'Sweeter Than a Swisher': amount and themes of little cigar and cigarillo content on Twitter

Ganna Kostygina,[1] Hy Tran,[1] Yaru Shi,[2] Yoonsang Kim,[1] Sherry Emery[1]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/tobaccocontrol-2016-053094).

[1]National Opinion Research Center, The University of Chicago, Chicago, Illinois, USA
[2]School of Public Health, University of Illinois at Chicago, Chicago, Illinois, USA

Correspondence to
Dr Ganna Kostygina, Senior Research Scientist, Health Media Collaboratory, NORC at the University of Chicago, 55 E Monroe Street, 3150L, Chicago, IL 60603, USA; kostygina-anna@norc.org

Received 31 March 2016
Revised 14 July 2016
Accepted 15 July 2016

ABSTRACT

Objective Despite recent increases in little cigar and cigarillo (LCC) use—particularly among urban youth, African-Americans and Latinos—research on targeted strategies for marketing these products is sparse. Little is known about the amount or content of LCC messages users see or share on social media, a popular communication medium among youth and communities of colour.

Methods Keyword rules were used to collect tweets related to LCCs from the Twitter Firehose posted in October 2014 and March–April 2015. Tweets were coded for promotional content, brand references, co-use with marijuana and subculture references (eg, rap/hip-hop, celebrity endorsements) and were classified as commercial and 'organic'/non-commercial using a combination of machine learning methods, keyword algorithms and human coding. Metadata associated with each tweet were used to categorise users as influencers (1000 and more followers) and regular users (under 1000 followers).

Results Keyword filters captured over 4 372 293 LCC tweets. Analyses revealed that 17% of account users posting about LCCs were influencers and 1% of accounts were overtly commercial. Influencers were more likely to mention LCC brands and post promotional messages. Approximately 83% of LCC tweets contained references to marijuana and 29% of tweets were memes. Tweets also contained references to rap/hip-hop lyrics and urban subculture.

Conclusions Twitter is a major information-sharing and marketing platform for LCCs. Co-use of tobacco and marijuana is common and normalised on Twitter. The presence and broad reach of LCC messages on social media warrants urgent need for surveillance and serious attention from public health professionals and policymakers. Future tobacco use prevention initiatives should be adapted to ensure that they are inclusive of LCC use.

INTRODUCTION

Little cigars and cigarillos (LCCs) are an understudied domain in tobacco control and are particularly interesting because of the strategic and targeted marketing used to promote these products to youth and communities of colour.[1–8] Although smoking rates in the USA have decreased, the recent declines in cigarette consumption are offset by sharp increases in the consumption of other tobacco products including cigarillos.[9] These products are increasingly aggressively marketed on the internet (eg, through social media) and at the point of sale and may serve as a means to introduce youth to tobacco.[8 10] By definition, cigarillos are slimmer versions of a large cigar and weigh 3–10 lb per 1000 cigars; little cigars weigh not more than 3 lb per 1000 cigars; they resemble cigarettes but are wrapped in tobacco leaf rather than paper.[11] Cigar smoke contains the same toxic and carcinogenic constituents found in cigarette smoke and may cause oral, laryngeal, oesophageal, lung cancer, heart disease, aortic aneurysm, etc.[11] Users often inhale LCC smoke and thus absorb it into their lungs and bloodstream.[11]

Although the US 109 Food and Drug Administration's decision in 2016 extended regulatory authority to all tobacco products, including LCCs, these products are not currently subject to many of the regulations on cigarette sales and advertising.[12 13] Unlike cigarettes, for example, cigars can still be sold in flavours and in packs of fewer than 20. This lack of regulation could provide an opportunity for the industry to market cigars more aggressively in the USA. Nationally, cigar smoking is the second most common form of tobacco use among youth.[9 14] Use of these products is also disproportionately high among youth, young adults and people of colour.[8] Furthermore, LCCs are often used as vehicles for marijuana consumption (a process called 'blunting' where the tobacco is hollowed out of a cigar and replaced with marijuana).[15] Blunt smokers often identify themselves as marijuana users but not as tobacco users, which may have led to underestimates of population of LCC consumption.[15–17] Thus, in recent years, there has been an increased scientific interest in the relationships between tobacco and marijuana use among youth and young adults in regard to the direction of uptake pathways. Since both substances are typically smoked, tobacco and marijuana use may support and reinforce use of each other.[18] Recent evidence suggests the emergence of a reverse gateway mechanism, where marijuana use precedes tobacco smoking and can lead to nicotine dependence.[19–22] Marijuana users may be an emerging target for the tobacco industry marketing.[23]

Characterising the role of new media platforms in tobacco product marketing and counter marketing is critically important as these platforms largely remain under the radar of tobacco control policymakers and are not currently covered by the advertising restrictions that apply to outdoor and television advertising. In fact, social media have become a major marketing platform for tobacco products.[24–26] New communication technologies offer alternative means for gathering and managing information, which are not present in traditional media and provide high brand visibility for tobacco products.[27] The emergence of new technologies has resulted in profound changes in the media

To cite: Kostygina G, Tran H, Shi Y, *et al*. *Tob Control* 2016;**25**:i75–i82.

landscape and has led internet users to encounter a vast amount of online information exposure, including social exposure on social-networking sites, such as Twitter. Twitter is particularly important as this platform is disproportionately popular among hard-to-reach populations traditionally at risk for tobacco use, such as youth and communities of colour.[28] Use of this platform is increasing. According to the 2015 Pew Research Center Report, 23% of online adults use Twitter, compared to 16% in 2012.[28] In 2015, roughly 38% of all Twitter users used the site daily.[28] Approximately 28% of Twitter users are black, 28% are Hispanic and 32% are aged 18–29, which is consistent with the fact that youth, African-Americans and Latinos in general use social media at higher rates than the general population in the USA. Twitter popularity among these groups is growing,[28] and it has come to play a major role in the life experience of American youth and ethnic minorities.

