Inez Wibowo
Professor Anagha Kulkarni
CSC 699 Independent Study
May 24, 2019

Smoking Norms on Social Media with Topic Modeling

In this report, we present results from Topic Modeling on a set of smoking queries using Gensim library and pyLDAvis on the social media platform Reddit. We focus on a specific time period that overlap with the qualitative data we've received from StreetWyse about smoking norms in a few bay area zip codes. This topic modeling results provides a quantitative look at smoking norms and a deeper look into the key terms we have become familiar with and uncovers additional topics not previously considered. We will first describe how the queries were selected, from previously annotated list of "smoking" terms on Twitter. Then how these terms were modified and chosen to inform queries for our search on Reddit. Lastly, we will review the results of the topic modeling Gensim library, and interpreting the pyLDAvis graphs to better understand which terms are the most helpful to understanding the 10 topics that were generated within the observed time frame of January 1, 2018 - February 28, 2018. We found that topic modeling on a rich set of data like Reddit has allowed us to better understand existing keywords like "cigarette smoking" as well as identify the rise of new terms such as "vaping mods".

Within the time frame of January 1st through February 28th, we found 10,010 posts on Reddit related to smoking from a set of smoking keywords. This is our document, "corpus", which we then fed into the LDA Gensim topic modeling algorithm. Originally, we began with a majority of traditional cigarette smoking keywords. However, there has been a recent shift in the tobacco/nicotine market, which is the fact that e-cigarettes had become a more popular product in 2017 compared to traditional cigarettes among middle and high school students (Wang et al. 1). Therefore, we wanted to also capture postings on Reddit with e-cigarette terms. We limited the LDA request to group 10 topics, in order to lay a foundation for understanding the model as well as the topics.

The methodology we used to choose the original set of smoking keywords was to use an online thesaurus and focused on traditional "cigarette smoking". We annotated the results between

two members of the team and independently chose a handful of queries based on relevancy and activity on Twitter during the observed time frame. The overlapping accepted keywords were then appended with additional e-cigarette queries and was used in this semester to implement on Reddit.

Using Reddit's Pushshift API, we were able to list of a mix of unigrams and polygrams that are relevant to smoking and e-cigarettes. The following is a list of the keywords we used to query Reddit: "cigarette smoking", "smoking cigarette", "vaping", "ecig", "e-cig","juul", "chain smok", "parliament lights", "lucky cig", "laramie cig", "kamel red", "cancer stick", "camel crush", "lung dart", "fake smoker", "cowboy killers","burn one", "kamel red", "marlboro menthol", "marlboro lights", "lucky strikes", and "joe camel"

As was expected, we do see from Gensim's topic modeling results distribution that "cigarette", "ecig" and "vaping" terms were prevalent in the document because we had searched for these exact keywords. However, we discovered "adjacent" terms, and topics, previously not considered in our research. For example, Topic ID 4, which we have identified as having a "Vaping Mods" theme. The top terms that was found helpful to understand this topic were "mod", "com", "batteri", "vapor", "amp", "tank", "kit". These terms aptly describe the various apparatus Vaping Mods offer to be customized and yield high level of nicotine without coughing or taste of traditional tobacco products (Stumacher). It has been referred to as the "third-generation cigarette" (e-cigarettes: Facts). The largest vaping mod manufacturer is Juul, which is a keyword in our search, which could explain the inclusion of vaping modification in the corpus. However, without topic modeling we would not have been able to see the relatively high level of prevalence of "Vaping Mods" in relative to other more tradition smoking and vaping topics in the document.

