Heliyon 5 (2019) e02523

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

Exploring the association between problem drinking and language use on Facebook in young adults

Davide Marengo [a], Danny Azucar [b], Fabrizia Giannotta [c,*], Valerio Basile [d], Michele Settanni [a]

[a] Department of Psychology, University of Turin, Turin, Italy
[b] Department of Management, University of Turin, Italy
[c] Division of Public Health, School of Health, Care and Social Welfare, Mälardalen University, Västerås, Sweden
[d] Department of Computer Science, University of Turin, Turin, Italy

ARTICLE INFO

Keywords:
Psychology
Linguistics
Problem alcohol drinking
Social media
Digital footprints
Data mining
Text analysis

ABSTRACT

Recent literature suggests that variations in both formal and content aspects of texts shared on social media tend to reflect user-level differences in demographic, psychosocial, and behavioral characteristics. In the present study, we examined associations between language use on Facebook and problematic alcohol use. We collected texts shared on Facebook by a sample of 296 adult social media users (66.9% females; mean age = 28.44 years (SD = 7.38)). Texts were mined using the closed-vocabulary approach based on the Linguistic Inquiry Word Count (LIWC) semantic dictionary, and an open-vocabulary approach performed via Latent Dirichlet Allocation (LDA). Then, we examined associations between emerging textual features and alcohol-drinking scores as assessed using the AUDIT-C questionnaire. As a final aim, we employed the Random Forest machine-learning algorithm to determine and compare the predictive accuracy of closed- and open-vocabulary features over users' AUDIT-C scores. We found use of words about family, school, and positive feelings and emotions to be negatively associated with alcohol use and problematic drinking, while words suggesting interest in sport events, politics and economics, nightlife, and use of coarse language were more frequent among problematic drinkers. Results coming from LIWC and LDA analyses were quite similar, but LDA added information that could not be retrieved only with LIWC analysis. Furthermore, open-vocabulary features outperformed closed-vocabulary features in terms of predictive power over participants' AUDIT-C scores ($r = .46$ vs. $r = .28$, respectively). Emerging relationships between text features and offline behaviors may have important implications for alcohol screening purposes in the online environment.

1. Introduction

Alcohol use disorder is one of the most common psychosocial disorders in the general population and is associated with personal, societal, and economic costs including poor health outcomes, increased risk of communicable diseases, and criminal behaviors (Esser, 2014; Haberstick et al., 2014; Simons et al., 2014). The World Health Organization's '*2014 Global Status Report on Alcohol and Health*' evidences that in 2012 3.3 million, or 5.9%, of all global deaths were attributable to alcohol consumption (WHO, 2014). WHO's report also highlights that alcohol consumption contributed to over 200 disease and injury-related health conditions, liver cirrhosis, and cancers (WHO, 2014). In order to prevent these negative outcomes of problem drinking, screening for the early identification of heavy drinkers and implementation of preventive interventions for those found to be at risk are required. However, even though many standardized instruments to screen for alcohol misuse exist (e.g., AUDIT-C), performing these screenings remains challenging as many individuals with excessive alcohol use do not seek routine or preventive health care, and oftentimes under-report sensitive or socially undesirable behaviors when screened by healthcare professionals (Gnambs and Kaspar, 2015; Moreno et al., 2012). Further research also highlights that individuals with alcohol misuse rarely seek treatment (Cunningham and Breslin, 2004) and this may be attributed to difficulties in accessing care, reluctance to do so due to social stigmas, and even failure of clinicians to accurately identify the early signs of problem behaviors (Drummond et al., 2004). These barriers to screening individuals for alcohol misuse challenge researchers and public health professionals to shift the emphasis from problem-focused screenings to more

* Corresponding author.
E-mail address: fabrizia.giannotta@mdh.se (F. Giannotta).

https://doi.org/10.1016/j.heliyon.2019.e02523
Received 24 June 2019; Received in revised form 26 August 2019; Accepted 23 September 2019

preventive methods through the creation of predictive models based on the analysis of collateral data, such as survey data on other individual characteristics (e.g., Kumari et al., 2018; García del Castillo Rodríguez et al., 2013; Rosenström et al., 2018) and features of individual activity on social media (e.g., Kosinski et al., 2013; Moreno et al., 2012).

Traditionally, in behavioral science research, alcohol abuse is detected with the use of surveys or interviews with a limited number of people (Paulhus and Vazire, 2007). Recently, however, social network sites (SNS), such as Twitter and Facebook, have provided additional insight about people's real time psychosocial characteristics by allowing researchers to observe individuals' natural online behavior through an analysis of users' written text, shared media, and expressions of interest toward online content (e.g., Facebook Likes). SNS have seen sharp and steady increases in use since the early 2000's, and Facebook specifically currently ranks as the most used SNS worldwide, with over 2 billion monthly active Facebook users (Statista, 2019a) and 1.56 billion users accessing their account on a daily basis (Statista, 2019b). User behaviors on SNS, together with the information users share on their individual profiles are digitally recorded and this data can be, and is, collected and analyzed by social media developers, or authorized third parties. These recorded digitally mediated behaviors have been referred to as "digital footprints", "digital records", or "digital traces" (e.g., Bai et al., 2014; Farnadi et al., 2016; Settanni and Marengo, 2015; Youyou et al., 2015), and represent an extensive source of naturally emerging ecological data of online human activity with connections to offline personal characteristics, attitudes, and behaviors (e.g., Kosinski et al., 2013; Markowetz et al., 2014; Strapparava and Mihalcea, 2017). In essence, digital footprints are recordings of users' activity on social media, such as time and frequency of posting behaviors, as well as the actual posted content (e.g., texts, pictures and videos). Specifically, recent meta-analyses evidence that digital footprints can be successfully mined to gain insights about user's individual characteristics, such as personality (Azucar et al., 2018), intelligence, and well-being (Settanni et al., 2018).