Prior research has shown that there is an influx of tobacco and nicotine product promotion on social media, with ~5 million messages about electronic cigarettes/vaping products posted on Twitter, by 1.2 million unique accounts over a 1 year period.[29] Social exposure to this content contributes to normalisation and glamorisation of smoking and may influence the spread of smoking behaviours via individual's social networks (followers/friends and followers of followers/friends of friends).[30] Social media use provides greater speed of information retrieval and higher level of media control, making it easier for consumers to actively search for, produce, block and retransmit tobacco-related information, including product marketing.[27 31] Consequently, selective information exposure and transmission processes may allow social media users to establish an information filter 'bubble' in which tobacco use is portrayed as a normal acceptable behaviour and becomes part of shared, in-group experience.[31] Thus, tobacco-related messages on social media may lead to tobacco use initiation through such mechanisms as social learning or modelling of behaviours[32] and socialisation into peer groups.[33–37] Indeed, portrayal of tobacco and alcohol misuse is becoming a common activity to network about on Twitter.[25 29 38 39]

In addition to product promotion, social-networking sites are also used by the tobacco industry and their allies to influence public opinion on tobacco control policy decisions.[40] Therefore, there is an urgent need to develop communication theory and technology, as well as programming infrastructure for active engagement with user-generated tobacco-related content on social media (ie, surveillance, labelling, filtering) to enable potential regulation of commercial advertising messages on these platforms as the methodological base for systematic audit of the tobacco-related content on these sites is lacking. Discovering how LCCs are marketed online and on social media has important and direct relevance to potential FDA regulations for these products.

Our study fills these research gaps by using cutting-edge statistical and computational methodologies to analyse LCC-related Twitter posts. While a recent study by Step *et al* (2016) analysed a sample of 288 LCC-related tweets, to the best of our knowledge, there has been no prior comprehensive systematic research on the magnitude of LCC message exposure and sharing on Twitter or marketing strategies used to promote tobacco products on this social-networking site.[41] For the purpose of this study, we collected data on the amount and variety of LCC-related information that smokers and non-smokers are exposed to and post on Twitter and conducted analyses to identify major sources and themes of LCC content. We used the message content and related metadata to investigate product preferences (eg, brand and flavour), behaviour (purchase and use context), social norms (eg, subculture frames, peer group references) and product marketing strategies.

METHODS

Data acquisition and processing

The present study is based on tweets filtered by 70 keywords related to LCCs over the period of 3 months (October 2014, March–April 2015). We purchased LCC-related tweets from Gnip (http://www.gnip.com), the official Twitter data provider. A tweet was included in the data set if it matched one or more of the keyword rules (eg, brands: swisher OR swishers, swisha OR swishas, splitarillo OR splitarillos; product names, including slang terms: rello OR rellos, rillo OR rillos, blunt-'james blunt'-'emily blunt'-st_blunt-'too blunt'-'be blunt') (see online supplementary appendix 1 for a complete list of search rules). Keyword rules were chosen based on the trends (eg, through use of http://www.topsy.com that showed the volume of relevant Tweets over the past 30 days as well as examples of actual Tweets containing the searched keyword), prior literature and research team expert consensus based on knowledge of LCC-related terminology and brands.[42] We used Boolean rules rather than individual keywords to make our search filter more efficient, minimise the amount of irrelevant tweets captured and reduce the number of duplicates. The Gnip Historical Powertrack delivered a collection of posts (in .json format) containing one or more search terms; the resultant data were stored in a NoSQL database, MongoDB, and cleaned using python programming language to create analytic data.

Training samples and machine learning classifiers

Relevance

To assess whether the captured tweets were relevant to LCCs, accurately measure the volume of the social conversation about LCCs and determine trends, we estimated the retrieval precision (the proportion of the data relevant to the LCC topic) and retrieval recall (the amount of all relevant conversation captured) of the keyword filters used to gather data.[43] For this purpose, we first built a machine learning classifier based on a human-coded training sample. Two coders rated a random sample of 5124 tweets (the sample was stratified by the search rule) as relevant and non-relevant to LCCs. The two coders achieved notably high agreement ($\alpha=0.95$) on an overlap sample of 600 tweets. This human-coded sample was used to train the machine learning classifier to clean the entire corpus of the tweets. Machine learning is data-driven analytic approach in which computational systems develop algorithms based on a training set (a subset of the data) to determine prediction of outcomes in a separate, test data set.[44] The goal of supervised learning is generalisation to unseen data,[45] that is, developing a model that allows to map unseen observations to one of the human labels.[46] If a model performs well in predicting outcomes for the test data set, it may predict well for the rest of the database. Hence, this approach allows to reliably automate large data classification. After comparing several machine learning methods including Naïve Bayes algorithm, logistic regression and linear support vector machine (SVM) classifier, linear SVM with L1-norm regularisation was selected due to its high performance. Ten-fold cross-validation was utilised to test the accuracy of the classifier.[47] Classifier accuracy was 0.95, classifier recall (sensitivity) was 0.96 and classifier precision (positive predictive value) was 0.96 (F1=0.96). The machine classifier performance was further tested with additional human coding of 1040 tweets (test data set) to confirm that the good classifier performance is not a coincidence due to the parameter set-up in

the classifier training but a good fit of the whole population, we took an additional random sample of the raw data to check the accuracy of machine classifier result against human labels (95%). In addition to classifier precision and recall, we estimated retrieval precision and recall following the suggestions by Kim *et al*.[43] Retrieval precision was approximated by classifier precision (96%). Computation of retrieval recall involves non-retrieved (although relevant) tweets, that is, LCC-relevant tweets that do not contain the LCC keyword rules in the denominator. To determine retrieval recall, we randomly sampled 4000 tweets that do not contain the LCC keyword rules (ie, tweets relevant to other products) from our database, which contains a total number of over 21 million tobacco and smoking-related tweets, and found that 4% of these 4000 non-retrieved tweets were relevant to LCC. Therefore, our retrieval recall of the keyword filter was estimated to be 87%, suggesting that our keyword filters retrieved about 87% of all relevant tweets in our larger tobacco-related corpus.

