Twitter for marijuana infodemiology

Víctor D. Cortés
Department of Industrial Engineering, University of Chile
Santiago, Chile
vcortes@ing.uchile.cl

Juan D. Velásquez
Department of Industrial Engineering, University of Chile
Santiago, Chile
jvelasqu@dii.uchile.cl

Carlos F. Ibáñez
Department of Psychiatry and Mental Health, University of Chile
Santiago, Chile
cibanez@hcuch.cl

ABSTRACT

Today online social networks seem to be good tools to quickly monitor what is going on with the population, since they provide environments where users can freely share large amounts of information related to their own lives. Due to well known limitations of surveys, this novel kind of data can be used to get additional real time insights from people to understand their actual behavior related to drug use. The aim of this work is to make use of text messages (tweets) and relationships between Chilean Twitter users to predict marijuana use among them. To do this we collected Twitter accounts using a location-based criteria, and built a set of features based on tweets they made and ego centric network metrics. To get tweet-based features, tweets were filtered using marijuana-related keywords and a set of 1000 tweets were manually labeled to train algorithms capable of predicting marijuana use in tweets. In addition, a sentiment classifier of tweets was developed using the TASS corpus. Then, we made a survey to get real marijuana use labels related to accounts and these labels were used to train supervised machine learning algorithms. The marijuana use per user classifier had precision, recall and F-measure results close to 0.7, implying significant predictive power of the selected variables. We obtained a model capable of predicting marijuana use of Twitter users and estimating their opinion about marijuana. This information can be used as an efficient (fast and low cost) tool for marijuana surveillance, and support decision making about drug policies.

KEYWORDS

marijuana, opinion mining, social network analysis, text mining, web content mining

ACM Reference format:
Víctor D. Cortés, Juan D. Velásquez, and Carlos F. Ibáñez. 2017. Twitter for marijuana infodemiology. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 7 pages.
https://doi.org/10.1145/3106426.3106541

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3106541

1 INTRODUCTION

Every day, users on different platforms generate large amounts of information where they assume the responsibility of content creation. People use these platforms to express themselves and behave as they naturally do in their daily lives. They disclose personal information, interests, activities, relationships, and interactions with others. In addition, they spend much of their time immersing themselves in these environments. For this reason, it is interesting to analyze data which represents how they actually are and behave. Although such data has certain known limitations [2, 15], it could be used to complement existent monitoring systems making them able of identifying new trends in people behavior.

In Chile, marijuana has been given particular attention due to extensive debate and relevance. In recent years, the perceived risk of the drug has declined, reflecting a social norm in favor of use, which can be the source of the significant rise in domestic consumption. Indeed, in 2014, the annual prevalence rate was 11.3% [22], and based on United Nations Office On Drugs and Crime (UNODC), Chile has the highest marijuana prevalence rate in Latin America.

The National Service for the Prevention and Rehabilitation of Drug and Alcohol (SENDA) conducts studies to collect information on the extent of use and risk perception, among other variables. These studies have a very time-consuming cycle (every two years). This obstructs continuous monitoring of the evolution of the prevalence and prevents early detection of abrupt changes. Thus the opportunity focuses on the creation of complementary sources to improve and enrich the quality of information, in addition to increasing the frequency of data collection. Such data must be transformed into information to support decision making.

In this paper, we use Twitter as the data source and build a set of Chilean users in order to collect their personal information, to track their relationships and to store their tweets (text messages). Text mining techniques [13] were applied to extract valuable information from tweets. Based on social network analysis metrics and real marijuana use data, patterns were obtained about marijuana use on Twitter.

2 RELATED WORK

The idea of exploring online social networks to supplement traditional means of gathering information is not new. Several prior studies have attempted to extract valuable information from web user generated content, some trying to get it through manually content coding and others under the web content mining concept [24]. Both approaches were used by researchers trying to obtain insights about what people comment about drugs [10, 23] and other health related [21] issues in social media. Both approaches conclude about the real utility of social media to such purpose.

In [8] tools are introduced to use data from online social networks and use of Twitter is suggested because of its free public data. Indeed, several studies have used Twitter as a source of input data for marijuana content analysis. Many of them filtered tweets using keywords to decrease the total number of cases, but the main goal was to get a marijuana-related set [12]. Later step is to code tweets to gain knowledge about the content of tweets. [6] reveals that the most common theme of tweets stating that people wants/plans to use marijuana and [7] explored the effects of a novel way to inhale marijuana. In some prior studies, the text was coded by membership into categories. The most frequent were perception or sentiment about marijuana and its varieties [6, 19]. Besides, it was confirmed the association between Twitter-based features and current marijuana use, even if this was done by mean of a survey [5]. These features born from the way marijuana content spread through Twitter.