When it comes to alcohol use, there is a handful of studies that have attempted to link digital footprints to problem drinking. For example, alcohol displays (e.g. wall, tagged pictures, profile pictures, and bumper stickers) on Facebook, the number of Facebook friends, and number of status updates referring to alcohol use have been related to problematic drinking among college students (Moreno et al., 2012; Moreno and Whitehill, 2014) and to the total number of uploaded pictures depicting alcohol use (Beullens and Schepers, 2013). These studies, although very innovative and informative, have two main limitations. First, their reliance on human coders to categorize the material obtained from SNS, an approach that while functional in demonstrating an association between online and offline alcohol-related behaviors in small sample conditions, is not easily scalable for the analysis of larger datasets. Second, their exclusive focus on alcohol-related content, without considering content that does not directly refer to alcohol. This is an important limitation as SNS can be seen as a "virtual" environment where people show "virtual behaviors", i.e. writing comments, answering to posts, posting pictures or emoji on different themes, and these online behaviors, in the same way as off-line behaviors, might be indicative of individual life-styles, including problematic alcohol use. In other words, it is possible that online behaviors that do not directly refer to alcohol consumption could be related, even if not in causal way, to alcohol drinking (e.g. posting about night life events could be correlated to higher likelihood to drink alcohol). To overcome these limitations, the use of automated, data-driven analytic approaches to extract and analyze digital footprints has been suggested (e.g. Schwartz et al., 2013). However, this approach has been rarely used in alcohol and substance use research. Among the few exceptions, Kosinski and colleagues explored the use of 'Likes' expressed by Facebook users to predict use (vs. no use) of different substances including alcohol (Kosinski et al., 2013). Zhou et al. (2016) developed a procedure to identify illicit drug use of Instagram users based on the analysis of time of posting. They also showed how illicit drug users expressed common interests in online content, such as celebrities and comedians (Zhou et al., 2016). More recently, a study by Curtis et al. (2018) investigated links between topics discussed in Twitter texts and excessive alcohol use at the county level. The authors analyzed social media data using a method commonly referred to as *differential language analysis* which is a type of open-vocabulary analysis introduced recently by Kern et al. (2014, 2016), which does not rely on *a priori* word or category judgments, and aims to find distinct sets of language features (e.g. words, n-grams, and topics) that distinguish groups of people based on specific sets of characteristics (Schwartz et al., 2013). The authors found excessive alcohol use at the county level to be positively related to frequency of Twitter topics about drinking behaviors and consequences of alcohol drinking, sport and music events, going out on Saturday night, and university tasks, and to be negatively related to topics about religion and church, and use of informal language and Internet slang (Curtis et al., 2018). These findings provide useful insights concerning the association between language use on social media and alcohol use, albeit at an aggregate level. In the present study, we go a step further and employ a similar approach at the individual level, which is more informative to gain insight about the association between language use on social media and individual problem drinking. At the individual level, the association between language use on SNS and alcohol drinking behaviors may be interpreted in light of existing differences in lifestyle, and cognitive and behavioral characteristics of problem drinkers vs. low-risk individuals, which may be reflected in what and how they write on SNS. Here, we examine data collected from a sample of Italian adult Facebook users and investigate the links between self-report problematic alcohol drinking and language use in their profile posts. Use of language in Facebook texts was examined using both traditional closed-vocabulary analysis with Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2007), and open-vocabulary analysis performed via topic modeling with Latent Dirichlet Allocation (LDA) (Blei et al., 2003) with the aim to investigate if the language used on social media is related to problem alcohol use. We hypothesize that a significant relationship exists between risk of problematic alcohol drinking and the way users communicate on social media, both in terms of content and style. We expect this relationship to reflect possible existing associations between alcohol use and users' psychosocial and behavioral characteristics, which in turn are reflected in the textual content shared online by users. As a final aim, we test the feasibility of predicting users' risk of problem drinking by mining the language use features extracted from participants' Facebook profiles, comparing closed- and open-vocabulary features based on their accuracy of predictions of users' problem drinking scores.

2. Materials and methods

2.1. Procedure and participants

In order to achieve the study aims, we proceeded as follows: 1. we recruited an online sample of Facebook users using a snowball procedure; 2. After obtaining participants' informed consent, we administered a self-report instrument (i.e. the AUDIT-C) to measure problem drinking and collected textual data from participants' individual social media profiles; 3. We applied closed- and open-vocabulary analyses to collected texts to extract linguistic features; 4. We examined associations between extracted linguistic features and problem drinking scores data; 5. We developed and tested machine-learning models to predict problem drinking based on linguistic features.

Participants consisted of adult volunteers from Italy, recruited online using a snowball sampling procedure. An initial seed consisting of 20 university students was asked to disseminate the research among their Facebook friends by sharing a link to the research page, which included the online informed consent form and a questionnaire. Participants were required to sign an informed consent, answer the questionnaire, and provide authorization to the researchers to access their Facebook posts using the Facebook Application Programming Interface (API). The research was approved by the Institutional Review Board of the

University of Turin, Italy.

Out of 512 individuals who answered the online questionnaire, 216 persons were not included in the final sample because they failed to provide the researchers with the authorization to access their status updates. This yielded a final sample of 296 participants (66.9% females) with a mean age of 28.44 (SD = 7.38). In order to investigate possible biases due to significant differences between actual participants and the 216 respondents who were not included in the final sample, we conducted chi-square test on gender, and t-tests on age and the self-report problem drinking measure: no significant differences emerged between the two groups (Gender: $\chi^2(1) = 3.13$, $p = .08$; Age: t (510) = 0.39, $p = .69$; AUDIT-C: t (510) = -0.12, $p = .90$).

2.2. Self-report alcohol drinking

Alcohol drinking behaviors were assessed using the Alcohol Use Disorders Identification Test - Consumption (AUDIT-C), which is a brief (3-item) validated alcohol screen that can reliably identify persons who are hazardous drinkers or have active alcohol use disorders (including alcohol abuse or dependence) (Bradley et al., 2007). The AUDIT-C is scored on a scale from 0-12, with each of the 3 questions having 5 answer choices, ranging from 0 to 4. The higher the score, the more likely it is that a person is participating in high risk drinking behavior. Average AUDIT-C score was 3.54 (2.09); Cronbach's alpha was adequate (alpha = .81).

