Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

DUI: the drug use insights web server

Zachary Prince, Deeptanshu Jha, and Rahul Singh*

Department of Computer Science, San Francisco State University, San Francisco, CA, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Substance abuse constitutes one of the major contemporary health epidemics. Recently, the use of social media platforms has garnered interest as a novel source of data for drug addiction epidemiology. Often however, the language used in such forums comprises slang and jargon. Currently, there are no publicly available resources to automatically analyse the esoteric language-use in the social media drug-use sub-culture. This lacunae introduces critical challenges for interpreting, sensemaking and modeling of addiction epidemiology using social media.

Results: Drug-Use Insights (DUI) is a public and open-source web application to address the aforementioned deficiency. DUI is underlined by a hierarchical taxonomy encompassing 108 different addiction related categories consisting of over 9,000 terms, where each category encompasses a set of semantically related terms. These categories and terms were established by utilizing thematic analysis in conjunction with term embeddings generated from 7,472,545 Reddit posts made by 1,402,017 redditors. Given post(s) from social media forums such as Reddit and Twitter, DUI can be used foremost to identify constituent terms related to drug use. Furthermore, the DUI categories and integrated visualization tools can be leveraged for semantic- and exploratory analysis. To the best of our knowledge, DUI utilizes the largest number of substance use and recovery social media posts used in a study and represents the first significant online taxonomy of drug abuse terminology.

Availability: The DUI web server and source code are available at: http://haddock9.sfsu.edu/insight/

Contact: rahul@sfsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Substance abuse destroys millions of lives in multifarious interconnected ways that result in cascading detrimental social impact (Murthy *et al.*, 2017). Data collection and analysis for modeling the epidemiology of substance use is challenging; the use of clinical reports and surveys for gathering information is prone to the problem of underreporting owing to the hesitancy of users to be open. Furthermore, (predefined) surveys often fail to capture user-specific context. Given these problems, semi-anonymous social media sources such as Reddit and Twitter, containing unfiltered self-reported information can be a valuable source of data for substance-use epidemiology. However, current data collection and analysis methods utilizing social media are stymied by the informal and colloquial nature of social media discourse (Sloane *et al.*, 2015). In part this challenge is reflected by the inability of the two most popular computerized text analysis tools LIWC (Pennebaker *et al.*, 2015) and Empath (Fast *et al.*, 2016) to adequately describe the particular nature of drug use-related discourse even though these methods are extremely useful for analyzing generic text (Supplementary Section S1). The DEA drug list (DEA 2018) and NOSLANG dictionary (NOSLANG 2021) focus on drug use and can serve as putative alternatives. However, these sources primarily consist of a list of drug terms and crucially, do not provide a vocabulary to study other contextual aspects concomitant with drug use. For example, none of these compendiums have a list of terms associated with acquisition and administration of drugs or addiction recovery of an individual. For further details see Supplementary Section S1 and Table S1 and S2.

DUI – an online system for analyzing addiction related textual communications fills this lacuna. For a given set of posts, the most significant application of DUI is to identify terms related to substance-use. DUI also includes a number of visualization tools to support experiential exploration and analysis of the data (Figure 1). Underlying DUI is a data-driven drug use and recovery-focused taxonomy consisting of 108 different categories and sub-categories manifesting different aspects of drug use sub-culture. These categories encompass over 9,000 terms including formal terms, slangs, misspellings, and phonetic substitutions. Example categories present in DUI include: *acquisition*, *administration*, *effects*, *withdrawal symptoms*, *recovery*, and *relapse*. Readers are referred to Supplemental section S9 for a complete list of categories. The taxonomy underlying DUI is updated annually (Supplemental section S8). DUI also allows users to create new

