Machine Learning Analysis of Inner Experiences in Reports of Psychoactive Substances

---

Author:
Alexander Apers
6272932
a.p.apers@students.uu.nl

Bachelor's Thesis in Artificial Intelligence
Utrecht University
7.5 ects

Utrecht University

Supervisor:
Dr. Denis Paperno

Second Reader:
Dr. Rick Nouwen

August 2020

Abstract

A better understanding of psychedelic experiences becomes increasingly important as research shows their therapeutic potential (Carhart-Harris and Goodwin, 2017). One possible approach to investigate these experiences is the analysis of written reports that document individuals' experiences from the first-person point of view. An attempt at analysing such reports, described in Coyle et al. (2012), tries to predict which psychedelic substance is described based on a Bag of Words representation of the reports. However, this approach doesn't provide enough meaningful information about the subjective inner experiences that users have. To overcome this limitation, the present research proposes an approach that predicts dimensions of the experiences that express information about the subjective inner state. A small subset of 120 reports in the Erowid Experience Vaults was annotated using a 5 point Likert scale for each of 8 dimensions that were identified as relevant based on Altered States of Consciousness literature. To predict these scores, seven different report representation methods and two types of regression techniques were used to create a total of 14 different models. Reports were represented using Sentiment Analysis, Bag of Words, Word2Vec and Doc2Vec approaches. The latter 3 approaches were also supplemented with Sentiment Analysis information. The most important finding is that all 14 regression models performed better compared to baseline null models in predicting dimensions relating to inner experiences. Furthermore, the models that adhered to word order and used Word2Vec/Doc2Vec techniques made better predictions compared to simple Bag of Words models. Finally, models that were supplemented with Sentiment Analysis information performed better compared to the counterpart models without this information. Additionally, challenges and suggested improvements for this new approach and some observations regarding the data are discussed. Future research could focus on investigating the influence of more variables, e.g. set and setting, on the inner experience. A more in-depth understanding of psychoactive experiences could help pharmacological and neuroscience research form hypotheses for new investigations.

Contents

1 Introduction

For thousands of years, humans in a multitude of cultures across the globe have used psychedelic substances. Most likely they were used in therapeutic, ritualistic and religious settings (Schultes and Hofmann, 1980). In contemporary Western civilisation, psychedelics and the altered states of consciousness they elicit don't seem to occupy a role of much importance. However, their potential value to society is currently being reassessed. As a whole host of both legal and illegal psychedelic retreats and ceremonies are now accessible for the general public, the need to shed more light on this phenomenon becomes greater. With claims of healing experiences and long-lasting improved mental health, it is no wonder that their popularity is on the rise. Scientific research into these substances finds itself in a Renaissance after a period of Dark Ages due to criminalisation and substance scheduling in most Western countries. The research into psychedelics continues with promising results of its therapeutic potential (Carhart-Harris and Goodwin, 2017). Although the neuroscience behind the phenomenon is being investigated, rigorous investigations of the inner experiences people to have, seem to be lacking. We are largely unaware of the psychological mechanisms that could explain psychedelics' therapeutic effects.

As a better understanding of psychedelic experiences is becoming increasingly important, one possible approach is the analysis of written reports where the experiences users have are described firsthand. Fortunately, there is no shortage of posts on the internet where experiences are shared in a narrative textual format. Information about how psychedelic substances tend to be experienced can potentially be extracted when reports are correctly analysed. Instead of using more primitive techniques for this, such as Dictionary approaches where documents are scored based on word occurrences (Oxman et al., 1988), more sophisticated Machine Learning techniques can also be used (Coyle et al., 2012). This provides a more data-driven approach to the study of psychedelic experiences alongside existing pharmacological and neuroscience research. Additionally, when techniques are developed to automate the analysis of written reports, new substances that become important to investigate can quickly be assessed based on only a few reports. This might help form hypotheses for more informed and focused research.

A first attempt at using Machine Learning techniques to analyse psychedelic reports is outlined in Coyle et al. (2012). They concluded that Machine Learning techniques were a suitable approach for investigating reports of drug experiences. However, their specific approach had some limitations which didn't allow for a more in-depth analysis of users' inner experiences. Instead of predicting which psychedelic substance is described in a report, a greater understanding regarding the experience can be attained by predicting a set of dimensions that quantify a user's inner experience based on a report. Additionally, the comprehension and prediction of inner experiences may benefit from the supplementation of information that captures the general sentiment in a report. The reason for this is that it is likely that a user's emotional state in a report which is quantified by a Sentiment Analysis score, correlates to some extent with certain dimensions of the inner experiences.

The objective of the present research was to take the first steps in the development of this new approach. The main research question was whether various existing computational methods could make better than baseline predictions of inner subjective effects based on written reports which describe psychoactive experiences. The computational methods that were used for the models included various report representations, two regression techniques and Sentiment Analysis information. The difference in performance between computational methods was also investigated. The inner subjective effects were quantified by a set of dimensions relating to common characteristics of

psychoactive experiences. Literature on Altered States of Consciousness was consulted to identify the relevant dimensions that capture psychoactive experiences characteristics. A subset of reports which described psychoactive experiences was annotated by quantifying for each report on a scale of 1 through 5 how present each dimension was in that report.

Three main hypotheses were identified. The first was that the performance of predicting the inner experience could be improved compared to baseline null models using Machine Learning techniques. The second was that more sophisticated models would perform better in their predictions compared to simpler models. The difference between sophisticated and simple models in this context is discussed later. The final hypothesis was that supplementing Sentiment Analysis information could improve performance over counterpart models without this information.

2 Background on Psychedelic Research

We start with a short introduction on the relevant terms, definitions and concepts regarding psychedelic research.

Psychoactive substances are chemical substances that act primarily upon the central nervous system when consumed (Department of Health Australia, 2005). They alter brain function, which results in temporary changes in perception, mood, behaviour, cognition, emotional state and consciousness (Department of Health Australia, 2005; Swanson, 2018). These temporary deviations from normal waking consciousness are often referred to as Altered States of Consciousness (ASC) in research (Studerus et al., 2010). Psychoactive substances are used for several different purposes of which a few will be named. They can be used recreationally to alter consciousness, they can be used for ritualistic, spiritual or shamanic purposes, for research purposes, or they can be used as medication due to their therapeutic value. Examples of medical use include narcotics for controlling pain and stimulants for the treatment of attention disorders.

Furthermore, their use as medication for other mental disorders is currently being investigated further. In particular, a subset of psychoactive substances, namely psychedelics, appear to have efficacy in the treatment of addictions (Winkelman, 2015). This is in contrast to some other psychoactive substances (e.g. alcohol) which have habit-forming properties. Additionally, psychedelic substances are known to have effects which counter anxiety and depression (Carhart-Harris and Goodwin, 2017; Thomas et al., 2017). Furthermore, OCD and PTSD symptoms can be significantly improved by psychedelic substances (Carhart-Harris and Goodwin, 2017; Thomas et al., 2017; Mithoefer et al., 2018). When looking at more physical healing effects, psychedelics have been shown to exercise strong anti-cancer and anti-inflammatory effects through immunomodulation (Szabo, 2015). The range of healing effects appears to be wide, spanning both mental and physical healing. Even though all of these healing effects need to be studied further, they show promising potential.

To find out more about the possible psychological effects that one can come across in written reports of psychoactive experiences we look at the literature on ASC. To study ASC more rigorously and to shine more light on their commonalities between different users and different methods of entering them, the OAV questionnaire was developed (Studerus et al., 2010). It is used to measure multiple relevant dimensions of ASC and compare different types of ASC along a set of di-

mensions. Currently the OAV measures the following dimensions: 'Experience of Unity', 'Spiritual Experience', 'Blissful State', 'Insightfulness', 'Disembodiment', 'Impaired Control and Cognition', 'Anxiety', 'Complex Imagery', 'Elementary Imagery', 'Audio-Visual Synesthesiae' and 'Changed Meaning of Percepts'. This indicates what directions ASC experiences might take from the phenomenological first-person perspective that is used in written reports.

3 Automated Text Analysis

We proceed with a discussion of automated text analysis and its applications for psychological investigations.