2.3. Data collection and feature extraction via closed- and open-vocabulary analyses

2.3.1. Data collection and preprocessing

Facebook textual data was collected using the Facebook Graph Application Programming Interface (API), and included users' status updates and associated comments posted by participants during the previous year of Facebook activity. Overall, 28,595 posts (status updates and comments) were collected, with an average of 160.18 posts (SD = 81.94), and an average word count of 1,652.67 words (SD = 840.25) per participant. Prior to performing text analyses, posts (status updates and comments) published by the same author were integrated into single text corpora, as suggested by many authors when collected data consists of short SNS texts (e.g., Hong and Davison, 2010; Weng et al., 2010).

2.3.2. Closed-vocabulary analysis

We performed closed-vocabulary analysis on participants' textual data utilizing the Italian version of the Linguistic Inquiry and Word Count (LIWC) software (Alparone et al., 2004; Pennebaker et al., 2007). LIWC's semantic dictionary allows the scoring of text corpora on 84 distinct categories assessing emotional, cognitive and structural language components. LIWC-categories have shown strong connections with a wide range of psychosocial characteristics, including cognitive skills, emotional distress, personality, and personal concerns (for a review, see Tausczik and Pennebaker, 2010). For a more in depth description of the LIWC dictionary, see the LIWC software documentation (Pennebaker et al., 2007).

2.3.3. Open-vocabulary analysis

Open-vocabulary analysis was performed by implementing a topic modeling approach via Latent Dirichlet Allocation (LDA) (Blei et al., 2003). When applied to a collection of documents, LDA provides estimates of the probability that emerging topics appear in each document. Given the small size of our study sample, in order to obtain high quality topics (Tran et al., 2013), LDA analyses were first performed on a large corpus of tweets, the TWITA corpus (Basile and Nissim, 2013), posted in Italy over the course of year 2014 (The TWITA corpus is available for download: http://valeriobasile.github.io/twita/about.html). LDA analyses were performed on a random sample of 6,402,174 Tweets posted over the year by a random sample of 55,206 Twitter users. Single tweets of each individual user were combined together, resulting in 55,206 text documents; text corpora were pre-processed as follows: We converted all text to lowercase and removed all Italian 'stopwords' (i.e., very frequent words with low specificity), punctuation, and numbers. In order to identify the optimal number of topics to be retained, we trained a set of competing LDA models with 100, 200, 300, 400, 500, and 1000 topics. The performance of the competing LDA models was compared by examining the quality of emerging topics using the perplexity statistic (Wallach et al., 2009) and visual inspection of topics. The coherence of LDA-derived topic-words association were visually examined by two human judges using word clouds. Based on perplexity and semantic coherence, we chose 300 as the final number of topics. As a last step, the final model was applied to participants' Facebook data in order to obtain topic proportion scores for the new documents. All analyses were performed in Mallet (version 2.08RC3, McCallum, 2002).

Analyses were conducted on Italian text corpora, however for clarity purposes, results were translated to English using the Google Translate web service, and checked for correctness by bilingual translators.

2.4. Analytic strategy

As a first step, we examined associations between participants' AUDIT-C scores and closed- and open-vocabulary features by computing Pearson's correlation coefficients. Next, we examined the predictive power of language use features over their level of risk problem drinking. Analyses were performed with Random Forests machine-learning algorithm (Breiman, 2001) in Weka 3.8.2 (Eibe et al., 2016). Instead of performing predictions using the whole set of features as in conventional regression and classification trees, the Random Forest (RF) algorithm bootstraps subsamples of features and observations, an approach that allows for the handling of very large sets of features (even more than there are observations) remaining robust against overfitting and collinearity problems (Breiman, 2001). One of the key features of the RF algorithm is that it's nonparametric, which means that the algorithm does not impose specific distributional assumptions on the structure of the data. Other relevant advantages of the RF algorithm with respect to other approaches, such as multivariate regression and classification trees, are that it allows for the use of both categorical and continuous independent variables, and that it permits to account for interactions and nonlinear relationships between predictors (Janitza et al., 2013). For the purpose of this paper, analyses were performed using an 80/20 split validation approach for training and testing the predictive models. We examined three distinct models, each including a distinct set of features: 1) Closed-vocabulary features (i.e., LIWC features); 2) Open-vocabulary features (LDA-Topics); and finally 3) a model combing both closed- and open-vocabulary features (LIWC + LDA-Topics). For each set of features, we ran the RF algorithm using 100, 500, 1000, 5000, 10000 trees, then, we selected the best performing model for each set of features. For each model, accuracy in predicting risk of problem drinking was evaluated by inspecting: (i) Pearson correlation between observed and predicted scores, as a measure of the model predictive power, and (ii) Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) statistics, to estimate the mean error committed using the model to predict the AUDIT-C scores.

3. Results

3.1. Associations between language use and alcohol use

Closed-vocabulary. We found 24 LIWC categories presenting a significant correlation with AUDIT-C score (see Table 1). AUDIT-C scores showed significant negative correlations with categories concerning family members, emotions (i.e., affective processes, feelings, positive emotions, and optimism), personal concerns and behaviors (i.e., sex, touching physical states/factors, symptoms/sensations, and occupation), social relationship (i.e., social words and both singular and plural

Table 1
LIWC categories showing significant correlations with AUDIT-C scores.