A little has been done in automatic processing of marijuana-related content [11, 26]. There is evidence of work done in the semi-supervised machine learning realm, developing a model capable of classifying tweets associated to recreational marijuana use [26]. In the supervised machine learning realm, prior research indicates the ability to train algorithms to classify marijuana-related tweets by categories. These categories involve the type/source of communication and sentiment [11].

Another method used to obtain information from social networks is the application of Opinion Mining techniques [11]. This is a subfield of text mining to extract opinions from documents which can be specifically applied to the content generated by Web users. [3] examines recent literature regarding the sub-field, explains the main problems of extraction and suggests how to solve them. There are continuous efforts to improve models performing this task, especially in the Spanish context [27].

The Social Network Analysis corresponds to building a set of metrics from the connections between people. For instance, it is possible to build complex networks based on five friend nominations for each subject surveyed. From this, the position within the structure of the network affects behavior [14] and use by part of the group, and interactions between different roles inside the network can predict use [17].

Online communities related to drug use were also explored. In [9] the connection between online network features and consumption by young adults was examined. In [25] the association between the presence of alcohol content and other drugs in online social networking, perceived norms, and marijuana use among young adults was evaluated.

The reviewed bibliography suggests a lot of research made exploring the presence of marijuana content in Twitter, manual and automatic categorization of tweets (sources and sentiments), and the relationship between marijuana-related content and current marijuana use. Nevertheless, we notice a lack of work done connecting prior research. Besides, social network analysis has never been used directly in the Twitter relationships to predict current marijuana use.

3 METHOD

Our research considered Twitter as a source of information, a microblogging service. In accordance with [18], Twitter has emerged as a new medium in the spotlight because of recent happenings. Information spreads through groups of users due to their following relationships. Twitter provides access to its data via two media, REST API and Streaming API. Both need a special credential to access them. We needed historical data from users, which means data that users generated since they created their account. For this reason, the first one was used to provide access and extract Twitter data. Every type of REST API query has a limit number of calls per credential, restricting the capacity of getting data. This was solved using many credentials in a queue system.

Twitter users can follow others and can be followed by others. This relationship requires no reciprocation, in other words, a user can follow any other user, and the other users don't need to follow back. This connection means that the user receives all the messages (called tweets) from those the user follows. These messages have a strict limit of 140 characters, promoting brevity of expression, and can be cited by other users. These new tweets are named retweets. For clarity, thinking of any k user, all users following k will be called k's "followers", and all users followed by k will be called k's "friends".

The first step was to select a set of Chilean users because this group is the natural target segment of SENDA. To build this set and better understand how we modeled data, we narrowed user data into three types: (1) information about the user; (2) tweets posted by each user; and (3) user network formed by its connections with other users.

Thinking of users and their connections as a graph, the user collection algorithm operated as a classic graph traversal algorithm, called "search in width," where all adjacent nodes for each node in the network are added. This procedure was adapted slightly to resemble the Focused Web Crawler [16], which runs through graphs in the same way, but only adjacent nodes to those elements that meet certain criterion are added. In the case of this study, the criterion is that users are Chileans, which information is contained in the user data. In practice, we matched user location field with a keywords list containing "chile", regions, cities, and communes with more than 30.000 population. There also was implemented a strict soft criterion based on previous iterations of the algorithm to avoid adding places of similar name in other countries.

Once the user database is established, it was extracted the set of tweets posted by each account. This process took place during January 2016. Only tweets related to marijuana were saved, the way of doing this was to identify keywords in the text. The list of keywords was derived from three different sources: expert knowledge, literature and a survey of the current use of words. The total set of words was filtered to confirm its use in Twitter and disambiguate its context of use. To generate this, a group of tweets for each word was extracted from a local repository. Topic Modeling [4] was applied in order to identify different contexts of use of the words (clusters) and verify their use related to marijuana. Then, ambiguous words were filtered, considering only those tweets that contain the string "fum" ("fumar" is Spanish for "to smoke"). This rule was the most simplified way to perform the disambiguation process, and this