3.1 Challenges

Nowadays, there is a vast amount of textual data on the Internet. Social media, online forums, blogs, news articles, and the like hold a lot of information which can potentially be used to uncover interesting patterns. The field of automated text analysis is concerned with training models with the ultimate goal of automatically classifying, rating or extracting useful information from documents. The main challenge to overcome in achieving this is the complex character and the unstructured qualities of natural language data. This complexity creates an information overload in which a correct semantic understanding becomes very difficult. The semantics of words and sentences are very context-dependent and nuanced computational processing of semantic contexts are not always reliable. Furthermore, there might be misspellings in the data which complicate a correct understanding even further.

Even though the analysis of textual data is relatively difficult to automate due to these challenges, the possible benefits include automatically extracting relevant information from vast amounts of data. This can overcome a lot of the limitations on resources that arise from manual text analysis using human analysers. Unfortunately, human analysers can be biased, fallible and they have limited processing power. Although these challenges still need to be overcome when designing a training data set, once the training stage is over, the automated models are expected to outperform human analysers in the long run.

3.2 Applications for Psychological Investigations

The methods in the field of automated text analysis are increasingly being used for psychological investigations (Iliev et al., 2015). It appears that these methods are suitable for analysing the psychological effects of a particular group of experiences through textual data. The group of experiences that are of interest in this investigation, namely psychoactive experiences, are well documented online. For many psychoactive substances, there are so many reports that it quickly becomes unfeasible to read them all. Automated text analysis could potentially highlight important patterns in narrative reports of these experiences. These kinds of analyses have many possible applications, some of which we will discuss now.

The first application of automating the analysis of psychoactive experience reports could be to facilitate research that investigates how the psychological effects correlate with neuroscientific data that is collected. Another application is the use of pre-trained models to perform quick analyses of a new psychoactive substance based on a few reports. This could be used to make initial safety

assessments before more thorough investigations can take place. It could also be used to develop hypotheses in clinical and pharmacological research on psychoactive substances. Additionally, a better understanding of the range of possible experiences that can occur may lead to safer, more informed research. Other variables could also be included in such analyses. This would, for example, open the door to investigating the influence of factors such as genetic predisposition to mental illness and the set and setting of an experience, without having to set up large clinical studies. Or when such studies do become necessary, preliminary analysis of textual reports could lead to more focused research.

4 Literature Review

In this section, we first examine previous relevant literature that used similar approaches as the present research. Then, we discuss some literature about the Machine Learning techniques that were used.

4.1 Previous Research

First attempts at studying written reports of psychoactive substance experience have used either qualitative methods or dictionary approaches (Coyle et al., 2012; Oxman et al., 1988). However, such approaches have obvious limitations. Qualitative methods are time-consuming and difficult to automate, while dictionary approaches rely on previously obtained categories of terms that are used for scoring documents (Coyle et al., 2012). Such dictionaries would need to be updated regularly to keep up with changing terminology of psychoactive reports.

To overcome these limitations, Coyle et al. (2012) describe an approach that uses Machine Learning techniques to analyse psychedelic reports. A Bag of Words representation of the reports was used to train a Random Forest Classifier to predict which of 10 psychedelic substances was described in the report. Since an accuracy of 51.5% was attained, it was shown that different reports describing the same substance appear to have consistencies which allow for correct classification. Coyle et al. (2012) concluded that Machine Learning techniques were a suitable approach for investigating reports of psychedelic experiences. It is important to note that Coyle et al. (2012) have used the same online source for the data as the present study. This source is discussed in the Data section.

Nevertheless, the specific approach of Coyle et al. (2012) has some limitations. While their research showed which terms in the reports were most important for making classification decisions, many of these terms are not very informative. Some terms provide hardly any information at all, e.g. 'back', 'day', 'tell'. Other terms give clues relating to the external context in which a substance was used, e.g. 'party', 'club', 'house'. However, these individual terms don't provide meaningful information about a user's internal subjective experience.

Another limitation is that only psychedelic substances were included. With the applications of the automated analysis of psychoactive reports as described in the Applications for Psychological Investigations section, the models shouldn't be trained only on reports of psychedelic experiences which is the case in Coyle et al. (2012). Rather, using both psychedelic and non-psychedelic psychoactive reports allows the models to recognise and characterise a wider range of possible experiences.

A similar approach which also includes non-psychedelic psychoactive substances and uses a Multinomial Naive Bayes classifier is described in Strapparava and Mihalcea (2017). However, instead of predicting which individual substance is described in a report, only 4 substance type categories are distinguished from each other. All substances are grouped into 'Empathogens' e.g. MDMA, 'Hallucinogens' (Psychedelics) e.g. LSD, 'Sedatives' e.g. alcohol and 'Stimulants' e.g. cocaine. This grouping decreases the complexity of the classification task significantly which results in a relatively high overall F1-score of 88% compared to a baseline of 61% which indicates the percentage of the majority class. The paper proceeds with specific analyses of dominant word classes and dominant emotions for each of the 4 substance classes. Since the approach is very similar to Coyle et al. (2012), this paper also inherits some of the same limitations. The most important of these limitations is that it is difficult to extract meaningful information about subjective inner states solely based on the occurrence of terms that are informative for classification.

A different approach, described in Bedi et al. (2014), analysed speech during and after MDMA experiences. This kind of analysis offers a more direct vantage point for investigating the altered state as the speech is recorded during and shortly after the experience as opposed to some time afterwards. Distributional semantics were used to analyse semantical proximity of words. Specifically, Latent Semantic Analysis was used to represent the meaning of a word as a vector in a high dimensional space. The cosine between the corresponding vectors of two words represents semantic proximity. The semantic proximity of spoken words to preselected words that capture the subjective states during MDMA intoxication was measured and compared to 2 controls, a placebo and a methamphetamine condition. Additionally, they used the semantic proximity to relevant terms as features to predict drug condition with Support Vector Machines.

They found that speech during an MDMA experience had greater semantic proximity to the word 'empathy' compared to placebo. Furthermore, they found that speech after MDMA had greater semantic proximity than placebo to the concepts 'friend', 'support', 'intimacy', and 'rapport'. Furthermore, classifiers could distinguish between speech on MDMA and placebo with 88% accuracy and between MDMA and methamphetamine with 84% accuracy. These findings suggest that automated semantic speech analyses are sensitive to subtle alterations in the state of consciousness as expressed by speech since it can accurately discriminate between 2 different substances (Bedi et al., 2014).

4.2 Machine Learning Techniques

The report representation that was used in Coyle et al. (2012), namely Bag of Words, has some limitations. The Bag of Words approach represents reports using a simple term-occurrence counting dictionary. To make such a dictionary, the complete vocabulary of all reports has to be determined first. After this, high dimensional sparse vectors are produced to represent reports. However, this approach loses high-level semantic meanings in textual data (Zhao and Mao, 2018). One of the reasons for this is that it assumes that every term in a report is independent (Kim et al., 2017). Using this approach, semantically similar reports may be mapped to very different vectors due to different word usage (Zhao and Mao, 2018). Another limitation is that a Bag of Words representation loses word order information which might be necessary to correctly understand negations or multiple term expressions.

One possible technique to overcome some of these limitations is Word Embeddings. This technique takes into account that different terms can have similar meanings since it creates a distributed

representation of each term (Mikolov et al., 2013). This is achieved by predicting target words based on context words in the neighbouring window of the target words. Another improvement over the Bag of Words method is that the number of dimensions can be manually determined. This leads to lower-dimensional and denser vector representations. The drawback of this is that the interpretability of the Bag of Words method is lost. The performance of a set of clinical text classification tasks was improved in Shao et al. (2018) when using the Word2Vec technique described in Mikolov et al. (2013) compared to Bag of Words. However, Word Embeddings are not guaranteed to improve performance in every context. One example is described in Yogarajan et al. (2020). For very domain-specific texts that tend to use particular terms, Bag of Words approaches might perform better compared to Word Embeddings.

To extend the notion of Word Embeddings to an entire report, the Doc2Vec technique was developed (Le and Mikolov, 2014). Instead of predicting target words based on context words alone, this technique also takes document context into account. This is achieved by adding a vector representation of documents to the Word2Vec neural network. This yields both Word Embeddings and Document Embeddings. The vector representations of new documents that did not appear in the training data can be inferred from the trained model.

Another interesting technique which could add valuable information to report representations is Sentiment Analysis. Sentiment Analysis quantifies how positive or negative the general sentiment is in sentences or documents (Mäntylä et al., 2018). Supervised Machine Learning techniques are used to predict the corresponding sentiment analysis score of a sentence or document. However, many challenges are associated with this. These challenges include irony, sarcasm, negations, word ambiguity and multipolarity, which involves multiple polarities of sentiment in a single text.