LIWC categories	r	p
Family	-0.24	$p < .01$
Pronouns	-0.22	$p < .01$
Present (Time)	-0.21	$p < .01$
Self	-0.19	$p < .01$
Sex	-0.19	$p < .01$
Cognitive processes	-0.18	$p < .01$
Symptoms/sensations	-0.18	$p < .01$
Affective processes	-0.18	$p < .01$
Positive emotions	-0.18	$p < .01$
Time	-0.18	$p < .01$
Feelings	-0.17	$p < .01$
Social	-0.16	$p < .01$
3rd person singular verbs (male)	-0.16	$p < .01$
Touching	-0.16	$p < .01$
Introspection	-0.16	$p < .01$
Future (Time)	-0.15	$p < .01$
1st person singular	-0.15	$p < .05$
Have	-0.14	$p < .05$
3d person plural	-0.14	$p < .05$
Physical states/factors	-0.13	$p < .05$
Swear words	0.13	$p < .05$
Negations	-0.12	$p < .05$
Optimism	-0.12	$p < .05$
Occupation	-0.11	$p < .05$

pronouns), cognitive processes, use of negations and time-related words (i.e., present, future). In turn, only use of swear words showed a significant positive correlation with problem alcohol drinking.

Open vocabulary. Overall, LDA analyses allowed us to identify language features showing a significant relationship with alcohol use as measured by the AUDIT-C. Of the 300 topics scored for the analysis, a relevant portion (n = 45, 15%) showed significant correlations with AUDIT-C scores. Figs. 1 and 2 report word cloud renderings for the 10 topics whose scores showed the strongest positive and negative correlations with participant AUDIT-C scores, while results for all topics with significant associations are reported in Table A1 in the Appendix. In the presented word clouds, top words are rendered with font size increasing proportionally with their relevance to the specific topic (i.e., word-by-topic frequency count), while color is used to distinguish between positive (blue) and negative (red) correlations. Based on the most correlated topics, AUDIT-C scores showed a positive association with frequency of use of topics indicating use of swearing (e.g., Topic 74, top words: *fuck, shit, ass*), interest in politics (e.g., Topic 100, top words: *government, reforms, renzi*, i.e., surname of former Italian prime minister) and sport events (e.g., Topic 257, top words: *Brazil, Germany, match*), use of words about nightlife (e.g., Topic 157, top words: *club, events, staff*) and online/social-media activity (e.g., Topic 119, top words: *social, web, facebook*). In turn, AUDIT-C scores showed a negative association with frequency of use of topics related to expressions of friendliness (e.g., Topic 84, top words: *thank you, dear, hug*), family (e.g., Topic 58, top words: *Christmas, mom, grandma)*, school (e.g., Topic 171, top words: *school, tomorrow, class, homework*), expressions of love (e.g., Topic 206, top words: *heart, love, soul*), and life in general (e.g., Topic 292, top words: *life, day, now*).

3.2. Prediction of AUDIT-C scores

Results of application of the RF algorithm using both closed- and open-vocabulary features as predictors of risk of AUDIT-C scores are reported in Table 2. Number of trees varied based on the specific set of features used for performing the prediction. Examination of correlations computed between observed and predicted values indicated that the model including only the LDA-derived topic scores provided the strongest prediction over the AUDIT-C scores (R = .462, MAE = 1.493, RMSE = 1.960) while the model employing only LIWC features provided the weakest results (R = .285, MAE = 1.569, RMSE = 2.019). Interestingly, the model including both LIWC and LDA features did not improve over the model including only the LDA features (R = .452, MAE = 1.513, RMSE = 1.955). As depicted by the MAE statistics, the prediction error was quite low for all models. Average relative errors, computed as MAE/variable range x 100, were between 13.07% and 12.44%, respectively for the prediction based on the LIWC and LDA features.

4. Discussion

The present study aimed to investigate the associations between language use on SNS and problematic alcohol drinking using both a closed (LIWC) and an open-vocabulary (LDA) analysis. We investigated Facebook content consisting of texts shared by participants over the course of one year of social media activity and our interests were twofold. First, we were interested in finding relations between style and content of the language used online and offline alcohol use behavior. We did find many relations. Second, we were interested in understanding if the information coming from two different types of analysis of textual material could complement each other. We found that results coming from the LIWC and LDA analyses were quite similar, but that LDA added information that could not be retrieved only with LIWC analysis. As a final aim, we tested and compared the predictive power of LIWC and LDA features over users' risk of problem drinking scores. Results showed LDA features improved over LIWC features in terms of predictive power.

Consistent with findings on other substance use and health-related behaviors (e.g., Allem et al., 2018; Muralidhara and Paul, 2018), both the style and the content of the language used on SNS can be related to alcohol drinking and to problematic drinking as well. Our study shows that certain topics and type of language's style are consistently related to alcohol drinking independently of the type of content analysis used, i.e., closed- or open-vocabulary analysis. For instance, low frequency of family-related words, and positive feelings in general, including love, are negatively associated with both alcohol use and problematic drinking, which is consistent with literature showing the protective role close and positive family ties exert on problem drinking (Catanzaro and Laurent, 2004; Ciairano et al., 2006; Smorti and Guarnieri, 2015). Moreover, when it comes to the style of the language, writing with a coarse language i.e. using swear words, as it emerged from both type of analyses, is positively related to risky alcohol drinking, which is consistent with what has recently been found only using a closed vocabulary analysis on online forum data (Kornfield et al., 2018). To summarize, LIWC and LDA work in quite a similar way to individuate aspects emerging from user-generated text on Facebook that are negatively associated with alcohol use.

Despite some similarities, however, the two content analyses approaches have some peculiarities. First, through the closed-vocabulary approach, it emerged that the more individuals drink the less they use words that indicated cognitive processes – a result which is consistent with literature indicating the presence of significant alterations in cognitive processes amongst alcohol users (Field et al., 2008). However, it is the open-ended analysis (LDA) that provides more unique information. This is not surprising, as this analysis is not limited by specific categories but open to all the possible combinations. Specifically, the open-ended analysis shines a light on the strong connections between writing on topics that are *per se* neutral, such as politics, sports, nighttime and holiday leisure activities and social media, and problematic alcohol use. These results synthesize findings in different studies evidencing the association between alcohol drinking, nightlife (Sunderland et al., 2014), and music and sport events (Lloyd et al., 2013). Moreover, the LDA gives some hints about the style of writing highlighting that those who use words indicating personal thoughts (e.g., *dear, love, eyes, moments, and life*) are also less likely to engage problematic drinking.