marihuana	cannabis	weed
mariguana	marijuana	prensada
porro (f)	thc	pito (f)
caño (f)	yerba (f)	sativa
sacate uno	canabis	macoña
de la buena (f)	hierba (f)	mota (f)
ganjah	cuete (f)	prensao
ganja	faso (f)	paraguaya (f)
de la wena (f)	cogollo (f)	bongazo
ganya	hachis	pitito (f)
matacola	hierva (f)	paragua (f)
marihuanita	troncho (f)	la verde (f)
canabica	cogollito (f)	pitits
cogoyo (f)	marimba (f)	paraguayo (f)
huiro (f)	bless (f)	yerva (f)
sacateuno		

Table 1: Marijuana keywords

word was used due to its frequent presence on marijuana contexts. The final list is shown in Table 1. The disambiguation process was done in keywords placed with (f) sign.

3.1 Tweet labeling

There were randomly selected a set 1,000 tweets to be labeled about marijuana use. This number of tweets has associated a 95% level of confidence and an error of 3.1%. It is well known that it is not easy to count on people who can manually tag all text, so the following methodology was used to minimize people involved in the tagging process. The dataset of tweets was classified at first by two coders. Only when the previous step was done, tweets in disagreement were labeled by a third coder. This person assumed the responsibility of choosing the correct category for those tweets. Requirements to be a coder were to know Twitter rules and practices and be aware of the terms and context of marijuana use. The positive class was composed by tweets which were clearly identified marijuana use from their authors. Otherwise, they were part of the negative class.

3.2 User survey

A direct survey was done to Twitter users, with the purpose of getting real marijuana use labels per user and then train classification algorithms with them regarding marijuana use in Twitter users. Once a database of Chilean users was built, some of them were randomly selected to be sent the survey. They were notified using direct mention via Twitter and redirected to a website which contained the survey. The survey included the following questions:

- When was the last time you consumed marijuana?
- How old are you?

3.3 Models

In order to obtain models of marijuana use per user and opinions in tweets related to marijuana, the process was done incrementally. In other words, tweets were analyzed first and then the result of this procedure was used at the user level. This process was implemented in this way because the tweets are generated by the users themselves.

First, the polarity of feelings for tweets was calculated, and classify them with respect to marijuana use (binary variable). Then, tweets, user information, and social network analysis were combined to determine the user age (numeric variable) and marijuana use in last year (binary variable). This process is visualized in Figure 1.

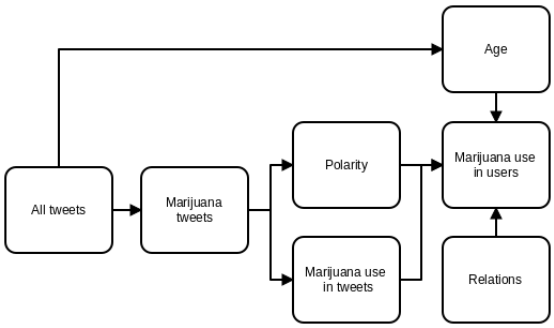

Figure 1: Macro-process of marijuana classification

Applying Data Mining in texts, variables usually arise from the same words represented in matrix form. To do this, we followed the process showed in Figure 2. First, we took raw text and normalize it, transforming all characters in their canonical decomposition. Then, we performed tokenization, and Twitter elements treatment to minimize sparsity from hashtags, mentions and HTTP links. Lemmatization was done only in case of polarity. Finally, n-grams and TF-IDF were applied in different ways to better performance of classification and regression algorithms.

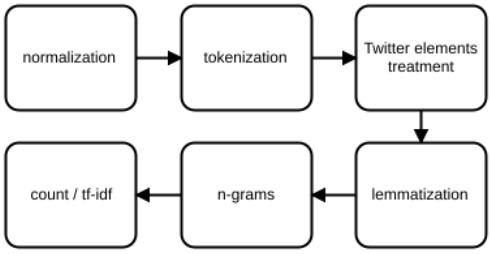

Figure 2: Text treatment

In evaluating performance, for each classification model, some metrics derived from the confusion matrix were evaluated, Precision, Recall and F-Measure. The first has priority over the other two because SENDA prioritizes recall fewer elements but having the certainty of the predicted classes. Cross-validation technique was also used in training the algorithms.

