A Computational Analysis of the Language of Drug Addiction

Carlo Strapparava
FBK-irst,
Trento, Italy
strappa@fbk.eu

Rada Mihalcea
University of Michigan,
Ann Arbor, USA
mihalcea@umich.edu

Abstract

We present a computational analysis of the language of drug users when talking about their drug experiences. We introduce a new dataset of over 4,000 descriptions of experiences reported by users of four main drug types, and show that we can predict with an F1-score of up to 88% the drug behind a certain experience. We also perform an analysis of the dominant psycholinguistic processes and dominant emotions associated with each drug type, which sheds light on the characteristics of drug users.

1 Introduction

The World Drug Report globally estimated that in 2012, between 162 million and 324 million people, corresponding to between 3.5 per cent and 7.0 per cent of the world population aged 15-64, had used an illicit drug (United Nations Office, 2014). Moreover, in recent years, drug users have started to share their experiences on Web forums.[1] The availability of this new and very large form of data presents new opportunities to analyse and understand the "drug use phenomenon." Recent studies have shown how by processing these data with language processing techniques, it is possible to perform tasks of toxicovigilance, e.g., finding new drugs trends, adverse reactions, geographic and demographic characterizations (Chary et al., 2013). Other studies have also focused on the phenomenon of intoxication (Schuller et al., 2014). However, despite the interest around these topics, as far as we know, textual corpora of drug addicts experiences are not yet available.

In this paper we introduce a corpus that can be exploited as a basis for a number of computational explorations on the language of drug users. One of the most controversial and interesting issues in addictionology studies is to understand why drug consumers prefer a particular type of drug over another. Actually differentiating drugs with respect to their subjective effects can have an important impact on clinical drug treatment, since it can allow clinicians to better characterize the patient in therapy, with regard to the effect they seek through the drugs they use.

The paper is organized as follows. We first review the related work, followed by a description of the dataset of drug addict experiences that we constructed. Next, we present a classification experiment on predicting the drug behind an experience. We then present specific analyses of the language of drug users, i.e. their psycholinguistic processes and the emotions associated with an experience. Lastly, we conclude the paper and present some directions for future work.

2 Related Work

An important research on texts from social media was the platform PreDOSE (Cameron et al., 2013), designed to facilitate the epidemiological study of prescription (and related) drug abuse practices, or its successors: eDrugTrends[2] and iN3.[3] Another significant work was that of Paul and Dredze (2012; 2013). They developed a new version of Blei's LDA, factorial LDA, and for each drug, they were able to collect multiple topics (route of administration, culture, chemistry, etc.) over posts collected from the website www.drugs-forum.com. The main directions

[1] www.erowid.org: 95000 unique visitor per day; www.drugs-forum.com: 210000 members with 3.6 million unique visitor per month; www.psychonaut.com: 46000 members.

[2] http://medicine.wright.edu/citar/edrugtrends

[3] http://medicine.wright.edu/citar/nida-national-early-warning-system-network-in3-an-innovative-approach

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 136–142,
Valencia, Spain, April 3-7, 2017.

of research on the state of consciousness are focused on alcoholic intoxication and mostly performed on the Alcohol Language Corpus (Schiel et al., 2012), only available in German: for example, speech analysis (Wang et al., 2013; Bone et al., 2014) and a text based system (Jauch et al., 2013) were used to analyse this data. Regarding alcohol intoxication detection, (Joshi et al., 2015) developed a system for automatic detection of drunk people by using their posts on Twitter. (Bedi et al., 2014) performed their analysis on transcriptions from a free speech task, in which the participants were volunteers previously administered with a dose of MDMA (3,4-methylenedioxy-methamphetamine). Even if this is an ideal case study for analyzing cognitively the intoxication state, it is difficult to replicate on a large scale. Finally, as far as we know, the only attempt to classify and characterize experiences over different kinds of drugs was the project of (Coyle et al., 2012). Using a random-forest classifier over 1,000 random-collected reports of the website `www.erowid.org` they identified subsets of words differentiated by drugs.

