Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models

Michael J. Paul and **Mark Dredze**
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
{mpaul,mdredze}@cs.jhu.edu

Abstract

Multi-dimensional latent text models, such as *factorial LDA* (f-LDA), capture multiple factors of corpora, creating structured output for researchers to better understand the contents of a corpus. We consider such models for clinical research of new recreational drugs and trends, an important application for mining current information for healthcare workers. We use a "three-dimensional" f-LDA variant to jointly model combinations of drug (marijuana, salvia, etc.), aspect (effects, chemistry, etc.) and route of administration (smoking, oral, etc.) Since a purely unsupervised topic model is unlikely to discover these specific factors of interest, we develop a novel method of incorporating prior knowledge by leveraging user generated tags as priors in our model. We demonstrate that this model can be used as an exploratory tool for learning about these drugs from the Web by applying it to the task of extractive summarization. In addition to providing useful output for this important public health task, our prior-enriched model provides a framework for the application of f-LDA to other tasks.

1 Introduction

Topic models aid exploration of the main thematic elements of large text corpora by revealing latent structure and producing a high level semantic view (Blei et al., 2003). Topic models have been used for understanding the contents of a corpus and identifying interesting aspects of a collection for more in-depth analysis (Talley et al., 2011; Mimno, 2011). While standard topic models assume a flat semantic structure, there are potentially many dimensions of a corpus that contribute to word choice, such as sentiment, perspective and ideology (Mei et al., 2007; Paul and Girju, 2010; Eisenstein et al., 2011). Rather than studying these factors in isolation, **multi-dimensional** topic models can consider multiple factors jointly.

Paul and Dredze (2012b) introduced *factorial LDA* (f-LDA), a general framework for multi-dimensional text models that capture an arbitrary number of factors (explained in §3). While a standard topic model learns distributions over "topics" in documents, f-LDA learns distributions over combinations of multiple factors (e.g. topic, perspective) called tuples (e.g. (HEALTHCARE,LIBERAL)). While f-LDA can model factors without supervision, it has not been used in situations where the user has prior information about the factors.

In this paper we consider a setting where the user has prior knowledge about the end application: mining recreational drug trends from user forums, an important clinical research problem (§2). We show how to incorporate available information from these forums into f-LDA as a novel hierarchical prior over the model parameters, guiding the model toward the desired output (§3.1).

We then demonstrate the model's utility in exploring a corpus in a targeted manner by using it to automatically extract interesting sentences from the text, a simple form of extractive multi-document summarization (Goldstein et al., 2000). In the same way that topic models can be used for aspect-specific summarization (Titov and McDonald, 2008; Haghighi and Vanderwende, 2009), we use f-LDA to extract snippets corresponding to fine-grained information patterns. Our results demonstrate that our multi-dimensional modeling approach targets more informative text than a simpler model (§4).

Proceedings of NAACL-HLT 2013, pages 168–178,
Atlanta, Georgia, 9–14 June 2013.

2 Analyzing Drug Trends on the Web

Recreational drug use imposes a significant burden on the health infrastructure of the United States and other countries. Accurate information on drugs, usage profiles and side effects are necessary for supporting a range of healthcare activities, such as addiction treatment programs, toxin diagnosis, prevention and awareness campaigns, and public policy. These activities rely on up-to-date information on drug trends, but it is increasingly difficult to keep up with current drug information, as distribution and information-sharing of novel drugs is easier than ever via the web (Wax, 2002). For the third consecutive year, a record number of new drugs (49) were detected in Europe in 2011 (EMCDDA, 2012). About two-thirds of these new drugs were synthetic cannabinoids (used as legal marijuana substitutes), which led to 11,000 hospitalizations in the U.S. in 2010 (SAMHSA, 2012). Treatment is complicated by the fact that novel substances like these may have unknown side effects and other properties.

Accurate information on drug trends can be obtained by speaking directly with users, e.g. focus groups and interviews (Reyes et al., 2012; Hout and Bingham, 2012), but such studies are slow and costly, and can fail to identify the emergence of new drug classes, such as mephedrone (Dunn et al., 2011). More recently, researchers have begun to recognize clinical value in information obtained from the web (Corazza et al., 2011). By (manually) analyzing YouTube videos, Drugs-Forum (discussed below), and other social media websites and online communities, researchers have uncovered details about the use, effects, and popularity of a variety of new and emerging drugs (Morgan et al., 2010; Corazza et al., 2012; Gallagher et al., 2012), and comprehensive drug reviews now include non-standard sources such as web forums in addition to standard sources (Hill and Thomas, 2011).

Organizing and understanding forums requires significant effort. We propose automated tools to aid in the exploration and analysis of these data. While topic models are a natural fit for corpus exploration (Eisenstein et al., 2012; Chaney and Blei, 2012), and have been used for similar public health applications (Paul and Dredze, 2011), online forums can be organized in many ways beyond topic. Guided by domain experts, we seek to model forums as a combination of drug type, route of intake (oral, injection, etc.) and aspect (cultural settings, drug chemistry, etc.) A multi-dimensional topic model can jointly capture these factors, providing a more informative understanding of the data, and can be used to produce fine-grained information such as the effects of taking a particular drug orally. Our hope is that models such as f-LDA can lead to exploratory tools that aide researchers in learning about new drugs.