3.4 Measurement

The measures detailed below correspond to the definition of features that were used in the model. In summary, were used metrics from social network analysis, some based on tweets and age.

3.4.1 Social Network Analysis features. These types of variables were considered for their applicability in the structure of relationships in Twitter, which share similarities with traditional connections between people. The latter has been shown to be significant predictors of substance use [1, 17]. The variables used in this study are derived purely from the work done in [14].

Before going deeper into social network analysis features some definition must be done. Let it be: V, the set of n users or vertexes in the graph, $\{1, 2, ..., n\}$; F_k, the set of n_k followers of the user k, $\{1, 2, ..., n_k\}$; A_k, the set of m_k friends of the user k, $\{1, 2, ..., m_k\}$; N_k, the union between followers of the user k and k itself, this is called neighborhood of user k, $F_k \cup \{k\}$; and M, the group of w users with at least one tweet classified as marijuana use, $\{1, 2, ..., w\}$. All these definitions are built considering the only relationship between Chilean users, not the whole Twitter network.

Below, the group of features is listed:

- **Neighborhood density** is defined as the number of existing connections divided by the total possible connections.

$$density_k = \frac{2 * (|F_k| + \sum_{i \in F_k} \sum_{j \in F_i} b_j^{N_k})}{|N_k| * (|N_k| - 1)}$$

Where:

$$b_j^{N_k} = \begin{cases} 1 & \text{if } j \in N_k \\ 0 & \text{if } j \notin N_k \end{cases}$$

- **In-degree** is a popularity indicator defined by the total number of nominations received (followers).

$$in_k = |F_k|$$

- **Reach centrality** measures the percentage of the social network that can be achieved by three or fewer nominations.

$$reach_k = \frac{|\bigcup_{j \in F_k} F_j \cup F_j - \{k\}|}{|V|}$$

- **Neighborhood use** performs sum of all adjacent nodes that report consumption and normalizes it by totaling all of them, without counting the node itself.

$$m_k = \frac{\sum_{j \in A_k} c_j}{|A_k|}$$

$$c_j = \begin{cases} 1 & \text{if } j \in M \\ 0 & \text{if } j \notin M \end{cases}$$

- **Polarity neighborhood** is an average of the average polarity of adjacent users.

$$p_k = \frac{\sum_{j \in A_k} d_j}{|A_k|}$$

Where p_k is polarity related to user k.

- **Distance to the first consumer** measures hops between nodes to reach the first consumer.

$$dist_k = \begin{cases} 1 & \text{if } A_k \cap M \neq \emptyset \\ 2 & \text{if } A_k \cap M = \emptyset \wedge \bigcup_{j \in A_k} A_j \cap M \neq \emptyset \\ 3 & \text{if not} \end{cases}$$

- **External nominations** considers the number of nominations outside the network under study (Chilean users). The more external nominations a node has, the less it will be embedded in the social network.

$$outnom_k = friends_k - out_k$$

Where, $out_k = |A_k|$ and $friends_k$ is total of Twitter followers of k.

3.4.2 Tweet-based features.

- **Mentions about marijuana use** are the number of emitted tweets related to marijuana use counted for each user.
- **User's polarity** is the average polarity of each emitted tweet by a user. It is assumed that users that did not tweet about marijuana get a zero polarity, i.e. neutral.

3.4.3 Age. Age is directly asked in the user survey as an integer, but because age is not part of user data, it must be estimated for new users. This estimation is done based on the perception that individuals change the lexicon occupied throughout their life and whole generations share lexical items [20]. This part is not to encapsulate the texts into categories, but to try to associate them with values. The age algorithm uses the lexicon of last 200 tweets made of each user.

3.5 Polarity

In most of the process is essential sentiment polarity calculation. This was performed using text classification too, but this time the problem was faced as a three-class classification. It was employed the TASS corpus, a corpus of tweets tagged for Sentiment Analysis in Spanish [27]. For global polarity in a tweet, TASS corpus included six possible labels: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE). Due simplicity purposes in training we randomly selected 2499 tweets and only used three classes that contained the previous ones: positive (P+ and P), neutral (NEU and NONE), and negative (N+ and N). Positive class indicates that a text speaks positively, while a negative class indicates a negative sentiment. Neutral class indicates no sentiment or opposite ideas in the text. Tweets were preprocessed as well as other tweets, except that there was included a lemmatization step. Besides, the same performance metrics were evaluated.