5 Hypotheses

First, we recall the main research question of whether various existing computational methods could make better than baseline predictions of inner subjective effects based on written reports which describe psychoactive experiences. Additionally, two sub-questions are formulated. How does the performance of models using different computational methods vary? And does the supplementation of Sentiment Analysis information improve performance? This section establishes the 3 main hypotheses that provide possible answers to these questions based on the discussed literature.

The two most similar research papers to the approach of the present research, Coyle et al. (2012) and Strapparava and Mihalcea (2017), found it was possible to predict which substance was described in a report with higher accuracy compared to baseline. Additionally, Bedi et al. (2014) found that alterations of mental states can be detected in speech. The textual output of users could be used for the more general task of gauging their subjective mental states. Even though the present research used a different approach which did not involve substance classification tasks, the findings from these papers were used to form the first main hypothesis. This hypothesis was that the performance of predicting the inner experience could be improved compared to baseline null models.

Since both very simple e.g. Bag of Words, and slightly more sophisticated techniques e.g. Word Embeddings, were used in this research to represent reports, it was likely that their performance would vary. Even though a Bag of Words approach might perform better in domain-specific con-

texts due to the use of jargon, it was expected that psychoactive experience reports were very liberal with terminology. Therefore, it was likely that models which used more sophisticated techniques for document representation to overcome the described limitations of a Bag of Words representation would perform better compared to a simple Bag of Word representation. This effect is in line with Shao et al. (2018) who found an increase in performance when using Word Embeddings compared to a Bag of Words approach.

Furthermore, it was likely that models that used Sentiment Analysis information would perform better compared to counterpart models without this information. The reason for this is that psychoactive substances are known to elicit alterations in the emotional state such as intensification and a broadening of emotional range (Swanson, 2018). Furthermore, Strapparava and Mihalcea (2017) have already identified that some basic emotions are more dominant in some substance classes compared to others. Therefore, sentiment analysis information was expected to correlate with at least some dimensions of ASC. The dimensions which have strong sentiment inherent in the formulation of the dimension were expected to have the strongest correlation. For example, a dimension such as 'anxiety' would likely be correlated with negative sentiment while a dimension as 'blissful state' would likely be correlated with positive sentiment.

6 Data

To investigate the feasibility of predicting inner experiences using computational methods a data set had to be created which coupled reports of psychoactive experiences with annotated scores which quantify the inner experience. This section discusses the data that was used for this, how it was collected and how it was annotated. Then, we look at analyses that were performed on the annotated data. Finally, a few points that surfaced when reading the reports are addressed.

6.1 Approach

This subsection gives a short preview of the steps that were taken to create the data set. The steps are described in more detail in the following subsections. The first step was to collect a text corpus of reports which described experiences of relevant psychoactive substances. Then, it needed to be established which dimensions expressed information about the subjective experiences of psychoactive substances. A subset of the most important dimensions had to be identified since annotation time was limited and more dimensions take more time to annotate. When these dimensions were identified, the annotation process was started and each document was assigned scores of 1 through 5 for each dimension. To obtain an indication of the subjectivity of this process, 3 volunteers were asked to annotate a small portion of the reports. The variance between annotators was analysed. Additionally, to achieve a better understanding of the dimensions, a correlation matrix was calculated to describe how dimensions were correlated with each other. To better understand how a typical experience varied between different substances, the median value for each dimension was calculated using the reports of each substance.

6.2 Collection

The reports that were used were collected from the Erowid Experience Vaults. These reports came tagged with the substance that was being described. The entire corpus consisted of 5432 reports

ranging 14 different psychoactive substances. These substances were chosen for the level of scientific comprehension of their effects and their general popularity. The selected substances are MDMA, Psilocybin, LSD, Cannabis, Salvia divinorum, Cocaine, Heroin, DMT, Ketamine, 2C-B, 5-MeO-DMT, Ayahuasca, Alcohol and Kratom. The Erowid Experience Vaults distinguishes multiple different species of Psilocybin mushroom as well as different concentration levels of Salvia divinorum. For this analysis, however, no distinction is made and different species or concentration levels are grouped together. For the 'MDMA' class both reports of 'MDMA' and 'Ecstasy' were included. The smaller subset of the corpus that was manually annotated consisted of 120 reports. Since the larger corpus included a portion of reports which were unsuitable since they were not subjective reports of drug experiences, these reports were disregarded during annotation. Types of reports that were disregarded included recipes, advice, retrospective summaries of long term drug use and warnings about adverse health effects. Also, some reports that used a third-person perspective were disregarded since they missed the first-person perspective.

Considering that not all 14 psychoactive substances were represented in the corpus in equal amounts, the 120 reports for annotation were selected to represent each substance using the same proportions as the corpus. As a result of this, some substances only had a handful of reports annotated. Table 2 shows the composition of the 120 reports. The 120 reports were split into a train (100 reports) and a test set (20 reports).

Table 1: Number of Analysed Reports of Each Substance

Psychoactive Substance	**Training Set**	**Test Set**
Salvia divinorum	17	3
Psilocybin	16	3
Cannabis	14	3
MDMA	13	3
LSD	13	3
Cocaine	5	1
5-MeO-DMT	4	1
DMT	4	1
Kratom	3	1
Ketamine	3	1
2C-B	2	0
Heroin	2	0
Alcohol	2	0
Ayahuasca	2	0

6.3 Dimension Selection

To decide which dimensions of the experiences described in the reports were important to analyse, the dimensions as described in 'Psychometric Evaluation of the Altered States of Consciousness Rating Scale (OAV)' and its corresponding questionnaire (Studerus et al., 2010) were taken as a basis. These dimensions are 'Experience of Unity', 'Spiritual Experience', 'Blissful State', 'Insightfulness', 'Disembodiment', 'Impaired Control and Cognition', 'Anxiety', 'Complex Imagery', 'Elementary Imagery', 'Audio-Visual Synesthesiae' and 'Changed Meaning of Percepts'. However, for this research, only dimensions relating to the inner subjective experience were relevant. Fur-

thermore, some dimensions were slightly rephrased to make annotation more accessible to other annotators. For example, 'Experience of Unity' was changed to 'Feeling of connectedness', 'Spiritual Experience' was changed to 'Transcendental Experience' and 'Blissful State' was changed to 'Experience of Euphoria'. Finally, dimensions that describe the healing effects some psychoactive substances tend to have and the perceived dangerousness of the experience were added. This resulted in the following 8 dimensions.

Table 2: The 8 dimensions used for annotation

Dimension	**Typical 1**	**Typical 5**	**Short Name**
How transcendental was the experience?	ordinary	mystical experiences	transcendental
How much physical discomfort was experienced?	ordinary	severe pains/vomiting/etc.	discomfort
How much ego dissolution was experienced?	ordinary	'ego death'	egodissolution
How anxiety-producing was the experience?	ordinary	panic attacks	anxiety
How euphoric was the experience?	ordinary	ecstatic	euphoria
How connected did users feel?	disconnected	one with everything	connected
Did user experience physical/mental healing?	no healing	life-changing	healing
Did user think the experience was dangerous?	completely safe	extremely dangerous	dangerous

6.4 Annotation Process

Each of the 120 reports was read and subsequently annotated. For each of the 8 dimensions of ASC that were identified as relevant, a score of 1 through 5 was assigned to the report. A score of 1 was assigned to a dimension in a report when the characteristic associated with the dimension was completely absent in the report. In theory, a report of an everyday, healthy, waking state of consciousness would hypothetically score 1's for each dimension. A score of 5 indicates that the characteristic associated with the dimension is very much present in the report. Scores of 2, 3 and 4 respectively fill this gap in roughly equal parts. While it is important to note that technically a Likert scale produces ordinal type data, it can be assumed that the data is approximately cardinal. This assumption is typical and allows the use of regression techniques to model the ordinal data. To get an indication of the subjectivity of the annotation process, 3 volunteers were asked to annotate a set of 5 selected reports. None of these annotators was a trained psychologist. All of them reported some level of familiarity with psychoactive substances and their effects. For each dimension, the variance in the scores was assessed to quantify the subjectivity of each dimension. In total, the scores of 4 different annotators were included in the variance analysis.