Of the 10 topics that were returned by the Gensim LDA model (full list in Appendix 2), utilizing pyLDAvis for visualization, we can view the top terms that is found helpful to understand the topic and adjust for relevance, a metric from 0 to 1 (λ,), which allows the user to modulate lift (λ = 0) and probability (λ = 1). The meaning of "relevance" is, as defined by Sievert and Shirley,"… a weighted average of the logarithms of a term's probability and lift..." of the term within the topic. "Lift" was defined in Taddy (2011) as a way to rank the terms within the topic by the ratio of the topic's probability within the topic, and to its probability across the entire document. "Probability" as defined by Blei and Laffert (2009) provide users with extra information about the usage of terms within the topics (Sievert and Shirley 3). The impact on the terms listed by modulating the relevance value is that by not only using Probability but also accounting for Lift, which on its own, decreases the

rankings of globally frequent terms and gives high ranking to very rare terms in one topic, is that we find the optimal list of terms to help our understanding of each topic.

The below table (Table 1) is a highlight of the top five topics that have the highest prevalence within the document. We also provide annotated themes associated to each topic to better understand the topic (Sievert and Shirley 1). As discussed above, “Smoking sentiment” and “Vaping sentiment” was expected, however “Vaping Mods” topic presented a new smoking norm that could inform future research.

Table 1

Topic ID	Topic percentage	Themes
1	29.30%	Smoking sentiment
2	28.30%	Vaping sentiment
3	19.00%	Assets and Metadata
4	11.30%	Vaping Mods (Modification)
5	3.60%	Sites used to buy and sell

Looking into the five most prevalent topics within the document, when accounting for probability, in Table 2, we have provided each topic’s Top 10 terms that are most useful in interpreting the topic and thus most relevant.

Table 2

	Top terms for Topic 1: Smoking Sentiment	Terms	Top terms for Topic 2: Vaping sentiment	Terms	Top terms for Topic 3: Assets and Metadata	Terms	Top terms for Topic 4: Vaping Mods	Terms	Top terms for topic 5: Sites used to buy and sell	Terms
1	1.60%	smoke	5.00%	vape	4.00%	tr	3.10%	mod	11.00%	comment
2	1.10%	like	1.90%	use	4.00%	com	2.30%	com	9.00%	xymarket
3	1.00%	time	1.50%	like	4.00%	lti	2.00%	batteri	7.60%	paypal
4	0.90%	feel	1.40%	get	3.50%	imgur	1.60%	vapor	1.80%	game

5	0.90%	day	1.20%	new	3.50%	png	1.50%	amp	1.80%	amazon
6	0.80%	get	1.10%	look	3.30%	origin	1.40%	tank	1.70%	usa
7	0.70%	go	1.00%	coil	2.30%	aegi	1.40%	kit	1.60%	gc
8	0.60%	want	1.00%	juic	1.70%	misc	1.30%	imgur	1.40%	card
9	0.60%	year	1.00%	good	1.60%	anvil	1.20%	coil	1.40%	usa
10	0.60%	high	0.80%	tri	1.50%	dynam	1.10%	rda	1.10%	Offer