Factor	Components
Drug	ALCOHOL AMPHETAMINES BETA-KETONES CANNABINOIDS CANNABIS COCAINE DMT DOWNERS DXM ECSTASY GHB HERBAL ECSTASY KETAMINE KRATOM LSA LSD NOOTROPICS OPIATES PEYOTE PHENETHYLAMINES SALVIA TOBACCO
Route	INJECTION ORAL SMOKING SNORTING
Aspect	CHEMISTRY (Pharmacology, TEK) CULTURE (Culture, Setting, Social, Spiritual) EFFECTS (Effects) HEALTH (Health, Overdose, Side effects) USAGE (Dose, Storing, Weight)

Table 1: The three factors of our model (details in §3.1). The forum tags shown in parentheses are grouped together to form aspects.

2.1 Corpus: Drugs-Forum

Our data set is taken from `drugs-forum.com`, a site active for more than 10 years with over 100,000 members and more than 1 million monthly readers. The site is an information hub where people can freely discuss recreational drugs with psychoactive effects, ranging from coffee to heroin, hosting information and discussions on specific drugs, as well as drug-related politics, law, news, recovery and addiction. With current information on a variety of drugs and an extensive archive, Drugs-Forum provides an ideal information source for public health researchers (Corazza et al., 2012).

Discussion threads are organized into numerous forums, including drugs, the law, addiction, etc. Since we are modeling drug use, we focus on the drug forums. Each thread is assigned to a specific forum or subforum (drug) and each thread has a user specified tag, which can indicate categories like "Effects" as well as routes of administration like "Oral." We organized the tags and subforum categorizations into factors and components, as shown in Table 1. We make use of these tags in §3.1.

3 Multi-Dimensional Text Models

Clinical researchers are interested in specific information about drug usage, including **drug** type, **route** of administration, and other **aspects** of drug use (e.g. dosage, side effects). Rather than considering these factors independently, we would like to model these in a way that can capture interesting interactions between all three factors, because the effects and other aspects of drugs can vary by route of administration. Oral consumption of drugs often produces longer lasting but milder effects than injection or smoking, for example. Many mephedrone users report nose bleeds and nasal pain as a health effect of snorting the drug: this could be modeled as the triple (MEPHEDRONE,SNORTING,HEALTH), a particular combination of all three factors.

To this end, we utilize the multi-dimensional text model **factorial LDA** (f-LDA) (Paul and Dredze, 2012b), which jointly models multiple semantic *factors* or dimensions. In this section we summarize f-LDA, then we describe an extension which incorporates user-generated metadata into the model (§3.1).

In a standard topic model such as LDA (Blei et al., 2003), each word token is associated with a latent "topic" variable. f-LDA is conceptually similar to LDA except that rather than a single topic variable, each token is associated with a K-dimensional **vector** of latent variables. In a three-dimensional f-LDA model, each token has three latent variables—drug, route, and aspect in this case.

In f-LDA, each document has a distribution over all possible K-tuples (rather than topics), and each K-tuple is associated with its own word distribution. Under this model, words are generated by first sampling a tuple from the document's tuple distribution, then sampling a word from that tuple's word distribution. In our three-dimensional model, we will consider **triples** such as (CANNABIS,SMOKING,EFFECTS).

Formally, each document has a distribution $\theta^{(d)}$ over triples, and each token is associated with a latent vector $\vec{z}$ of size K=3. (We'll describe the model in terms of the three factors we are modeling in this paper, but f-LDA generalizes to K dimensions.) The Cartesian product of the three factors forms a set of triples and the vector $\vec{z}$ references three discrete components to form a triple $\vec{t} = (t_1, t_2, t_3)$. The cardinality of each dimension (denoted Z_k) is the number of drugs, routes, and aspects, as shown in Table 1. Each triple has a corresponding word distribution $\phi_{\vec{t}}$. The graphical model is shown in Figure 1.

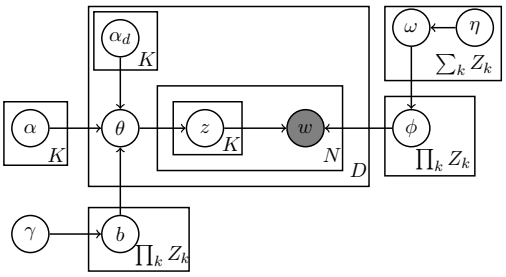

Figure 1: The graphical model for f-LDA augmented with priors η learned from labeled data (§3.1). In this work, $K = 3$.