Content coding

A similar iterative process of combining human coding and machine learning was used to classify all collected tweets based on the themes of interest to informing policy and public health, namely, commercial/promotional content and co-use of marijuana and tobacco content.

First, we classified the relevant tweets as either organic or commercial. Organic tweets were those deemed non-sponsored; they reflected individual opinions or experiences or linked to non-promotional content. Commercial tweets were defined by the presence of any of the following: branded promotional messages; URLs linking to commercial websites; usernames indicating affiliations with commercial sites or user's Twitter page consisting only of promotional tweets (ie, spammer accounts). Two human coders reviewed all tweets posted by a sample of 3000 Twitter accounts (intercoder reliability was high: α=93%). Human codes were used to train linear SVM machine classifier. Classifier accuracy was validated using 10-fold cross-validation. Classifier accuracy was 0.97, classifier recall (sensitivity) was 0.92 and precision (positive predictive value) was 0.92 (F1=0.92). Machine classifier was further tested with additional human coding of tweets posted by 343 accounts; 97% of the test data set was correctly classified.

In addition, we classified the relevant tweets as those referencing co-use of tobacco and marijuana and those referencing tobacco use only (ie, LCCs). Co-use tweets contained any reference to marijuana. Specifically, these posts were defined by the presence of any references to using LCCs for the purpose of making blunts (ie, hollowed-out cigars filled with marijuana leaf), any terms referring to marijuana strains, marijuana slang terms such as loud, green, purp and mid and any references to being under the influence ('high') due to marijuana use. Tobacco use only tweets contained references to LCC use exclusively. Two human coders rated a sample of 2670 tweets (inter-coder reliability was high: α=95%). Resultant codes were used to train the linear SVM classifier; 10-fold cross-validation was applied to assess classifier performance. Classifier accuracy was 0.98, classifier recall (sensitivity) was 0.99 and precision (positive predictive value) was 0.99 (F1=0.99). Machine classifier was further tested with additional human coding of 185 tweets; 98% of the test data set was correctly classified.

Metadata

Metadata associated with each tweet were used to examine the characteristics of accounts tweeting about LCCs. Thus, such user-level information as the number of followers was analysed to measure the potential reach of collected LCC-related tweets. We defined potential reach as the total times tweets were posted or the sum of followers for all tweets.

Furthermore, we utilised account metadata to categorise users as influencers (1000 and more followers) and regular users (under 1000 followers). As tweets posted by influencers have greater potential reach compared to messages posted by regular users, we assessed whether there were substantive differences in LCC-related content posted by these groups.

Since we analyse the entire population of LCC-related tweets posted over the 3-month period rather than a sample, we directly interpret the proportions of tweets posted by the two groups of users without statistical hypothesis testing. Furthermore, due to large data size, any null hypothesis will be rejected and p value from any statistical test will be close to 0.

Keyword algorithms

Keyword algorithms were used to search the tweets to assess the frequency with which they mentioned specific brands, price-related, music-themed promotion and other content related to marketing strategies targeting youth and vulnerable populations that may be of interest to informing policymaking, public health and communication research. More specifically, we used keyword strings to quantify the amount of content featuring promotional offers (ie, using strings including money, deal, %, $, save, promo, dollars, discount, coupon, code, price, cost), brand references, popular memes (ie, using strings such as 'hits blunt', 'pass the twitter blunt'), lyrics and subculture references (eg, rap/hip-hop lyrics, celebrity). These themes were selected to help inform future policymaking and interventions to prevent LCC use among youth and other populations at risk.

Topic modelling

To conduct additional exploratory analyses of the meaningful trends in LCC-related conversation, we used latent Dirichlet allocation (LDA) topic modelling—a form of machine learning and a natural language processing tool for identifying patterns of themes or topics in a corpus of unlabelled documents.[48] This is an unsupervised method to discover topics occurring in documents, that is, tweets. A topic may be defined as a cluster of words that frequently appear together. The R package 'mallet' was used to generate the topics by LDA. The number of topics was set to 200, and retweets were excluded from the analysis because retweets may dominate the amount of posts and obscure the topic patterns. The R package 'wordcloud' was used to generate word clouds that visualise topics using the top 200 terms ranked by LDA-generated weights per topic. Given a topic, weights indicate the relative importance of terms; the higher the weight of a term, the more likely a document (ie, tweet) containing this term will belong to this topic. Within a word cloud, a larger font size indicates greater weight and the same colours indicate approximately the same weights. Therefore, word clouds provide a relative gage of how important a word is within a given topic. This visualisation allows the reader to see the most important terms, as well as the less important ones to aid interpretation. On the automatic discovery of 200 topics, the research team reviewed the word clouds displaying the topics and assigned labels for each substantive topic based on its salient terms. After identifying meaningful patterns, the LCC-related topics were grouped into larger categories or archetypes.[48–50]

