Text Classification for Automatic Detection of Alcohol Use-Related Tweets

A Feasibility Study

Yin Aphinyanaphongs, Bisakha Ray, Alexander Statnikov, Paul Krebs
NYU Langone Medical Center
New York, NY
yin.a@nyumc.org

***Abstract*—We present a feasibility study using text classification to classify tweets about alcohol use. Alcohol use is the most widely used substance in the US and is the leading risk factor for premature morbidity and mortality globally. Understanding use patterns and locations is an important step toward prevention, moderation, and control of alcohol outlets. Social media may provide an alternate way to measure alcohol use in real time. This feasibility study explores text classification methodologies for identifying alcohol use tweets. We labeled 34,563 geo-located New York City tweets collected in a 24 hour period over New Year's Day 2012. We preprocessed the tweets into stem/ not stemmed and unigram/ bigram representations. We then applied multinomial naïve Bayes, a linear SVM, Bayesian logistic regression, and random forests to the classification task. Using 10 fold cross-validation, the algorithms performed with area under the receiver operating curve of 0.66, 0.91, 0.93, and 0.94 respectively. We also compare to a human constructed Boolean search for the same tweets and the text classification method is competitive with this hand crafted search. In conclusion, we show that the task of automatically identifying alcohol related tweets is highly feasible and paves the way for future research to improve these classifiers.**

Keywords—social media, twitter, text classification, text categorization, alcohol use

I. INTRODUCTION

The use of social media to monitor substance use trends remain largely unexplored. Surveillance efforts for alcohol use have traditionally consisted of various national and state-level surveys conducted via in-person or telephone interview, such as the National Survey on Drug Use and Health (NSDUH) [1] and the Youth Risk Behavior Survey (YRBS). [2] Accurate and timely data are necessary to understand alcohol use trends, establish national and regional health goals, and inform prevention campaigns. Surveys, however, cannot provide continual and location-based information. Systems that can monitor social media streams, such as Twitter, may hold potential for supplementing survey methods by allowing access to real time data that can be visualized geographically and across time in novel ways that increase the depth of knowledge regarding how populations engage in substance use.

Among social media platforms, Twitter offers unique potential to serve as a tool for tracking substance use. Twitter is a micro-blogging service (with posts limited to 140 characters) founded in 2006 through which users can send messages to a set of followers. It has over 600 million users worldwide with 46% of users logging on daily. In a recent Pew Research survey conducted August-September, 2013, 18% of US adults use Twitter. [3] A higher percentage of Blacks/African-Americans (29%) use Twitter compared with Whites (16%) and Hispanics (16%). Of Twitter subscribers, 31% are 18-29 and 19% are 30-49.1 Interestingly, there are relatively no differences in use by education level, gender, or income suggesting that the data cuts across socioeconomic definitions. In addition, 88% of Twitter users allow public access to their tweets thus minimizing limits to tweet access. [4]

Few studies have examined methods of using social media to track alcohol use. Culotta et al. matched alcohol sales to alcohol use trends detected from Twitter. [5] In addition, Chary et al. estimated underage alcohol consumption from Twitter data. [6] Myslín et al. used a machine learning process to classify tweets to examine sentiment (opinions) regarding use of tobacco products. [7] Their algorithms, however, were not trained to identify instances of use. More work is needed to capture the potential of social media to support alcohol use surveillance.

The data points available via a Twitter stream offer a number of possibilities for visualizing and explaining data. For instance, time stamps and geolocation offer potential to examine trends in specific locales and times, which may potentially be correlated with public health data on accidents and crime. Such information may offer more precise predictions where and when risky behaviors occur, to target resources including public health messaging, and even policing and safety services.

Building applications that use this data relies on a technological foundation that can identify tweets correctly. This feasibility paper explores state of the art machine learning based text classification methodologies for identifying alcohol use tweets. This paper makes several key contributions:

(1) Defines a novel classification task for identifying alcohol use.

(2) Describes a process for labeling tweets that identify alcohol use.

(3) Establishes baseline classification results for this task.

