½-1 page summary:

Abstract:

Keywords: psychedelics, NLP, psychedelic therapy

Outline

17-20 pages:

Introduction 6-8 pages
Methods 6-8 pages
Discussion 6-8 pages
Conclusion 1 pages

Introduction

Psychedelic substances

The substances studied in this paper have seen various classifications throughout the history of research and recreational use. Today, the substances are most commonly referred to as psychedelic drugs, meaning "mind manifesting" (Nichols, 2004) or "mind revealing" (Carhart-Harris & Goodwin, 2017). This term is often used interchangeably with hallucinogenic (referring to the albeit rare presence of hallucinations), psychotomimetic ("psychosis mimicking") and entheogenic ("god within") drugs (Nichols, 2004). For simplicity, the term "psychedelic" is used in this paper, as seems to be the standard for current scientific research.

History of psychedelics

The natural occurrence and effects of psychedelic substances have likely played a key role in development of various religions and philosophic thought throughout history (Nichols, 2004). There is anthropological and ethnopharmacological evidence of psychedelic substance use in ancient India (a substance called *soma*), ancient Greece, Aztec and Maya Cultures (e.g. *teonanacatl* and *ayahuasca*), Australia, Tanzania, and by Native North Americans (*Peyote* cacti) to name a few (Nichols, 2016). Considering the described effects of psychedelic substances, it is not hard to imagine ancient peoples gazing at a bonfire, sunset or starry night sky after ingestion, discussing with awe and wonder (and, perhaps, existential dread) what divine powers lie behind.

More recently in the West, lysergic acid diethylamide (LSD) became almost synonymous with the '60s and '70s hippie culture, with musicians and the festival scene embracing the

substance due it heightening feelings of connectedness, love, freedom and creativity (Nichols, 2016). Psychedelic culture has inspired new genres of psychedelic music (DeRogatis, 2003), art (Grunenberg & Harris, 2005), literature (Dickins, 2012) and film (Gallagher, 2004). It was also the anti-war, anti-government, anti-convention aspects of this movement that partly led to the passage of the U.S Controlled Substances Act of 1970, under which several known psychedelics (e.g. LSD, psilocybin and mescaline) became Schedule 1 controlled substances, immediately halting all research on these substances (Nichols, 2016). Following a two-decade hiatus, psychedelic research was once again active during the '90s in Germany, Switzerland and U.S.A. Combining the period before 1970 and the period from the '90s until today, a foundational body of research on psychedelics now exists in the fields of neuroscience, psychology and psycho-pharmacology (Carhart-Harris & Goodwin, 2017).

The effects of psychedelics

What characteristics does a psychedelic experience contain to be deemed *mind manifesting* or *mind revealing*? A psychedelic experience can well be categorized as a pharmacologically induced altered state of consciousness (ASC). Charles Tart defines these states as "alternate patterns or configurations of experience, which differ qualitatively from a typical baseline state." (Garcia-Romeu & Tart, 2013). According to Jaffe (1990), what distinguishes psychedelics from other drugs is "their capacity reliably to induce states of altered perception, thought, and feeling that are not experienced otherwise except in dreams or at times of religious exaltation." (Nichols, 2016).

Generally, the effects of psychedelics can be categorised accordingly: somatic, perceptual, psychic and cognitive (Nichols, 2004; Preller & Vollenweider, 2016). Although effects vary from user to user, common reported somatic effects include dizziness, tremors, increased heartrate, nausea (sometimes resulting in vomiting) and blurred vision, particularly when the first effects of the substance set in, known as the "come-up" phase.

The perceptual, psychic and cognitive effects are really at the core of what makes the experience "psychedelic". Simply put, these effects are alterations in visual, auditory and haptic perception, alterations of cognitive faculties such as memory, attention, learning, creativity and language production, all leading to potential experiences of depersonalization and dreamlike feelings (Nichols, 2004; Preller & Vollenweider, 2016).

Effects of visual perception include perceiving different or more intense colors, alterations in shapes, perceived movement and/or animacy of objects and difficulty in gauging distances. Auditory perception is often altered in intensity; some sounds are perceived louder or sharper, and misperceptions of sound can also occur. Haptic perception is sometimes altered in the somatosensory sense; the texture, shape and weight of objects are altered, sensations of warmth and cold at hands and feet; and sometimes in the sense of proprioception, whereas balance and sense of body size (such as arm length) is altered (Preller & Vollenweider, 2016).

The cognitive effects of psychedelics are mainly observed in working memory, attention, learning, creativity and language production. Users find it harder to connect thoughts after ingesting LSD (Preller & Vollenweider, 2016), perhaps a result of impairment in working memory and attention, which is also found in studies with psilocybin (Carter et al., 2005). Furthermore, performance in tests such as Raven's Progressive Matrices and the Stroop test is impaired, reaction time is consistently slowed, and attentional tracking is impaired as well. It is found that response inhibition is decreased whilst under the effects of psilocybin (Gouzoulis-Mayfrank et al., 2002) The perceptual and cognitive changes are, perhaps, what leads to a reported increase in creativity, although there are methodological limitations in studying this claim (Preller & Vollenweider, 2016).

Assessing the subjective effects of psychedelics has mostly been done through introspective questionnaires, starting with the Hallucinogen Rating Scale (HRS) (Strassman et al., 1994), followed by Dittrich's Abnormal Mental States (APZ) questionnaire (Dittrich, 1994), which was further edited to the OAV scales (Bodmer et al., 1994) and the five-dimensional altered states ofconsciousness (5D-ASC) questionnaire (Dittrich et al., 2006). The OAV scales were revised by Studerus et al. (2010), increasing dimensionality to 11 lower-order scales: experience of unity, spiritual experience, blissful state, insightfulness, disembodiment, impaired control and cognition, anxiety, complex imagery, elementary imagery, audio-visual synesthesia, and changed meaning of percepts (Nichols, 2016).

Treatment with psychedelics

Both before the Controlled Substances Act of 1970 and since the revival of psychedelic research, a primary motivation of the field has been the potential therapeutic and medicinal benefits of treatment with psychedelic substances.

Some success has been achieved with psychedelic-assisted treatment for anxiety and depression. Grob et al. (2011) found a reduction in anxiety (STAI trait anxiety subscale scores) at 1 and 3 months post treatment with psilocybin, and an improvement of mood at 6 months. Gasser et al. (2014) found positive outcomes in the long-term when applying LSD-assisted psychotherapy in patients with anxiety related to a life-threatening disease or diagnosis. Sanches et al. (2016) found an improvement in depressive symptoms lasting up to 21 days following a single dose of DMT (in the form of Ayahuasca) (Nichols et al., 2017).

Alongside a plethora of anecdotal reports and single-subject studies (Nichols, 2016), psilocybin has been administered in patients with obsessive-compulsive disorder (OCD), showing improvements within 24 hours post treatment (Moreno et al., 2006), though there are methodological concerns with these findings (Nichols et al., 2017).

Perhaps the most succesful application of psychedelics is in treating various forms of addiction. A sum of these have focused on alcoholism, e.g. treating with LSD (Chwelos et al., 1959) and psilocybin (Bogenschutz et al., 2015). In a meta-analysis of six studies using LSD to treat alcoholism (Krebs & Johansen, 2012), a significant decrease in alcohol misuse was found in patients who were administered varying doses of LSD compared to control groups.

Another study focused on opioid addiction and found less usage of heroin in the LSD-receiving group compared to the control group post treatment, significant at all examined time windows up to a year post treatment (Savage & McCabe, 1973). Johnson et al. (2014) conducted a pilot study on treating tobacco addiction with cognitive behavioural therapy (CBT) alongside administering psilocybin. They found confirmed abstinence in 80% of participants (n = 15) at 6 months post quitting smoking, and a follow-up study (Johnson et al., 2017) found abstinence in 67% of participants at 12 months post quitting smoking. This is to be compared with the current most effective medication for smoking cessation Veranicline, which has an approx. 35% succes rate at 6 months post quit-date (Nichols et al., 2017).