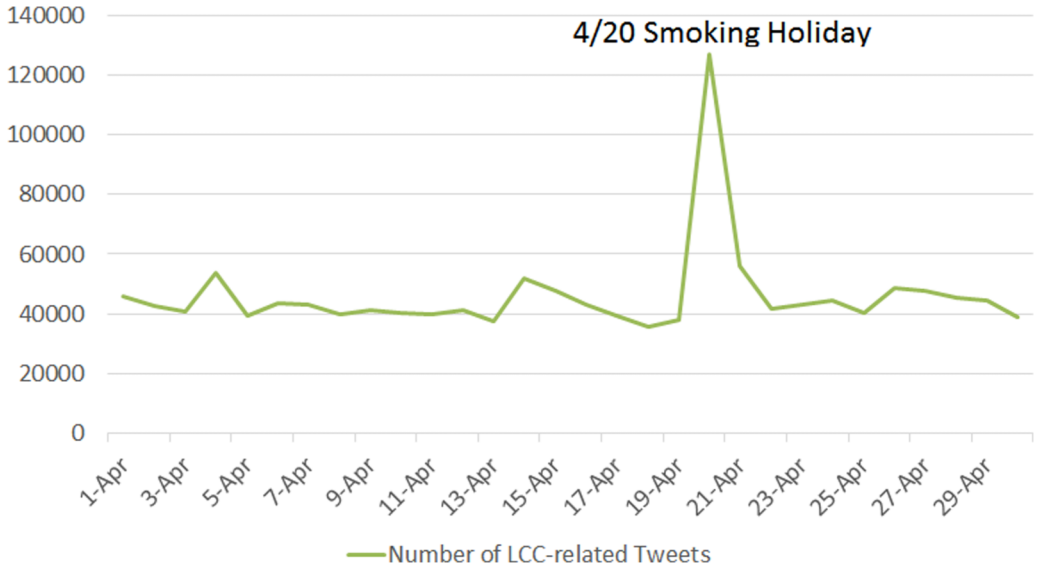


Figure 1 Frequency of little cigar and cigarillo-related tweets over time, 1 April–30 April 2015. LCC, little cigar and cigarillo.

RESULTS

During the data collection period (October 2014, March–April 2015), our keyword filters captured over 4.5 million tweets, of which 4 372 293 were classified as LCC-relevant tweets. These tweets were posted by 1 849 322 individual Twitter users. Our analyses revealed that 1 836 557 accounts posting about LCCs (99%) were organic, and <1% (N=12 765) of accounts were obviously commercial. The overwhelming majority of LCC-related tweets (3 636 176 or 83%) contained references to marijuana. The frequency of LCC-related tweets over time for the month of April 2015 to help illustrate temporal trends in posts is shown in figure 1. There was a sharp increase in LCC-related tweets on the 4/20 'smoking holiday'. An example of a tweet sent on this holiday was from a cigar and cigarillo brand Executive Branch, owned by Snoop Dogg (a famous rapper). Snoop Dogg retweeted the tweet promoting his cigarillo brand, as well as Snoopdogg rolling papers: 'RT @ExecBranch: Only way to celebrate #snoop420 right is with some @snoopdogg rolling papers! Get yours now on [link redacted]'.

Almost one-third of all tweets about LCCs (29%) were memes. An internet meme can be defined as an element of a culture (ie, activity, concept, catchphrase or piece of media) which spreads, often as mimicry, from person to person via the internet.[51–53] LCC-related memes were predominantly humorous tweets containing links to or embedding images, videos or vines referencing blunt smoking behaviour. One of the most popular memes were the 'hits blunt' tweets, containing questions a smoker would ask after smoking a blunt. An example of a popular retweet featuring this meme was: '*Hits blunt* If I go see the Grand Canyon do I actually see the Grand Canyon? #PSAT [link redacted]' (retweeted 4323 times). Other examples of popular retweets featuring blunt-related memes were

> When you're so high you roll your homie up into the blunt [link redacted] (10406 retweets or RTs)
>
> When you hit the blunt too hard [link redacted] (7468 RTs)
>
> *Girl hits blunt once* Changes twitter name to Flower Child *wears huff socks**listens to Bob Marley* (5778 RTs)
>
> Joints or Blunts? [link redacted] (6300 RTs)

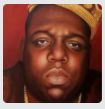
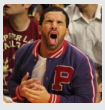
Table 1 shows the characteristics of LCC-related tweets collected, presented in total and separately for influencers versus regular users. We found that ~17% of account users posting the LCC content were influencers, these users had a potential reach

Table 1 Characteristics of influencer and regular user tweets

Tweet type	Total	Influencer tweets (≥1000 followers)	Regular user tweets (<1000 followers)
LCC tweets	4 372 293	1 035 515 (23%)	3 336 778 (77%)
Users	1 849 322	318 893 (17%)	1 530 429 (83%)
Reach	2 379 854 236	1 868 926 085 (79%)	510 928 151 (21%)
Mentions			
Brands	300 813 (7%)	88 256 (9%)	212 557 (6%)
Promotions	113 637 (3%)	33 845 (3%)	79 792 (2%)
Marijuana	3 636 176 (83%)	817 732 (79%)	2 818 444 (85%)
Retweets	2 121 928 (49%)	444 311 (43%)	1 677 617 (50%)
Memes	1 276 834 (29%)	232 897 (23%)	1 043 937 (31%)

LCC, little cigar and cigarillo.

Table 2 Examples of frequently retweeted accounts on little cigars and cigarillos