Our research is also related to the broad theme of latent user attribute prediction, which is an emerging task within the natural language processing community, having recently been employed in fields such as public health (Coppersmith et al., 2015) and politics (Conover et al., 2011; Cohen and Ruths, 2013). Some of the attributes targeted for extraction focus on demographic related information, such as gender/age (Koppel et al., 2002; Mukherjee and Liu, 2010; Burger et al., 2011; Van Durme, 2012; Volkova et al., 2015), race/ethnicity (Pennacchiotti and Popescu, 2011; Eisenstein et al., 2011; Rao et al., 2011; Volkova et al., 2015), location (Bamman et al., 2014), yet other aspects are mined as well, among them emotion and sentiment (Volkova et al., 2015), personality types (Schwartz et al., 2013; Volkova et al., 2015), user political affiliation (Cohen and Ruths, 2013; Volkova and Durme, 2015), mental health diagnosis (Coppersmith et al., 2015) and even lifestyle choices such as coffee preference (Pennacchiotti and Popescu, 2011). The task is typically approached from a machine learning perspective, with data originating from a variety of user generated content, most often microblogs (Pennacchiotti and Popescu, 2011; Coppersmith et al., 2015; Volkova et al., 2015), article comments to news stories or op-ed pieces (Riordan et al., 2014), social posts (originating from sites such as Facebook, MySpace, Google+) (Gong et al., 2012), or discussion forums on particular topics (Gottipati et al., 2014). Classification labels are then assigned either based on manual annotations (Volkova et al., 2015), self identified user attributes (Pennacchiotti and Popescu, 2011), affiliation with a given discussion forum type, or online surveys set up to link a social media user identification to the responses provided (Schwartz et al., 2013). Learning has typically employed bag-of-words lexical features (ngrams) (Van Durme, 2012; Filippova, 2012; Nguyen et al., 2013), with some works focusing on deriving additional signals from the underlying social network structure (Pennacchiotti and Popescu, 2011; Yang et al., 2011; Gong et al., 2012; Volkova and Durme, 2015), syntactic and stylistic features (Bergsma et al., 2012), or the intrinsic social media generation dynamic (Volkova and Durme, 2015). We should note that some works have also explored unsupervised approaches for demographic dimensions extraction, among them large-scale clustering (Bergsma et al., 2013) and probabilistic graphical models (Eisenstein et al., 2010).

3 Dataset

A corpus of drug experiences was collected from the user forum section of the `www.erowid.org` website. The data collection was performed semi-automatically, considering the most well-known drugs and those with a large number of reports. The corpus consists of 4,636 documents, any user ID removed, split into four main categories according to their main effects (U.S. Department of Justice, 2015): (1) **Empathogens** (EMP), covering the following substances: MDA, MDAI, MDE, MBDB, MDMA; (2) **Hallucinogens** (HAL), which include 5-MeO-DiPT, ayahuasca, peyote, cacti (trichocerus pachanoi, peruvianus, terschekcii, cuzcoensis, bridgesi and calea zachatechichi), mescaline, cannabis, LSD, belladonna, DMT, ketamine, salvia divinorum, hallucinogen mushrooms (psilocybe cubensis, semilanceata, 'magic mushrooms'), PCP, 2C-B and its derivatives (2C-B-FLY, 2C-E, 2C-I, 2C-T-2,2C-T-7); (3) **Sedatives** (SED), which include alcohol, barbitures, buprenorphine, heroin, morphine, opium, oxycodone, oxymorphone, hydrocodone, hydromorphone, methadone, nitrous-

oxide, DXM (dextromethorphan) and benzodiazepines (alprazolam, clonazepam, diazepam, flunitrazepam, flurazepam, lorazepam, midazolam, phenazepam, temazepam); (4) **Stimulants** (STI), including cocaine, caffeine, khata edulis, nicotine, tobacco, methamphetamines, amphetamines.

In the scientific literature about drug users, "purists" (i.e., consumers of only one specific substance) are rare. Nonetheless, when collecting the data, we decided to consider only reports describing one single drug in order to avoid the presence of a report in multiple categories, as well as to avoid descriptions of the interaction of multiple drugs, which are hard to characterize and still mostly unknown. Table 1 presents statistics on the dataset, while Table 2 shows excerpts from experiences reported for each drug type.[4]

Drug type	Number reports	Total words
EMP	399	378,478
HAL	2,806	3,494,223
SED	954	692,121
STI	480	449,596

Table 1: Corpus statistics.

4 Predicting the Drug behind an Experience

To determine if an automatic classifier is able to identify the drug behind a certain reported experience, we create a document classification task using Multinomial Naïve Bayes, and use the default information gain feature weighting associated with this classifier. Each document corresponds to a report labelled with its corresponding drug category. Only minimal preprocessing was applied, i.e., part-of-speech tagging and lemmatization. No particular feature selection was performed, only stopwords were removed, keeping nouns, adjectives, verbs, and adverbs. Since the major class in the experiment was the hallucinogens category, we set the baseline corresponding to its percentage: 61%. In evaluating the system we perform a five-fold cross-validation, with an overall F1-score (micro-average) of 88%, indicating that good separation can be obtained by an automatic classifier (see Table 3). Not surprisingly, the hallucinogen experiences are the easiest to classify, probably due to the larger amount of data available for this drug.