Importantly, a note should be made on the importance of context when administering psychedelic substances. As such, many of the promising results come from studies that provide extensive psychological support before, during and after partaking in the experiment. This includes several preparation meetings with professionals, creating an optimal therapeutic environment with lighting, music and aesthetics, and always having someone to compassionately support the participant before, during and after the study. This is important to ensure a correct "*set*" and "*setting*"; set is the mindset (e.g. expectations and assumptions) of the user, and setting is the physical environment surrounding the user (Carhart-Harris et al., 2018). Administering psychedelics without this psychological support may limit the effects of the treatment or potentially worsen the condition of a patient. The potential of psychedelic treatment lies in leveraging the psychedelic state by combining it with methods such as cognitive therapy and attentional-bias training (Carhart-Harris & Goodwin, 2017).

Differences between substances

Psychedelics are commonly associated with three types of chemical structures; tryptamines (e.g. psilocybin, N,N-dimethyltryptamine (DMT) and 5-MeO-DMT), phenethylamines (e.g. mescaline and 2C-B) and lysergamides (e.g. LSD) (Nichols, 2016).

It is widely accepted that psychedelic effects are mainly caused by agonist or partial agonist activity at the serotonin 5-HT2A receptor, which has given rise to the name "serotonergic psychedelics" (Nichols, 2016). Some define psychedelic substances as "compounds with appreciable serotonin 2A receptor agonist properties that can alter consciousness in a marked and novel way" (Carhart-Harris & Goodwin, 2017). It is important to note that all psychedelic effects are not necessarily explained by this activity; tryptamines such as 5-MeO-DMT, LSD and psilocybin have also been found to act as 5-HT1A agonists (Coyle et al., 2012), LSD has a strong interaction with the five dopamine receptors (D_1, D_2, D_3, D_4 and D_5), mescaline hits an adrenergic receptor (α_{2C}) (Ray, 2010) and DMT binds to σ-1 receptors (Fontanilla et al., 2009) and inhibits the serotonin transport (SERT) (Cozzi et al., 2009). Although the evidence is inconclusive, psychedelic effects are still thought to emerge as a result of direct receptor actions (Nichols, 2016).

Whether differences in receptor binding profiles result in different subjective effects has scarcely been studied. Holze et al. (2022) compared the acute effects of LSD and psilocybin in a double-blind, placebo-controlled study and found no significant difference of the substances at high doses, rated on the 5D-ASC scale. There was a significant difference between low dosage psilocybin and low dosage LSD, although this could be due to faulty dosage calculation rather than actual differences in drug effects, which also poses a challenge for psychedelic drug comparison (although research has been done on dosing, see e.g. Griffiths et al. (2011) & Hasler et al. (2004)). To this authors knowledge, the difference in acute effects of psychedelic substances such as mescaline, psilocybin, LSD, DMT, and Ayahuasca have not been studied in a framework such as the one used by Holze et al. (2022).

Griffiths et al. (2019) studied perceived God encounters across different psychedelics (LSD, psilocybin, Ayahuasca and DMT) using a survey-based approach and the Mystical Experience Questionnaire (MEQ30). They found no difference between psilocybin and LSD, substantial difference between Ayahuasca and psilocybin/LSD, little difference between Ayahuasca and DMT and substantial difference between DMT and psilocybin/LSD. The results are likely influenced by differences in demographics between groups and the surrounding culture for substances like Ayahuasca, which is often used in a social spiritual/religious context. Though the importance of context is paramount, the authors suggest that N,N-dimethyltryptamine (DMT, also the psychedelic compound in Ayahuasca) "produces a unique profile of effects that is phenomenologically distinct from two widely used classic psychedelics (psilocybin and LSD)" (Griffiths et al., 2019).

Furthermore, anecdotal evidence supporting differences in subjective effects of various psychedelic substances has been extensively reported (Zamberlan et al., 2018). Perhaps the largest and most renowned works on effects of psychedelic drugs, in which the authors report on the subjective effects of over 200 psychoactive substances, are *PiHKAL* (Shulgin & Shulgin, 1992) and *TiHKAL* (Shulgin & Shulgin, 1997).

Using self-reports to study the subjective effects

While a firm understanding of the difference in acute effects between psychedelic substances has yet to be established, it would be reasonable to explore different methods to further our knowledge of the matter. An increasing number of papers have been published which include an analysis of already available, retrospectively written experience reports from the Erowid Experience Vaults (https://www.erowid.org/experiences/) (Tagliazucchi, 2022; Hase et al., 2022; Coyle et al., 2012; Zamberlan et al., 2018).

A recent review by Tagliazucchi (2022) focused on the effects of psychedelic drugs on speech organization and the semantic content of experience reports from Erowid. A point is made in the paper that, despite difficulties in extracting meaningful and quantitative data from unstructured texts, the information embedded in written reports outweigh that of questionnaires, and that "natural language reports obtained during the acute effects of psychedelics open a new dimension of analysis beyond the possibilities of psychometric

questionnaires". Hase et al. (2022) compared linguistic profiles of experience reports from Erowid between six psychedelic drug categories using latent semantic analysis (LSA) and various rating scales from the LIWC2015 library. They found language profiles of drug categories to differ in, among others, emotional intensity and analytical thinking. Coyle et al. (2012) used a random-forest classifier on a subset of the experience reports from Erowid, where they achieved a 51.1% estimated accuracy on a 10-class classification task, classifying drug class based on report. Zamberlan et al. (2018) used a dataset of binding affinity profiles for various psychedelics combined with experience reports from Erowid, and found a correlation between similarity of binding affinity profiles of drugs and the semantic similarity of written reports associated with the drugs.

These examples show that many have turned to the field of natural language processing to analyze unstructured written reports. Though the mentioned papers have achieved some success, the use of natural language processing as a tool to study subjective experiences between psychedelic substances remains to be demonstrated (Tagliazucchi, 2022). Natural language processing (NLP) has already seen many applications in adjacent fields such as psychiatry, where it has been used to detect linguistic differences in schizophrenia spectrum disorders (Tang et al., 2021), predict suicide risk/ideation (Cook et al., 2016) and discover symptoms that are sometimes missed by professionals (Rezaii et al., 2022).

The current study

To further the research on the subjective effects of psychedelic substances, the approach of this paper is to use novel NLP methods on retrospective written reports of psychedelic experiences, acquired from the Erowid Experience Vaults. Building on the success of Coyle et al. (2012) in predicting drug class, this framework utilizes topic modeling to interpret the difference between classes. The methods of analysis include calculating TF-IDF (term frequency-inverse document frequency) to assess word importance across six psychedelic substance categories: LSD, psilocybin (mushrooms), mescaline, Ayahuasca, DMT and 2C-B. Furthermore, per substance topic modeling is performed using the cutting-edge BERTopic topic model (Grootendorst, 2022), giving insight into the content of subjective reports across psychedelic substances. The work presented here contributes to the overall understanding of the subjective effects of psychedelic substances and the efficacy of applying NLP methods in this field.

Methods

Procedure

Dataset

The dataset of written reports was acquired from the Erowid Experience Vaults (https://www.erowid.org/experiences/). The vaults are an "attempt to collect, catalog, and

publish the wide variety of experiences people have with psychoactive plants and chemicals" (https://www.erowid.org/experiences/exp_about.cgi). As of today (FOOTNOTE), the vaults contain approx. 39.000 experience reports, spanning across experiences with ~950 different substances. Each entry consist of a title, author title, the substance type and a report on the writer's experience whilst influenced by the substance. Some reports also include information on dose, body weight, route of administration, year of experience, gender and age at the time of experience. To be viewable in the vaults, reports must be read and approved by at least two members of the Erowid Reviewing Crew. Reports are rejected if they are believed to be falsified, impossible to read or do not address the effects of the substance (Experience Report Reviewing: The Good, the Bad and the Ugly, 2002).