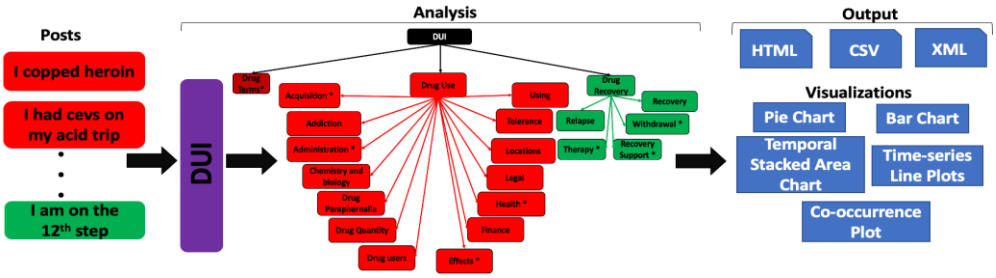

Figure 1. The operational semantics of DUI: given one or more social media posts, DUI identifies terms reflecting substance use and classifies them using over 108 categories and sub-categories. DUI provides its output as html, csv and xml files. Furthermore, it incorporates a number of visualization tools to analyze the categories and constituent terms.

categories by leveraging its underlying term embedding (Supplementary section S6).

2 The DUI System

2.1 Data Collection

DUI was developed by utilizing 7,472,545 posts made by 1,402,017 unique redditors in 117 recreational drug use (RDU) subreddits and 29 drug addiction recovery (DAR) subreddits during the period December 31, 2010 to June 27, 2020 (Supplementary Section S2).

2.2 Determination of Categories and Seed Terms

We used the DEA report (DEA 2018) and thematic analysis (Braun and Clarke 2006) to develop the categories present in DUI (Supplemental section S3). The drug classes in the DEA report were used to initiate the creation drug categories in DUI. The terms in the DEA report for each drug category were then used as "seed terms" to identify semantically related terms and expand the categories. To create contextual categories representing activities concomitant with substance use and recovery, such as, acquisition, ingestion, and relapse, that are not present in any existing compendium, we used thematic analysis - a qualitative data analysis technique used for identifying common patterns (topics or categories) present in a text corpus (Supplementary Section S3). We used a set of 504 random posts consisting of 252 posts from RDU subreddits and 252 posts from DAR subreddits to identify themes (and their representative terms). The expansion of these themes were used as contextual categories and supplemented the categories derived from the DEA report (Supplementary Section S3).

Once the initial categories and their seed terms were finalized, the continuous bag-of-words and skip-gram term embedding obtained by applying the word2vec algorithm (Mikolov *et al.*, 2013) on the set of 7,472,545 Reddit posts (Supplemental section S4) was used to expand the categories as follows (Supplementary section S5 and Fig S1): (1) each distinct pair of seed terms (s_a, s_b) in a category was combined with the term "drug" to create a query triplet <s_a , s_b , "drug">. This triplet was subsequently used to search the embedding and identify similar *n*-grams. The query triplet allowed us to ameliorate the problem of polysemy associated with many terms that are encountered in drug-use discourse and focus the search (Supplemental section S5 and Table S8), (2) only *n*-grams whose cosine similarity to the query triplet exceeded an empirically defined threshold of 0.4 and which occurred in at least 9% of the retrieved results were included after manual verification (Supplemental section S5), and (3) the process was iterated until the number of new terms that could be identified became small. Supplemental section S9 presents the final categories and terms in DUI.

2.3 Creation of novel categories

DUI utilizes the embedding to support creation of novel categories dynamically. Given seed terms that are positively and negatively related to a new category, DUI calculates a category vector by taking the mean of the vectors of the terms in this set. Subsequently, the top terms most similar to the category vector are returned (Supplementary section 6) using the skip-gram and CBOW embedding.

2.4 Implementation and Usage

DUI was developed using Python 3 and Django, along with CSS/JavaScript, SQL Lite and HTTPD. Social media posts can be uploaded to DUI as zipped folder of text or csv files to undergo the following types of time-stamped or non-time-stamped analyses (Supplementary section S10): (1) single user analysis, (2) multiple user analysis and comparison, and (3) follow-up analysis (of single or multiple subject) by uploading prior results from DUI. After correcting for any misspellings, the posts are processed to determine the categories and relevant terms (Supplemental section S7). The software also provides integrated visualization support for exploration and interpretation of the results (Supplementary section S10). The results from DUI can be downloaded in html, csv, or xml formats. Due to the semantic complexity of substance use discourse (Supplemental section S10.4), it is recommended that the output from DUI be used for drawing conclusions by keeping the discourse context and complexity in consideration.