4 RESULTS

At the end of the extraction period, the total number of users was 1,505,367, however, the number of valid users for analysis was 1,361,285, due to blockage of information from some of them. Figure 3 shows the number of tweets created for each year. It is to be recalled that only marijuana-related tweets were stored.

Table 2 summarizes the agreement measures among the first two taggers in tweet labeling. The raw agreement between them was equal to 96%. Cohen's Kappa coefficient, dedicated to reflecting the level of agreement between two people, shows a ratio equal to 0.79. The final proportion (after third labeling) of positive cases was equal to 10.5%.

The Twitter user survey was conducted in the period from February 9th, 2016 to March 6th of that year. In that period it was answered by a total of 209 people. After the crossing with the database, 204 valid user cases were obtained, reflecting a response rate

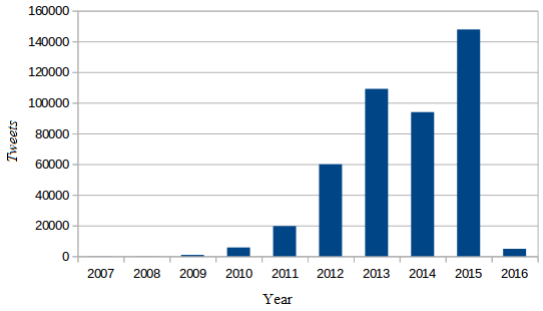

Figure 3: Tweets per year

Category	Raw agreement	Cohen's kappa
Marijuana use	0.96	0.79

Table 2: Agreement between taggers

of 0.3%. Using the annual consumption of marijuana (11.3%) of 2014 as the heterogeneity rate of the data, an error of 4.35% and a confidence level of 95% is obtained.

The two questions addressed in the survey yielded some statistics. Figure 4 shows age box plots of the sample and the national prediction for 2016 (based on the Chilean CENSUS 2012,) where the absence of ages at the older segment is clearly seen. On the other hand, 42.1% of users confessed marijuana use in the last year and 34.3% in the last month.

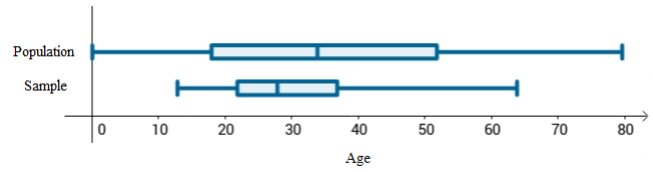

Figure 4: Age statistics

Evidence of marijuana use in tweets and users is approached as a binary classification problem, whilst polarity was tackled as a three class problem. There are a lot of algorithms capable of performing these tasks. Algorithms were chosen which had the best performance measure for classification tasks. In each task, several algorithms were evaluated, and the one with the greatest precision was selected. The algorithms are not the only ones that vary in the testing process. There are a number of parameters that can be modified, but the assessment procedure will not be detailed. The performance of each algorithm implicitly incorporates these modifications, only being named by the algorithm.

There was evaluated the usefulness of three algorithms for marijuana use classification in tweets: (1) Naive Bayes with unigrams and binary vector attributes; (2) Voted Perceptron with unigrams and log-normalized vectors; and (3) Support Vector Machines with unigrams to trigrams and binary vectors. Table 3 shows the performance of winner algorithm. In it, the values of Precision, Recall, and F-Measure are evaluated. As was pointed out earlier, the consumer (positive) class Precision was prioritized, therefore the model of Support Vector Machines was chosen.

Class	Precision	Recall	F-Measure
Not use (0)	0.924	0.989	0.955
Use (1)	0.804	0.302	0.417
Weighted	0.911	0.917	0.899

Table 3: SVM performance

Best performance in polarity classification was achieved by the Support Vector Machines algorithm, with trigrams and TF-IDF calculation in the vector representation of documents. Performance metrics are shown in Figure 4, where the weighted precision and recall were near to 0.59 and 0.58, respectively. The neutral class got the worst precision and recall values. To better visualize the polarity classification performance, the confusion matrix is showed in Figure 5. In there it can be seen how false positives are distributed.