Previous studies (e.g., Schwartz et al., 2013) have demonstrated that an open-vocabulary analysis can give more insights when it comes to the study of personality than a conventional close-vocabulary analysis, i.e. LIWC. When it comes to alcohol use, results emerging from our predictive analyses seem to confirm this result. Indeed, open-vocabulary features

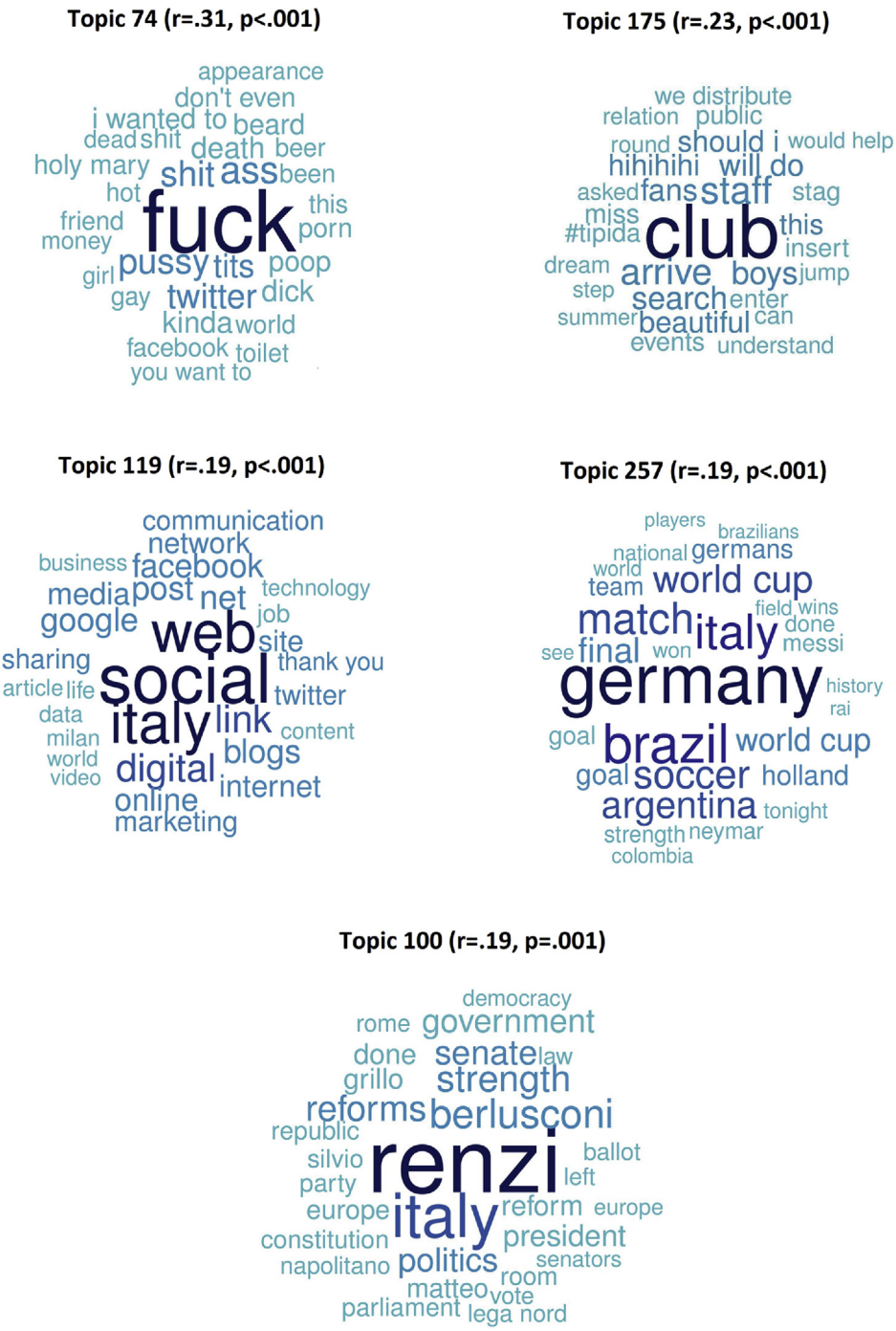

Fig. 1. Word clouds of LDA topics with top positive correlations with AUDIT-C scores.

outperformed closed-vocabulary features in terms of predictive power over the AUDIT-C scores; further, when combined together in a single model, closed- and open-vocabulary features did not improve over the model including only open-vocabulary features in terms of predictive accuracy. Overall, results show that, based solely on the analysis of open-vocabulary features, the predictive accuracy that can be obtained over the AUDIT-C score is moderate ($r \sim .46$), comparable in effect-size to that observed when correlating the AUDIT-C to a breath alcohol concentration test (a Pearson's correlation $r \sim .46$, Barry et al., 2015), and similar to the correlation emerging between AUDIT scores collected on patients attending consecutive visits in clinical settings ($r \sim .49$, Sahker et al., 2017). However, the accuracy of prediction does not reach the standard for screening instruments. Still, our findings indicate that the linguistic features of SNS text contain valuable information that can be used to predict individuals' risk of problematic drinking with a remarkable degree of accuracy. Taking into account the findings from the meta-analyses conducted by Azucar and colleagues and Settanni and colleagues, we expect that better performances in the prediction of alcohol drinking could be reached by employing the linguistic features in combination with other data sources as predictors in the tested models (Azucar et al., 2018; Settanni et al., 2018). In particular, we can hypothesize that information about linguistic features of the texts shared on social media could contribute to the development of screening procedures, in combination with data extracted from other sources of information available online, such as Likes, reactions, posted pictures and/or videos. It is worthy to note that achieving higher levels of predictive

Fig. 2. Word clouds of LDA topics with top negative correlations with AUDIT-C scores.

Table 2
Results of prediction models for the AUDIT-C using the Random Forests algorithm.