Paraphrased Post	Categories and terms identified using DUI
Okay so today is my 61st day in *recovery*. I was deep into dealing and using *acid* and *DMT*. I was eventually put into *rehab*, which I quickly signed myself out of. Now Im in *IOP* (*intensive out patient*) and have been going to *NA meetings* regularly since.	**drug_recovery.recovery:** recovery; **drug_recovery.recovery_support.aa**: meetings, na, na_meetings; **drug_recovery.recovery_support.recovery programs**: out_patient, iop, iop_intensive; **drug_recovery.recovery_support.rehab**: rehab; **drug_terms.hallucinogens:** acid, dmt

Table 1. Different categories and terms identified using DUI for a paraphrased Reddit post. The DUI category hierarchy is shown with sub-categories delineated by a period. Categories have been bolded and separated using semi-colons.

2.5 Data Visualization and Analysis

DUI provides a number of integrated visualization tools. These include a *bar chart* (Figure 2 (A)) and a *pie chart (*Supplementary Figure S6) both of which can be used to visualize the presence and frequency of different DUI categories in the posts. The semantic correlations in the data can be examined through a *category correlation plot* (Figure 2 (B)) which helps identify co-occurring categories. Users can also drill down into the categories and generate *n*-gram-level *bar charts* and *pie charts* for each category (Supplementary Figure S8). If time-stamped posts from the subject(s) are uploaded, then DUI provides tools for temporal analysis though various *time series plots* (Figure 2(C) and Figure 2(D)), and a *stacked area chart* (Supplementary Figure S10). These tools can be used, among others, to perform longitudinal analysis of the post contents. Finally, to visualize the posts from multiple users across different categories, DUI provides a side-by-side panel view of all the aforementioned visualization functionalities.

3 Results

In Table 1 we present the categories and terms identified by DUI for a (paraphrased) post. For this post, DUI identified terms related to a number of distinct drug-use and recovery categories. Other examples

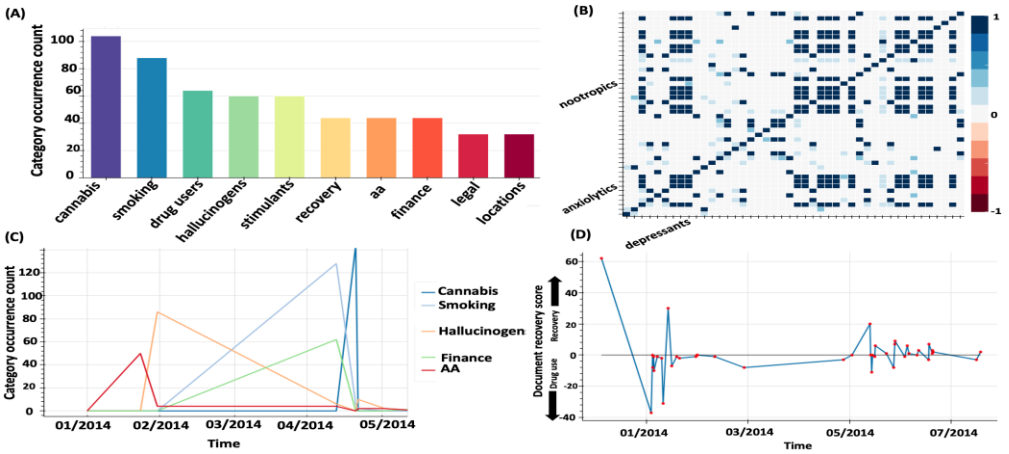

Figure 2. Visualization and data analysis functionality in DUI: (A) the bar chart displays the categories present in posts of a user sorted by their frequency of occurrence, (B) the category co-occurrence plot displays the correlation between different categories present in user posts, (C) the time series chart can be used to visualize thematic changes in posts over time. This example captures a case where the frequency of the category AA (alcoholics anonymous) decreases while the frequency drug use categories (hallucinogens, cannabis, and smoking) increases in posts of a user. D) A longitudinal plot of the difference between terms related to the frequency of recovery and drug use. In this case, a user had a high number of recovery-related terms in the posts. Subsequently however, the number of such terms decreased, while the number of drug-use terms increased, possibly indicating a relapse.

