encapsulating both inductive and deductive reasoning can work on unstructured data. This may add an additional dimension to existing text mining method by assessing social media, forums and other online platforms for trends and insights (Cole-Lewis, et al., 2015). In other words, researchers and public health professionals can obtain valuable information from social media platforms or any online user-generated contents that are related to public health even when the volume of data is big and scattered. Moreover, these findings scraped from massive and diverse online forums can guide the public in their decision-makings related to health issues. Findings of this study can be used to create and

design relevant public health campaigns to facilitate changes in public health behaviour. Results show that enjoyment, cost and convenience in using e-cigarettes are important motivation factors to encourage people to quit smoking. It is an opportunity for public health professionals to actively engage in social media, forums and other online platforms conversations to monitor trends and address misunderstandings related to e-cigarette use.

This study contributes to public health research field by identifying several motivational factors for initiation of e-cigarette use. We assessed and compared the results with previous findings using surveys and interviews. To the best of our knowledge, it is the first study which investigated the differences of reasons for e-cigarette use identified from various research methods. Another major contribution made by this study is the use of predictive modelling and probabilistic machine learning on unstructured data on e-cigarette use. It provided an opportunity for public health advocates to create a balanced conversation across social media, forums and other online platforms.

5.2 Limitations and future research

There is a wide range of opinions regarding e-cigarette use and the data collected in this study was from Reddit, Vapingunderground and e-cigarette-forum. However, social media platforms such as Facebook can also be useful data for further analysis. Another limitation of this study is the problems of bots, fake accounts and spams from the data sources. Hence, future studies could be conducted on data sources with higher level of accuracy (Kim, Miano, Chew, Eggers, & Nonnemaker, 2017). It would be challenging but meaningful to develop filters to rule out unreliable data.

References

Barbieri, D., Braga, D., Ceri, S., Della Valle, E., Huang, Y., Tresp, V., Rettinger, A. & Wermser, H. (2010). Deductive and inductive stream reasoning for semantic social media analytics. *IEEE Intelligent Systems*, 25(6), 32-41.

Bullen, C., Howe, C., Laugesen, M., McRobbie, H., Parag, V., Williman, J., & Walker, N. (2013). Electronic Cigarettes for Smoking Cessation: a Randomised Controlled Trial. *The Lancet*, 382(9905), 1629-1637.

Chabi, A. H., Kboubi, F., & Ahmed, M. B. (2012). Thematic analysis and visualization of textual corpus. *International Journal of Information Sciences and Techniques*, 2(1), 53-63.

Chen, A., Zhu, S.-H., & Conway, M. (2015). What Online Communities Can Tell Us About Electronic Cigarettes and Hookah Use: A Study Using Text Mining and Visualization Techniques. *Journal of Medical Internet Research, 17*(9), e220.

Cole-Lewis, H., Pugatch, J., Sanders, A., Varghese, A., Posada, S., Yun, C., Schwarz, M. & Augustson, E. (2015). Social Listening: A Content Analysis of E-Cigarette Discussions on Twitter. *Journal of Medical Internet Research, 17*(10), e243.

Cooper, M., Harrell, M. B., & Perry, C. L. (2016). Comparing Young Adults to Older Adults in E-cigarette Perceptions and Motivations for Use: Implications for Health Communication. *31*(4), 429-438.

Etter, J.-F. (2016). Throat Hit in Users of the Electronic Cigarette: An Exploratory Study. *Psychology of Addictive Behaviours, 30*(1), 93-100.

Etter, J.-F., & Bullen, C. (2011). Electronic Cigarette: Users Profile, Utilization, Satisfaction and Perceived Efficacy. *Addiction, 106*(11), 2017-2028.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods, 5*(1), 80-92.

Gallaire, H., Minker, J., & Nicolas, J. M. (1984). Logic and databases: A deductive approach. *ACM Computing Surveys (CSUR), 16*(2), 153-185.

Grana, R. A., Popova, L., & Ling, P. (2014). A Longitudinal Analysis of E-cigarette Use and Smoking Cessation. *JAMA Internal Medicine, 174*(5), 812-813.

Grant, A., & O'Mahoney, H. (2016). Portrayal of Waterpipe (Shisha, Hookah, Nargile) Smoking on Twitter: A Qualitative Exploration. *Public Health, 140*(1), 128-135.

