Das et al. *EPJ Data Science* (2025) 14:40
https://doi.org/10.1140/epjds/s13688-025-00554-w
EPJ Data Science
a SpringerOpen Journal
RESEARCH
Open Access

Dream content discovery from social media using natural language processing

Anubhab Das[1], Sanja Šćepanović[2,3*], Luca Maria Aiello[4,5], Remington Mallett[6], Deirdre Barrett[7] and Daniele Quercia[2,8]

Handling Editor: Robert West

*Correspondence:
sanja.scepanovic@nokia-bell-labs.com
[2]Nokia Bell Labs, Broers Building, Cambridge, CB3 0FA, United Kingdom
[3]Uehiro Oxford Institute, University of Oxford, St Ebbe's Street, Oxford, OX1 1PT, United Kingdom
Full list of author information is available at the end of the article

Abstract

Dreaming is a fundamental but not fully understood part of human experience. Traditional dream content analysis practices, while popular and aided by over 130 unique scales and rating systems, have limitations. Often based on retrospective surveys or lab studies, and sometimes on in-home dream reports collected over some days, they struggle to be applied on a large scale or to show the importance and connections between different dream themes. To overcome these issues, we conducted data-driven mixed-method analysis identifying topics in free-form dream reports through natural language processing. We applied this analysis on 44,213 dream reports from Reddit's `r/Dreams` subreddit, where we uncovered 217 topics, grouped into 22 larger themes: the most extensive collection of dream topics to date. We validated our topics by comparing it to the widely-used Hall and van de Castle scale. Going beyond traditional scales, our method can find unique patterns in different dream types (like nightmares or recurring dreams), understand topic importance and connections (like finding a greater predominance of indoor location settings in Reddit dreams than what was in general stipulated by previous work), and observe changes in collective dream experiences over time and around major events (like the COVID-19 pandemic and the recent Russo-Ukrainian war). We envision that the applications of our method will provide valuable insights into the complex nature of dreaming and its interplay with our waking experiences.

Keywords: Dreams; Dream content analysis; Neural topic modeling; Reddit

1 Introduction

Dreaming is a fundamental human experience and a cornerstone of sleep psychology, yet its underlying mechanisms remain elusive. The fascination with the contents and meaning of our dreams dates back to early human civilizations [1], but despite significant progress in dream research, fundamental questions about the physiological and psychological functions of dreaming remain unanswered, leaving us to ponder the question: *why do we dream* [2]? One step closer towards answering this question is to understand the nature of *what we dream*. This question is important not only in that it could help us understand the fundamental function of dreams but also as it offers a window into our psyche and what is prominent in people's minds in a given time. Dream content is composed of fragments of waking life experiences and events [3], but these are not veridical replays [4, 5].

One popular approach to investigating dream content systematically is through *content analysis* [3]. This is a family of methods that analyze quantitatively the elements present in dreams to answer specific questions, such as whether depressed individuals experience more rejection in their dreams [6], how dream content changes from teenage years to adulthood [7], or whether in the times of collective health crises there is a shift in the medical symptoms people dream about [8]. Although these studies may seem narrow in scope, they collectively provide a critical foundation for addressing the overarching question of why we dream. The importance of content analysis is evidenced by the development of over 130 scales and rating systems for dream content analysis [9]. Early scales tended to score based on the raters' subjective interpretation of dream symbolism and rarely reported inter-rater reliability [10]. Dream research became more systematic with the development of the *Hall and van de Castle method* [10], a quantitative system that scores dream reports based on the frequency and type of characters, interactions, activities, emotions, settings, and objects present in the dream. This method relies solely on the dream reports and does not use any additional information provided by the dreamer. Studies using the Hall and van de Castle scales have revealed consistent patterns and norms in dream content across different groups of people, such as gender, age, culture, and personality. For instance, women tend to dream more about family members, friends, and indoor settings, while men tend to dream more about strangers, violence, and outdoor settings. Moreover, research using these scales has demonstrated that dream content correlates with an individual's waking concerns and interests, such as work, relationships, hobbies, and fears [7].