Top mentioned celebrity and community accounts	Number of followers	Co-use	Age warning	Sample tweets/mentions
 **KOE** @wizkhalifa	23 800 000	Yes	No	RT @wizkhalifa: Hearing Drake say roll up a cone means the world may stop smoking blunts soon. Thank goodness. RT @wizkhalifa: Forever smokin joints over blunts.
 **Snoop Dogg** @SnoopDogg	13 200 000	Yes	No	RT @SnoopDogg: Sometimes I hit a blunt to relax from hittin a blunt RT @ExecBranch: Only way to celebrate #snoop420 right is with some @snoopdogg rolling papers! Get yours now on [link redacted]
 **Weed Tweets™** @stillblazingtho	1 210 000	Yes	Yes (18+)	RT @stillblazingtho: It's legal to get drunk and act like an idiot but God forbid you smoke a blunt and eat a pizza. RT @stillblazingtho: Fav if you need a blunt [link redacted]
 **Stoner Nation™** @ThaStonerNation	483 000	Yes	No	RT @ThaStonerNation: In a terrible mood. *hits the blunt* 'Dude, I love life' RT @ThaStonerNation: I love how one blunt can change my mood from 0-420 real quick.
 **Marijuana Posts™** @MarijuanaPosts	460 000	Yes	No	RT @MarijuanaPosts: If smoking weed offends you: 1. I'm sorry. 2. It won't happen again. 3. 1 &2 are lies. 4. *hits blunt* 5. *hits blunt* RT @MarijuanaPosts: RT if you need a blunt rn
 **Rappers Said** @RappersSaid	445 000	Yes	No	RT @RappersSaid: When Tupac said "I smoke a blunt to take the pain out and if I wasn't high, I'd probably try to blow my brains out."
 **High Ideas** @ReallyHighIdeas	369 000	Yes	No	RT @ReallyHighIdeas: RT to pass the Twitter blunt and get everyone high af [link redacted] RT @ReallyHighIdeas: I need a blunt the size of a burrito.
 **Stoner Vines** @StonerVines	302 000	Yes	No	RT @StonerVines: How to pass the blunt like a boss [link redacted] RT @StonerVines: The never ending blunt [link redacted] RT @StonerVines: He hit the blunt [link redacted]
 **Happy Campers** @HappyCampersTHC	234 000	Yes	No	RT @HappyCampersTHC: How to deal with stress: 1. Roll a blunt 2. Smoke it 3. Repeat steps as necessary RT @HappyCampersTHC: Obama hit the blunt once now he's like [link redacted]
 **Rapper Reactions** @RapperReact	224 000	Yes	Yes (18+)	RT @RapperReact: You may think you're hard but you're not "Rihanna rolling a blunt on her bodyguard's head at Coachella" hard

of 1 868 926 085 impressions (or 79% of total potential reach). Influencers were ~30% more likely to mention specific LCC brands and 33% more likely to post promotional messages. Regular users were more likely to retweet messages, to post marijuana-related content and to post tweets featuring memes.

Table 2 lists examples of popular influencer accounts whose tweets referencing LCCs were frequently retweeted, and it also includes examples of popular retweets of messages posted by these accounts.

Influencer users included three major groupings: rap or hip-hop celebrities such as Snoop Dogg, Wiz Khalifa, Drake; rap community accounts (eg, Rappers Said, Rapper Reactions); and marijuana user or 'stoner' communities (eg, Stoner Nation, Happy Campers, Weed Tweets, Marijuana Posts, High Ideas, Stoner Beauties, Intelligent Stoners, Life as a Stoner, Stoner Chicks, StonerXpress, Stuff Stoners Like, etc). These influencer accounts had a large number of followers (for instance, Stoner Nation/@TheStonerNation was followed by over 480 000 users) and posted content on tobacco and marijuana co-use (ie, blunt use), however, it is noteworthy that most of these accounts did not feature any age restrictions or health warnings.

Topic modelling

Figure 2 illustrates the results of the exploratory topic modelling analysis. As mentioned above, we set the number of topics to be 200. The resultant topics could be generally grouped into four major categories or archetypes: (1) product-related messages, including references to specific brands, flavours, cigarillos or blunt size, quality and burning speed; (2) marijuana references, including co-use of tobacco and marijuana; marijuana slang terms and strains; (3) smoking behaviour—purchase, intention to smoke and buy and (4) normative and cultural context references, including memes, rap/hip-hop lyrics, birthday, subcultures, work, school, celebrities, music, etc. Topic modelling captured a number of hip-hop lyrics quotes referencing LCCs, for example, lyrics by such rappers as Wiz Khalifa, Drake, Chief Keef, Tupac, were captured as individual topics. Although retweets were excluded from the analysis, memes represented a significant proportion of the topics (eg, 'hits blunt' and 'when you hit the blunt too hard' appeared among the 200 topics). This grouping appears to be a sizeable part of the corpus. We generated word clouds from the weights of top 200 terms within each LCC-related topic, and figure 2 shows sample word