II. Materials and Methods

A. Corpus Construction

In this study, we purchased the full twitter stream from the provider Gnip [8] for a 24 hour period from 12/31/12 at noon to 1/1/13 at noon centered on a 5 mile radius from the center of Manhattan. This geo-located data was provided in json format with metadata regarding each tweet. [9]The corpus contained 34,563 total tweets.

B. Corpus Labels

To initially label tweets, two authors (YA and PK) independently coded a random subsample of 3000 tweets using a draft coding protocol. Both authors held a consensus meeting to discuss labels which did not agree, and to refine the coding protocol.

To confirm the quality of the coding guidelines, both authors blindly labeled 2000 additional tweets. Because of the widely varying prevalence in classes, we calculated Siegel & Castellan's bias adjusted kappa. The resulting kappa was calculated at 0.85. [10] This high kappa suggested that the guidelines and task were sufficiently generalizable. Corpus statistics are listed in Table I and a visualization of tweet frequency over time is illustrated in Figure 1. Finally authors YA and PK independently labeled the remaining 32,563 tweets.

Tweets were considered indicators of alcohol use if they referred to: the act of drinking, intention to drink, location at a bar or liquor store, mention of a specific brand, drinking paraphernalia (e.g. flask), alcohol-related hashtags, or consequences from drinking (e.g. drunk, wasted, tipsy). Tweets were excluded if they referred to others drinking, were from advertisers, were not in English, or indicated location at an establishment that also served food (e.g. Bars and grills were excluded). Table II exemplifies included and excluded tweets in line with these criteria.

TABLE I. Corpus Statistics

Tweet	Label
Number of Tweets	34,563
Number of Tweets discussing alcohol use	725 (2.1%)

C. Tweet Preprocessing

We relied on several preprocessing steps used successfully in other twitter classification studies. [11,12] For each tweet, we removed stopwords, screen names (e.g. @britney), and urls. We then produced 4 encodings of the tweets as shown in Table III using the libshorttext program.[13]

TABLE II. Selected Example Labeled Tweets

Tweet	Label
Out of work having drink happy new years	Positive
Dear boyfriend, please come back from the gym already, the champagne is dying to be popped.	Positive
I'm at beer authority	Positive
Moet and Ciroc on deck	Positive
Yes, that was a huge flask I was carrying	Positive
Casually drunk already on NYE	Positive
#alcoholic #vodka	Positive
Are these guys pounding beers in the booth?	Negative
Just added weyerbacher old heathen on tap. See our full beer menu	Negative
I'm at Black Bear Bar and Grill	Negative

TABLE III. Encoded Datasets

Encoding Name	*Stemmed?*	*Unigram or Bigram*	*Number of Features*
stem_uni	Yes	Unigram	28,436
stem_bi	Yes	Bigram	141,572
uni	No	Unigram	32,974
bi	No	Bigram	150,400

D. Learning Algorithms

We use one baseline text classification and three state of the art text classification algorithms. We chose the baseline algorithm to establish the general difficulty of the task and three of the most recent state of the art classification algorithms.

- Naïve Bayes. This algorithm directly applies Bayes theorem to the classification task. This algorithm assumes that the probability distribution of a feature is independent of another feature, given the class labels. We used the Multinomial Naïve Bayes [14] implementation in the mallet package. [15]
- Linear Support Vector Machine. We employed a linear Support Vector Machine (SVM) classification algorithm. The linear SVM's calculate maximal margin hyperplane(s) separating two or more classes of the data. Linear SVMs demonstrated superior text classification performance compared to other methods [16], and this motivated our use of them. We used a linear SVM classifier implemented in libSVM v3.18. [17] We used the default penalty parameter of 1.
- Bayesian Logistic Regression. We employed Bayesian logistic regression. This algorithm uses a Laplace prior to constrain the coefficients in high dimensions and uses cyclic coordinate descent for the convex optimizer.