One would expect that triples that have components in common should have similar word distributions: (CANNABIS,SMOKING,EFFECTS) is expected to have some commonalities with (CANNABIS,ORAL,EFFECTS). f-LDA models this intuition by sharing parameters across priors for triples which share components: all triples with CANNABIS as the drug include cannabis-specific parameters in the prior, and all triples with SMOKING as the route have smoking-specific parameters. Formally, $\phi_{\vec{t}}$ (the word distribution for tuple $\vec{t}$) has a Dirichlet($\hat{\omega}^{(\vec{t})}$) prior, where for each word w in the vector, $\hat{\omega}_w^{(\vec{t})}$ is a log-linear function:

$$\hat{\omega}_w^{(\vec{t})} \triangleq \exp\left(\omega^{(B)} + \omega_w^{(0)} + \omega_{t_1 w}^{(\text{drug})} + \omega_{t_2 w}^{(\text{route})} + \omega_{t_3 w}^{(\text{aspect})}\right) \quad (1)$$

where $\omega^{(B)}$ is a corpus-wide precision scalar (the bias), $\omega_w^{(0)}$ is a corpus-specific bias for word w, and $\omega_{t_k w}^{(k)}$ is a bias parameter for word w for component t_k of the kth factor. That is, each drug, route, and aspect has a weight vector over the vocabulary, and the prior for a particular triple is influenced by the weight vectors of each of the three factors. The ω parameters are all independent and normally distributed around 0 (effectively L2 regularization).

The prior over each document's distribution over triples has a similar log-linear prior, where weights for each factor are combined to influence the distribution. Under our model, $\theta^{(d)}$ is drawn from Dirichlet($\mathbf{B} \cdot \hat{\alpha}^{(d)}$), where $\cdot$ denotes an element-wise product between $\mathbf{B}$ (described below) and $\hat{\alpha}^{(d)}$, with

$\hat{\alpha}^{(d)}_{\vec{t}}$ for each triple $\vec{t}$ defined as:

$$\hat{\alpha}^{(d)}_{\vec{t}} \triangleq \exp\Big(\alpha^{(B)} + \alpha^{(\mathcal{D},\text{drug})}_{t_1} + \alpha^{(d,\text{drug})}_{t_1} + \alpha^{(\mathcal{D},\text{route})}_{t_2} + \alpha^{(d,\text{route})}_{t_2} + \alpha^{(\mathcal{D},\text{aspect})}_{t_3} + \alpha^{(d,\text{aspect})}_{t_3}\Big) \quad (2)$$

Similar to the ω formulation, $\alpha^{(B)}$ is a global bias parameter, while the $\alpha^{\mathcal{D}}$ vectors are corpus-wide weight vectors and α^{d} are document-specific weight vectors over the components of each factor. Structuring the prior in this way models the intuition that if a triple with a particular component has high probability, other triples containing that component are likely to also have high probability. For example, if a message discusses triples of the form (CANNABIS,*,EFFECTS), it is more likely to discuss (CANNABIS,*,HEALTH) than (COCAINE,*,HEALTH), because the message is about cannabis.

Finally, $\mathbf{B}$ is a 3-dimensional array that encodes a sparsity pattern over the space of possible triples. This is used to accommodate triples that can be generated by the model but are not supported by the data. For example, not all routes of administration may be applicable to certain drugs, or certain aspects of a drug may happen to not be discussed in the forum. Each element $b_{\vec{t}}$ of the array is a real-valued scalar in $(0, 1)$ which is multiplied with $\hat{\alpha}^{(d)}_{\vec{t}}$ to adjust the prior for that triple. If the b value is near 0 for a particular triple, then it will have very low prior probability. The b values have Beta(γ_0,γ_1) priors ($\gamma < 1$) which encourage them to be near 0 or 1, so that they function as binary variables.

Posterior inference and parameter estimation consist of a Monte Carlo EM algorithm that alternates between an iteration of collapsed Gibbs sampler on the $\vec{z}$ variables (E-step), and an iteration of gradient ascent on the α and ω hyperparameters (M-step). See Paul and Dredze (2012b) for more details.

3.1 Tags and Word Priors

In an unsupervised setting, there is no reason f-LDA would actually infer parameters corresponding to the three factors we have been describing. However, the forums include metadata that can help guide the model: the messages are organized into forums corresponding to drug type (factor 1), and some threads are tagged with labels corresponding to routes of administration and other aspects (factors 2 and 3). Tags for aspects are manually grouped into components: e.g. USAGE (tags: Dose, Storing, Weight). Table 1 shows the factors and components in our model.

COCAINE	SNORTING	HEALTH
η (Prior over ω)		
coke	snort	kidney
cocaine	snorting	hcv
crack	snorted	pains
cola	nose	symptoms
blow	nasal	guidelines
lines	drip	diet
ω (Prior over ϕ)		
coke	snort	symptoms
cocaine	snorting	long-term
crack	snorted	depression
cola	passages	disorder
rocks	nostril	schizophrenia
coca	insufflating	severe

COCAINE SNORTING HEALTH
ϕ (Posterior)
nose
cocaine
coke
blood
water
pain

Figure 2: Example of parameters learned by f-LDA. The highest weight words in the ω and η vectors for three components are shown on the left. These are combined to form the prior for the word distribution ϕ. The tripling of (COCAINE,SNORTING,HEALTH) results in high probability words about nose bleeds and nasal damage.

One could simply use these tags as labels in a simple supervised model—this will be our experimental baseline (§4.1). However, this approach has limitations in that most documents are missing labels (less than a third of our corpus contains one of the labels in Table 1) and many messages discuss several components, not just the one implied by the tag. For example, a message tagged "Side effects" may talk about both side effects and dosage. While a supervised classifier may attribute all words to a single tag, f-LDA learns per-token assignments.