Table 4 shows a sample of the most informative features for the four categories. For example, we can observe that those using emphatogens are more "night"-oriented, while those addicted to sedatives and stimolants are "day"-oriented. Instead, the use of hallucinogens seems to be associated with a perceptual visual experience (i.e., `see#v`).

5 Understanding Drug Users

5.1 Psycholinguistic Processes

To gain a better understanding of the characteristics of drug users, we analyse the distribution of psycholinguistic word classes according to the Linguistic Inquiry and Word Count (LIWC) lexicon – a resource developed by Pennebaker and colleagues (Pennebaker and Francis, 1999). The 2015 version of LIWC includes 19,000 words and word stems grouped into 73 broad categories relevant to psychological processes. The LIWC lexicon has been validated by showing significant correlation between human ratings of a large number of written texts and the rating obtained through LIWC-based analyses of the same texts.

For each drug type T, we calculate the dominance score associated with each LIWC class C (Mihalcea and Strapparava, 2009). This score is calculated as the ratio between the percentage of words that appear in T and belong to C, and the percentage of words that appear in any other drug type but T and belong to C. A score significantly higher than 1 indicates a LIWC class that is dominant for the drug type T, and thus likely to be a characteristic of the experiences reported by users of this drug.

Table 5 shows the top five dominant psycholinguistic word classes associated with each drug type. Interestingly, descriptions of experiences reported by users of empathogens are centered around people (e.g., Affiliation – which includes words such as club, companion, collaborate; We; Friend). Hallucinogens result in experiences that relate to the human senses (e.g., See, Hear, Perception). The experiences of users of sedatives and stimulants appear to be more concerned with mundane topics (e.g., Money, Work, Health).

To quantify the similarity of the distributions

[4]Note that each report is annotated with a set of metadata attributes, such as gender, age at time of experience, dose and number of views; these attributes are not used in the experiments reported in this paper, but we plan to use them for additional analyses in the future.

Drug Type	Example
EMP	I found myself witnessing an argument between a man and a woman whom I've never met. I felt empathetic towards both of them, recognizing their struggle, he meant well, but couldn't find the right words, she, obviously cared a great deal for him but was doubtful of his intentions. The Argument escalated and I became very disturbed...I had to open my eyes again. My heart rate was up, my breathing was heavy, I had found a window to my own fears, to see what frustrates you the most, and not be able to do anything about it.
HAL	After watching TV for a bit I looked around the room and was suddenly jerked awake, I felt vibrant, alive and aware of my entire physical body. The friction of blood in my veins, the movement of my diaphragm, the tensing of muscles, the clenching of my heart. I looked down at my hands and was acutely aware of the bones within, I could feel the flesh sliding over the bone internally while my normal sense of touch was reduced so every thing felt like cold chrome.
SED	Feeling kind of nausea, but I'm not worried about throwing up. Shooting great pool, I'm making several shots in a row. I'm so happy right now, I would like to be like this all day. I'm begining to notice that I'm having slight audio hallucinations, like hearing small noises that aren't there. Also some slight visual hallucinations, thinking I see something move nearby but nothing alive is even close to me.
STI	I get up in the morning for work and do about two lines while I'm getting ready and somehow manage to make it through work without a line. Not that I don't want to only because of the fear of getting caught. I can say that it takes the edge off things at work though. Through the evening I do a line whenever I feel like it. At bedtime I tell myself over and over that it's time to go to sleep. Sometimes I sleep but if I can't I know I have my friend to help me through the next day.

Table 2: Sample entries in the drug dataset.

	Prec.	Rec.	F1
EMP	0.84	0.71	0.77
HAL	0.93	0.92	0.92
SED	0.86	0.86	0.86
STI	0.73	0.85	0.78
micro-average			0.88

Table 3: Naïve Bayes classification performance.

EMP	experience#n good#a pill#n people#n about#r drug#n night#n start#v
HAL	see#v experience#n trip#n look#v back#r say#v try#v down#r as#r
SED	day#n drug#n start#v about#r try#v good#a hour#n still#r effect#n
STI	day#n drug#n coke#n good#a try#v start#v about#r want#v really#r

Table 4: Most informative features (words and parts-of-speech).

of psycholinguistic processes across the four drug types, we also calculate the Pearson correlation between the dominance scores for all LIWC classes. As seen in Table 6, empathogens appear to be the most dissimilar with respect to the other drug types. Hallucinogens instead seem to be most similar to stimulants and sedatives.