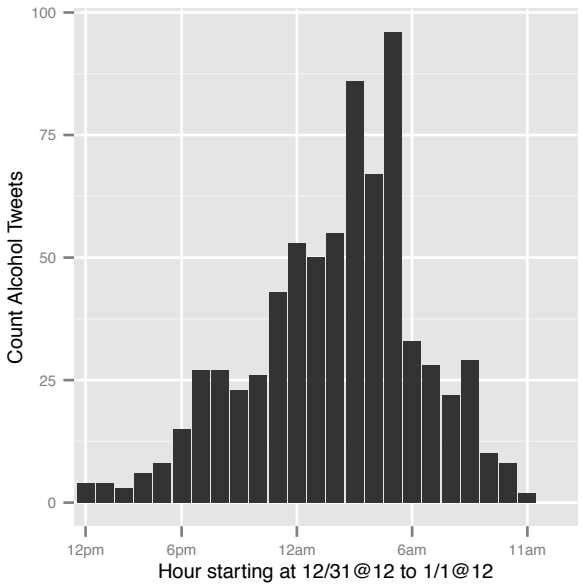

Fig 1: Counts By Hour of Alcohol Tweets

This algorithm demonstrated superior performance in text classification benchmarks. We used the bbrtrain [18] implementation for this study. We used the autosearch option to optimize the regularization parameter. This option does a grid search using 10 fold cross validation across the lamba parameters of 0.01 to 316 in multiples of the square root of 10.

- Random forests. We employed the random forest implementation in the fest [19] program. Random forests [20] are an ensemble classification method. The method produces a classification tree at each iteration. This classification tree is built from a random subset of the data, and at each node in the tree, a random subset of predictor variables are selected. Multiple trees are constructed in this fashion until at test time, the classification of this individual trees are combined to form a final prediction. We use the default settings of the fest package that produces 100 trees with a maximum depth of 1000.

E. Keyword Comparison

We compared the machine learning models to a simple keyword based approach for identifying tweets. Based on our definition, we asked PK to look at our guidelines and generate a single word Boolean keyword set that would provide a relative baseline for this classification task. To simplify the comparison, we chose to compare the keywords to the "unigram" encoding in one 25% holdout split of the data. Table IV shows the keyword "OR" search that was used.

TABLE IV. KEYWORD "OR" SEARCH

drink OR drinker OR drinks OR drinking OR drank OR wine OR champagne OR alcohol OR alcoholics OR alcoholism OR beer OR beers OR bottle OR bottles OR pint OR pints OR cocktail OR cocktails OR bar OR brewery OR lounge OR pub OR liquor OR booze OR vodka OR tequila OR gin OR ciroc OR margarita OR margaritas OR shot OR shots OR ale OR whiskey OR lager OR tipsy OR drunk OR sober OR wasted OR pregame OR pregaming

F. Model Selection and Performance Estimation

We used 10 fold cross-validation [21]to provide unbiased performance estimates of the learning algorithms. The cross-validation procedure randomly divides the corpus into 10 non-overlapping subsets of documents maintaining the proportion of positive and negative documents in each subset. We then repeat 10 times: one subset of documents for testing and the other 9 subsets for training.

III. RESULTS

Table V illustrates the results of our feasibility experiments. Naïve Bayes performed the poorest across the encodings. Random Forests, Bayesian Logistic Regression and the Linear SVM demonstrated excellent performance. The standard deviations across 10 cross validations are tight and accurately represent the likely out of sample generalized area under the curve (AUC) performance.

TABLE V. 10 FOLD CROSS VALIDATION AREA UNDER THE CURVE (AUC) PERFORMANCE (STANDARD DEVIATION)

Encoding Name	***Naïve Bayes***	***Linear SVM***	***Bayesian Logistic Regression***	***Random Forests***
uni	0.66 (0.06)	0.91 (0.03)	0.93 (0.02)	0.94 (0.02)
bi	0.67 (0.03)	0.91 (0.03)	0.93 (0.03)	0.94 (0.02)
stem_uni	0.68 (0.02)	0.90 (0.04)	0.93 (0.01)	0.94 (0.02)
stem_bi	0.66 (0.06)	0.91 (0.03)	0.93 (0.02)	0.94 (0.03)

Having established the tight bounds on the performance estimates in Table V, Tables VI and VII compares a keyword search on one 75/25 split of the data. We reuse a method in our prior work to compare keywords to a machine learning anking. [22] Specifically, the keyword search does not rank documents and cannot generate an AUC. Thus to make the comparison, we take the fixed sensitivity or specifity of the keyword search and compare the results to the corresponding sensitivity and specificity of the ranked documents.