describing the results from applying DUI are presented in Supplemental Tables S12-S14. The process of creating novel categories in DUI, given a set of positive and negative seed terms, is explained through an example in Supplementary section S6 and Table S9.

	Number of terms identified	Number of True Positives	Number of False Positives	Accuracy (%)
DUI	**3,660**	**3,068**	**592**	**83.8**
DEA	1,469	160	1,309	10.8
NOSLANG	3,181	311	2,870	9.7

Table 2. Number of terms identified and accuracy comparison for DUI, DEA, and NoSlang dictionary for a random set of 500 Reddit posts. DUI identifies 3,068 terms correctly compared to 160 and 311 terms identified correctly by DEA and NoSlang. DUI's accuracy for the terms identified is 0.838 compared to 0.108 (DEA) and 0.097 (NoSlang).

Number of terms identified	Number of True Positives	Number of False Positives	Number of False Negatives	Precision	Recall
1,297	1,105	192	69	0.85	0.94

Table 3. Number of terms identified, true positives, false positives, and false negatives (terms missed) for applying for DUI on a set of 318 tweets. Out of the 1,297 terms manually identified to be relevant, DUI identified 1,105 terms correctly and 192 terms incorrectly. Of these only 69 substance-use-related terms were not present in the DUI taxonomy.

To comparatively determine the efficacy of DUI vis-à-vis existing resources, we experimentally compared it with the DEA term list (DEA 2018) and the NOSLANG dictionary (NOSLANG 2021). The DEA term list contains of slang and code terms made available by the United States Drug Enforcement Administration as a reference for law enforcement personnel. Similarly, the NOSLANG dictionary is a public resource containing slang terms for substance use and recovery. In this experiment, we randomly selected a set of 500 Reddit posts and used DUI, the DEA term list, and the NOSLANG dictionary to identify substance use and recovery related terms present in these posts. Subsequently, the results from each system were manually examined to determine if these terms were correctly or incorrectly identified (Table 2) and DUI was found to significantly outperform both the other resources.

To study the application of DUI to data from a social media platform that was not used in its creation, a set of 318 tweets from Twitter (Supplementary section 2) were analyzed and the results manually assessed (Table 3). The performance of DUI on this dataset was found to be comparable that on posts from Reddit. This result supports the hypothesis that substance-use language is content-driven and not platform dependent and underlines the broad applicability of DUI.

Acknowledgment

The authors thank the anonymous reviewers for their comments.

Funding

This work was funded by the National Science Foundation grant IIS-1817239 and the National Institutes for Health grant R25MD011714.

4 References

Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), pp.77-101.

DEA.gov. (2018) Slang Terms and Code Words: A Reference for Law Enforcement Personnel. The U.S. Drug Enforcement Administration https://www.dea.gov/sites/default/files/2018-07/DIR-022-18.pdf

Fast, E., Chen, B. and Bernstein, M.S., 2016, May. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4647-4657).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Murthy VH. Facing addiction in the United States: The surgeon general's report of alcohol, drugs, and health. Jama. 2017 Jan 10;317(2):133-4.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. 2015 Sep 15.

Sloane, R., Osanlou, O., Lewis, D., Bollegala, D., Maskell, S. and Pirmohamed, M., 2015. Social media and pharmacovigilance: a review of the opportunities and challenges. *British journal of clinical pharmacology*, *80*(4), pp.910-920.

NOSLANG (2021). https://www.noslang.com/ . 02/07/2021.