6.5 Data Analyses

Figure 1 shows how the dimension scores (1-5) differ between 4 different annotators. Each variable is plotted along with the mean variance between the 4 scores over 5 selected documents. The variables 'anxiety', 'dangerous' and 'discomfort' appear to have low variance; 'healing' and 'connected' appear to have intermediate variance while 'transcendental', 'euphoric' and 'ego dissolution' show a high amount of variance. A higher amount of variance suggests a larger degree of subjectivity in the annotation of that variable.

The dimensions 'ego dissolution', 'euphoria' and 'transcendental' show the most variance. These

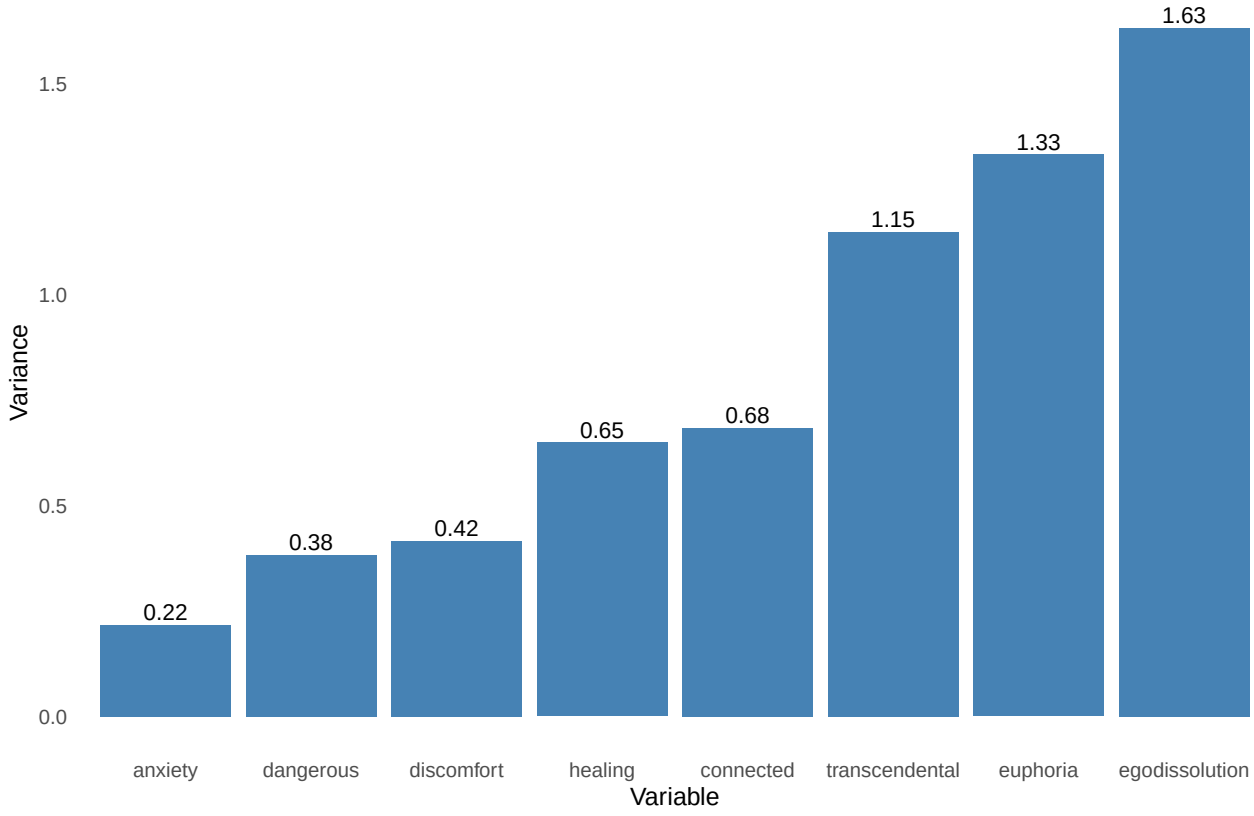

Figure 1: Variance analysis between 4 annotators

dimensions may be rather vague for inexperienced annotators making it unclear how these dimensions should be annotated. The dimensions with a higher degree of variance also appear to be dimensions which are relatively rare in everyday waking consciousness and occur more frequently in psychoactive experiences. The dimensions which show a smaller degree of variance are more common and familiar dimensions for most people. Examples are anxiety, discomfort and a feeling of connectedness.

Since this analysis used only 4 individuals annotating a very small set of 5 reports due to time and resource constraints, this analysis should not be over-interpreted. The uncertainty in the correlation scores is very large as both the selection of reports and the selection of annotators may have biases. Additionally, the inexperienced annotators who read psychoactive reports for the first time did not have much material to compare against which is necessary to make balanced annotation decisions.

Figure 2: Dimension correlation matrix

Figure 2 shows how the 8 different dimensions that were annotated correlate with each other over the 120 annotated reports. It is important to note that the analyses from now on are be based solely on one annotator due to limited resources.

The largest amount of correlation is between the dimensions 'transcendental' and 'ego dissolution'. Additionally, the dimension pairs 'connected - healing', 'transcendental - connected' and 'anxiety - dangerous' also show a strong correlation. 'Anxiety - euphoria' show the largest amount of negative correlation indicating a small contrary relationship between these dimensions.

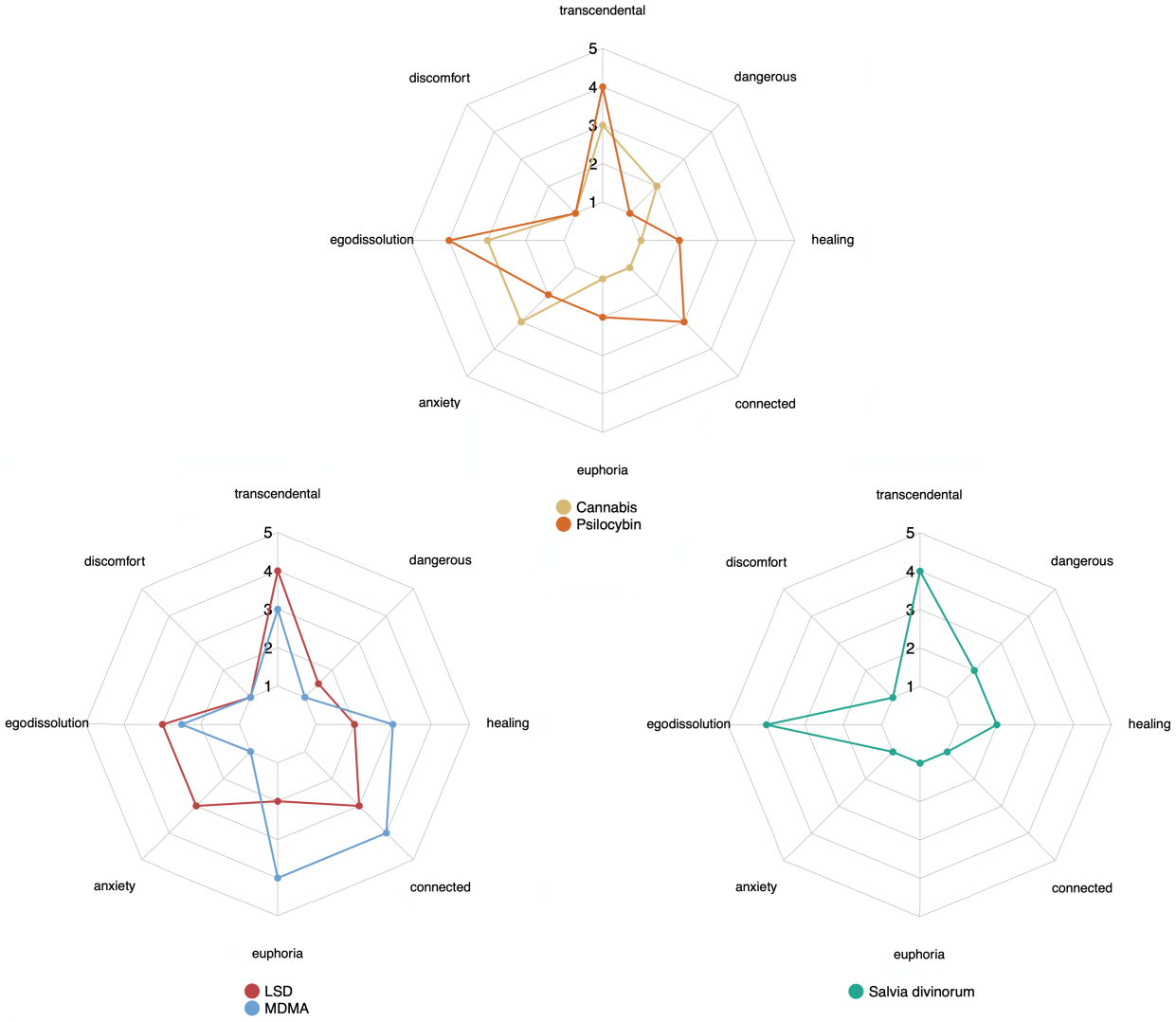

Figure 3: Compound Differences

The radar plots in figure 3 show what a typical experience with a particular substance looks like. For LSD 16 reports were used in this analysis, for MDMA and Cannabis 17 reports, for Psilocybin 19 reports and for Salvia divinorum 21 reports. Substances of which less than 15 reports were analysed were disregarded as too few analysed reports of a substance increases the chance for a biased subset of reports. This might lead to a skewed view of a substance. It is important not to over-interpret these analyses as they are based on only one annotator and a very limited amount of reports. A more thorough discussion of bias in this research will follow in the Discussion section. This analysis is meant as a very general overview to compare experiences from different substances.