Features	R	MAE	RMSE	n. trees
LIWC	.285	1.569	2.019	5000
LDA-Topics	.462	1.493	1.960	500
LIWC + LDA-Topics	.452	1.513	1.955	1000

Note. Prediction performed using 80/20 split cross-validation.

power will also pose some ethical issues linked to the possibility of using social media data to infer characteristics that individuals are not willing to share publicly (e.g. drinking habits). Inferred characteristics may also be used by third parties for a variety of purposes with both potential benefits and consequences for users (e.g., in hiring procedures or to target commercial ads or political messages). This highlights the need for more careful attention to the possible ethical challenges related to the use of data extracted from Facebook or other social media.

4.1. Limitations and strengths

Results from the present study should be understood in light of some limitations. First, the small size of the recruited sample may have limited our ability to implement more sophisticated, robust analytical approaches. Second, the use of snowball sampling may have introduced a potential self-selection bias and negatively influenced the representativeness of our sample. However, as noted by Kosinski et al. (2015), studies conducted using this kind of sampling are not necessarily affected by stronger biases than studies employing other recruitment approaches. A further limitation is related to the limited size of our sample of Facebook users. For this reason, open-vocabulary analyses were performed with a two-step approach, which involved the use of secondary data- i.e., collection of SNS posts of a large sample of Twitter users. However,

correlation and predictive analyses were performed in a small sample condition. While the cross-validation of the predictive model represents quite a strong proof of generalizability to similar populations, a replication of the study on a larger and more diverse sample will be useful to confirm our findings.

Notwithstanding these limitations, the present study has also strength. First and foremost, to our knowledge this is the first study to apply both a closed- and open-vocabulary analysis to the study of alcohol and problematic alcohol use. Contrary to previous studies, we used an automated analytic approach (as opposed to time-intensive manual approaches, Moreno et al., 2012; Ridout et al., 2012), which renders our approach replicable in large-scale situations, such as those provided by the ever-increasing SNS user population. Additionally, and in contrast with previous studies employing similar automated approaches with other digital footprints (e.g., "likes" on Facebook, see Kosinski et al., 2013), alcohol-related measures were assessed by administering an internationally validated instrument; the AUDIT-C. By using a validated instrument with known high-risk thresholds, we were able to highlight for the first time the existence of links between digital footprints on SNS and the risk of problematic drinking.

4.2. Conclusions

Since the introduction of the psychoanalytic theory, the use of language, both in terms of style i.e. word use, and content, has been largely investigated for its relation to personality and social processes (for a review see Pennebaker et al., 2003), and behaviors. SNS have become a modern playground in which individuals freely express and write about themselves; and much like with traditional expressions of language their resulting digital footprints might be linked to individual psychosocial characteristics and behaviors. The current study provides evidence for this idea, highlighting that digital footprints that refer to textual information might reveal problematic alcohol use among adult Facebook users. Given the importance of alcohol drinking as a public health issue, the approach presented in our study provides the foundation for an innovative and unobtrusive method to reach these at risk populations by way of social media.

Declarations

Author contribution statement

M. Settanni, D. Marengo: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

D. Azucar, F. Giannotta: Analyzed and interpreted the data; Wrote the paper.

V. Basile: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2019.e02523.

Appendix

Table A1
Topics showing significant correlations with AUDIT-C scores (N = 296).

r	p	Topic	Top Words
0.308	p<.001	74	fuck life ass pussy shit twitter death tits guy fuck wanted poop porno miss shit
0.234	p<.001	175	club fans staff arrive search boys beautiful should this miss enter events insert tag publish
0.194	p<.001	119	social italy web digital web post google facebook blog online media internet site network marketing
0.194	p<.001	257	germany brazil italy game world soccer argentina world cup goal holland german team goal neymar
0.193	0.001	100	renzi berlusconi reforms senate politics government president reformation grillo europe matteo republic camera silvio party
0.172	0.003	200	italy history country politics now world so today rome reason left times time to be journalists
0.171	0.003	38	done have to do well that being so maybe not even say that part milan
0.156	0.007	75	time xoxo next lignano evening alassio nut subscribe win information tell me indie pray rudeness
0.145	0.013	123	eli volta call usa known memory multiple mica sclerosis finally summer meantime beautiful cute internet
0.144	0.013	81	done earth ramazzotti eros world say you want I can born beautiful confirmation try understand
0.144	0.013	141	genova concordia liguria perugia century ship agi umbria lily lily agency port costa savona ligure
0.14	0.016	22	made cock ass shit life force ass you gotta balls cocks said you can jerk
0.137	0.018	5	done thanks so much to say it seems I think sense point case said problem
0.135	0.020	221	revenge today call cash step center I speak twitter opinion made death I'll work live dance pussy
0.132	0.023	185	italy politics grillo mafia country politicians voted party italians votes shame democracy say vote be
0.129	0.026	101	rome italy mafia milan euro mayor capital million arrested marine house case police video money
0.122	0.036	173	happened retweet occupied seats we sat big part friends twisted anonymous man seen strong words shit
0.122	0.036	187	life washing machine bice raffaele italian been edition memi photo passes niky inserted contest click
0.122	0.036	213	mery cuin sere miki ire kaety multifandom speleus alexia fiki odi last shalala hate
0.121	0.037	264	life history art culture exhibition today milan book film rome literature war books city world cinema
0.12	0.039	167	band mars for jared the gerard frank love day thank you leto letter shannon concert
0.119	0.041	177	reform government senate work law reforms italy room fees renzi workers costs euro employees public
0.116	0.046	110	instagram exchange made like reciprocate likes follow me we can call you want to reciprocate
0.115	0.048	237	work thanks problem case be know use copy serve mail use saw price site pay
-0.117	0.044	296	night facts isa update eunhyuk donghae tagged alexiara location instagram twitter curti seconds pfv sister
-0.119	0.041	39	life be strong man part world wants to give truth so many good words it
-0.119	0.041	48	love being made to life like that man italy that wife rethinked to have time
-0.122	0.036	76	sardinia cagliari Sardinian sea Sardinian bag Sardinian seas bombs zone island sassari luck work olbia