1.1 Limited scope and representativeness of existing dream content scales

Traditionally, dream researchers had to manually sift through large numbers of dream reports to gain insights into the range of dream topics [10]. Even with recent developments in automated dream analysis [11], most studies continued to rely on experimenter-driven content searches, i.e., supervised approaches for content analysis, such as the Hall and van de Castle method. These methods involve the use of predetermined categories that are often biased towards existing knowledge of dreams. On the other hand, dreams are often characterized as bizarre, involving impossible or improbable events that deviate from everyday experiences [12], which may not fit within the predetermined categories established in the literature. As a result, current approaches to dream analysis may miss important aspects of dream content that fall outside of these predetermined categories. In contrast to traditional methods of dream coding, *unsupervised theme discovery* from dream reports may provide a fresh perspective and a more comprehensive understanding of the categorization of dream content and its relationship with waking life events. Furthermore, previous studies predominantly relied on dream reports collected through retrospective surveys that are susceptible to memory biases, and laboratory studies that may be confounded by the strong impact of the laboratory setting on dream content [13]. Those studies that proactively collected in-home dream reports over a period of time are fewer and are limited by the number of dream reports that could be collected in such a way. Therefore, a more *ecologically valid approach* to studying dream content at the population level is necessary.

1.2 Unsupervised mixed-method analysis for dream content discovery

Our study addresses the identified research gap (i.e., limited scope and representativeness of dream content scales) by demonstrating an application of an unsupervised mixed-method approach for dream content discovery to new large-scale data that more closely approximates spontaneous dream recollections than most previous studies. To achieve this, we *i)* leveraged recent advances in AI for Natural Language Processing (NLP) and *ii)* used a crowd-sourced dataset of dream self-reports from the `r/Dreams` community on Reddit. Instead of pre-determining the concepts of interest, this method discovers topics emerging from the dream reports. Dream experiences shared on `r/Dreams` are reported by a various set of users when they felt the need to do so and not as part of a predetermined study plan, enabling us to collect a large set of dream reports and conduct an ecological study. We collected over 44K dream reports from more than 34K Reddit users over the past five years, and applied the BERTopic method [14] to automatically discover topics in each dream report. We then built a taxonomy revealing 22 *themes* that can be broken down into 217 specific *topics*. Confirming its validity, we found that most of the themes in our taxonomy align with the dream element categories present in the Hall and Van de Castle scale, but the specific topics inside those themes provide a description of dreams that is much more detailed. Going beyond what was possible with existing scales, our method also allowed to uncover the prevalence, importance, and relationships among specific topics and themes.

To demonstrate the applications enabled by this method, we used the metadata from the `r/Dreams` community, and classified dreams into four types: *nightmares, recurring dreams, vivid dreams*, and *lucid dreams*. Our analysis revealed that each type of dream has distinct characteristics and prominent topics. Notably, nightmares were associated with *scary* and *shadowy* imagery, and *sexual-assault* scenes. Vivid dreams featured rich expressions of *feelings* and topics inducing extreme emotions such as *pregnancy and birth*, *religious* figures, *war*, and *aliens*. Lucid dreams were characterized by topics of *control*, and by an overarching theme of *mental reflections and interactions*. For recurring dreams, we found that the most salient topics were *dating*, *sex*, and *cheating*, with recurring themes related to *school* and mentions of parts of the *human body*. Additionally, we investigated the relationship between dream topics and real-life experiences by studying the evolution of topics over the past five years. Our findings showed that the COVID-19 outbreak coincided with a gradual and collective shift in dream content. `r/Dreams` users started to dream less about *people and relationships*, *feelings*, *sight and vision*, *outdoor locations*, and *movement and action*, and more about *the human body, especially teeth and blood*, *violence and death*, *religious and spiritual* themes, and *indoor locations*. Similarly, after the war in Ukraine started, the topics about *soldiers* and *nuclear war* both peaked.