Class	Precision	Recall	F-Measure
Negative (-1)	0.584	0.572	0.578
Neutral (0)	0.47	0.532	0.499
Positive (1)	0.674	0.616	0.644
Weighted	0.585	0.577	0.58

Table 4: Performance of polarity classification

		Predicted		
		Negative	Neutral	Positive
Actual	Negative (-1)	433	208	116
	Neutral (0)	178	404	177
	Positive (1)	130	247	606

Table 5: Confusion matrix of polarity classification

Linear relationship measures and aggregated differences between the actual and predicted age were used to measure performance in regression. Specifically, the Pearson correlation and other errors were used. In this instance, three algorithms designed for data regression were evaluated: Linear Regression, M5P and the version of Support Vector Machines for regression. All were trained with unigrams and frequency log-normalized vectors, although the best model has a small variation. The performance measures are shown in Table 6, which include all options plus a version of SVM with binary vectors. It can be clearly seen that measures improve strictly top down. The best model is the latest version of SVM, having a Pearson correlation of 0.583 and mean absolute error of 6.28. In other words, the model is wrong on average about 6 years.

The classification of marijuana use by users is included with typical data mining models, since here text processing for building a set of attributes was not done. A group of 13 attributes, consisting of measures derived from the tweets and social environment of the user, were used to predict their marijuana use. There were

Model	Correlation	MAE	RMSE
Linear Regression	0.248	7.913	9.792
M5P	0.469	7.286	9.234
SVMreg log-normalized	0.526	6.573	8.503
SVMreg binary	0.583	6.280	8.151

Table 6: Age algorithms performance

compared three algorithms on their classification performance: Support Vector Machines, Logistic and Voted Perceptron. Three in their optimized versions yielded almost identical measures. Table 7 shows the set of performance measures for Logistic, used as a final model. These values are widely different from previous results since the performance measures are similar in scale. This is true for both classes and values of Precision and Recall. In short, individually 57.5% of marijuana accounts use will be recovered and 68.5% of all consumers will be effectively predicted.

Class	Precision	Recall	F-Measure
Not use (0)	0.718	0.803	0.758
Use (1)	0.685	0.575	0.625
Weighted	0.704	0.706	0.701

Table 7: User SVM performance

Table 8 shows normalized weights for each feature. The data indicate that the features with greater predictive power are users with use tweets, the density of the neighborhood, the percentage of marijuana users in the personal neighborhood and nominations outside the Chilean social network.

Feature	Normalized weight
Age	−0.06
Marijuana tweets	0.26
Use tweets	1.72
Polarity	0.36
Followers	0.00
Density	1.14
Reach Centrality	0.51
Neighborhood use	−5.83
Neighborhood polarity	−0.84
Users distance	−0.38
External nominations	−2.56
Intercept	2.77

Table 8: Feature weights

This classifier was used to predict marijuana use among different years in the Twitter data. It was selected a sample of accounts per year and calculate prevalence. To compare trends between the predicted data and SENDA data, the first was pondered and an offset to past of two years was applied. The result of this comparison is shown in Figure 5. The early years were discarded because of the few data for those years. Data show a Pearson correlation of 0.933.

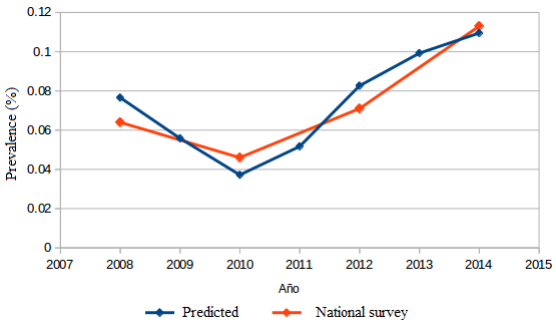

Figure 5: Comparison of trends in prevalence data

5 DISCUSSION

The availability of data is a critical point on any platform based on it. It follows that about 10% of users have their tracking relationships and/or list of tweets blocked for external agents, a non-significant percentage, but its future evolution should be observed.

Several interesting points emerge from the data obtained by the survey of Twitter users. The most important point is consistent with the belief that Twitter users are younger than the average population. The age distribution of the sample confirms this, as when compared to projections for 2016, based on Census 2012, the average age is lower and there is a lack of younger and older subjects, particularly at the older segment.

In the classification of marijuana use in tweets, the Recall value of the positive class is low, and the Precision value for the best is close to 0.8. All this is compensated by the good results for the negative class (not consumption). This means that the classifier will not recover all positive cases, and categorize fairly well, but overall it is a good classifier.