For every substance, the median score of each of the 8 dimensions was calculated and plotted. The 5 substances are distributed over 3 radar plots. In general, reports which describe MDMA experiences tend to be rather high in euphoria, a feeling of connectedness and healing. A typical

Cannabis report tends to lack these dimensions and shows higher levels of anxiety and perceived dangerousness. Reports of Psilocybin and LSD appear to have a similar shape over the 8 dimensions. Reports of Salvia divinorum tend not to show large amounts of anxiety, euphoria or connectedness but they focus more on ego dissolution and the transcendental qualities of the experience.

A general note on these analyses is that only general trends, as opposed to detailed findings, could be identified due to limitations of subjectivity and the small data set. This is why trends that were identified here need to be confirmed in larger investigations with more resources.

6.6 Remarks on Data

During the annotation process, it stood out that only a very small portion of a report contains the relevant information if the relevant information is explicitly stated at all. The result of this is that annotation tends to feel like finding a needle in a haystack. Many reports contain a lot of information about events that precede or succeed the experience. However, this does make a substantial portion of the corpus suitable for researching the influence of set and setting on the inner experience.

Other points that should be taken into account are more practical. The reports tend to vary significantly in length. Whereas some reports are too short to contain valuable information, some reports are very elaborate. A few reports use a poetic depiction of an inner experience instead of the more common objective reporting style, making it difficult to annotate accurately. Furthermore, it is important to note that the dose of the psychoactive substances that were used in an experience can vary significantly. The fact that it is almost always unclear what dosage was used makes it difficult to eliminate the effect of this variable on the experience.

7 Methodology

To test the formulated hypotheses regarding the ability of different computational methods to predict dimensions of the subjective experience, computational models had to be created. This section describes how the models were created and how their performance was tested.

7.1 Models

In total, 14 different models were used for training and testing. A model consists of two parts: a method for representing a document (feature extraction) and a regression technique to fit the train data and predict scores for the test data. A separate model was trained for each of the 8 dimensions that were annotated. The models represented the reports using a few different techniques. These techniques were: Bag of Words, Word2Vec Word Embeddings, Doc2Vec. Additionally, Sentiment Analysis was used as a representation and each of the mentioned techniques was also coupled with Sentiment Analysis information. To predict the scores based on these representations both Logistic and Linear regression were used.

The least complex report representation method was Bag of Words (BoW). This technique starts by building a vocabulary based on each report in the data set. Then it proceeds to create a high dimensional vector representation for each report where each vector component corresponds to the term frequency of a term in that report. However, as discussed, a BoW representation has obvious

limitations since it disregards word order and semantic similarity between different terms. To address the latter limitation of the BoW model, where different terms are assumed to be completely independent, a model which uses Word2Vec Word Embeddings (WE) (Mikolov et al., 2013) was constructed. The pre-trained Word2Vec model that was used was trained on a Google News data set and contains 300-dimensional vectors for 3 million English words (Google Model). To create a 300-dimensional vector representation for each report, the vector representation of every single term in that report was weighted according to its IDF value in the corpus and summed. Finally, the resulting vector was normalised. Terms which occurred in the document but were not present in the pre-trained Word2Vec model were disregarded. This technique significantly reduces the size of the representation vectors from the size of the vocabulary using BoW, which is 8408, to 300 using WE.

To address the other limitation of not taking into account word order, a Gensim Doc2Vec (Le and Mikolov, 2014) model was trained on 24788 reports from the Erowid Experience Vaults. The Doc2Vec algorithm creates a vector representation for each document in the training set in addition to a vector representation for each word in the vocabulary. Furthermore, it takes into account the word order of the document. To match the size of 300 features that were used in the WE approach, the Doc2Vec model was also instructed to use 300-dimensional vectors. The model was left to train over 40 epochs.

To investigate whether supplying information about the general sentiment in a document would improve performance, a sentiment score of each report in the data set was calculated. To calculate these scores, a general-purpose AllenNLP model that was adapted for sentiment analysis was used. The general-purpose model was pre-trained on the 'RoBERTa large' data set (Liu et al., 2019) and was adapted to sentiment analysis by fine-tuning it on the Stanford Sentiment Treebank. However, this sentiment analysis model only predicts scores for individual sentences. To extend the notion of sentiment analysis from sentences to an entire report, the mean sentiment score over all sentences in the report was calculated. A model was trained which used sentiment analysis as the sole predictor variable to set a baseline. Additionally, sentiment analysis was added as a feature to each of the models that were investigated.

The regression techniques that were used to fit the train data and predict scores for the test data were logistic regression and linear regression. The application of both of these techniques to our data is discussed now.

Logistic regression is generally used for modelling the probability of the outcome of a binary dependent variable. Since the current data set uses a scale of 1-5 as the dependent variable, an extension of the logistic regression model is necessary. One-Vs-Rest logistic regression, which produces a set of binary classifiers for each possible value of the dependent variable, was used. In our case, this creates 4 different decision boundaries. On these boundaries the probability of a data point falling in either of the neighbouring areas is equal. Since these separate classification problems are assumed to be independent, the linear decision borders can have different slopes. This divides the dimensional space into at most 5 separate areas, where if a data point falls in one area, it is classified with the corresponding score of that area.

It had to be taken into account that the annotated scores of the dimensions were unbalanced, i.e. some scores were more frequent compared to others. This is why the class weights were set to 'balanced' to account for the unbalanced distribution of dimension scores. This weighs each score

to the frequency with which it occurs in the data.

The linear regression technique is rather straightforward. A single linear function models the relationship between the independent variables and the dependent variable. This means that each feature in the report representation is assigned a coefficient, that indicates how much the dependent variable should be changed if the feature is increased by one unit, provided that all other features keep the same value. However, since the linear regression model treats the scores of the dependent variables as continuous, it is not restricted to integer values. Therefore, each real number the linear regression model outputs was automatically rounded to the nearest integer in {1, 2, 3, 4, 5}.

The following 14 models were used in the investigation. The first two models use solely sentiment analysis information coupled with both logistic and linear regression. The following models combine Bag of Words, Word Embeddings and Doc2Vec report representations with linear and logistic regression. Each of the latter 6 models is also coupled with sentiment analysis information, which brings the total up to 14 models.

7.2 Performance Assessment

One of the most commonly used metrics to evaluate the performance of regression tasks is Mean Squared Error (MSE). However, this metric is not very suitable since the ordinal character of the data allows for imbalance. Standard MSE doesn't take into account that some scores occur more frequently compared to others. To account for this, the macroaveraged analogue of MSE as described in Baccianella et al. (2009) is used. This weighs the error of each score according to the frequency of that score.

While MSE provides information about the magnitude of the error, the other metric that was used, Pearson's r [1], gives information about whether the predictions follow the same general trend as the actual scores. This correlation coefficient can range from -1 to 1 with -1 indicating perfect negative correlation, 0 indicating no correlation and 1 indicating perfect correlation. It is important to note that when one variable has a constant output, Pearson's r cannot be calculated since the standard deviation of that variable, used in the denominator, is 0.