(*continued on next column*)

Table A1 (*continued*)

r	p	Topic	Top Words
-0.125	0.032	144	juve conte juventus rome vidal iturbe italy morata marotta pogba coach tevez evra player market
-0.129	0.026	122	love wish wanna make you eyes person need be know vault will be can heart
-0.13	0.025	71	thank you beautiful picture sun beautiful sea beautiful like beautiful so beautiful beauty day pleasure
-0.13	0.025	121	darling kiss kisses night sweet love dreams hug goodnight heart I want good morning joy
-0.137	0.018	258	love idol want thank you life dream hope smile I'll be world my dear idols miss
-0.14	0.016	35	made go home days photo so can see tomorrow time day just kind week today
-0.142	0.014	87	would like treport emis emi killa nick profile inactive you might like arrive
-0.148	0.011	114	really want feel think would like to be told can today happy be hopeful yesterday I think beautiful
-0.156	0.007	239	good day thanks evening good morning hello easter good night wishes goodnight afternoon week greetings
-0.164	0.005	65	love life heart soul words happiness music emotions passion night thoughts moment thought pain woman
-0.164	0.005	216	life love made person so beautiful you know how beautiful you can be rest be afraid
-0.18	0.002	249	thank you heart congratulations family beautiful today wishes tonight good tomorrow beautiful we hope yesterday
-0.184	0.001	171	school tomorrow day go start class days start today tasks want monday come back anxiety
-0.189	0.001	84	thanks dear hello hug happy friends dear serene good night good day evening beautiful dearest
-0.192	p<.001	206	heart love life soul night eyes words world moon sweet smile sun sky sea stars
-0.219	p<.001	58	christmas made thank you tree dad gift house mom gifts wants mother beautiful grandma daughter
-0.22	p<.001	292	life be day so live person world you can all say love moment need heart

References

Allem, J.P., Dharmapuri, L., Unger, J.B., Cruz, T.B., 2018. Characterizing JUUL-related posts on twitter. Drug Alcohol Depend. 190, 1–5.

Alparone, F., Caso, S., Agosti, A., Rellini, A., 2004. The Italian Liwc2001 Dictionary. LIWC. net, Austin.

Azucar, D., Marengo, D., Settanni, M., 2018. Predicting the Big 5 personality traits from digital footprints on social media: a meta-analysis. Personal. Individ. Differ. 124, 150–159.

Bai, S., Gao, R., Hao, B., Yuan, S., Zhu, T., 2014. Identifying Social Satisfaction from Social media arXiv preprint arXiv:1407.3552.

Barry, A.E., Chaney, B.H., Stellefson, M.L., Dodd, V., 2015. Evaluating the psychometric properties of the AUDIT-C among college students. J. Subst. Use 20 (1), 1–5.

Basile, V., Nissim, M., 2013. Sentiment analysis on Italian tweets. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bradley, K.A., DeBenedetti, A.F., Volk, R.J., Williams, E.C., Frank, D., Kivlahan, D.R., 2007. AUDIT-C as a brief screen for alcohol misuse in primary care. Alcohol Clin. Exp. Res. 31 (7), 1208–1217.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Beullens, K., Schepers, A., 2013. Display of alcohol use on Facebook: a content analysis. Cyberpsychol., Behav. Soc. Netw. 16 (7), 497–503.

Catanzaro, S.J., Laurent, J., 2004. Perceived family support, negative mood regulation expectancies, coping, and adolescent alcohol use: evidence of mediation and moderation effects. Addict. Behav. 29 (9), 1779–1797.

Ciairano, S., Settanni, M., van Schuur, W., Miceli, R., 2006. Adolescent substance use, resources and vulnerabilities: a cross-national and longitudinal study. Suchttherapie 52 (4), 253–260.

Cunningham, J.A., Breslin, F.C., 2004. Only one in three people with alcohol abuse or dependence ever seek treatment. Addict. Behav. 29 (1), 221–223.

Curtis, B., Giorgi, S., Buffone, A., Ungar, L.H., Ashford, R., Hemmons, J., Summers, D., Hamilton, C., Schwartz, H.A., 2018. Can Twitter be used to predict county excessive alcohol consumption rates? PLoS One 13 (4), e0194290.

Drummond, C., Oyefeso, A., Phillips, T., Cheeta, S., Deluca, P., Perryman, K., Winfield, H., Jenner, J., Cobain, K., Galea, S., Saunders, V., Fuller, T., Pappalardo, D., Baker, O., Christoupoulos, A., 2004. Alcohol Needs Assessment Research Project (ANARP). The National Needs Assessment for England. Department of Health and the National Treatment Agency, London.

Eibe, F., Hall, M.A., Witten, I.H., 2016. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Esser, M.B., 2014. Prevalence of alcohol dependence among US adult drinkers, 2009–2011. Prev. Chronic Dis. 11.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M., De Cock, M., 2016. Computational personality recognition in social media. User Model. User-Adapted Interact. 1–34.

Field, M., Schoenmakers, T., Wiers, R.W., 2008. Cognitive processes in alcohol binges: a review and research agenda. Curr. Drug Abuse Rev. 1 (3), 263–279.

García del Castillo Rodríguez, JA, López-Sánchez, C., Quiles Soler, M.C., García del Castillo-López, A., Gázquez Pertusa, M., Marzo Campos, J.C., Inglés, C.J., 2013. Predictive models of alcohol use based on attitudes and individual values. J. Drug Educ. 43 (1), 19–31.

Gnambs, T., Kaspar, K., 2015. Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. Behav. Res. Methods 47 (4), 1237–1259.

Haberstick, B.C., Young, S.E., Zeiger, J.S., Lessem, J.M., Hewitt, J.K., Hopfer, C.J., 2014. Prevalence and correlates of alcohol and cannabis use disorders in the United States: results from the national longitudinal study of adolescent health. Drug Alcohol Depend. 136, 158–161.