In polarity classification, all metrics seem not to be good enough, but it is important to remember that we are talking about a three-class classification problem, hence the algorithm has more choices to fail. Although the accuracy was near to 0.58, using the only bag of words, proves to be good in comparison to the work done in [27]. Moreover, the confusion matrix lets us see that the distribution of miss failed predictions is similar in wrong predicted classes. This minimizes the possibility of failing to negative class when the positive class is true, and the contrary way too.

The prediction result of age is within expectations. Theoretically, the age of the users will depend on the lexicon they use to communicate. This agrees with the literature used as a basis point [20]. The regression yields good results in the segments of low and middle-aged, but not for older ones. Supporting the idea that users stabilize their lexicon from and after a certain age.

In the case of the consumer classifier, performance metrics were relatively good, all bordering 0.7. This is not a powerful classifier, but it demonstrates the capacity to predict marijuana use only using Twitter based features. Based on the relative weights of each variable, beliefs are confirmed and some results in the literature are replicated. Consumers are concentrated in younger age segments. The popularity of a person increases their risk of using marijuana. The behavior of the environment directly influences the behavior

of people. The publication of any message related to marijuana is linked to a more likely use. However, some measures of closeness to other marijuana users lack sense. Because this implies that a person, surrounded by marijuana users and people that favorable comment about marijuana, is less prone to use marijuana. This is contradictory to previous results of the work that inspired the use of these variables [14].

6 CONCLUSIONS

This study proposes the use of information generated on Twitter to replicate the behavior of Chilean users regarding marijuana. The algorithm involves the combination of various algorithms and procedures to obtain the desired results. The first one, and the most important consists of a Twitter account classifier according to marijuana use, which can be used to calculate the aggregate use. The second corresponds to a methodology to calculate the polarity of the opinion regarding marijuana. Both can be used for real-time monitoring of the prevalence and opinion about the drug.

The relationships between Twitter users are an important set of features of the model, and without them, the predictive power of the marijuana classifier would greatly diminish. This contradicts the results obtained in other studies about social networks outside the virtual context, implying that usage statements of friends diminish the user's own use. The cause of this contradiction can be a kind of rejection to people openly publishing their own marijuana use. Nevertheless, this supports studies indicating that the behavior of an individual is strongly affected by its peers.

There is a vast amount of research on the fields of epidemiology and sentiment analysis on several topics, but never from an approach with the potential to generate real-time information about marijuana use, which can be used to continuously monitor the evolution of the substance's usage behavior. We hope to continue with this work through the involvement of other kinds of media, like images, and replicate our results to alcohol use.

ACKNOWLEDGMENTS

This work was supported partially by the FONDEF IT16I10055 and the Millennium Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16).