A small MSE and high correlation coefficient indicate that the predicted scores are close to the actual scores, while also following the same general trend as the actual scores. This means that when actual scores increase, the predicted scores are also higher and vice versa. To establish a baseline level for both of these metrics two null models were fabricated. The first null model predicted random scores for each dimension while the second null model predicted the most frequently occurring score for each dimension.

8 Results

The following results have been obtained from the computational experiments as described in the Methodology section.

[1] The rounded outputs of the linear regression models were used in the calculation of Pearson's r for the sake of fair comparison with logistic regression models.

8.1 Sentiment Analysis

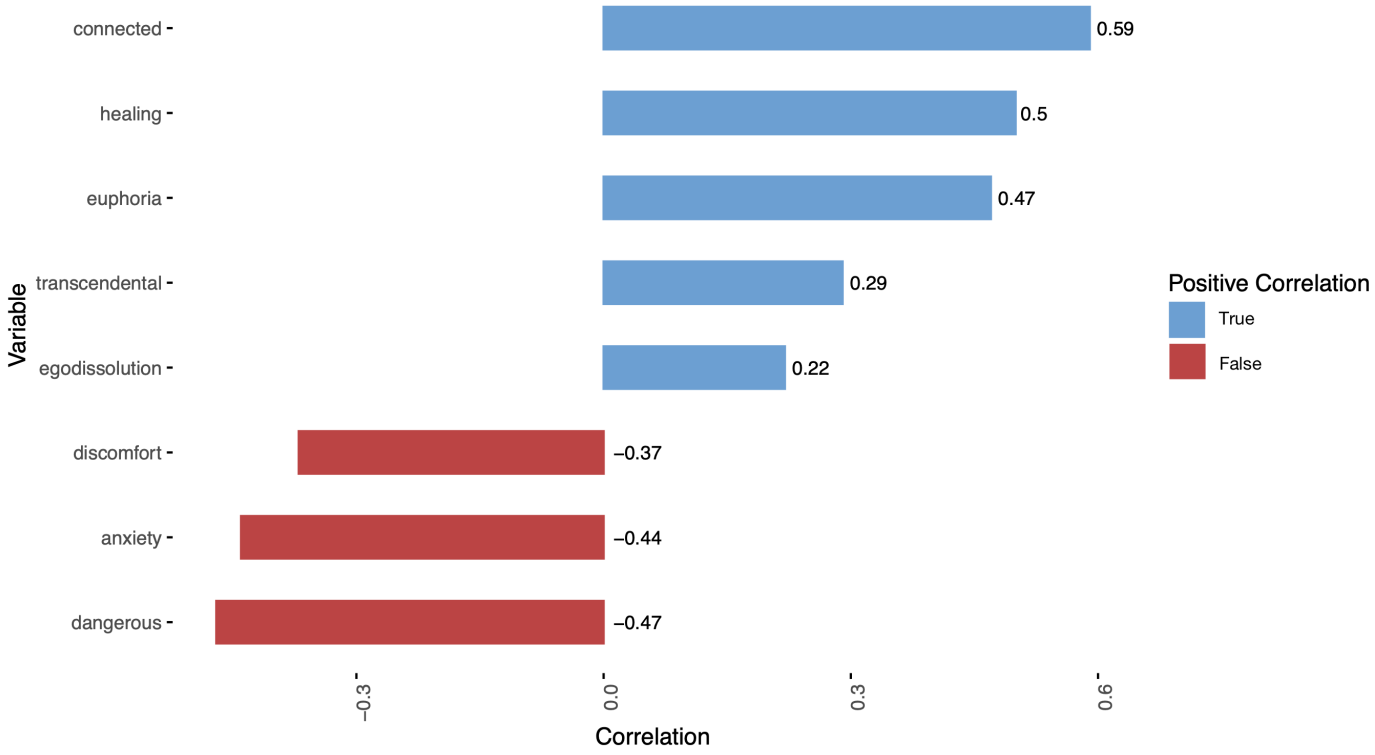

Figure 4: Sentiment dimension correlation

To investigate whether adding the sentiment analysis score of a report aids prediction of the 8 dimensions it first had to be established which dimensions show correlation with positive sentiment. Additionally, this analysis gives an overview of how positively or negatively each dimension tends to be experienced.

Figure 4 shows the Pearson's r correlation coefficient between the sentiment analysis scores of reports and each of the 8 dimension scores that were assigned during annotation. 'Connected', 'healing' and 'euphoria' tend to be the most positively experienced. 'Transcendental' and 'ego dissolution' are experienced on average as reasonably positive while 'discomfort', 'anxiety' and 'dangerous' are inversely correlated with positive sentiment.

8.2 Model Results

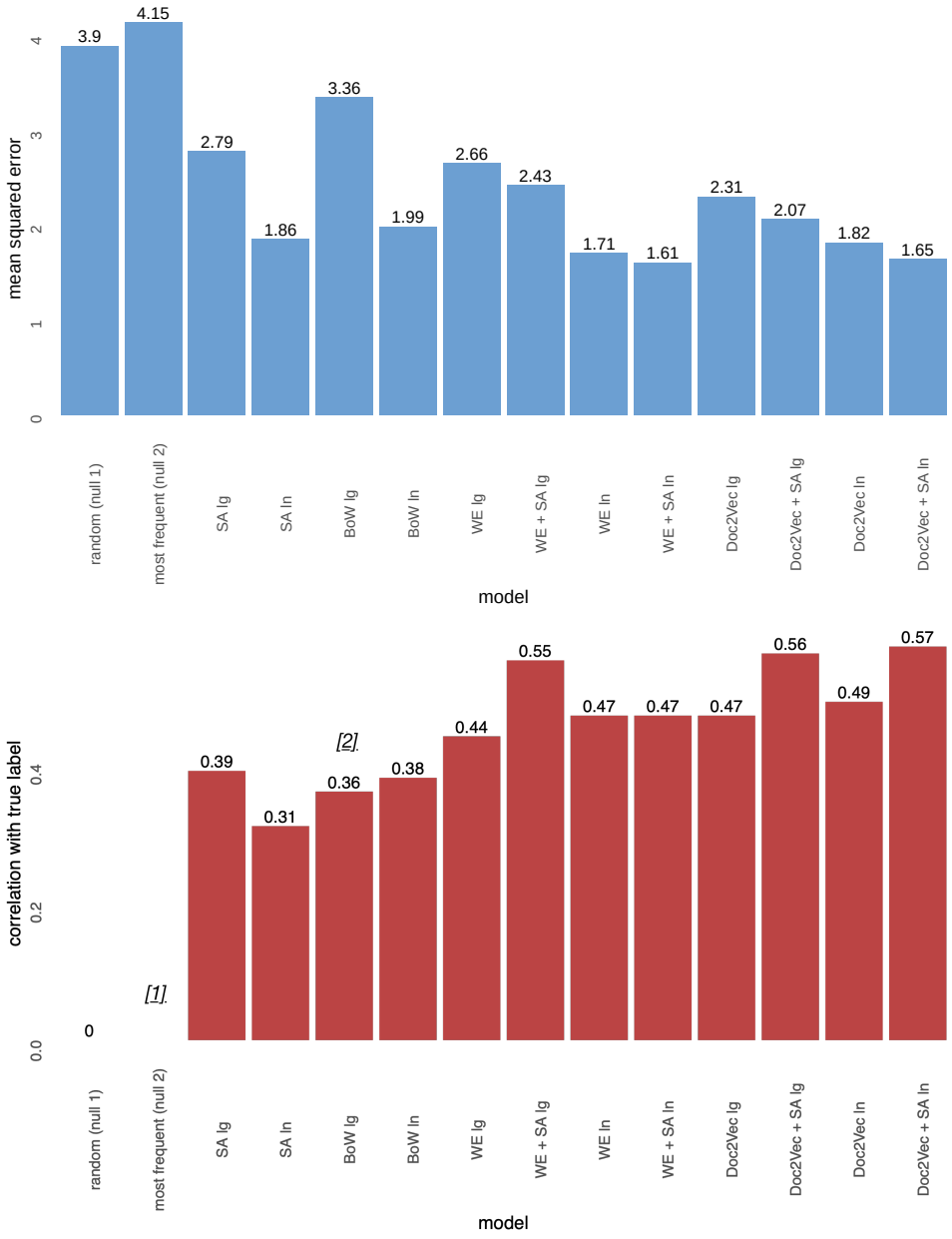

Figure 5: Model Performance

Abbreviations SA: sentiment analysis; BoW: bag of words; WE: word embeddings; lg: logistic regression; ln: linear regression.