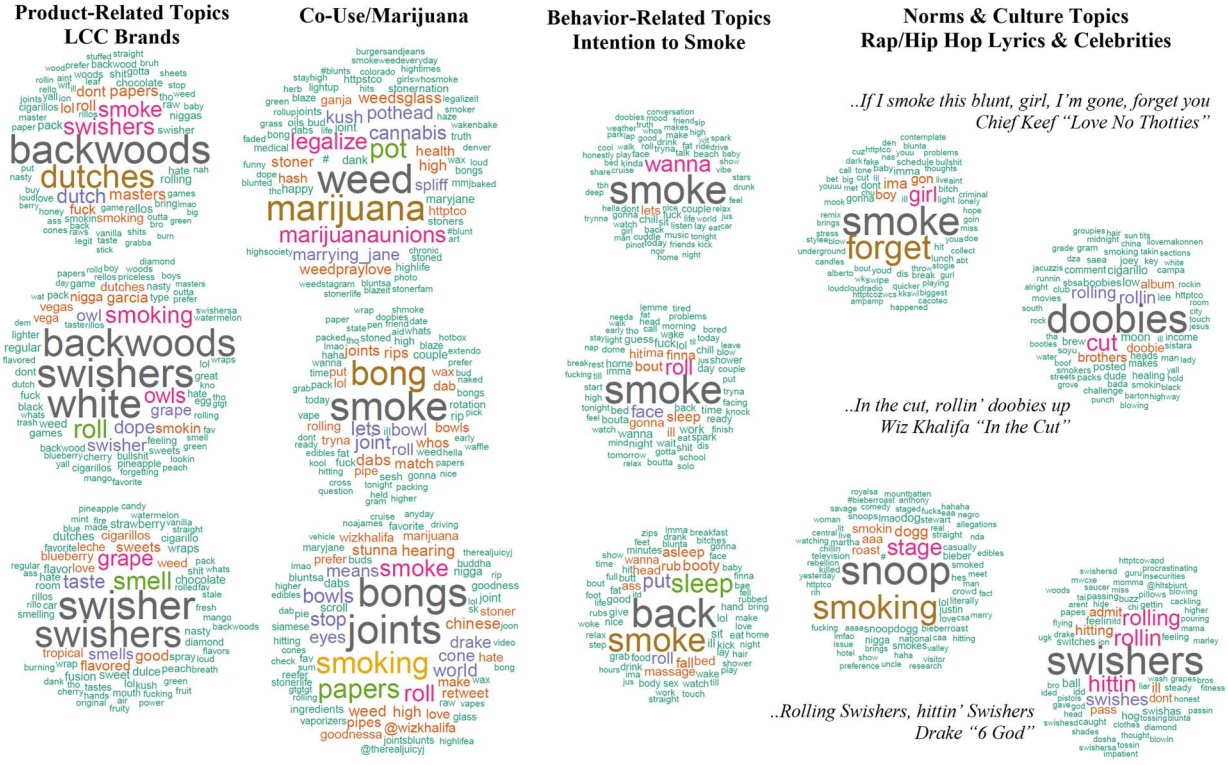


Figure 2 Major archetypes of LCC-related topics. Word clouds generated from the weights of top 200 terms within a topic. Terms with larger weights are depicted in larger font size. Terms with approximately the same weights are of the same colours. LCC, little cigar and cigarillo.

clouds for each of the four archetypes. Overall, the majority of LCC-related topics appeared to be organic as very few topics featured explicit promotional messages.

SUMMARY AND DISCUSSION

The role of social media platforms, such as Twitter, in tobacco control is an emerging area of research. Our work builds on a growing body of the literature that uses mixed-method computational approaches to assess health trends from digital media[24–26 38 39 54] and adds to prior literature by demonstrating that Twitter appears to be a major messaging and information-sharing social platform for LCCs. In fact, our study is first to quantify the overall presence of LCC-relevant content on Twitter. The amount of tobacco-related messages has grown rapidly in recent years; for instance, estimates of e-cigarette-related content range from 1.7 million tweets over a 5-year period between 2008 and 2013[55] to nearly 5 million posts captured in 2013,[29] suggesting a sharp increase in the popularity of e-cigarettes and marketing efforts. Our study provides further evidence that there is an influx of tobacco-related conversation on Twitter. Our keyword filters captured over 4 372 293 LCC-related tweets, which is more than twice as many tweets over a 3-month period as the number of posts about e-cigarettes over a 5-year time frame described by Kim *et al*. The majority of the LCC-related conversation was organic and referenced LCC and marijuana co-use in the form of blunts. The ostensibly organic nature of the conversation and the large proportion of memes (nearly 30%) may indicate that open discussion of LCC use appears to be a normalised, generally accepted/popular activity on Twitter. LCCs are becoming part of popular culture. Although the majority of LCC-related topics appeared to be organic, with few topics featuring overt commercial content, we found that influential users, such as rap/hip-hop celebrities, were more likely to mention specific LCC brands and post promotional messages about LCCs. These findings indicate that prior strategies used to identify commercial content on social media may not be effective in capturing integrated industry promotional tactics, for example, across tobacco and music industries. Further, these findings suggest that tobacco promotion via brand references/product placement is blurring the boundary between rap music, the hip-hop culture and the tobacco industry, which also allows for active consumer engagement, for example, through event promotion, contests stimulating organic user-generated content. Such a focus on developing personal two-way relationships with consumers allows product promoters to integrate into organic user social networks and foster high-speed viral/word-of-mouth information dissemination.[56–59] These findings are consistent with LCC and cigarette promotion strategies starting in the 1970s, such as integration with lifestyle trends, music-themed promotion and use of celebrities.[60 61]

We found that ~83% of LCC-related tweets referenced marijuana co-use. This finding confirms that LCCs cannot be considered without the context of marijuana use and may be primarily used for rolling blunts. For instance, recent data indicate that up to 90% of marijuana users are concurrent tobacco smokers and that tobacco use among marijuana users may be seriously underreported.[19 62] Marijuana users are becoming an emerging target for the tobacco industry marketing and numerous accounts including celebrity accounts, such as Snoop Dogg, rapper and 'stoner' community accounts on Twitter post and retweet promotional content on tobacco. Further research is needed to understand the relationships between tobacco and marijuana promotion and use among at-risk populations and the direction of uptake pathways.

Our study has several limitations and raises questions for future research. As with any social media research, our conclusions depend on the validity of data collected through our search filters. However, we achieved high amount of accuracy

based on testing classifier precision and recall, as well as the retrieval recall of the keyword filter. Our full list of keyword rules is disclosed in online supplementary appendix 1. Another potential limitation has to do with the definition of influencer accounts. Our definition was based on the number of user followers only; future studies should seek to further develop this definition and measure additional features that may have an impact on tweeter influence (eg, number and frequency of posted tweets, number of likes/retweets by followers). Furthermore, while media sources, including social-networking sites, are an important source of influence on tobacco users, individual users of LCCs and social media have complex motives and predispositions that may mediate or moderate behavioural outcomes.

Future research and health campaigns need to address cultural engagement, for example, music-themed promotion, and social media presence of the tobacco industry. Our findings have direct implications for future FDA regulations of LCCs and related products, particularly with respect to marketing restrictions on social media. There is an urgent need for surveillance, monitoring and regulation of social media content relevant to tobacco. Content of Twitter posts is currently not subject to any regulation in regard to health risk disclosure or restriction for underage users. New strategies are needed to protect youth and address the transformation of tobacco advertising into transcendental branding, where the boundaries between marketing and entertainment are indistinguishable. Traditional efforts to restrict the amount of tobacco advertising, its placement and content to protect vulnerable populations cannot address the industry's integrated marketing approaches.