The data was collected by scraping each report site using the package *rvest* (Wickham, 2022) for R (R Core Team, 2020). Only title, substance and report were scraped, as these were the only parameters available for all entries, which means that parameters like dose and route of administration are ignored in the analysis. Many entries were combination substances, i.e. experiences with taking multiple substances at the same time. These were removed to assure only single-substance entries remained. The final dataset contains 4219 reports, describing experiences from six substance groups: LSD, psilocybin (mushrooms), mescaline, Ayahuasca, DMT and 2C-B.

LSD (1130 reports)

This group contains experiences where the writer has ingested lysergic acid diethylamide (LSD), a synthesized lysergamide. Taken orally as a blotter under the tongue.

Psilocybin (mushrooms) (1783 reports)

This group contains experiences where the writer has ingested mushrooms (or "magic mushrooms") which contain the psychedelic compound psilocybin. Mushroom type was not accounted for, and approx. 45 different types of fungi are included in this category.

Mescaline (355 reports)

This group contains experiences where the writer has ingested 3,4,5-trimethoxyphenethylamine (mescaline) in the form of mescaline-containing cacti (such as *San Pedro* or *peyote*) or synthesized mescaline.

Ayahuasca (120 reports)

This group contains experiences where the writer has ingested the brew Ayahuasca, which contains N,N-dimethyltryptamine (DMT) and a monoamine oxidase inhibitor (MAOI) (Nichols, 2016).

DMT (652 reports)

This group contains experiences where the writer has smoked N,N-dimethyltryptamine (DMT).

2C-B (179 reports)

This group contains experiences where the writer has ingested 4-Bromo-2,5-dimethoxyphenethylamine (2C-B), a psychedelic phenethylamine derivative belonging to the 2C-family, which effects have not yet been fully studied in humans, but is structurally close to mescaline (Papaseit et al., 2018).

Whilst inspecting the reports, some where found to be written in Italian. The *cld3* package (Ooms, 2022) was used to detect all non-english reports which were subsequently excluded.

Analysis

TF-IDF weighted word clouds

As a first measure of analyzing the experiences of different psychedelic substances, the term frequency-inverse document frequency (TF-IDF) was calculated for all six substance groups in R using the *tidytext* package (Silge & Robinson, 2016). This statistic measures the importance of each word in a document (experience report), importance being high when the word occurs often in the document but rarely in other documents. In this case, words with a high score are more defining for the reported experience (KARABIBER FOOTNOTE).

The term frequency per word is calculated as thus:

$$TF = \frac{no.\ of\ times\ the\ term\ appears\ in\ the\ document}{no.\ of\ terms\ in\ the\ document}$$

The inverse document frequency is per word is calculated as thus:

$$IDF = log(\frac{no.\ of\ documents\ in\ the\ corpus}{no.\ of\ documents\ in\ the\ corpus\ containing\ the\ term})$$

TF-IDF is calculated by multiplying the term frequency with the inverse document frequency.

$$TF - IDF = TF * IDF$$

A TF-IDF score is calculated for each word in each document. The per-substance-group TF-IDF scores were visualized in six word clouds (figure 1) created in R with the *ggwordcloud* package (Le Pennec & Slowikowski, 2019).

BERTopic topic modelling

- What is BERTopic?
  - What older methods does it replace?
  - What are the benefits of BERTopic compared with these older methods?
  -

WRITE ON STOPWORDS HERE, ARGUE WHY REMOVED!!!!

Results

TF-IDF weighted word clouds

Interpret it and talk about differences between substance

LSD

The LSD group contains important words which can be sub-grouped into psychic/mystical effects (e.g. *epiphanies, thesis, dreams, extraordinary, flashbacks, philosophers, theory*), somatic effects (e.g. *pain, headaches, migraine, itching*) and narrative experience words (what writers were doing whilst they were affected, e.g. *tunes, temple, chair, yoga, people*).

Psilocybin (mushrooms)

The psilocybin group is mostly dominated by words related to preparation and route of administration (e.g. *swallow, eating, truffles, seeds, ounce, milk, cakes*). Some words also relate to effects of the substance, both somatic (e.g. *migraines, pain, cluster, sore, submit, veils*) and cognitive/psychic (e.g. *OCD, neuron*).

Mescaline

The majority of the most important words for the mescaline group relate to the preparation and route/method of administration (e.g. *swallow, powdered, ingestion, intestinal, inflection, chew, cuttings, buttons, cut, trays, powder, drinks, extraction, boiled, sift*). This is perhaps because the preparation process is more complicated with cacti.

Ayahuasca

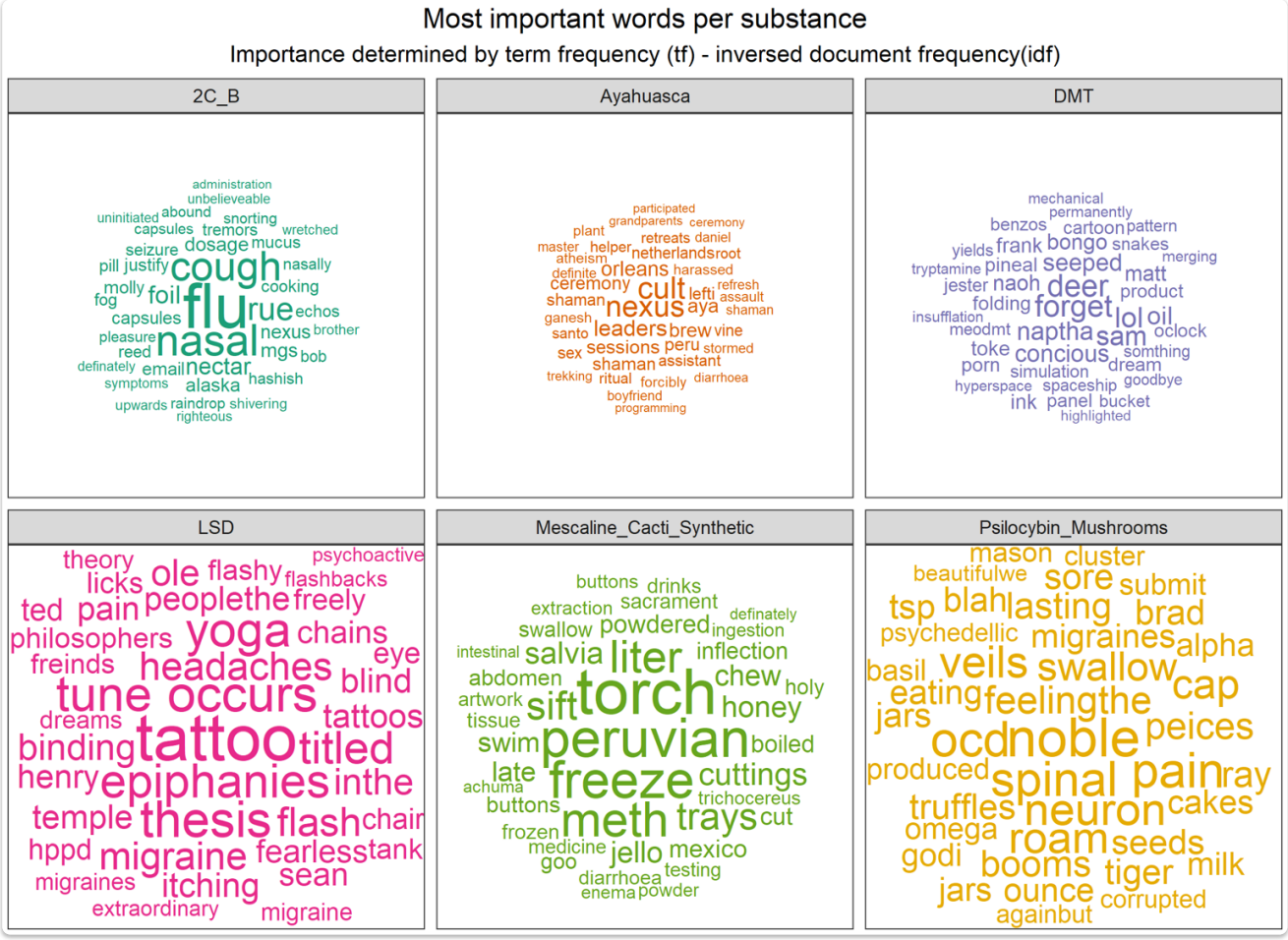
The Ayahuasca group is dominated by words pertaining to the ritual or traditions around the brew (e.g. *cult, shaman, ceremony, sessions, ritual, leaders, retreats*). Locations are also present such as *Peru*, which carries cultural weight. Some words also suggest uncomfortable experiences (e.g. *harassed, assault, forcibly*).