[1] since the output of this model is a constant value, Pearson's correlation coefficient could not be calculated.

[2] since the output of one dimension was constant, this dimension was omitted from the metric calculation and only the remaining 7 dimensions were used.

Figure 5 shows the two performance metrics of each model. These were used to evaluate model performance. The two null models, which are meant as a baseline for the metrics, perform the worst as expected. The random model predictions, for which repeated random sampling is used, show zero correlation with the actual scores and the MSE is 3.9. This is a relatively large error. The second null model, which predicts the most frequent score for each dimension has an even higher MSE score of 4.15. The reason that this null model performs slightly worse when looking purely at MSE is that most dimensions have a most frequent score of 1. This means that when the actual score is a 5 the predicted label of 1 gives a large error.

All other models that were tested performed better compared to the two null models as shown by both metrics. The two models that use sentiment analysis information as the sole predictor variable provide a baseline for sentiment analysis information. Interestingly, when coupled with linear regression, this model shows both lower MSE (1.86) and correlation (0.31) compared to the logistic regression model (2.79, 0.39 respectively).

The Bag of Words models perform similarly to the models that use only sentiment analysis information. The Bag of Words models that were coupled with sentiment analysis information showed exactly the same performance as the regular Bag of Words models. For this reason, they are omitted from the results figure.

The models that use linear regression consistently show lower MSE compared to logistic regression. Models that don't use sentiment analysis information also show a slight increase in correlation when switching from logistic to linear regression. The Word Embeddings models that do use sentiment analysis information show a decrease in correlation when switching from logistic to linear but the Doc2Vec models that use sentiment analysis information show a slight increase.

When analysing the difference in results between different report representations, it appears that the Word Embeddings models perform better compared to the Bag of Words and baseline sentiment analysis models as indicated by lower MSE and higher correlation scores. Generally, Doc2Vec slightly increases performance over the Word Embeddings models. All Doc2Vec models show more correlation with the actual scores compared to their Word Embeddings counterpart models.

Apart from the Bag of Words model, each other model in which sentiment analysis information is added performs better compared to its counterpart without sentiment analysis information. Furthermore, all models supplemented with sentiment analysis information, except for the Bag of Words models, perform better compared to the baseline models which only use sentiment analysis information as the predictor variable. Both of these statements are evident because of lower MSE and higher correlation scores.

9 Discussion

9.1 Discussion of the Results

When combining the variance analysis of the 8 dimensions and how they were correlated with positive sentiment, we notice specifically that there tends to be more agreement between annotators on dimensions which are correlated to negative sentiment. Some dimensions that had a relatively strong correlation with positive sentiment showed more mid-level agreement. An exception is 'euphoria' which shows a greater degree of variance. A possible explanation for this based on inspection

of the data is that some experiences started positively with euphoric characteristics but were later overshadowed by different, sometimes even negative effects. The annotation form questions don't state how to annotate in such cases. For this reason, annotators might have chosen different ways to resolve this which may have led to less agreement.

Combining the dimension correlation analysis with correlation to positive sentiment, we notice that generally, dimensions with correlation to negative sentiment appear to be slightly positively correlated to each other. This also holds for the dimensions with positive sentiment correlation. However, there are some exceptions. An example is 'euphoria' which shows a small negative correlation to 'transcendental' and 'ego dissolution'. The most notable correlation is between 'ego dissolution' and 'transcendental' showing that these dimensions appear to go hand in hand. It is possible that on the one hand, a strong ego dissolution effect tends to be experienced as a transcendental experience, while on the other hand, the absence of ego dissolution is an indicator for reports which lack transcendental elements.

The sentiment analysis correlation with each of the 8 dimensions appears to follow our intuitive ideas about their sentiment. This confirms our expectation that the correlation of the dimensions follows the sentiment inherent in the formulation of the dimension. However, it should be noted that the correlations coefficients show only moderate scores. This could be a result of the more complex and multipolar character of sentiment in the reports which is lost by averaging sentiment over all sentences. Many reports show both elements of positive and negative sentiment which produces more noise on the correlation coefficients. Furthermore, after reading and analysing the reports, it stood out that in a large portion of the reports the experience cannot be characterised by the simple dichotomy of positive and negative sentiment. Even though there is often some form of sentiment present, it is usually more nuanced and complex. One explanation for this is that the studied type of reports isn't necessarily aimed at evaluating the described experience. Another explanation is that users have come to new insights because of the experience which allows them to reflect on matters from a new perspective. Since psychedelics are known to alter perception, things that are typically perceived as negative could now be perceived as positive or vice versa.

When looking at the performance difference between the different models we notice that all 3 of the main hypotheses are confirmed. However, the limitation of using small data sets for training (100 examples) and testing (20 examples) produces uncertainty in the results that should be taken into account. Nonetheless, the fact that all models show improved performance over the baseline null models suggests that it is possible to make better than chance predictions regarding dimensions that quantify inner experiences. The models which used the sophisticated report representations Word2Vec and Doc2Vec improved performance, as indicated by both metrics, over the standard Bag of Words approach. This is in line with the findings of Shao et al. (2018). The models which used Sentiment Analysis information improved performance in all cases except for the Bag of Words models. One explanation for this is that the high dimensionality of this representation makes it very difficult for the Sentiment Analysis feature to make an impact on the predictions. The other report representations were lower-dimensional and didn't have this problem.

The difference in performance between logistic and linear regression can be easily explained by understanding the mechanics of these regression techniques. Linear regression assumes the data is approximately cardinal and fits the training data in such a way as to minimise MSE. This is reflected in lower MSE on the test data. Logistic regression predicts the class with the highest probability for each example in the test data. The one-versus-rest scheme that was used builds

separate models to distinguish each class from the remaining classes. Despite its effort to minimise error in the described way, this technique has no explicit incentive to minimise MSE. It doesn't understand that predicting a score of 5 that should have been a 1 should yield greater error compared to predicting a score of 2.

With only one predictor variable in the Sentiment Analysis models, logistic regression allows for more flexible predictions compared to linear regression due to the possibility of unequal placements of thresholds. The relation between Sentiment Analysis scores and the predicted values may be non-linear. Combined with linear regression's incentive to minimise MSE this may explain why linear regression models coupled only with Sentiment Analysis information perform better when looking at MSE but worse when looking at correlation with the true labels.

9.2 Limitations of Retrospective Reports

There are some inherent limitations in using retrospectively written reports. As opposed to analysing direct speech as in Bedi et al. (2014), the analysed reports in this investigation might have been influenced by inaccurate recall of users. It is unknown how long after the experience the reports were written. Furthermore, since the experiences were not blinded they are open to expectancy effects. Also, it cannot be guaranteed that the reports are truthful. In quite a few reports multiple substances are used at the same time which leads to the possibility of 'contaminated' experiences.

Since the reports are publicly available, it might be possible that people are very specific in deciding what they share. Even though anonymous submission is possible it might be easier to talk about positive experiences compared to difficult experiences. Furthermore, very personal experiences are less likely to be shared. In contrast to this, people who had very difficult experiences might be more likely to share their story to process the experience. In summary, there is a certain bias in only analysing online reports.

On the contrary, an important advantage of using free-form reports as opposed to self-report measures is that they don't limit the user in expressing their experience. However, the downside to this is that it becomes more difficult to automatically analyse the reports.

9.3 Suggested Improvements

The most important improvement to the annotation process would be to obtain scores of multiple annotators for each report. This way biases are more likely to cancel out. Additionally, it would be preferable that annotation is done by trained psychologists who have a better understanding of the subjective inner experience and its descriptive language. There are a few options for combining the scores of multiple annotators to infer a good estimate of ground truth. The most basic option is to take the mean score. However, strictly speaking, this violates the ordinal character of the data. Another option is to calculate the median value to better bypass the effect of outlier scores, or use the mode which will use the majority's opinion for each report. A comparison of more sophisticated methods for combining scores of multiple annotators can be found in Lakshminarayanan and Teh (2013).

Another improvement to the annotation process lies in better substantiation and formulation of the dimensions. Instead of one sentence formulations of the dimensions, annotators could be supplied

with a more detailed description of the characteristics of each dimension based on the relevant literature. Another option which could improve agreement between annotators could be to use more binary dimensions alongside the Likert scales, e.g. 'Did the user experience anxiety?'. For each dimension, it should be figured out separately whether it is best to use a Likert scale or a binary option. The number of points on the Likert scale can also be altered from the standard 5 point scale. Using a 3 point scale might increase agreement between annotators over a 5 point scale.