What this paper adds

- Despite recent increases in cigarillo use, research on targeted strategies marketing these products is sparse. Little is known about the amount or content of little cigar and cigarillo (LCC) messages users see or share on social media.
- This study reveals that Twitter is a major information-sharing and marketing platform for LCCs and co-use of tobacco and marijuana is common and normalised on Twitter.
- Tobacco use prevention initiatives should be adapted to ensure they are inclusive of LCC use and social media.

Contributors GK, SE, HT, YK and YS together designed the study; YS and HT conducted cleaning and pre-processing of the data; HT, GK, YK and YS conducted data analysis; SE, GK and HT contributed to data interpretation; GK wrote the first draft; SE and YK revised the draft; the final version of the paper has been reviewed and approved by all four coauthors.

Funding This study was funded by National Cancer Institute (U01CA154254).

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Sterling K, Berg CJ, Thomas AN, *et al*. Factors associated with small cigar use among college students. *Am J Health Behav* 2013;37:325–33.
2 Sterling KL, Moore RS, Pitts N, *et al*. Exposure to celebrity-endorsed small cigar promotions and susceptibility to use among young adult cigarette smokers. *J Environ Public Health* 2013;2013:520286.
3 Richardson A, Ganz O, Vallone D. The cigar ambassador: how Snoop Dogg uses Instagram to promote tobacco use. *Tob Control* 2014;23:79–80.
4 Kozlowski LT, Dollar KM, Giovino GA. Cigar/cigarillo surveillance: limitations of the U.S. Department of Agriculture system. *Am J Prev Med* 2008;34:424–6.
5 Borawski EA, Brooks A, Colabianchi N, *et al*. Adult use of cigars, little cigars, and cigarillos in Cuyahoga County, Ohio: a cross-sectional study. *Nicotine Tob Res* 2010;12:669–73.
6 Singer M, Mirhej G, Page JB, *et al*. Black 'N Mild and carcinogenic: cigar smoking among inner city young adults in Hartford, CT. *J Ethn Subst Abuse* 2007;6:81–94.
7 Patterson F, Lerman C, Kaufmann VG, *et al*. Cigarette smoking practices among American college students: review and future directions. *J Am Coll Health* 2004;52:203–10.
8 Cantrell J, Kreslake JM, Ganz O, *et al*. Marketing little cigars and cigarillos: advertising, price, and associations with neighborhood demographics. *Am J Public Health* 2013;103:1902–9.
9 Centers for Disease Control and Prevention, Tobacco Use Among Middle and High School Students—United States, 2013. *MMWR Morb Mortal Wkly Rep* 2014;63:1021–26.
10 Richardson A, Ganz O, Vallone D. Tobacco on the web: surveillance and characterisation of online tobacco and e-cigarette advertising. *Tob Control* 2015;24:341–7.
11 National Cancer Institute, Cigars: Health Effects and Trends, Smoking and Tobacco Control Monograph No. 9, USDoA. Economic Research Service, Editor. 1998: TTB, Tobacco Statistics. Bethesda, MD: US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, 1998.
12 Zeller M. Three years later: an assessment of the implementation of the Family Smoking Prevention and Tobacco Control Act. *Tob Control* 2012;21:453–4.
13 Food and Drug Administration, Deeming Tobacco Products To Be Subject to the Federal Food, Drug, and Cosmetic Act, as Amended by the Family Smoking Prevention and Tobacco Control Act; Restrictions on the Sale and Distribution of Tobacco Products and Required Warning Statements for Tobacco Products. 2016: Fed Regist. p. 28974–29106. https://federalregister.gov/a/2016-10685
14 Substance Abuse and Mental Health Services Administration, Results from the 2011 National Survey on Drug Use and Health: Detailed Tables, 2012. http://www.samhsa.gov/data/sites/default/files/Revised2k11NSDUHSummNatFindings/Revised2k11NSDUHSummNatFindings/NSDUHresults2011.htm
15 Soldz S, Huyser DJ, Dorsey E. The cigar as a drug delivery device: youth use of blunts. *Addiction* 2003;98:1379–86.
16 Delnevo CD, Bover-Manderski MT, Hrywna M. Cigar, marijuana, and blunt use among US adolescents: are we accurately estimating the prevalence of cigar smoking among youth? *Prev Med* 2011;52:475–6.
17 Lee JP, Battle RS, Lipton R, *et al*. 'Smoking': use of cigarettes, cigars and blunts among Southeast Asian American youth and young adults. *Health Educ Res* 2010;25:83–96.
18 Ramo DE, Liu H, Prochaska JJ. Tobacco and marijuana use among adolescents and young adults: a systematic review of their co-use. *Clin Psychol Rev* 2012;32:105–21.
19 Timberlake DS, Haberstick BC, Hopfer CJ, *et al*. Progression from marijuana use to daily smoking and nicotine dependence in a national sample of U.S. adolescents. *Drug Alcohol Depend* 2007;88:272–81.
20 Ream GL, Benoit E, Johnson BD, *et al*. Smoking tobacco along with marijuana increases symptoms of cannabis dependence. *Drug Alcohol Depend* 2008;95:199–208.
21 Agrawal A, Scherrer JF, Lynskey MT, *et al*. Patterns of use, sequence of onsets and correlates of tobacco and cannabis. *Addict Behav* 2011;36:1141–7.
22 Patton GC, Coffey C, Carlin JB, *et al*. Reverse gateways? Frequent cannabis use as a predictor of tobacco initiation and nicotine dependence. *Addiction* 2005;100:1518–25.
23 Kostygina G, Huang J, Emery S. TrendBlendz: how Splitarillos use marijuana flavours to promote cigarillo use. *Tob Control* Published Online First 6 Apr 2016. doi:10.1136/tobaccocontrol-2015-052710
24 Huang J, Kornfield R, Emery SL. 100 million views of electronic cigarette YouTube videos and counting: quantification, content evaluation, and engagement levels of videos. *J Med Internet Res* 2016;18:e67.
25 Huang J, Kornfield R, Szczypka G, *et al*. A cross-sectional examination of marketing of electronic cigarettes on Twitter. *Tob Control* 2014;23(Suppl 3):iii26–30.
26 Cranwell J, Murray R, Lewis S, *et al*. Adolescents' exposure to tobacco and alcohol content in YouTube music videos. *Addiction* 2015;110:703–11.
27 Cappella JN, Kim HS, Albarracín D. Selection and transmission processes for information in the emerging media environment: psychological motives and message characteristics. *Media Psychol* 2015;18:396–424.
28 Pew Research Center, Mobile Messaging and Social Media 2015. 2015. http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/
29 Binns S, Tran H, Kornfield R, *et al*. A year of electronic cigarette promotion on Twitter. Presented at: *State and Community Tobacco Control Research Initiative Conference*. Chicago, Illinois, September 2014.