DMT

The most important words for experiences in the DMT group are related to altered perception and psychic effects (e.g. *forget, concious, dream, hyperspace, simulation, merging, pattern*). There are also words related to route/method of administration (e.g. *product, insufflation*) and substance type (e.g. *tryptamine, benzos, oil*). A few specific words give insight into the narrative experiences, such as *bongo, snakes, deer, jester*. The word *pineal* is also present, perhaps due to the false notion that enough DMT is naturally produced in the pineal gland to cause psychedelic effects (Nichols, 2018).

2C-B

The most important words in the 2C-B group are largely related to route/method of administration (e.g. *nasal, foil, pill, capsules, mucus, dosage, snorting*) and somatic effects of the substance (e.g. *cough, tremors, seizure, shivering*). Other substances (e.g. *molly and nexus*) are also important in some documents, perhaps due to comparison between substances.



BERTopic topic modeling

LSD
min_cluster_size = 10, min_topic_size = 5, min_samples=2
Overfit. 15 topics, alot of names in top terms etc.

min_cluster_size = 50, min_topic_size = 30, min_samples=10
4 topics, very similar tho..

min_cluster_size = 50, min_topic_size = 20, min_samples=10

Classification?

Discussion

Limitations

Dataset

language
how much time has past since trip (when writing)
writing abut their state before and after taking the drug is not meaningful for the effects of the drug?
if they actually took this drug
the limitations of memory
quality of data generally, i.e. grammar errors fuck up NLP, some reports in italian
There seems to be low amount of reports for some substances, i.e. skewed topic model quality?

TF-IDF word clouds

Using TF-IDF weighting assures that unique words are very important in the document, therefore names (such as "jon", "sean", "alaska" or "sam") were assigned high importance, which is not meaningful in understanding the effects of the substance.

Including entactogens/empathogens such as MDMA

Conclusion

Notes

(Coyle et al., 2012) Quantitative Analysis of Narrative Reports of Psychedelic Drugs

- Use this for inspiration, but make sure to do it differently than them.

- Their hypothesis:
  i) there would be detectable differences between reports for different drugs as indicated by the ability of a classifier to accurately predict drug class;
  ii) drugs with similar effects would have similar reports as indicated by classifier confusion and class means;
  iii) inspection of the discriminating variables would allow insight into the differences between these drugs.

- **Use a LLM (BERT) to classify and extract informative parameters**
  - I have to read up on this - maybe ask Rebekah?
    - Then use the "transform" package to inform me about what affected the classification decision!
  - COMBINE LDA AND BERT WITH BERTopic
    - https://www.pinecone.io/learn/bertopic/
    - https://www.youtube.com/watch?v=fb7LENb9eag

- ## Latent Semantic Analysis (LSA)
  - What is this? Read about it!
  - A description can be found in (Tagliazucchi, 2022)
  - Used by Hase et al. (2022) as well

DECISION:
TextBlob package was used to detect and remove non-english reports.

DECISION:
There seems to be low amount of reports for some substances, but this should go into the limitations segment. Should I include/disclude for this reason?

DECISION: (Idea)
To reduce the sections of trip reports BEFORE the onset of the drug, maybe only include reports which have the timer (T + 0.00), and then include after a couple hours?
IDEA FROM REBEKAH: Feed the reports to GPT-3, and get it to split the text where it thinks the trip report actually starts.

DECISION:
I only include reports for ISOLATED drugs, that is NO combinations of drugs allowed.

DECISION:
I only include drugs that are serotonin 2A agonists, because I have to define in some way what a psychedelic drug is (as not to include caffeine, cocaine, MDMA etc.)

Notes for stop-word list

Remove all words from the substance list!
from `unique(df_subset$substance`

Continue stopword list:
("Dose", "dose", "mg", "")

26. Oct - Data acquired!

... now what?

Okay, so I think that I'm going to make a few things:

1. Use LDA and visualise with LDAvis.
2. Try to classify with the LDA topics.
3. Use BERT to classify and extract information with extra packages?

Perhaps read PiHKAL and TiHKAL by Alexander Shulgin. Gives anecdotal backing to why the differences between drugs is interesting to understand.

11. Nov - Mail to Rebekah part 2.

Hi Rebekah,

Here's an update on my bachelor project :-)

I have now successfully done the following:

Dataset:

- Scraped the Erowid reports (~40K reports)
- Subset them to about 5600 reports in 6 substance categories.
- Preprocessed the text so it is ready for analysis.

Wordcloud:

- Calculated the TF-IDF to use as weighting.
- Made six wordclouds - one for each substance category to visualize differences between substance.

Topic modeling:

- Did LDA topic modeling with six topics, as to see if they magically correlated with the six actual labels.
  - None of the topics were meaningfully connected to any of the substances, bummer
- Ran some tuning algorithms to see which amount of topics (k) would yield the most informative clusters.
  - With 1-15 topics, information kept increasing so perhaps i will try 1-50 topics over night...
- Tried out LDAvis even though topics were bad. Really cool tool!

So far I haven't done any classification, and I don't think I can use the topic model for anything meaningful in this regard.

I haven't started actually writing up the paper yet, but I definitively have more room, so perhaps I'll try a supervised classification method also? Maybe using a LLM like BERT or otherwise. What do you think?

29. Nov - What to do?

Topic modelling:

- I have made 6 dtm's, one for each substance.
- Run tuner algorithm for all
- Run LDA for lowest recommended no. of topics

Today:

- Try to interpret the models, see if they make sense!

BERTopic:

- Both topics and classification?

Removing stop words ¶

At times, stop words might end up in our topic representations. This is something we typically want to avoid as they contribute little to the interpretation of the topics. However, removing stop words as a preprocessing step is not advised as the transformer-based embedding models that we use need the full context in order to create accurate embeddings.

Instead, we can use the `CountVectorizer` to preprocess our documents **after** having generated embeddings and clustered our documents. Personally, I have found almost no disadvantages to using the `CountVectorizer` to remove stopwords and it is something I would strongly advise to try out:

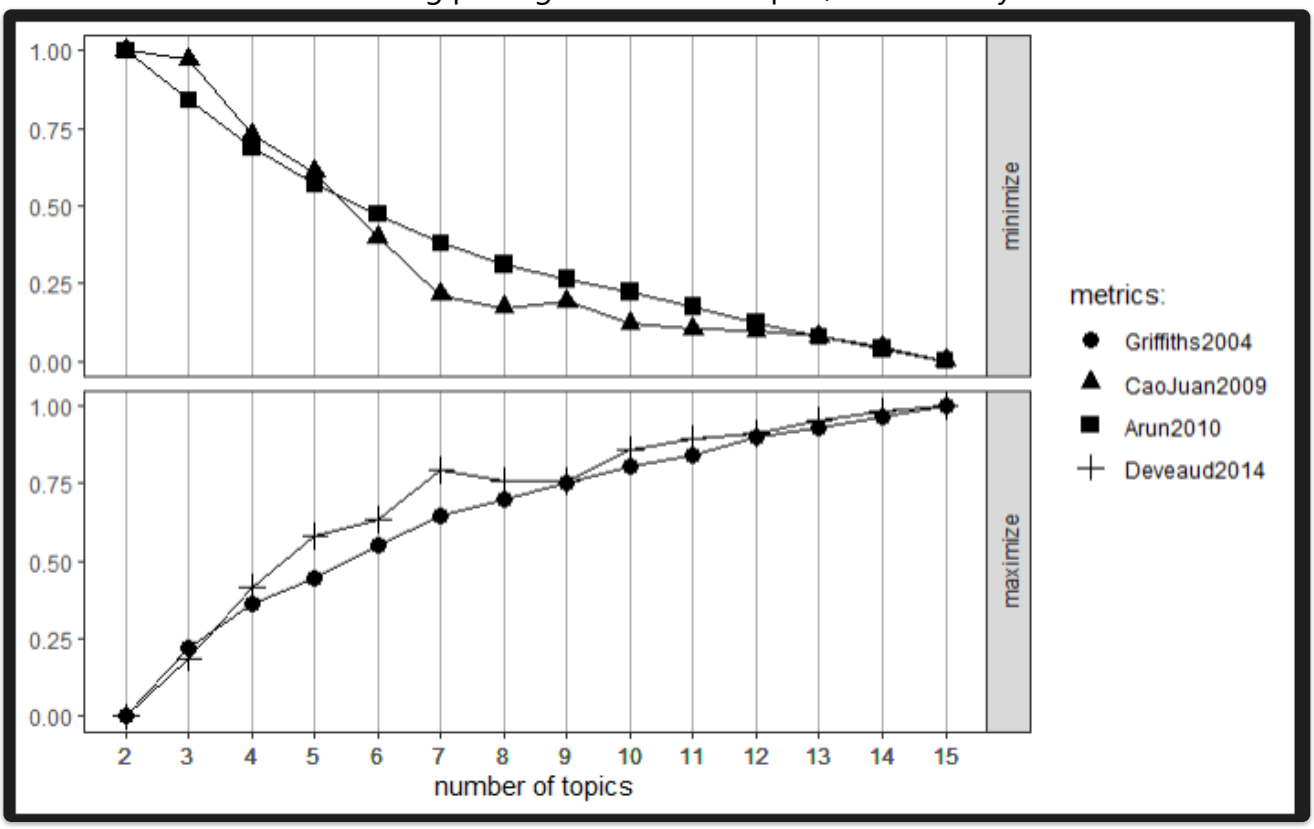
```
from bertopic import BERTopic
from sklearn.feature_extraction.text import CountVectorizer

vectorizer_model = CountVectorizer(stop_words="english")
topic_model = BERTopic(vectorizer_model=vectorizer_model)
```

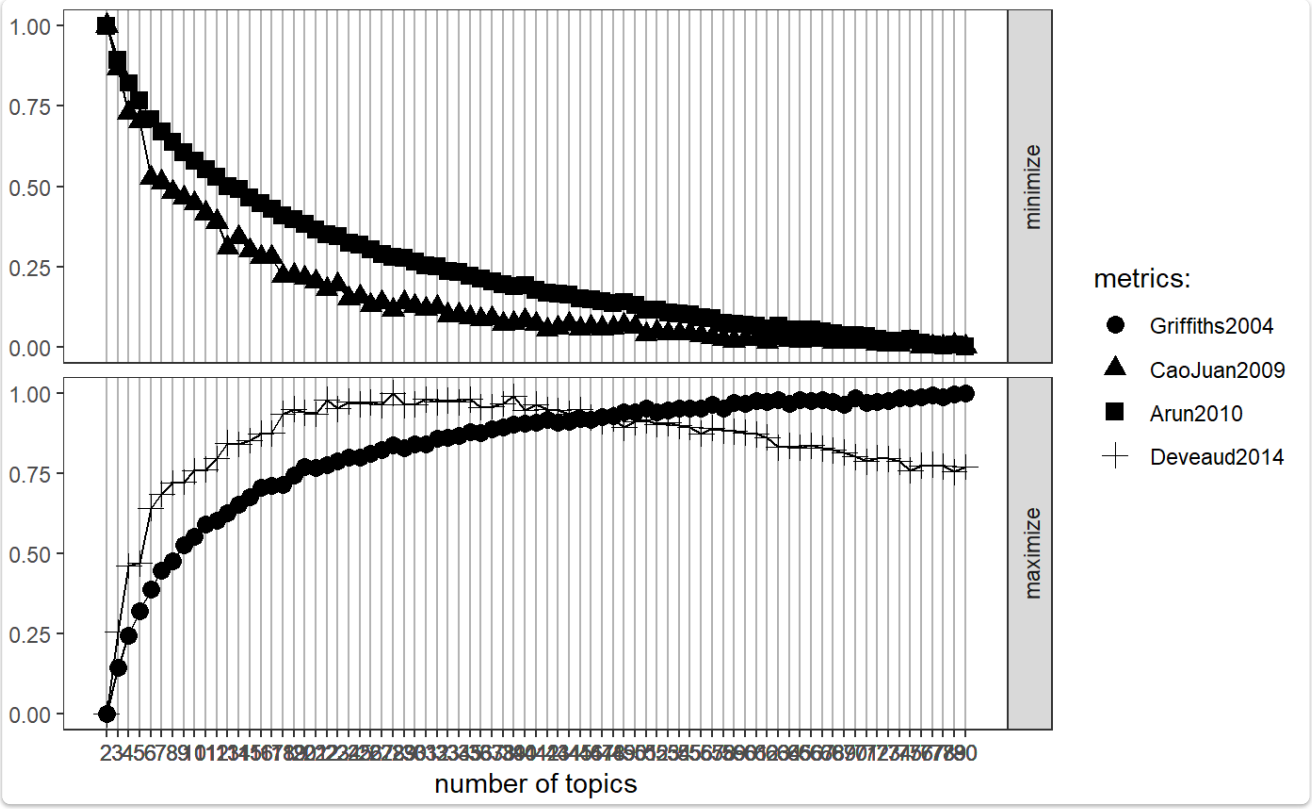
Appendices

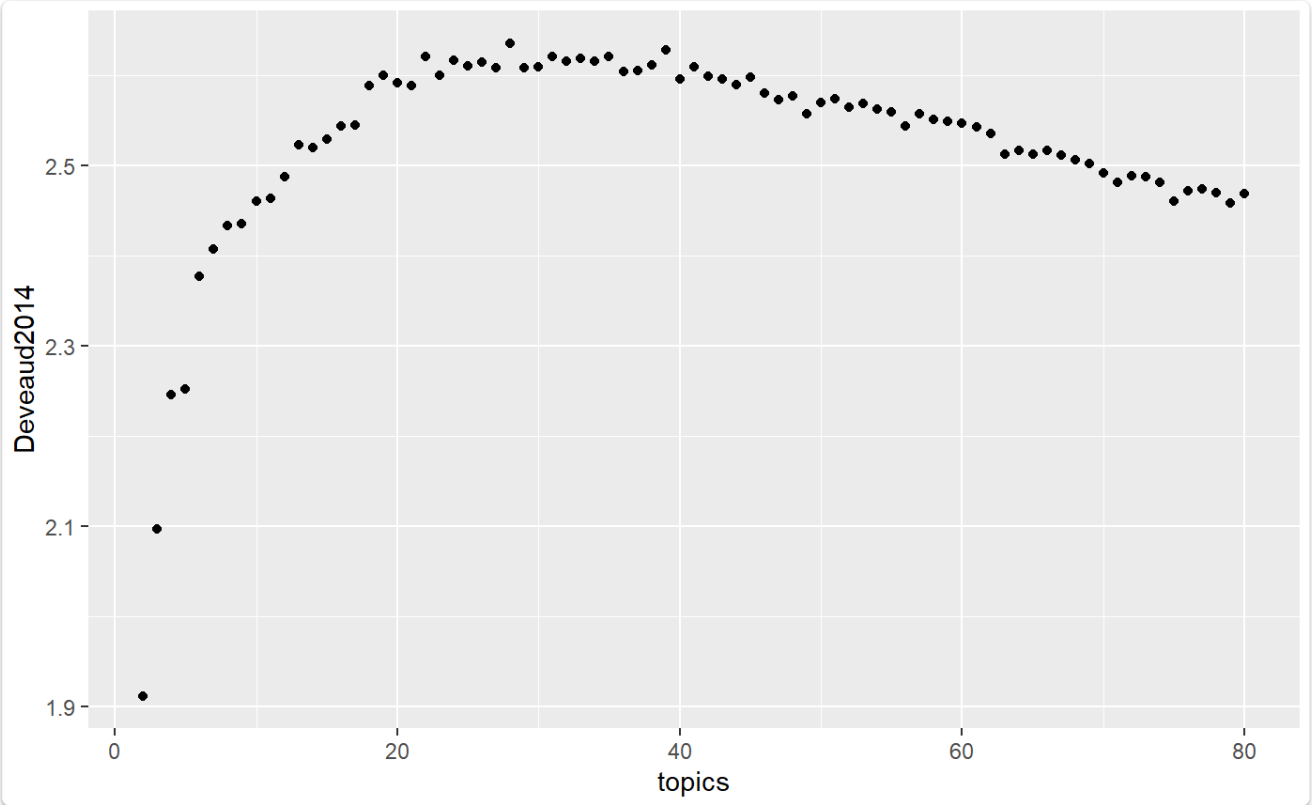
One model for all text:

TUNING TOPICS with ldatuning package. From 1-15 topics, it is not very useful.

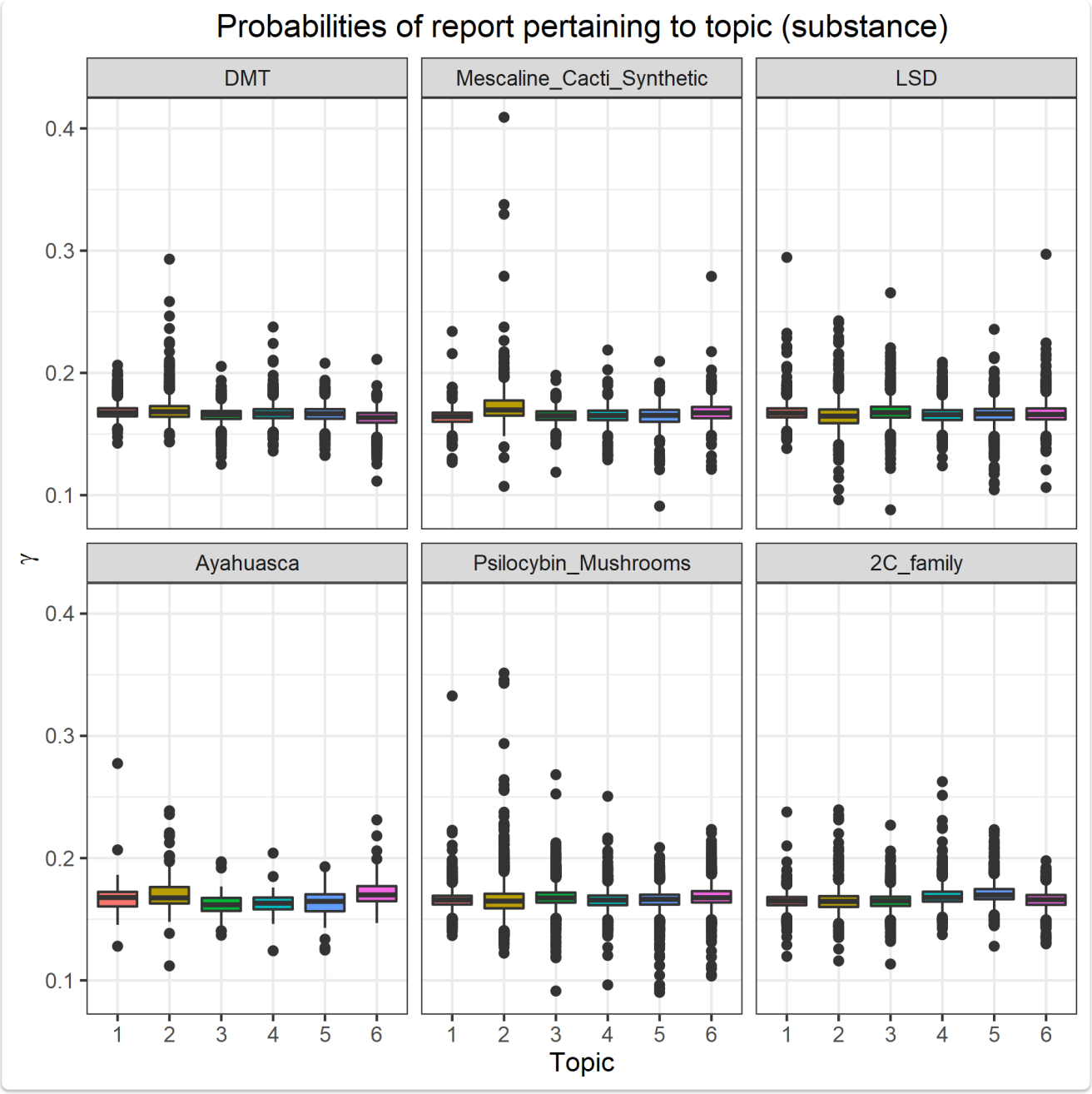


With 2-15 topics, the metric peaks at around 7 for Deveaud and minimizes here as well for CaoJuan.





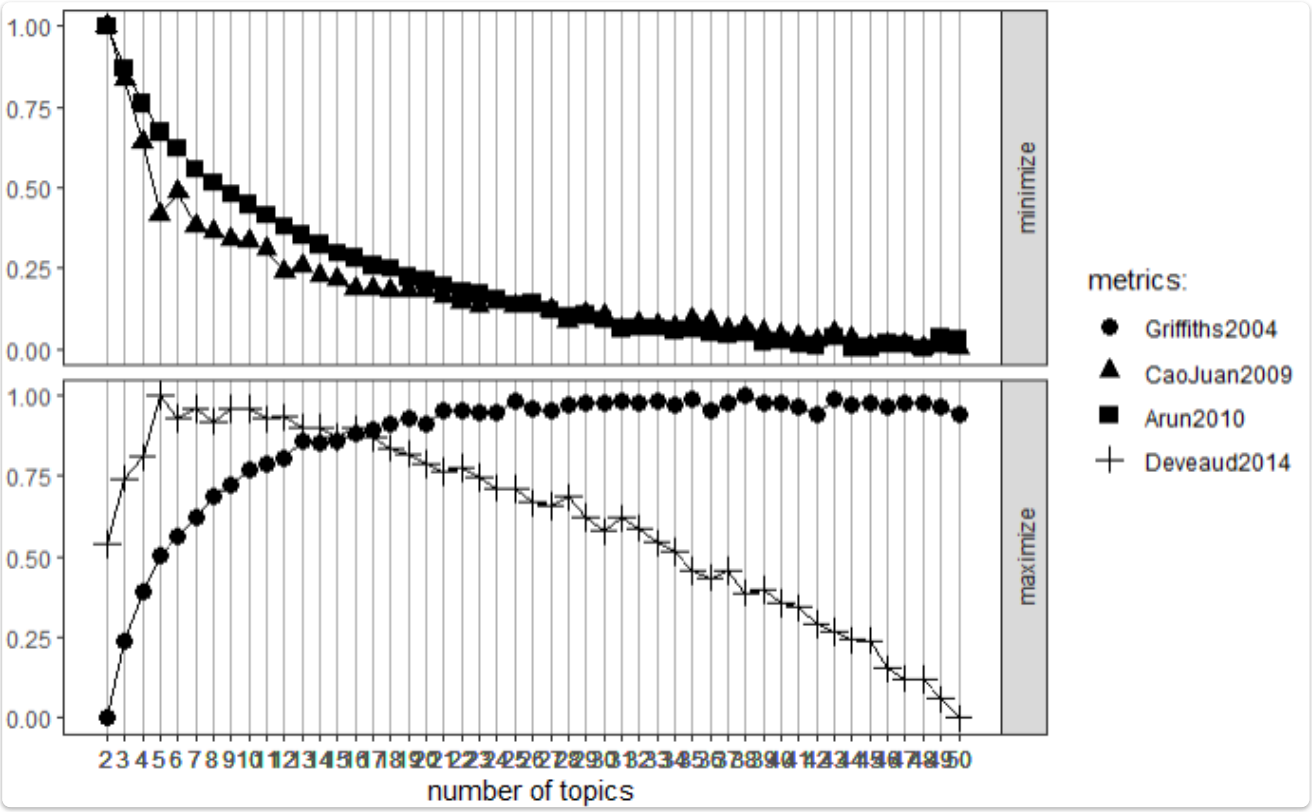
With 80 topics. Optimal no. of topics is 22-49 for Deveaud, perhaps i'll try with 28 to see if it is meaningful.