Hong, L., Davison, B.D., 2010. Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social media Analytics. ACM.

Janitza, S., Strobl, C., Boulesteix, A.L., 2013. An AUC-based permutation variable importance measure for random forests. BMC Bioinf. 14 (1), 119.

Kern, M.L., Eichstaedt, J.C., Schwartz, H.A., Dziurzynski, L., Ungar, L.H., Stillwell, D.J., Ramones, S.M., Seligman, M.E., 2014. The online social self an open vocabulary approach to personality. Assessment 21 (2), 158–169.

Kern, M.L., Park, G., Eichstaedt, J.C., Schwartz, H.A., Sap, M., Smith, L.K., Ungar, L.H., 2016. Gaining insights from social media language: methodologies and challenges. Psychol. Methods 21 (4), 507–525.

Kornfield, R., Toma, C.L., Shah, D.V., Moon, T.J., Gustafson, D.H., 2018. What do you say before you relapse? How language use in a peer-to-peer online discussion forum predicts risky drinking among those in recovery. Health Commun. 33 (9), 1184–1193.

Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. 110 (15), 5802–5805.

Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D., 2015. Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. Am. Psychol. 70 (6), 543.

Kumari, D., Kilam, S., Nath, P., Swetapadma, A., 2018. Prediction of alcohol abused individuals using artificial neural network. Int. J. Inf. Technol. 10 (2), 233–237.

Lloyd, B., Matthews, S., Livingston, M., Jayasekara, H., Smith, K., 2013. Alcohol intoxication in the context of major public holidays, sporting and social events: a time–series analysis in Melbourne, Australia, 2000–2009. Addiction 108 (4), 701–709.

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., Schlaepfer, T.E., 2014. Psycho-informatics: big data shaping modern psychometrics. Med. Hypotheses 82 (4), 405–411.

McCallum, A.K., 2002. Mallet: A Machine Learning for Language Toolkit.

Moreno, M.A., Christakis, D.A., Egan, K.G., Brockman, L.N., Becker, T., 2012. Associations between displayed alcohol references on Facebook and problem drinking among college students. Arch. Pediatr. Adolesc. Med. 166 (2), 157–163.

Moreno, M.A., Whitehill, J.M., 2014. Influence of social media on alcohol use in adolescents and young adults. Alcohol Res. Curr. Rev. 36 (1), 91.

Muralidhara, S., Paul, M.J., 2018. # Healthy selfies: exploration of health topics on Instagram. JMIR Public Health Surveill. 4 (2).

Paulhus, D.L., Vazire, S., 2007. The self-report method. Handb. Res. Methods Personal. Psychol. 1, 224–239.

Pennebaker, J.W., Booth, R.J., Francis, M.E., 2007. Linguistic Inquiry and Word Count: LIWC [Computer Software]. liwc. net, Austin, TX.

Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G., 2003. Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. 54 (1), 547–577.

Ridout, B., Campbell, A., Ellis, L., 2012. 'Off your Face (book)': alcohol in online social identity construction and its relation to problem drinking in university students. Drug Alcohol Rev. 31 (1), 20–26.

Rosenström, T., Torvik, F.A., Ystrom, E., Czajkowski, N.O., Gillespie, N.A., Aggen, S.H., Krueger, R.F., Kendler, K.S., Reichborn-Kjennerud, T., 2018. Prediction of alcohol use disorder using personality disorder traits: a twin study. Addiction 113 (1), 15–24.

Sahker, E., Lancianese, D.A., Arndt, S., 2017. Stability of the alcohol use disorders identification test in practical service settings. Subst. Abus. Rehabil. 8, 1.

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., Ungar, L.H., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 8 (9), e73791.

Settanni, M., Azucar, D., Marengo, D., 2018. Predicting individual characteristics from digital traces on social media: a meta-analysis. Cyberpsychol., Behav. Soc. Netw. 21 (4), 217–228.

Settanni, M., Marengo, D., 2015. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. Front. Psychol. 6.

Smorti, M., Guarnieri, S., 2015. The parental bond and alcohol use among adolescents: the mediating role of drinking motives. Subst. Use Misuse 50 (12), 1560–1570.
Simons, J.S., Wills, T.A., Neal, D.J., 2014. The many faces of affect: a multilevel model of drinking frequency/quantity and alcohol dependence symptoms among young adults. J. Abnorm. Psychol. 123 (3), 676.
Statista, 2019a. Number of Monthly Active Facebook Users Worldwide as of 1st Quarter 2019 (In Millions). www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.
Statista, 2019b. Number of Daily Active Facebook Users Worldwide as of 1st Quarter 2019 (In Millions). www.statista.com/statistics/346167/facebook-global-dau/.
Strapparava, C., Mihalcea, R., 2017. A computational analysis of the language of drug addiction. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2, pp. 136–142.
Sunderland, M., Chalmers, J., McKetin, R., Bright, D., 2014. Typologies of alcohol consumption on a Saturday night among young adults. Alcohol Clin. Exp. Res. 38 (6), 1745–1752.
Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29 (1), 24–54.
Tran, N.K., Zerr, S., Bischoff, K., Niederée, C., Krestel, R., 2013. Topic cropping: leveraging latent topics for the analysis of small corpora. In: International Conference on Theory and Practice of Digital Libraries. Springer, Berlin Heidelberg.
Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D., 2009. June). Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning(pp. 1105-1112). ACM.
Weng, J., Lim, E.P., Jiang, J., He, Q., 2010. Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (Pp. 261-270). ACM.
World Health Organization, 2014. Global Status Report on Alcohol and Health. World Health Organization, Geneva, Switzerland.
Youyou, W., Kosinski, M., Stillwell, D., 2015. Computer-based personality judgments are more accurate than those made by humans. Proc. Natl. Acad. Sci. 112 (4), 1036–1040.
Zhou, Y., Sani, N., Lee, C.K., Luo, J., 2016. Understanding Illicit Drug Use Behaviors by Mining Social media arXiv preprint arXiv:1604.07096.