We will instead use the tags to inform the priors over our f-LDA word distribution parameters. We do this with a two-stage approach. First, we use the tags to train parameters of a related but simplified model. We then use the learned parameters as priors over the corresponding f-LDA parameters.

In particular, we will place priors on the ω vectors, the Dirichlet hyperparameters which influence the word distributions. Suppose that we are given a vector $\eta^{(0)}$ which is believed to contain desirable values for $\omega^{(0)}$, the weight vector over words in the corpus, and similarly we are given vectors $\eta^{(f)}_i$ over the vocabulary for the ith component of factor f, which are believed to be good values for $\omega^{(f)}_i$. One option

is to fix ω as η, forcing the component weights to match the provided weights. However, in our case η will only be an approximation of the optimal component parameters since it is estimated from incomplete data (only some messages have tags) and the η vectors are learned using an approximate model (see below). Instead, these weight vectors will merely guide learning as prior knowledge over model parameters ω. While f-LDA assumes each ω is drawn from a 0-mean Gaussian, we alter the means of the appropriate ω parameters to use η.

$$\omega_w^{(0)} \sim \mathcal{N}(\eta_w^{(0)}, \sigma^2); \omega_{iw}^{(k)} \sim \mathcal{N}(\eta_{iw}^{(k)}, \sigma^2) \quad (3)$$

Recall that $\omega_w^{(0)}$ are corpus-wide bias parameters for each word and $\omega_{iw}^{(k)}$ are component-specific parameters for each word. This yields a hierarchical prior in which η parameterizes the prior over ω, while ω parameterizes the prior over ϕ (the word distributions). The resulting ω parameters can vary from the provided priors to adapt to the data. An example of learned parameters is shown in Figure 2, illustrating the hierarchical process behind this model.

Learning the Priors In various applications, priors can come from many different sources, such as labeled data (Jagarlamudi et al., 2012). We learn the prior means η from tagged messages. However, these parameters imply a latent division of responsibility for observed words: some are present because of the tag while others are general words in the corpus. As a result, they must be estimated.

We learn these parameters from the tagged messages using SAGE, which model words in a document as combinations of background and topic word distributions. Eisenstein et al. (2011) present SAGE models for Naive Bayes (one class per document), admixture models (one class per token), and admixture models where tokens come from multiple factors. We combine the first and third models, such that a document has multiple factors which are given as labels across the entire document—the drug type and the tag, which could correspond to a component of either the route or aspect factors. We posit the following model of text generation per document:

$$P(\text{word } w|\text{drug} = i, \text{factor} f = j) \quad (4)$$
$$= \frac{\exp(\eta_w^{(0)} + \eta_{iw}^{(\text{drug})} + \eta_{jw}^{(f)})}{\sum_{w'} \exp(\eta_{w'}^{(0)} + \eta_{iw'}^{(\text{drug})} + \eta_{jw'}^{(f)})}$$

This log-linear model has a similar form as Eq. 1, but with two factors instead of three, and it is a distribution rather than a Dirichlet vector. As in SAGE, we fix $\eta^{(\mathbf{0})}$ to be the observed vector of corpus log-frequencies over the vocabulary, which acts as an "overall" weight vector, while parameter estimation yields $\eta_i^{(f)}$, the logit parameters for the ith component of factor f.[1] These parameters are then used as the mean of the Gaussian priors over ω.

Standard optimization methods can be used to estimate these parameters. The partial derivative of the likelihood with respect to the parameter $\eta_{iw}^{(\text{drug})}$ is:

$$\frac{\partial}{\partial \eta_{iw}^{(\text{drug})}} = \sum_f \sum_{j \in f} c(i,j,w) - \pi(i,j,w) c(i,j,*) \quad (5)$$

where $c(i,j,w)$ is the number of times word w appears in documents labeled with i (drug) and j (tag), and $\pi(i,j,w)$ denotes the probability given by (4). The partial derivative of each $\eta_j^{(f)}$ is similar.

4 Experiments with Topic Modeling for Extractive Summarization

Our corpus consists of messages from `drugs-forum.com` (§2.1). The site categorizes threads into many forums and subforums, including some on specific drugs, which are categorized hierarchically. We treated higher-level categories with pharmacologically similar drugs as a single drug type (e.g. OPIOIDS, AMPHETAMINES); for others we took the finest-granularity subforum as the drug type. We selected 22 popular drugs and from these forums we crawled 410K messages. We selected a subset of tags to form components for the route and aspect factors. (Some tags were too general or infrequent to be useful.) A list of the tags and drugs used appears in Table 1. We also included a GENERAL component in the latter two factors to model word usage which does not pertain to a particular route or aspect; the prior parameters η for these components were simply set to $\mathbf{0}$.

We wish to demonstrate that our modified f-LDA model can be used to discover useful information in the text. One way to demonstrate this is by using the model to extract relevant snippets of text from the

[1] SAGE models sparsity on the weights via a Laplacian prior. Such sparsity is not modeled in f-LDA, so we ignore this here.

forums, which will form the basis of our evaluation experiments. Our goal is not to build a complete summarization system, but rather to use the model to direct researchers to interesting messages.