30 Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. *Stat Med* 2013;32:556–77.
31 Himelboim I, Smith M, Shneiderman B. Tweeting apart: applying network analysis to detect selective exposure clusters in Twitter. *Commun Methods Meas* 2013;7:195–223.
32 Fishbein M, Yzer MC. Using theory to design effective health behavior interventions. *Commun* 2003;13:164–83.
33 Mead EL, Rimal RN, Ferrence R, *et al*. Understanding the sources of normative influence on behavior: the example of tobacco. *Soc Sci Med* 2014;115:139–43.
34 Sussman S, Pokhrel P, Ashmore RD, *et al*. Adolescent peer group identification and characteristics: a review of the literature. *Addict Behav* 2007;32:1602–27.
35 Moran MB, Sussman S. Translating the link between social identity and health behavior into effective health communication strategies: an experimental application using antismoking advertisements. *Health Commun* 2014;29:1057–66.
36 Christakis NA, Fowler JH, The collective dynamics of smoking in a large social network. *N Engl J Med* 2008;358:2249–58.
37 Slater MD. Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory* 2007;17:281–303.
38 Cavazos-Rehg PA, Krauss MJ, Sowles SJ, *et al*. "Hey Everyone, I'm Drunk." An evaluation of drinking-related Twitter Chatter. *J Stud Alcohol Drugs* 2015;76:635–43.
39 Cole-Lewis H, Pugatch J, Sanders A, *et al*. Social listening: a content analysis of E-cigarette discussions on Twitter. *J Med Internet Res* 2015;17:e243.
40 Step M, Bracken C, Trapl ES, *et al*. User and content characteristics of public tweets referencing little cigars. *Am J Health Behav* 2016;40:38–47.
41 Feng M, Szczypka G, Emery S. Winning Twitter, but losing the election: media campaign lessons from California's Prop 29. *World Association for Public Opinion Research (WAPOR) 68th Annual Conference*; Buenos Aires, Argentina, 2015.
42 Corey CG, Dube SR, Ambrose BK, *et al*. Cigar smoking among U.S. students: reported use after adding brands to survey items. *Am J Prev Med* 2014;47(Suppl 1):S28–35.
43 Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *J Med Internet Res* 2016;18:e41.
44 Murthy K. *Adaptive computation and machine learning: machine learning: a probabilistic perspective*. MIT Press, 2012.
45 Domingos P. A Few Useful Things to Know about Machine Learning. *Commun ACM* 2012;55:78–87.
46 Russell S, Norvig P. *Artificial intelligence: a modern approach*. 2nd edn. Prentice Hall, 2003.
47 Hastie T. *The elements of statistical learning: data mining, inference, and prediction*. 2nd edn. New York, NY: Springer, 2009.
48 Blei DMN, Andrew Y, Jordan MI, *et al*. Latent Dirichlet allocation. *J Mach Learn Res* 2003;3:993–1022.
49 McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet.cs.umass.edu
50 Evans MS. A Computational Approach to Qualitative Analysis in Large Textual Datasets. *PLoS One* 2014;9:e87908.
51 Dawkins R. *The selfish gene*. Oxford University Press, 1989.
52 Yang J, Leskovec J. Patterns of temporal variation in online media. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.
53 Weng L, Menczer F, Ahn Y. Predicting successful memes using network and community structure. *Association for the Advancement of Artificial Intelligence Conference*, 2014.
54 Richardson A, Vallone DM. YouTube: a promotional vehicle for little cigars and cigarillos? *Tob Control* 2014;23:21–6.
55 Kim AE, Hopper T, Simpson S, *et al*. Using Twitter data to gain insights into E-cigarette marketing and locations of use: an infoveillance study. *J Med Internet Res* 2015;17:e251.
56 Li C, Bernoff J. *Groundswell: winning in a world transformed by social technologies*. Boston, MA: Harvard Business Review Press, 2011.
57 Burmann C, Arnhold U. *User generated branding: state of the art of research*. Munster DE: LIT Verlag, 2008.
58 Keller KL. Building strong brands in a modern marketing communications environment. *J Market Commun* 2009;15:139–55.
59 Kozinets RV, De Valck K, Wojnicki AC, *et al*. Networked narratives: understanding word-of-mouth marketing in online communities. *J Market* 2010;74:71–89.
60 Kostygina G, Glantz SA, Ling PM, Tobacco industry use of flavours to recruit new users of little cigars and cigarillos. *Tob Control* 2016;25:66–74.
61 Hafez N, Ling PM. Finding the Kool Mixx: how Brown & Williamson used music marketing to sell cigarettes. *Tob Control* 2006;15:359–66.
62 Timberlake DS. A comparison of drug use and dependence between blunt smokers and other cannabis users. *Subst Use Misuse* 2009;44:401–15.