IV. DISCUSSION

Naïve Bayes performed the poorest across all groups. The linear SVM, Bayesian Logistic Regression, and random forests algorithms performed with high area under the curve.

Though we do not formally compare algorithms in this study, we find it unsurprising that the random forests performed well.

TABLE VI. PERFORMANCE COMPARISON TO KEYWORD SEARCH AT FIXED SENSITIVITY

Algorithm	*Fixed Sensitivity*	*Specificity*	*Area Under the Curve for This Fold*
Keyword	0.73	0.98	Not Applicable
Linear SVM		0.96	0.90
Bayesian Logistic Regression		0.97	0.93
Random Forests		0.98	0.94

TABLE VII. PERFORMANCE COMPARISON TO KEYWORD SEARCH AT FIXED SPECIFICITY

Algorithm	*Sensitivity*	*Fixed Specificity*	*Area Under the Curve for This Fold*
Keyword	0.73	0.98	Not Applicable
Linear SVM	0.64		0.90
Bayesian Logistic Regression	0.67		0.93
Random Forests	0.73		0.94

Ensemble methods demonstrated superior performance in many applications and this example further supports this finding. [23] More thorough statistical comparisons as in our prior work [24] would be necessary to make a more definitive statement.

The performance of the "OR" keyword query is interesting. The machine learning algorithms demonstrates comparable performance though the comparison is not exactly apples to apples. Specifically, the machine learning algorithms is trained on a 75% and tested on 25% of the data. In contrast, PK created the keyword query after labeling the collection. PK likely overfit to this corpus and thus the resulting high performance keyword query. A more consistent design is to design a keyword query for the one fold and then measure performance.

The previous observation also may explain why the keyword query works so well in comparison to the machine learning algorithms. One potential explanation is that the human produced keyword query relies on knowledge and can identify words about alcohol use that do not occur often in the corpus. For example, "ciroc" was identified by the reviewer as a brand of alcohol. This term only appears twice in the corpus as positive instances. A statistically based machine learning technique that is cross-validated will have difficulty learning a model to identify these tweets.

We also would comment on Figure 1 and the trend toward alcohol use tweets in the morning hours. Though we leave for future work inspection of the content of the tweets throughout the 24 hour period, we hypothesize that the peak is a result of consequences related to alcohol use. For example, in our definition of alcohol use, we consider being wasted or tipsy indicators of alcohol use.

Another comment is the relatively low volume of tweets in this 24 hour period. The absolute count of alcohol related tweets seems low in comparison to the number of people drinking on one of the heaviest drinking days of the year. The tweets in this corpus are geo-located and thus are a subset of all tweets sent. Researchers have estimated that 0.77% of all public tweets are geo-located. [25] Despite these low counts, researchers identifying trends in other areas such as the flu are able to correlate to validated publicly available data. [26] Only additional testing, external validation, and further use cases will determine whether these low counts are sufficient for a specific application.

Another point is that it is the trend between these tweets that may be important. Some applications may only require understanding trends of use. These tweets accumulate over time and broader pictures of geo-located alcohol use may develop over subsequent tweets.

We note that we were quite liberal in our classification guidelines in identifying tweets that were indicators of alcohol use. In this feasibility study, we wanted to address a broad spectrum of tweets that indicate alcohol use. In future studies, we may build classifiers for individual types of tweets

Though the classifier performs well, it is not perfect. The necessary performance is dependent on the specific application. For example, an application that may identify locations for alcohol moderation and cessation campaigns may allow a degree of false positives and false negatives as the use of the classifier over time will identify potential venues for intervention. Another application that may identify underage drinking and venues that are selling to minors will need classifiers that have limited false positives.

V. LIMITATIONS

The labeled corpus in this collection is limited to one 24 hour day for one specific task in one US city in one language. Whether these techniques will generalize to other days, languages, or other cities we leave for future work. We would hypothesize that these results would generalize as long as the language for alcohol use remains the same.