While we model all 22 drugs, our summarization experiments will focus on five drugs which have been studied only relatively recently: mephedrone and MDPV (β-ketones), Bromo-Dragonfly (synthetic phenethylamines), Spice/K2 (synthetic cannabinoids), and salvia divinorum. We will consider these drugs in particular because these are the five drugs for which technical reports were created by the EU Psychonaut Project (Schifano et al., 2006), an online database of novel and emerging drugs, whose information is collected by reading drug websites, including Drugs-Forum. Extensive technical reports were written about these five popular drugs, and we can use these reports to produce reference summaries for our experiments (§4.2).

Of these five drugs, only salvia has its own subforum; the others belong to subforums representing the broader categories shown in parentheses. We simply model the drug type as a proxy for the specific drug, as most of the drugs in each category have similar effects and properties. The first two drugs are both in the same subforum, so for the purpose of our model we treat mephedrone and MDPV as the single drug type, β-ketones. These two drugs are grouped together during summarization (§4.2), but the corresponding reference summaries incorporate excepts from the technical reports on both drugs.

4.1 Model Setup

Of the four drug types being considered for summarization, our data set contains 12K messages with one of the tags in Table 1 and 30K without. Of those without tags, we set aside 5K as development data. There are also over 300K messages (140K tagged) from the remaining 18 drug types: some of these messages are utilized when training f-LDA. Even though we only consider four drug types in our experiments, our intuition is that it can be beneficial to model other drugs as well, because this will help to learn parameters for the various aspects and routes of administration. Our model of the effects of mephedrone can be informed by also modeling the effects of other stimulants such as cocaine.

Each message was treated as a document, and we only used documents with at least five word tokens after stop words, low-frequency words, and punctuation were removed. The preprocessed data sets contained an average of 45 tokens per document.

Below, we describe two f-LDA variants as well as the baseline used in our experiments.

Baseline Our baseline model is a unigram language model trained on the subset of messages which are tagged. We treat the drug subforum as a label for the drug factor, and each message's tag is used as a label for either the route or aspect factor. For example, the word distribution for the pair (SALVIA,EFFECTS) is estimated as the empirical distribution from messages posted in the salvia forum and tagged with "Effects." We use add-λ smoothing where λ is chosen to optimize likelihood on the held-out development set.

This is a two-dimensional model, since we explicitly model pairs such as (MEPHEDRONE,SNORTING) or (SALVIA,EFFECTS). However, we also created word distributions for triples such as (SALVIA,ORAL,EFFECTS) by taking a mixture of the corresponding pairs: in this example, we estimate the unigram distribution from salvia documents tagged with either "Oral" or "Effects."

Factorial LDA Because f-LDA does not rely on tagged data (the tags are only used to create priors), we can run inference on larger sets of data. The drawback is that despite these priors, it is still mostly unsupervised and we want to be careful to ensure the model will learn the patterns we care about. We thus add some reasonable constraints to the parameter space to guide the model further.

First, we treat the drug type as an observed variable based on the subforum the message comes from, just as with the baseline. For example, only tuples of the form (SALVIA,$*$,$*$) can be assigned to tokens in the salvia forum. Second, we restrict the set of possible routes of administration that can be assigned to tokens in particular drug forums, since most drugs can be taken through only a subset of routes. For example, marijuana is typically smoked or eaten orally, but rarely injected. We therefore restrict each drug's allowable set of administration routes to those which are tagged (e.g. with "Oral" or "Snorting") in at least 1% of that drug's data. Similar ideas are used in Labeled LDA (Ramage et al.,

Reference Text	System Snippet
Mephedrone (β-ketones/Bath salts)	
It is recommended by users that Mephedrone be taken on an empty stomach. Doses usually vary between 100mg–1g.	• If it is SWIYs first time using Mephedrone SWIM recommends a 100mg oral dose on an empty stomach.
Reported negative side effects include: • Loss of appetite. • Dehydration and dry mouth • Tense jaw, mild muscle clenching, stiff neck, and bruxia (teeth grinding) • Anxiety and paranoia • Increase in mean body temperature (sweating/Mephedrone sweat and hot flushes) • Elevated heart rate (tachycardia) and blood pressure, and chest pains • Dermatitis like symptoms (Itch and rash)	• Neutral side effects: Lack of appetite, occasional loss of visual focus, [...] weight loss, possible diuretic. Negative side effects: Grinding teeth, "Cotton mouth", unable to acheive orgasm • Aside from his last session he has never experienced any negative symptoms at all, no raised heart beat, vasoconstriction , sweating, headaches, paranoia e.t.c nothing at all except sometimes cold hands the next day. • lot of people report that anxiety and paranoia are some of the side effects of taking mephedrone [...] is it also possible that alot of the chest pains people are experiencing is due to anxiety? • moisturize the affected areas of skin twice daily with E45 or a similar unperfumed dermatalogical lotion.
Salvia divinorum	
Sublingual ingestion of the leaf (quid): reduces intensity of effects and can taste disgusting. When Salvia is consumed as a smokeable formulation the duration of the trip lasts 30 minutes or less, whereas if Salvia is consumed sublingually the effects lasts for 1 hour or more.	• The taste of sublingual salvia is foul and it is easy to have a dud trip unless large amounts of it are used. • SWIM has heard from many other users that chewing the fresh leaves of the Salvia plant allow for a much longer and mellower trip. [...] SWIM has read that a trip this way can last anywhere from a half on hour or longer.
Dried leaves and/or salvia extract are smoked (using a butane lighter) either by pipe (considered to be the most effective but is considered to be quite painful) or water bong.	• 2. Use a water pipe. Its harsh and needs to be smoked hot so this should be self explanatory. 3. Use a torch style lighter [...] Salvinorin A has a VERY high boiling point (around 700 degrees F I believe) so a regular bic just wont do it
Salvia is appealing to recreational users because of intense, unique, hallucinatory effects. Brief hallucinations occur rapidly after administration and are typically very vivid. Users report weird thoughts, feelings of unreality, feelings of immersion in bizarre non-Euclidian dimensions/geometries, feelings of floating.	• He noticed very clear [closed eye visuals], which looked similar to patterns on a persian rug, or ethnic oriental design. SWIM felt as if he was moving around, that he had got up and run and fallen, and that falling had shattered the space around his body as if I'd fallen through many glass framed pictures [...] • I was aware of my body and my friends and my life below, but I was [...] standing outside of time and outside of space.