5.2 Emotions and Drugs

Another interesting dimension to explore in relation to drug experiences is the presence of various emotions. To quantify this dimension, we use a methodology similar to the one described above, and calculate the dominance score for each of six emotion word classes: anger, disgust, fear, joy, sadness, and surprise (Ortony et al., 1987; Ekman, 1993). As a resource, we use WordNet Affect (Strapparava and Valitutti, 2004), in which words from WordNet are annotated with several emotions. As before, the dominance scores are calculated for the experiences reported for each drug type when compared to the other drug types.

Table 7 shows the scores for the four drug types and the six emotions. A score significantly higher than 1 indicates a class that is dominant in that category. Clearly, interesting differences emerge from this table: the use of emphathogens leads to experiences that are high on joy and surprise, whereas the dominant emotion in the use of hallucinogens as compared to the other drugs is fear. Sedatives lead to an increase in disgust, while stimulants have a mix of anger and joy.

6 Conclusions

Automating language assessment of drug addict experiences has a potentially large impact on both toxicovigilance and prevention. Drug users are inclined to underreport symptoms to avoid negative consequences, and they often lack the self awareness necessary to report a drug abuse problem. In fact, often times people with drug misuse problems are reported on behalf of a third party (social services, police, families), when the situation is no longer ignorable.

In this paper, we introduced a new dataset

EMP		HAL		SED		STI	
Affiliation	1.76	See	1.81	Health	2.26	Money	2.25
We	1.63	Relig	1.72	Ingest	1.59	Ingest	1.75
Friend	1.46	Hear	1.44	Money	1.51	Work	1.64
Positive Emotions	1.41	Perception	1.24	Bio	1.50	Sexual	1.58
Sexual	1.34	Home	1.23	Swear	1.40	Swear	1.39

Table 5: Psycholinguistic word classes dominant for each drug type.

	EMP	HAL	SED	STI
EMP	1.00	0.34	0.03	0.15
HAL		1.00	0.80	0.83
SED			1.00	0.67
STI				1.00

Table 6: Pearson correlations of the LIWC dominance scores.

	EMP	HAL	SED	STI
Anger	1.09	0.91	1.01	**1.13**
Disgust	0.82	0.53	**2.68**	0.94
Fear	0.89	**1.26**	0.78	0.84
Joy	**1.26**	0.85	1.07	**1.11**
Sadness	1.08	0.95	0.96	1.09
Surprise	**1.46**	0.92	0.94	0.90

Table 7: Emotion word classes dominant for each drug type. Dominance scores larger than 1.10 are shown in bold face.

of drug use experiences, which can facilitate additional research in this space. We have described preliminary classification experiments, which showed that we can predict the drug behind an experience with a performance of up to 88% F1-score. To better understand the characteristics of drug users, we have also presented an analysis of the psycholinguistic process and emotions associated with different drug types.

We would like to continue the present work along the following directions: (i) Extend the corpus with texts written by people who supposedly do not ordinarily make use of drugs, using patient submitted forum posts when talking about ordinary medicines. The style of such patient submitted posts is expected to be similar to the one of drug experience reports, since both address writing about an experience with some particular substance; (ii) Explore the association between drug preferences and personality types. Following Khantzian's hypothesis (Khantzian, 1997), certain personalities may be more prone to a particular drug with respect to its subjective effects. Characterizing subjects by their potential drug preferences could enable clinicians, like in a reversed "recommender system," to explicitly warn their patients to avoiding particular kind of substances since they could become addictive.

The dataset introduced in this paper is available for research purposes upon request to the authors.

Acknowledgments

We would like to thank Samuele Garda for his insight and enthusiasm in the initial phase of the work. We also thank Dr. Marialuisa Grech, executive psychiatrist and psychotherapist at Serd (Service for Pathological Addiction) APSS, Trento, who helped us to better understand the drug consumption and drug-addicted world. This material is based in part upon work supported by the National Science Foundation (#1344257), the John Templeton Foundation (#48503), and the Michigan Institute for Data Science. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Michigan Institute for Data Science.

References

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland, June.

Gillinder Bedi, Guillermo A. Cecchi, Diego F. Slezak, Facundo Carrillo, Mariano Sigman, and Harriet de Wit. 2014. A window into the intoxicated mind? speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, 39(10):2340–8.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Pro-*

ceedings of the North American Association of Computational Linguistics, pages 327–337, Montreal, CA.

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on Twitter. In *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT)*, pages 1010–1019.