Huang, J., Kornfield, R., Szczypka, G., & Emery, S. L. (2014). A Cross-sectional Examination of Marketing of Electronic Cigarettes on Twitter. *Tobacco Control, 23*(3), 26-30.

Hummel, K., Hoving, C., Nagelhout, G. E., de Vries, H., van den Putte, B., Candel, M. J., Borland, R., & Willemsen, M. C. (2015). Prevalence and Reasons for Use of Electronic Cigarettes among Smokers: Findings from the International Tobacco Control (ITC) Netherlands Survey. *International Journal of Drug Policy, 26*(1), 601-608.

Kavuluru, R., & Sabbir, A. (2016). Toward Automated E-cigarette Surveillance: Spotting E-cigarette. *Journal of Biomedical Informatics, 61*(1), 19-26.

Kim, A., Hopper, T., Simpson, S., & Porter, L. (2015). Using Twitter Data to Gain Insights into E-cigarette Marketing and Locations of Use: An Infoveillance Study. *Journal of Medical Internet Research, 17*(11), e251.

Kim, A., Miano, T., Chew, R., Eggers, M., & Nonnemaker, J. (2017). Classification of Twitter Users Who Tweet about E-Cigarettes. *Journal of Medical Internet Research, 3*(3), e63.

Kim, Y., Huang, J., & Emery, S. (2016). Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. *Journal of Medical Internet Research, 18*(2), e41.

Kinouani, S., Castéra, P., Laporte, C., Pétrègne, F., & Gay, B. (2016). Factors and Motivations Associated with Use of E-cigarette among Primary Care Patients in a Prospective Cohort Study: E-TAC Study Protocol. *BMJ Open, 6*(1), 1-6.

Kong, G., Idrisov, B., Galimov, A., Masagutov, R., & Sussman, S. (2016). Electronic Cigarette Use Among Adolescents in the Russian Federation. *Substance Use & Misuse*, 1-8.

Kong, G., Morean, M. E., Cavallo, D. A., Camenga, D. R., & Krishnan-Sarin, S. (2015). Reasons for Electronic Cigarette Experimentation and Discontinuation among Adolescents and Young Adults. *Nicotine & Tobacco Research, 17*(7), 847-854.

Lazard, A., Saffer, A. J., Wilcox, G. B., Chung, A., Mackert, M. S., & Bernhardt, J. M. (2016). E-Cigarette Social Media Messages: A Text Mining Analysis of Marketing and Consumer Conversations on Twitter. *Journal of Medical Internet Research, 2*(2), e171.

Li, J., Newcombe, R., & Walton, D. (2015). The Prevalence, Correlates and Reasons for Using Electronic Cigarettes among New Zealand Adults. *Addictive Behaviours, 45*(1), 245-251.

Myslín, M., Zhu, S.-H., Chapman, W., & Conway, M. (2013). Using Twitter to Examine Smoking Behaviour and Perceptions of Emerging Tobacco Products. *15*(8), e174.

Sharma, R., Wigginton, B., Meurk, C., Ford, P., & Gartner, C. E. (2017). Motivations and Limitations Associated with Vaping among People with Mental Illness: A Qualitative

Analysis of Reddit Discussions. *International Journal of Environmental Research and Public Health, 14*(1), 7.

Simmons, V., Quinn, G. P., Harrell, P. T., Meltzer, L. R., Correa, J. B., Unrod, M., & Brandon, T. H. (2016). E-cigarette Use in Adults: A Qualitative Study of Users' Perceptions and Future Use Intentions. *Addiction Research & Theory, 24*(4), 313-321.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research (2nd ed.). Newbury Park, CA: Sage

Sumner, H. M., McQueen, A., Scott, M. J., & Sumner, W. (2014). Analysis of Comments in a Petition Defending Electronic Cigarettes. *Nicotine & Tobacco, 16*(11), 1503-1511.

Surís, J.-C., Berchtold, A., & Akre, C. (2015). Reasons to Use E-cigarettes and Associations with other Substances among Adolescents in Switzerland. *Drug and Alcohol Dependence, 153*(1), 140-144.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation, 27*(2), 237-246.

Torii, M., Tilak, S. S., Doan, S., Zisook, D. S., & Fan, J.-w. (2016). Mining Health-Related Issues in Consumer Product Reviews by Using Scalable Text Analytics. *Biomed Inform Insights, 8*(1), 1-11.

Zhan, Y., Liu, R., Li, Q., Leischow, S. J., & Zeng, 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.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems,* 25(6), 13-16.