Figure 3: Example snippets generated by f-LDA along with the corresponding reference text. For space, the references and snippets shown have been shortened in some cases. "SWIM" and "SWIY" stand for "someone who isn't me/you" and are used to avoid self-incrimination on the web forum.

2009), in which tags are used to restrict the space of allowed topics in a document.

We use f-LDA as a three-dimensional model which explicitly models triples, but we also obtain distributions for pairs such as (SALVIA,EFFECTS) by marginalizing across all distributions of the form (SALVIA,$*$,EFFECTS). We trained f-LDA on two different data sets, yielding the following models:

- **f-LDA-1:** We use the 12K messages with tags and fill the set out with 13K messages with tags uniformly sampled from the 18 other drugs, for a total of 25K messages.
- **f-LDA-2:** We use all 37K messages (many without tags) and fill the set out with 63K messages with tags uniformly sampled from the 18 other drugs, for a total of 100K messages.

All f-LDA instances are run with 5000 iterations alternating between a sweep of Gibbs sampling followed by a step of gradient ascent on the hyperparameters. While we do not use the tags as strict labels during sampling, we initialize the Gibbs sampler so that each token in a document is assigned to its label given by the tag, when available. In the absence of tags (in f-LDA-2), we initialize tokens

to the GENERAL components. We initialized ω to its prior mean (Eq. 3), while the variance σ^2 and the initialization of bias $\omega^{(B)}$ are chosen to optimize likelihood on the held-out development set.

We optimized the hyperparameters and sparsity array using gradient descent after each Gibbs sweep. We use a decreasing step size of $a/(t+1000)$, where t is the current iteration and a=10 for α and 1 for ω and the sparsity values. To learn priors η, we ran our version of SAGE for 100 iterations of gradient ascent (fixed step size of 0.1). See Paul and Dredze (2012a) for examples of parameters (the top words associated with various triples) learned by this model on this corpus.

4.2 Summary Generation

We created twelve reference summaries by editing together excerpts from the five Psychonaut Project reports ((Psychonaut), 2009). Each reference is matched to drug-specific pairs and triples. For example, a paragraph describing the differences in effects of salvia between smoking and oral routes was matched to distributions for (SALVIA,EFFECTS), (SALVIA,SMOKING,EFFECTS), (SALVIA,ORAL,EFFECTS). Descriptions of creating tinctures and blotters for oral consumption were matched to (SALVIA,ORAL,CHEMISTRY). We consider pairs in addition to triples because not all summaries correspond to particular routes or aspects.

For each tuple-specific word distribution (a pair or a triple), we create a "summary" by extracting a set of five text snippets which minimize KL-divergence to the target word distribution. We consider all overlapping text windows of widths {10,15,20} in the corpus as candidate snippets. Following Haghighi and Vanderwende (2009), we greedily add snippets one by one with the lowest KL-divergence at each step until we have added five.

We only considered candidate snippets within the subforum for the particular drug, and snippets are based on the preprocessed topic model input with no stop words. Before presenting snippets to users, we then map the snippets back to the raw text by taking all sentences which are at least partly spanned by the window of tokens. Because each reference may be matched to more than one tuple, there may be more than five snippets which correspond to a reference.

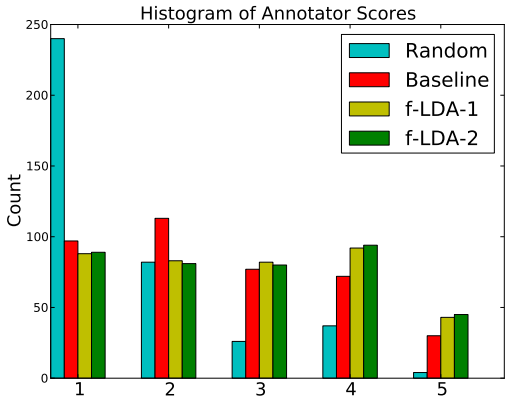

Figure 4: The distribution of annotator scores (§4.3.1). The "Random" counts have been scaled to fit the same range as the other systems, since fewer random snippets were shown to annotators.