Every word is assigned a topic. Seeing if these, when the real topic (substance) is known, show any pattern which informs us of succesful modeling. It clearly does not. TRY TUNING WITH >100 TOPICS, AND SEE IF USEFUL!!!

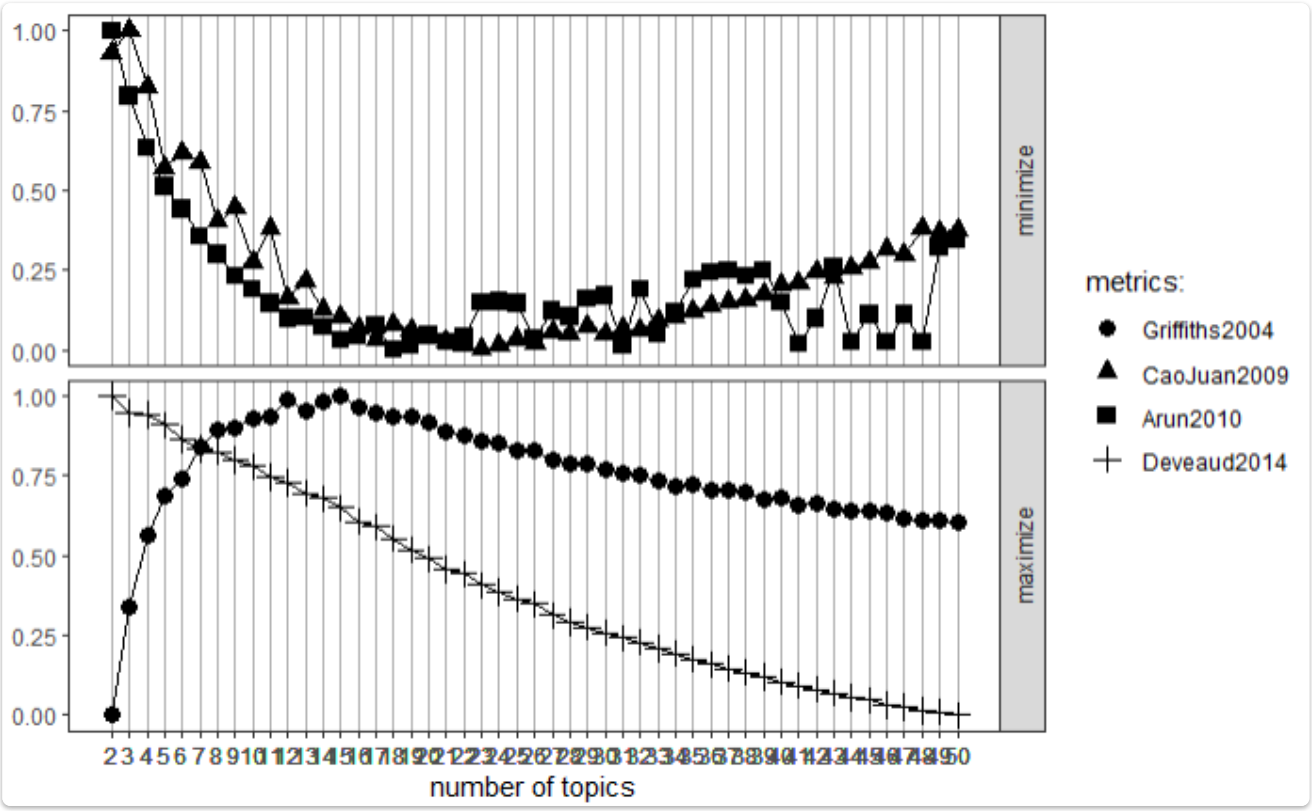
2-C family

Tuning 2C model