Daniel Bone, Ming Li, Matthew P. Black, and Shrikanth S. Narayanan. 2014. Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. *Computer Speech and Language*, 28:375—391.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 1301–1309.

Delroy Cameron, Gary A. Smith, Raminta Daniulaityte, Amit P. Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z. Watkins, and Russel Falck. 2013. PreDOSE: A semantic web platform for drug abuse epidemiology using social media. *Journal of Biomedical Informatics*, 46:985—997.

Michael Chary, Nicholas Genes, Andrew McKenzie, and Alex F. Manini. 2013. Leveraging social networks for toxicovigilance. *Journal of Medical Toxicology*, 9:184—191.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

Michael Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of Twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Jeremy R. Coyle, David E. Presti, and Matthew J. Baggott. 2012. Quantitative analysis of narrative reports of psychedelic drugs. *arXiv:1206.0312*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1365–1374.

Paul Ekman. 1993. Facial expression of emotion. *American Psychologist*, 48:384–392.

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1478–1488.

Neil Zhenqiang Gong, Ameet Talwalkar, Lester W. Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Shi, and Dawn Song. 2012. Predicting links and inferring attributes using a social-attribute network (SAN). In *The 6th SNA-KDD Workshop*.

Swapna Gottipati, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. 2014. An integrated model for user attribute discovery: A case study on political affiliation identification. In VincentS. Tseng, TuBao Ho, Zhi-Hua Zhou, ArbeeL.P. Chen, and Hung-Yu Kao, editors, *Advances in Knowledge Discovery and Data Mining*, volume 8443 of *Lecture Notes in Computer Science*, pages 434–446. Springer International Publishing.

Andreas Jauch, Paul Jaehne, and David Suendermann. 2013. Using text classification to detect alcohol intoxication in speech. In *Proceedings of the 7th Workshop on Emotion and Computing (in conjunction with the 36th German Conference on Artificial Intelligence)*, Koblenz, Germany, September.

Aditya Joshi, Abhijit Mishra, Balamurali AR, Pushpak Bhattacharyya, and Mark James Carman. 2015. A computational approach to automatic prediction of drunk-texting. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (short papers)*, Beijing, China, July.

Edward J. Khantzian. 1997. The self-medication hypothesis of substance use disorders: a reconsideration and recent applications. *Harvard Review of Psychiatry*, 4(5):231–44.

Moshe Koppel, Shlomo Argamon, and Anat Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 4(17):401–412.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the Association for Computational Linguistics (ACL 2009)*, Singapore.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217.

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. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 439–448.

Andrew Ortony, Gerald Clore, and Mark Foss. 1987. The referential structure of the affective lexicon. *Cognitive Science*, (11).

Michael J. Paul and Mark Dredze. 2012. Experimenting with drugs (and topic models): Multidimensional exploration of recreational drug discussions. In *Proceedings of AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*. AAAI Publications, November.

Michael J. Paul and Mark Dredze. 2013. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *Proceedings of HLT-NAACL 2013*, pages 168–178.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and Starbucks afficinados: User classification in Twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2011)*, pages 430–438.

James Pennebaker and Martha Francis. 1999. Linguistic inquiry and word count: LIWC. Erlbaum Publishers.

Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical Bayesian models for latent attribute detection in social media. pages 598–601.

Brian Riordan, Heather Wade, and Afzal Upal. 2014. Detecting sociostructural beliefs about group status differences in online discussions. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 1–6.

Florian Schiel, Christian Heinrich, and Sabine Barfusser. 2012. Alcohol language corpus: The first public corpus of alcoholized german speech. *Language Resources and Evaluation*, 46(3):503–521, September.

Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, Jarek Krajewski, Felix Weninger, and Florian Eyben. 2014. Medium-term speaker states - a review on intoxication, sleepiness and the first challenge. *Computer Speech and Language*, 28:346—374.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open vocabulary approach. *PLOS ONE*, 8(9):1–16, Sept.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon.

United Nations Office, editor. 2014. *World Drug Report*. United Nations, New York.

U.S. Department of Justice. 2015. *Drug of Abuse*. Drug Enforcement Administration - U.S. Department of Justice.

Benjamin Van Durme. 2012. Streaming analysis of discourse participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 48–58.

Svitlana Volkova and Benjamin Van Durme. 2015. Online bayesian models for personal analytics in social media.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI Conference on Artificial Intelligence*, pages 4296–4297.

William Yang Wang, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg. 2013. Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. *Computer Speech and Language*, 27:168—189.

Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. 2011. Like like alike: Joint friendship and interest propagation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 537–546.