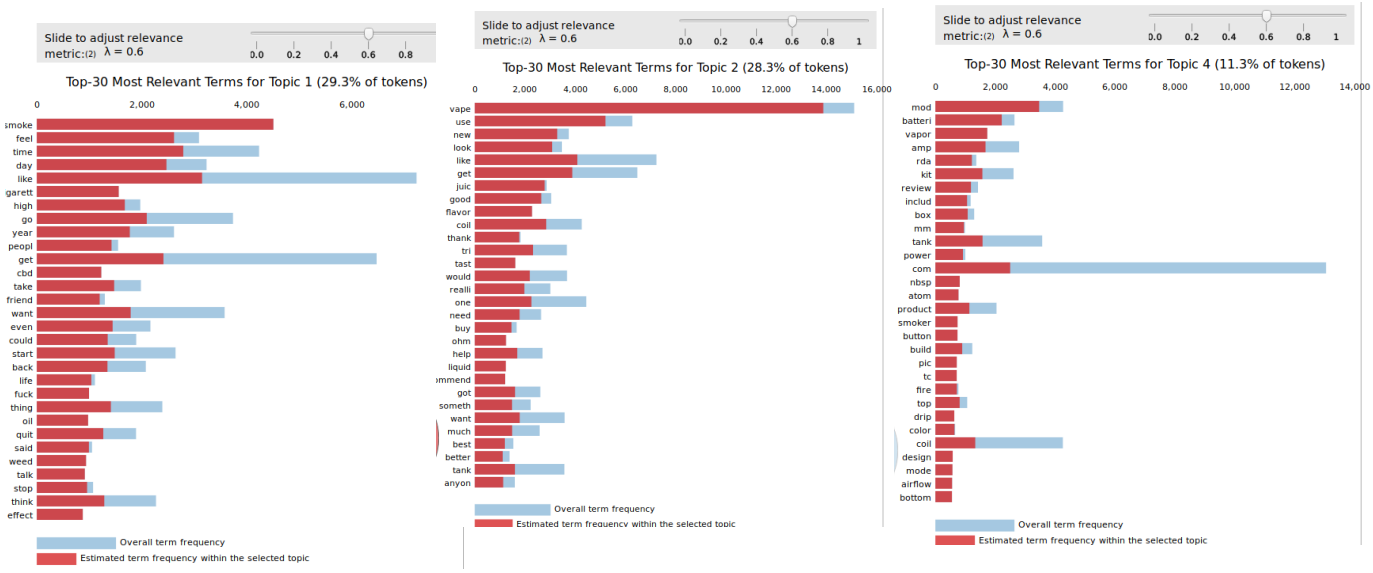
Although the above Table 2 utilized the relevance (λ) value as 1 for its rankings, it means that the terms are ranked for the probability of the term within the topic, and not accounting for the ratio of the topic within the entire corpus, such that would be provided if we included lift, we can already start to understand how the topics were grouped together. “Smoking sentiment” and “Vaping sentiment” has “smoke” and “vape” unigrams, respectively as the highest probability of the terms occurring in that topic. Additionally “mod” is the term with the highest probability for “Vaping Mods.”

To modify relevance value, we can use LDAvis tool to adjust the λ value slider to 0.6 as it is recommended by the creators of LDAvis, Sievert and Shelley. They have determined that this is an optimal lambda value that can be applied to any data set, to help aid interpretation of the topic. Table 3 shows the comparison between the relevance values. I’ve highlighted notable differences in the terms ranked.

Table 3

	Topic 1 : Smoking sentiment		Topic 2: Vaping Sentiment		Topic 4 : Vaping Mods	
	Terms λ = 1	Terms λ = 0.6	Terms λ = 1	Terms λ = 0.6	Terms λ = 1	Terms λ = 0.6
1	smoke	smoke	vape	vape	mod	mod
2	like	feel	use	use	com	batteri
3	time	time	like	new	batteri	vapor
4	feel	day	get	look	vapor	amp
5	day	like	new	like	amp	rda
6	get	cigarett	look	get	tank	kit
7	go	high	coil	juic	kit	review
8	want	go	juic	good	imgur	includ
9	year	year	good	flavor	coil	box
10	high	peopl	tri	coil	rda	mm

Figure 1



In the above Table 3, we've highlighted notable sorting shifts when the relevance value was modified to 0.6, 60% Probability and 40% Lift.

In "Smoking sentiment" topic, "cigarett" and "high" are now ranked higher and does provide more context to the topic. This is not only a smoking topic, but specifically a topic that includes Reddit posts about smoking cigarettes, as well as include terms about getting high – a term typically used for describing marijuana smoking.

In the "Vaping sentiment" topic collection, we see "juic" and "flavor" as being rated higher than previously when the lift was not included, when lambda was 1. These terms add to the understanding around vaping that the liquid which lends vaping the flavor is called "juice". A quick search on Google, will yield many online website pedaling "Vape Juice" or "E-Juice" with a variety of flavors for sale.