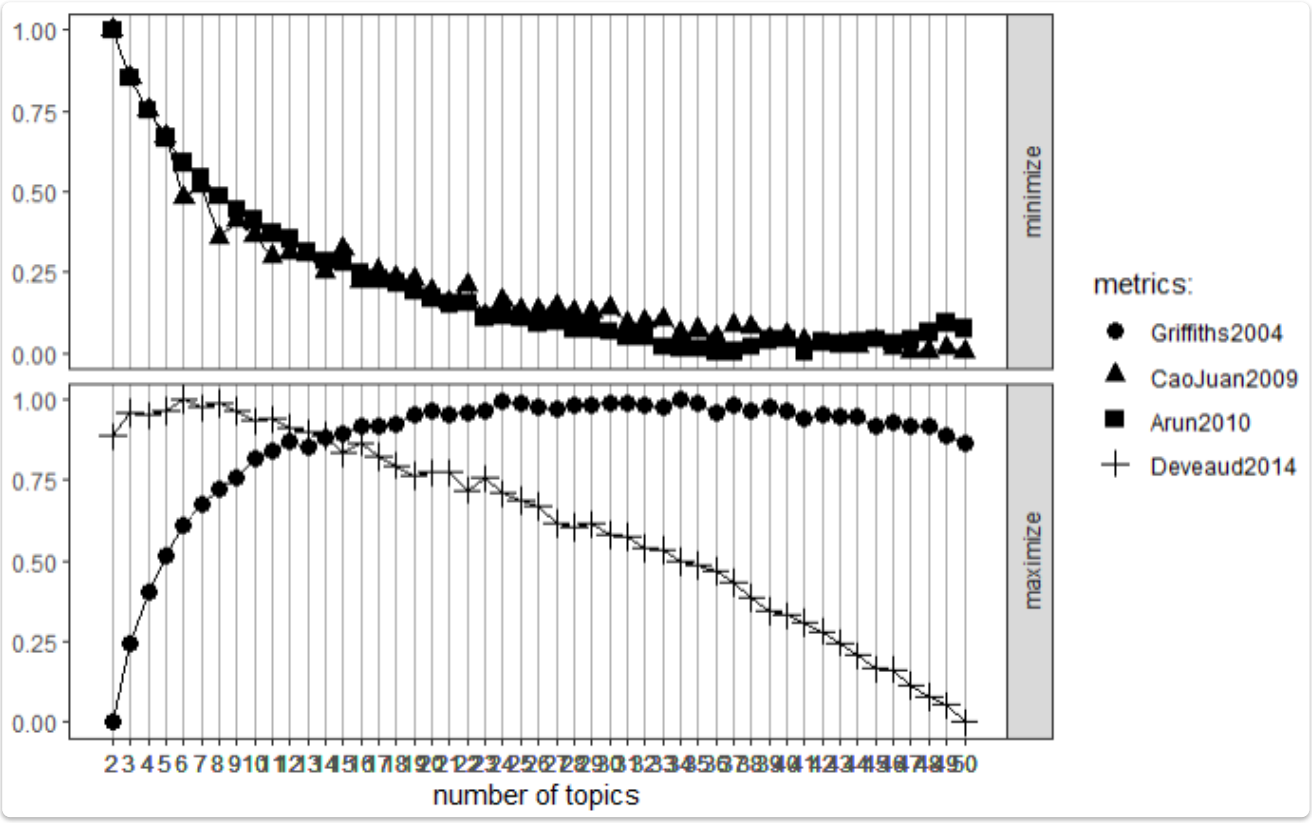
Ayahuasca

Tuning Ayahuasca model



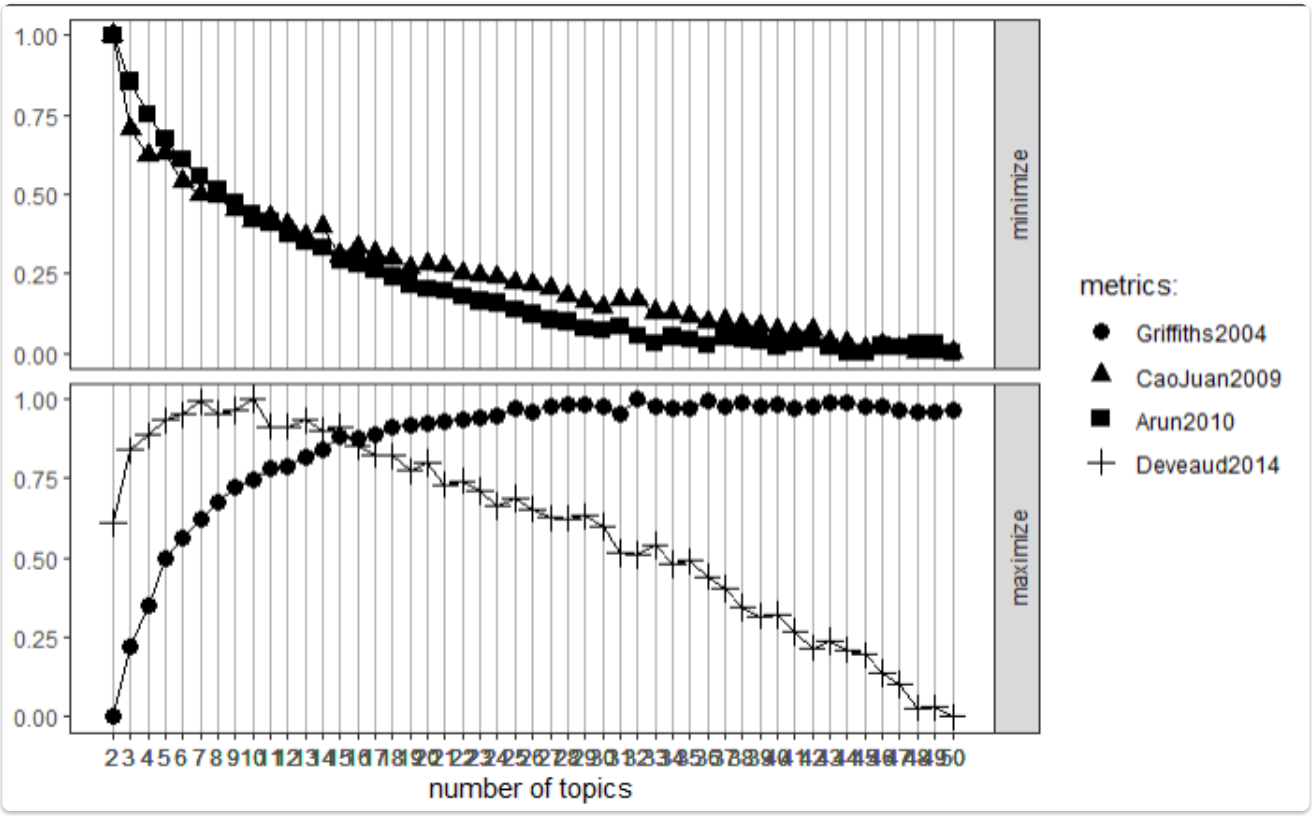
DMT

Tuning DMT model



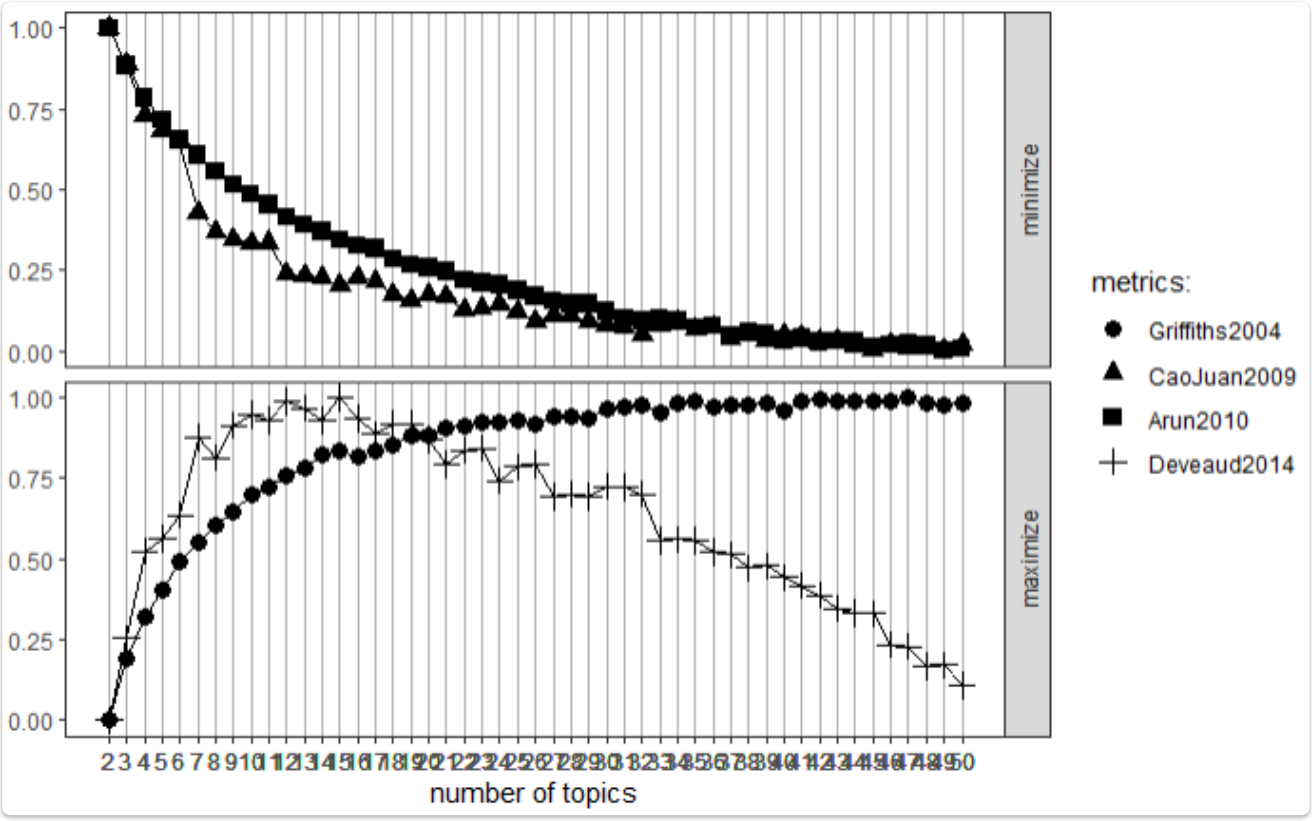
LSD

Tuning LSD model



Psilocybin mushrooms

Tuning mushroom model



Mescaline

Tuning Mescaline model