We purposely limited the number of algorithms and optimizations to establish feasibility of this task. We chose not to use additional classifiers or more computationally heavy performance estimation designs such as nested cross validation to identify parameters. We also avoided comparing feature selection algorithms that may have improved some of the classifiers performance. For this study, we limited

ourselves to a baseline and top performing text classifiers to establish feasibility of this task.

We chose to focus this paper on the feasibility of using machine learning text classification approaches to identify alcohol use. A point of future work is to apply a more robust study design comparing classifiers as in our prior work. [24]

Finally, we leave for future work efforts to externally validate these results against more traditional methods of surveillance using surveys.

VI. Conclusions

This feasibility study showed that it is indeed possible to classify alcohol use tweets using modern text classification algorithms. This work paves the way for future studies developing these capabilities.

References

1. SAMHSA. National Survey on Drug Use and Health. 2014; https://nsduhweb.rti.org/respweb/homepage.cfm. Accessed Mar 4, 2014.
2. CDC. Youth Risk Behavior Surveillance System. 2014; http://www.cdc.gov/HealthyYouth/yrbs/index.htm. Accessed Mar 1, 2014.
3. Pew Research Internet Project. Social Media Update 2013. 2013; http://www.pewinternet.org/2013/12/30/social-media-update-2013/. Accessed Feb 20, 2014.
4. Beevolve.com. An Exhaustive Study of Twitter Users Across the World. 2012; http://www.beevolve.com/twitter-statistics/. Accessed Mar 1, 2014.
5. Culotta A. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*: Springer; 2013:1-22.
6. Chary MA, Genes N, Manini A. Underage Alcohol Consumption; Using Social Media for Syndromic Surveillance. *New York City Epidemiology Forum First Annual Conference*2014:73.
7. Myslin M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research.* 2013;15(8):e174.
8. Gnip. The Source for Social Data - Gnip. *gnip.com* 2013; http://gnip.com. Accessed March 25, 2014.
9. Gnip. Gnip Data Format. 2014; http://support.gnip.com/sources/twitter/data_format.html. Accessed Jun 14, 2014.
10. Siegel S CN. *Nonparametric statistics for the behavioral sciences.* New York; McGraw Hill; 1988.
11. Kouloumpis E, Wilson T, Moore J. Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.* 2011.
12. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation.* 2010.
13. H.-F. Yu C-HH, Y.-C. Juan, C.-J. Lin. LibShortText: A Library for Short-text Classification and Analysis. 2013.
14. A.M. Kibriya EF, B. Pfahringer, and G. Holmes. Multinomial naive bayes for text categorization revisited. *Lecture notes in computer science.* 2004.
15. McCallum AK. MALLET: A Machine Learning for Langauge Toolkit. 2002; http://mallet.cs.umass.edu.
16. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms (The Springer International Series in Engineering and Computer Science)* [computer program]. Version 1: Springer; 2002.
17. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST).* Vol 2: ACM; 2011:27.
18. Genkin A, Lewis DD, Madigan D. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics.* Vol 492007:291-304.
19. Karampatziakis N. Fast Ensembles of Sparse Trees (FEST). 2014; http://lowrank.net/nikos/fest/. Accessed Jun 14, 2014.
20. Breiman L. Random forests. *Machine Learning.* Vol 452001:5-32.
21. Hastie T, Tibshirani R. The elements of statistical learning second edition. *Springer Series in Statistics*: Springer; 2009.
22. aphinyanaphongs Y. Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. *Journal of the American Medical Informatics Association.* Vol 122005:207-216.
23. Seni G, Elder JF. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predicitons.* Morgan and Claypool Publishers; 2010.
24. Aphinyanaphongs Y, Fu LD, Li Z, et al. A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *J Assn Inf Sci Tec*2014.
25. Semiocast. Twitter reaches half a billion accounts More than 140 millions in the U.S. 2012; http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. Accessed Jun 14, 2014.
26. David A. Broniatowski MJP, Mark Dredze. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PloS one.* 2013.