REFERENCES

[1] Mir M. Ali, Aliaksandr Amialchuk, and Debra S. Dwyer. 2011. The social contagion effect of marijuana use among adolescents. *PLoS ONE* 6 (2011), 1–6.
[2] J. W. Ayers, B. M. Althouse, and M. Dredze. 2014. Could behavioral medicine lead the web data revolution? *JAMA* 311, 14 (2014), 1399–1400.
[3] Jorge A Balazs and Juan D Velásquez. 2016. Opinion mining and information fusion: A survey. *Information Fusion* 27 (2016), 95–110.
[4] Megan R. Brett. 2012. Topic modeling: a basic introduction. *Journalof Digital Humanities* 2, 1 (2012), 12–16.
[5] E. P. Cabrera-Nguyen, P. Cavazos-Rehg, M. Krauss, L. J. Bierut, and M. A. Moreno. 2016. Young adults' exposure to alcohol- and marijuana-related content on Twitter. *J Stud Alcohol Drugs* 77, 2 (2016), 349–353.
[6] P. A. Cavazos-Rehg, M. Krauss, S. L. Fisher, P. Salyer, R. A. Grucza, and L. J. Bierut. 2015. Twitter chatter about marijuana. *J Adolesc Health* 56, 2 (2015), 139–145.
[7] P. A. Cavazos-Rehg, S. J. Sowles, M. J. Krauss, V. Agbonavbare, R. Grucza, and L. Bierut. 2016. A content analysis of tweets about high-potency marijuana. *Drug Alcohol Depend* 166 (2016), 100–108.
[8] Michael Chary, Nicholas Genes, Andrew McKenzie, and Alex F Manini. 2013. Leveraging social networks for toxicovigilance. *Journal of Medical Toxicology* 9, 2 (2013), 184–191.
[9] Stephanie H. Cook, José A. Bauermeister, Deborah Gordon-Messer, and Marc A. Zimmerman. 2013. Online network influences on emerging adults' alcohol and drug use. *Journal of Youth and Adolescence* 42, 11 (2013), 1674–1686.
[10] Hongying Dai and Jianqiang Hao. 2017. Mining social media data for opinion polarities about electronic cigarettes. *Tobacco Control* 26, 2 (2017), 175–180. DOI: http://dx.doi.org/10.1136/tobaccocontrol-2015-052818 arXiv:http://tobaccocontrol.bmj.com/content/26/2/175.full.pdf
[11] R. Daniulaityte, L. Chen, F. R. Lamy, R. G. Carlson, K. Thirunarayan, and A. Sheth. 2016. "When 'bad' is 'good'": identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health Surveill* 2, 2 (2016), e162.
[12] R. Daniulaityte, R. W. Nahhas, S. Wijeratne, R. G. Carlson, F. R. Lamy, S. S. Martins, E. W. Boyer, G. A. Smith, and A. Sheth. 2015. "Time for dabs": analyzing Twitter data on marijuana concentrates across the U.S. *Drug Alcohol Depend* 155 (2015), 307–311.
[13] Luis Dujovne and Juan D. Velásquez. 2009. Design and implementation of a methodology for identifying website keyobjects. *Knowledge-Based and Intelligent Information and Engineering Systems* (2009), 301–308.
[14] Susan T Ennett, Karl E Bauman, Andrea Hussong, Robert Faris, Vangie A Foshee, Li Cai, and Robert H DuRant. 2006. The peer context of adolescent substance use: Findings from social network analysis. *Journal of Research on Adolescence* 16, 2 (2006), 159–186.
[15] G. Eysenbach. 2011. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am J Prev Med* 40, 5 Suppl 2 (2011), S154–158.
[16] A. Gupta and P. Anand. 2015. Focused web crawlers and its approaches. In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*. 619–622. DOI: http://dx.doi.org/10.1109/ABLAZE.2015.7154936
[17] Kimberly Kobus and David B. Henry. 2009. Interplay of network position and peer substance use in early adolescent cigarette, alcohol, and marijuana use. *The Journal of Early Adolescence* (2009).
[18] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media. In *Proceedings of the 19th International Conference on World Wide Web*. USA, 591–600.
[19] F. R. Lamy, R. Daniulaityte, A. Sheth, R. W. Nahhas, S. S. Martins, E. W. Boyer, and R. G. Carlson. 2016. "Those edibles hit hard": exploration of Twitter data on cannabis edibles in the U.S. *Drug Alcohol Depend* 164 (2016), 64–70.
[20] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. (2013). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/5984
[21] Michael J. Paul and Mark Dredze. 2011. You are what you tweet: analyzing Twitter for public health. *ICWSM* 20 (2011), 265–272.
[22] SENDA. 2015. *Eleventh National Study of Drugs in National Population.* SENDA, Santiago.
[23] Lukas Shutler, Lewis S. Nelson, Ian Portelli, Courtney Blachford, and Jeanmarie Perrone. 2015. Drug use in the Twittersphere: a qualitative contextual analysis of tweets about prescription drugs. *Journal of Addictive Diseases* 34, 4 (2015), 303–310. DOI: http://dx.doi.org/10.1080/10550887.2015.1074505 arXiv:http://dx.doi.org/10.1080/10550887.2015.1074505 PMID: 26364675.
[24] Gino Slanzi, Gaspar Pizarro, and Juan D Velásquez. 2017. Biometric information fusion for web user navigation and preferences analysis: an overview. *Information Fusion* 38 (2017), 12–21.
[25] Sarah A. Stoddard, Jose A. Bauermeister, Deborah Gordon-Messer, Michelle Johns, and Marc A. Zimmerman. 2012. Permissive norms and young adults' alcohol and marijuana use: The role of online communities. *Journal of Studies on Alcohol and Drugs* 73, 6 (2012), 968–975.
[26] Q. Tian, J. Lagisetty, and B. Li. 2016. Finding needles of interested tweets in the haystack of Twitter network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 447–452. DOI: http://dx.doi.org/10.1109/ASONAM.2016.7752273
[27] Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José González-Cristóbal. 2013. TASS - workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural* 50, 0 (2013), 37–44.