Due to the small data set size, it was impossible to analyse the performance of the different models on specific dimensions. Performance metrics of the different individual dimensions varied inconsistently between the different models. It was quite common that a different report representation resulted in a big improvement in one dimension while giving up performance on other dimensions. Taking the mean metric scores over all 8 dimensions made them more robust to such fluctuations. Using bigger data sets with dimensions that have more agreement might dampen these fluctuations. This could allow for studying the individual difficulty of predicting each dimension's scores.

9.4 Future Research

Future research could focus on a few directions.

The first direction is to improve upon the report representation techniques and use more sophisticated Machine Learning techniques. The present study used very basic techniques as a starting point but these can be expanded upon further.

Additionally, to streamline the annotation process, further research could focus on developing methods for extracting the relevant parts in reports that contain information about the inner experience. This also includes automatically detecting and disregarding reports that don't contain sufficient relevant information about the inner experience.

Another interesting direction would be to add more dimensions that capture the set and setting of an experience. This way models can be trained to extract this information from a report. In general, more variables can be identified and added to the annotation process to investigate how they influence the inner experience. This can lead to a broader and more robust understanding of the psychological inner effects during psychoactive experiences and how sensitive they are to different variables.

If the first steps in the computational processing of psychoactive reports as described in the present research can be developed further, a better understanding of the differences between psychoactive substance experiences can be gained. Developing and using larger and more reliable data sets for the training of models could result in a data-driven approach to studying the psychological effects of psychoactive substances. This can potentially be used to identify which substances might have therapeutic values in the treatment of various mental disorders.

10 Conclusion

In conclusion, this research presents an approach to investigating subjective inner qualities of psychoactive experiences as documented in online reports. The most important finding is that Machine Learning techniques can make better predictions about dimensions relating to inner experiences

compared to baselines. Models that adhere to word order and use Word2Vec/Doc2Vec techniques make better predictions compared to simple BoW models. Adding Sentiment Analysis information to the models improves performance further. The approach detailed in this paper appears to be promising for the investigation of the inner experiences in psychoactive reports. However, challenges regarding subjectivity, dimension selection and the annotation process have to be overcome first before attempting to improve the models. Further improving this line of research might lead to a better understanding of the psychological effects of various psychoactive substances. This better understanding could have a wide range of applications. One example is the implementation of initial safety assessments based on narrative reports of new psychoactive substances. Furthermore, the psychological effects that are analysed from a report can be correlated with neuroscientific data on psychoactive experiences. Finally, when the general psychological effects that are produced by specific psychoactive substances are better understood, it could lead to a better understanding of their potential use in the treatment of mental disorders.

References

S. Baccianella, A. Esuli, and F. Sebastiani. Evaluation measures for ordinal regression. pages 283–287, 01 2009. doi: 10.1109/ISDA.2009.230.

G. Bedi, G. A. Cecchi, D. F. Slezak, F. Carrillo, M. Sigman, and H. de Wit. A window into the intoxicated mind? speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, 39 (10):2340–2348, 2014. doi: 10.1038/npp.2014.80. URL `https://doi.org/10.1038/npp.2014.80`.

R. L. Carhart-Harris and G. M. Goodwin. The therapeutic potential of psychedelic drugs: Past, present, and future. *Neuropsychopharmacology*, 42(11):2105–2113, 2017. doi: 10.1038/npp.2017.84. URL `https://doi.org/10.1038/npp.2017.84`.

J. R. Coyle, D. E. Presti, and M. J. Baggott. Quantitative analysis of narrative reports of psychedelic drugs. *arXiv preprint arXiv:1206.0312*, 2012.

Department of Health Australia. *Public Health Bush Book. Volume 2*. Public Health Strategy Unit Department of Health and Community Services, 2005. URL `https://hdl.handle.net/10137/7207`.

Erowid Experience Vaults. Erowid experience vaults. URL `https://erowid.org/experiences/`.

Google Model. Pre-trained word2vec model. URL `https://github.com/eyaler/word2vec-slim`.

R. Iliev, M. Dehghani, and E. Sagi. Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition*, 7(2):265–290, 2015. doi: 10.1017/langcog.2014.30.

H. K. Kim, H. Kim, and S. Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336 – 352, 2017. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2017.05.046. URL `http://www.sciencedirect.com/science/article/pii/S0925231217308962`.

B. Lakshminarayanan and Y. Teh. Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. 04 2013.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v32/le14.html`.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

T. Mikolov, G. Corrado, K. Chen, and J. Dean. Efficient estimation of word representations in vector space. pages 1–12, 01 2013.

M. C. Mithoefer, A. T. Mithoefer, A. A. Feduccia, L. Jerome, M. Wagner, J. Wymer, J. Holland, S. Hamilton, B. Yazar-Klosinski, A. Emerson, and R. Doblin. 3,4-methylenedioxymethamphetamine (mdma)-assisted psychotherapy for post-traumatic stress dis-

order in military veterans, firefighters, and police officers: a randomised, double-blind, dose-response, phase 2 clinical trial. *The Lancet Psychiatry*, 5(6):486–497, 2020/06/30 2018. doi: 10.1016/S2215-0366(18)30135-4. URL https://doi.org/10.1016/S2215-0366(18)30135-4.

M. V. Mäntylä, D. Graziotin, and M. Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16 – 32, 2018. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2017.10.002. URL http://www.sciencedirect.com/science/article/pii/S1574013717300606.

T. E. Oxman, S. D. Rosenberg, P. P. Schnurr, G. J. Tucker, and G. Gala. The language of altered states. *The Journal of Nervous and Mental Disease*, 176(7), 1988. URL https://journals.lww.com/jonmd/Fulltext/1988/07000/The_Language_of_Altered_States.2.aspx.

R. E. Schultes and A. Hofmann. Plants of the gods : origins of hallucinogenic use. 1980.

Y. Shao, S. Taylor, N. Marshall, C. Morioka, and Q. Zeng-Treitler. Clinical text classification with word embedding features vs. bag-of-words features. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2874–2878, 2018.

Stanford Sentiment Treebank. Stanford sentiment treebank. URL https://nlp.stanford.edu/sentiment/treebank.html.

C. Strapparava and R. Mihalcea. A computational analysis of the language of drug addiction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 136–142, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2022.

E. Studerus, A. Gamma, and F. X. Vollenweider. Psychometric evaluation of the altered states of consciousness rating scale (oav). *PLOS ONE*, 5(8):1–19, 08 2010. doi: 10.1371/journal.pone.0012412. URL https://doi.org/10.1371/journal.pone.0012412.

L. Swanson. Unifying theories of psychedelic drug effects. *Frontiers in Pharmacology*, 9, 2018.

A. Szabo. Psychedelics and immunomodulation: Novel approaches and therapeutic opportunities. *Frontiers in Immunology*, 6:358, 2015. ISSN 1664-3224. doi: 10.3389/fimmu.2015.00358. URL https://www.frontiersin.org/article/10.3389/fimmu.2015.00358.

K. Thomas, B. Malcolm, and D. Lastra. Psilocybin-assisted therapy: A review of a novel treatment for psychiatric disorders. *Journal of Psychoactive Drugs*, 49(5):446–455, 2017. doi: 10.1080/02791072.2017.1320734. URL https://doi.org/10.1080/02791072.2017.1320734. PMID: 28481178.

M. Winkelman. Psychedelics as medicines for substance abuse rehabilitation: Evaluating treatments with lsd, peyote, ibogaine and ayahuasca. *Current drug abuse reviews*, 7, 01 2015. doi: 10.2174/1874473708666150107120011.

V. Yogarajan, H. Gouk, T. Smith, M. Mayo, and B. Pfahringer. Comparing high dimensional word embeddings trained on medical text to bag-of-words for predicting medical codes. In N. T. Nguyen, K. Jearanaitanakij, A. Selamat, B. Trawiński, and S. Chittayasothorn, editors, *Intelligent Information and Database Systems*, pages 97–108, Cham, 2020. Springer International Publishing. ISBN 978-3-030-41964-6.

R. Zhao and K. Mao. Fuzzy bag-of-words model for document representation. *IEEE Transactions on Fuzzy Systems*, 26(2):794–804, 2018.