4.3 Experimental Results

Recall that the reports used as reference summaries were themselves created by reading web forums. Our hypothesis is that f-LDA could be used as an exploratory tool to expedite the creation of these reports. Thus in our evaluation we want to measure how useful the extracted snippets would be in informing the writing of such reports. We performed both human and automatic evaluation on the summaries generated by f-LDA (variants 1 and 2) as well as our baseline. We also included randomly selected snippets as a control (five per reference).

Example output is shown in Figure 3.

4.3.1 Human Judgments of Quality

Three annotators were presented snippets pooled from all four systems we are evaluating alongside the corresponding reference text. Within each set corresponding to a reference summary, the snippets were shown in a random order. Annotators were asked to judge each snippet independently on a 5-point Likert scale as to how useful each snippet would be in writing the reference text.

The distribution of scores is shown in Figure 4 and summarized in Table 2. Annotators generally agreed on the relative quality of snippets: the average correlation of scores between each pair of annotators was 0.49. Snippets produced by f-LDA were given more high scores and fewer low scores than the baseline, while the two f-LDA variants were rated comparably. The breakdown is more interesting when we compare scores for snippets that were matched

	Rand.	Base.	f-LDA-1	f-LDA-2
	Annotator Scores			
Mean	1.67	2.55	2.79	**2.81**
Pairs only	n/a	2.58	**2.79**	2.72
Triples only	n/a	2.50	2.80	**2.95**
	ROUGE			
1-gram	.112	.326	**.355**	.327
2-gram	.023	.072	**.085**	.084

Table 2: Summary quality evaluation across four systems.

to word distributions for pairs versus word distributions for triples. The gap in scores between f-LDA and the baseline increases when we look at the scores for only triples: f-LDA beats the baseline by a margin of 0.45 for snippets matched to triples and 0.21 for pairs. This suggests that we produce better triples by modeling them jointly. For triples, f-LDA-2 (which uses more data) beats f-LDA-1 (which uses only tagged data), while the reverse is true for pairs.

While some of the randomly selected control snippets happened to be useful, the scores for these snippets were much lower than those extracted through model-based systems. This suggests that exploring the forums in a targeted way (e.g. through our topic model approach) would be more efficient than exploring the data in a non-targeted way (akin to the random approach).

Finally, we asked two expert annotators (faculty members in psychiatry and behavioral pharmacology, who have used drug forums in the past to study emerging drugs) to rate the snippets corresponding to mephedrone/MDPV. The best f-LDA system had an average score of 2.57 compared to a baseline score of 2.45 and random score of 1.63.

4.3.2 Automatic Evaluation of Recall

The human judgments effectively measured a form of precision, as the quality of snippets were judged by their correspondence to the reference text, without regard to how much of the reference text was covered by all snippets. We also used the automatic evaluation metric ROUGE (Lin, 2004) as a rough estimate of summary recall: this metric computes the percentage of n-grams in the reference text that appeared in the generated summaries.

We computed ROUGE for both 1-grams and 2-grams. When computing n-gram counts, we applied Porter's stemmer to all tokens. We excluded stop words from 1-gram counts but included them in 2-gram counts where we care about longer phrases.[2]

Results are shown in Table 2. We find that f-LDA-1 has the highest score for both 1- and 2-grams, suggesting that it is extracting a more diverse set of relevant snippets. When performing a paired t-test across the 12 reference summaries, we find that f-LDA is better than the baseline with p-values 0.14 and 0.10 for 1-gram and 2-gram recall, respectively. f-LDA's recall advantage may come from the fact that it learns from a larger amount of data and it may learn more diverse word distributions by directly modeling triples. f-LDA-1 had slightly better recall (under ROUGE), while f-LDA-2 was slightly better according to the human annotators.

5 Conclusion

We have proposed exploratory tools for the analysis of online drug communities. Such communities are an emerging source of drug research, but manually browsing through large corpora is impractical and important information could be missed. We have demonstrated that topic models are capable of modeling informative portions of text, and in particular multi-dimensional topic models can target desired structures such as the combination of aspect and route of administration for each drug. We have presented an extension to factorial LDA tailored to a particular application and data set which was demonstrated to induce desired properties. As a technical contribution, this study lays out practical guidelines for customizing and incorporating prior knowledge into multi-dimensional text models.

Acknowledgments

We are grateful to Dr. Margaret S. Chisolm and Dr. Ryan Vandrey from the Johns Hopkins School of Medicine for providing the mephedrone/MDPV annotations, and Alex Lamb and Hieu Tran for assisting with the full annotations. We also thank Dr. Matthew W. Johnson for additional advice, and the anonymous reviewers for helpful feedback and suggestions. This research was partly supported by an NSF Graduate Research Fellowship.

[2] In both cases, ROUGE scores were higher when stop words were included. f-LDA beats the baseline by similar margins regardless of whether we include stop words.

References

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *JMLR*.