Lastly in "Vaping Mods" topic, "rda" was ranked higher, "com" which is part of the website address is no longer in the top 10, and "mm" and "box" has moved up into the top 10. "rda" is a Rebuildable Drip Atomizer, an RDA, in which the user builds their own vaping system, the coil, the wick, batteries, etc. (What are the difference between RBA, RDA, and RTA?). "mm" is a metric used to describe the diameter of the base of a vape tank, and "box" is typically larger kind of vaping device, and has higher levels of customization.

The above observations show the benefit of modulating for both Probability and Lift. The terms ranked higher in the relevance λ value of 0.6 ranked meaningful terms and lowered visibility of non-relevant terms such as “com”, or “go”, or “get”.

The pyLDAvis also provides an Intertopic Distance Map, Figure 2, which computes the distance between topic circles centers to show relationship between each topics. Smoking and Vaping sentiments show close correlation, while surprisingly Vaping mods is a bit further. Further iterations of the topic modeling with a wider set of keywords will need to be done to generate more graphs to confirm why Vaping Mods is not more closely correlated to Vaping than one would intuitively expect.

Figure 2

We have presented the results from topic modeling using Gensim library which yielded 10 topics, of which three were meaningful. We named them “Smoking sentiments”, “Vaping sentiments” and “Vaping Mods”. The Gensim LDA library allowed us to identify “Vaping Mods” as a new source of terms that could inform future queries for the smoking norms research.

With the assistance of pyLDAvis we were also able to visualize the data and gain better understanding of which terms contributed highly to helping us understand these top three terms. When comparing the top 10 terms with a different relevance metric, 100% Probability v. 60% Probability & 40% Lift, with the relevance slider as provided in the LDAvis interactive tool, we are able to gain

better insight in the form of new terms that can contribute to future smoking norms research as new keywords to explore, i.e. “juice”, or “rda”, etc.

We have laid the foundation for discovery of relevant terms and topics in this semester’s study. We have shown that the topic modeling using Reddit’s Pushshift, stemming using Snowball Stemmatizer, suppressing stop words with NLTK and appending observed stop words, has yielded a good start for a corpus to be ingested by the LDA model. The coherence score of 0.468 for 10 topics is considered quite good. We can improve this by testing different LDA algorithm like Mallet, or continuing to increase the topic numbers and plotting the topic to coherence score until we find the optimal value of topics for this search. There is also more work to do in terms of refining the corpus to include geolocation so that we can compare smoking/vaping norms between underrepresented communities and control communities in the SF Bay Area. Lastly, with this data collected, we hope we can help visualize smoking norms in various communities with a large number of topics in a compact and understandable way.

Works Cited

"E-cigarettes: Facts, stats and regulations" *Truth Initiative*, truthinitiative.org, 19 July, 2018, https://truthinitiative.org/news/e-cigarettes-facts-stats-and-regulations .

Sievert, Carson, and Kenneth Shirley. "LDAvis: A method for visualizing and interpreting topics" *Proceedings of the workshop on interactive language learning, visualization, and interfacess,* pages 63–70, Baltimore, Maryland, USA, June 27, 2014. Association for Computational Linguistics. https://www.aclweb.org/anthology/W14-3110

Stumacher, Richard. "Pod mods and vaping are creating a new generation of youths addicted to nicotine" *STAT, statnews.com*, 21 September,2018,https://www.statnews.com/2018/09/21/pod-mods-vaping-nicotine-addiction/

Wang, Teresa W et al. "Tobacco Product Use Among Middle and High School Students - United States, 2011-2017." *MMWR. Morbidity and mortality weekly report* vol. 67,22 629-633. 8 Jun. 2018, doi:10.15585/mmwr.mm6722a3,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5991815/

"What are the differences between RBA, RDA and RTA?", *Vaping.com*, 2019 Vaping.com, https://vaping.com/differences-rba-rda-rta