A. Chaney and D. Blei. 2012. Visualizing topic models. In *ICWSM*.

O. Corazza, F. Schifano, M. Farre, P. Deluca, Z. Davey, C. Drummond, M. Torrens, Z. Demetrovics, L. Di Furia, L. Flesland, et al. 2011. Designer drugs on the Internet: a phenomenon out-of-control? The emergence of hallucinogenic drug Bromo-Dragonfly. *Current Clinical Pharmacology*, 6(2):125–129.

Ornella Corazza, Fabrizio Schifano, Pierluigi Simonato, Suzanne Fergus, Sulaf Assi, Jacqueline Stair, John Corkery, Giuseppina Trincas, Paolo Deluca, Zoe Davey, Ursula Blaszko, Zsolt Demetrovics, Jacek Moskalewicz, Aurora Enea, Giuditta di Melchiorre, Barbara Mervo, Lucia di Furia, Magi Farre, Liv Flesland, Manuela Pasinetti, Cinzia Pezzolesi, Agnieszka Pisarska, Harry Shapiro, Holger Siemann, Arvid Skutle, Aurora Enea, Giuditta di Melchiorre, Elias Sferrazza, Marta Torrens, Peer van der Kreeft, Daniela Zummo, and Norbert Scherbaum. 2012. Phenomenon of new drugs on the Internet: the case of ketamine derivative methoxetamine. *Human Psychopharmacology: Clinical and Experimental*, 27(2):145–149.

Matthew Dunn, Raimondo Bruno, Lucinda Burns, and Amanda Roxburgh. 2011. Effectiveness of and challenges faced by surveillance systems. *Drug Testing and Analysis*, 3(9):635–641.

J. Eisenstein, A. Ahmed, and E. P. Xing. 2011. Sparse additive generative models of text. In *ICML*.

Jacob Eisenstein, Duen Horng "Polo" Chau, Aniket Kittur, and Eric P. Xing. 2012. Topicviz: Semantic navigation of document collections. In *CHI Work-in-Progress Paper*.

EMCDDA. 2012. 2012 annual report on the state of the drugs problem in Europe. *European Monitoring Centre for Drugs and Drug Addiction, Lisbon*.

Cathal T. Gallagher, Sulaf Assi, Jacqueline L. Stair, Suzanne Fergus, Ornella Corazza, John M. Corkery, and Fabrizio Schifano. 2012. 5,6-methylenedioxy-2-aminoindane: from laboratory curiosity to 'legal high'. *Human Psychopharmacology: Clinical and Experimental*, 27(2):106–112.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 40–48.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Simon L. Hill and Simon H. L. Thomas. 2011. Clinical toxicology of newer recreational drugs. *Clinical Toxicology*, 49(8):705–719.

Marie Claire Van Hout and Tim Bingham. 2012. Costly turn on: Patterns of use and perceived consequences of mephedrone based head shop products amongst Irish injectors. *International Journal of Drug Policy*.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *EACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July.

Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*.

David Mimno. 2011. Reconstructing Pompeian households. In *UAI*.

Elizabeth M. Morgan, Chareen Snelson, and Patt Elison-Bowers. 2010. Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 26(6):1405–1411. Online Interactivity: Role of Technology in Behavior Change.

Michael J. Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Michael J. Paul and Mark Dredze. 2012a. Experimenting with drugs (and topic models): Multi-dimensional exploration of recreational drug discussions. In *AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*.

Michael J. Paul and Mark Dredze. 2012b. Factorial LDA: Sparse multi-dimensional text models. In *Neural Information Processing Systems (NIPS)*.

M. Paul and R. Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*.

Psychonaut WebMapping Research Group (Psychonaut). 2009. Bromo-Dragonfly, MDPV, Spice, Mephodrone, and Salvia Divinorum reports. http://www.psychonautproject.eu/technical.php. Institute of Psychiatry, King's College London.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256.

J. Reyes, J. Negrón, H. Colón, A. Padilla, M. Millán, T. Matos, and R. Robles. 2012. The emerging of

xylazine as a new drug of abuse and its health consequences among drug users in Puerto Rico. *Journal of Urban Health*, pages 1–8.

SAMHSA. 2012. The DAWN report. http://www.samhsa.gov/data/2k12/DAWN105/SR105-synthetic-marijuana.pdf, December 4.

Fabrizio Schifano, Paolo Deluca, Alex Baldacchino, Teuvo Peltoniemi, Norbert Scherbaum, Marta Torrens, Magi Farrõ, Irene Flores, Mariangela Rossi, Dorte Eastwood, Claude Guionnet, Salman Rawaf, Lisa Agosti, Lucia Di Furia, Raffaella Brigada, Aino Majava, Holger Siemann, Mauro Leoni, Antonella Tomasin, Francesco Rovetto, and A. Hamid Ghodse. 2006. Drugs on the web: the Psychonaut 2002 EU project. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 30(4):640 – 646.

Edmund Talley, David Newman, Bruce Herr II, Hanna Wallach, Gully Burns, Miriam Leenders, and Andrew McCallum. 2011. A database of National Institutes of Health (NIH) research using machine learned categories and graphically clustered grant awards. *Nature Methods*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *International World Wide Web Conference (WWW)*, Beijing.

P.M. Wax. 2002. Just a click away: Recreational drug web sites on the Internet. *Pediatrics*, 109(6).