2 Results

Figure 1 outlines the framework of our study. It consists of the following stages: 1) *data preprocessing*, which includes collecting dream reports, ideally in an ecological setting, and using NLP methods for cleaning the content; 2) *topic modelling*, a three-step process, which is where we automatically discover topics and themes; and tag the dream reports with these labels. Once this stage is completed, various applications are supported, and we demonstrate three of those: 3) *building a dream topics taxonomy*, which allows to uncover the relationships between individual themes and topics, as well as the frequency of each

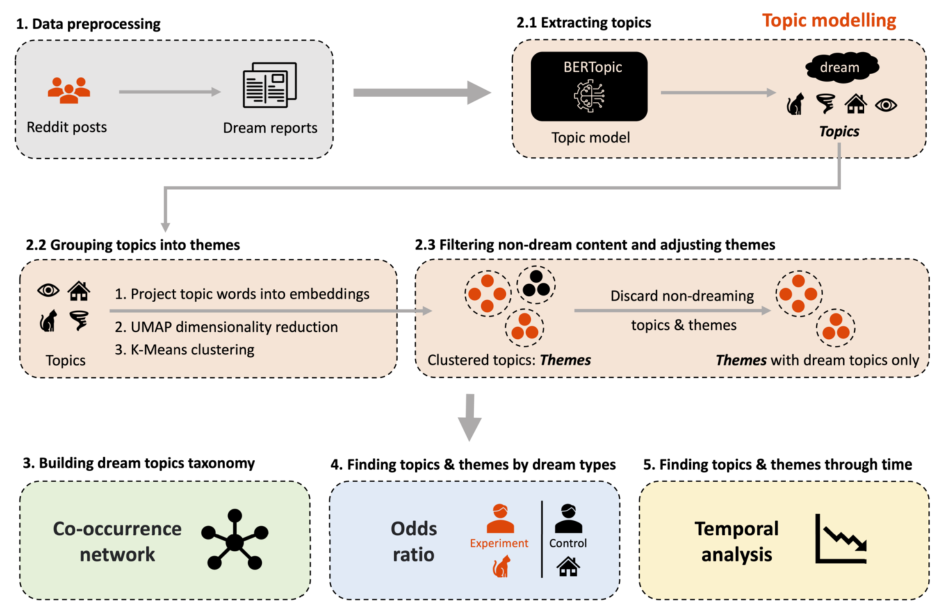

Figure 1 The framework of our study. *Stage 1.* involves data curation. *Stage 2.* consists of three steps that are the core of our proposed unsupervised mixed-method approach for dream content discovery: *Step 2.1* is topic extraction using unsupervised BERTopic method, *Step 2.2.* is grouping similar topics into themes, and *Step 2.3* is a semi-manual filtering and adjustment step to preserve only relevant dream-content. *Stages 3.-5.* demonstrate various applications enabled by our methodology, such as *3.* uncovering the topic taxonomy in a particular dream report collection, *4.* finding topics and themes that are specific to a given subset of dream reports in the given collection, and *5.* tracking topic and theme trends through time

of them in the dream collection; 4) *finding topics and themes by dream types*, which uses the proposed measure of topic or theme importance in a dream and odds-ratio analysis to discover topics that are specific to a dream type (or any other dream reports subset of choosing), and 5) *finding topics and themes through time*, which uses the proposed topic or theme importance in a dream to quantify the prevalence of dream topics and themes through time.

2.1 Reddit dream reports

In the established literature of dream analysis, dream reports are defined as *"the recollection of mental activity which has occurred during sleep"* [2]. To gather a dataset of such reports, we turned to Reddit, a social media platform organized in communities known as *subreddits*. Using the PushShift API [15], we collected data from `r/Dreams` — a subreddit where members share their dreams and engage informally in their interpretation, which is common in therapy contexts [16], as well as in providing social and emotional support to other members.

The `r/Dreams` subreddit was established in September 2008, and as of June 2022, it had accumulated 280K subscribers. We collected over 185K posts published on `r/Dreams` from March 2016 to September 2022; prior to 2016, the community was almost inactive (Fig. 2 (a)). Authors on `r/Dreams` annotate their posts with one or more tags selected from a fixed set of community-specific labels called *flairs*. Flairs on `r/Dreams` denote posts that contain dream reports of a given type (*short dream*, *medium dream*, *long dream*, *nightmare*, and *recurring dream*) or posts that contain discussions about dreams in general (e.g., *dream help*, *dream art*, or *question*). To ensure that we considered only posts

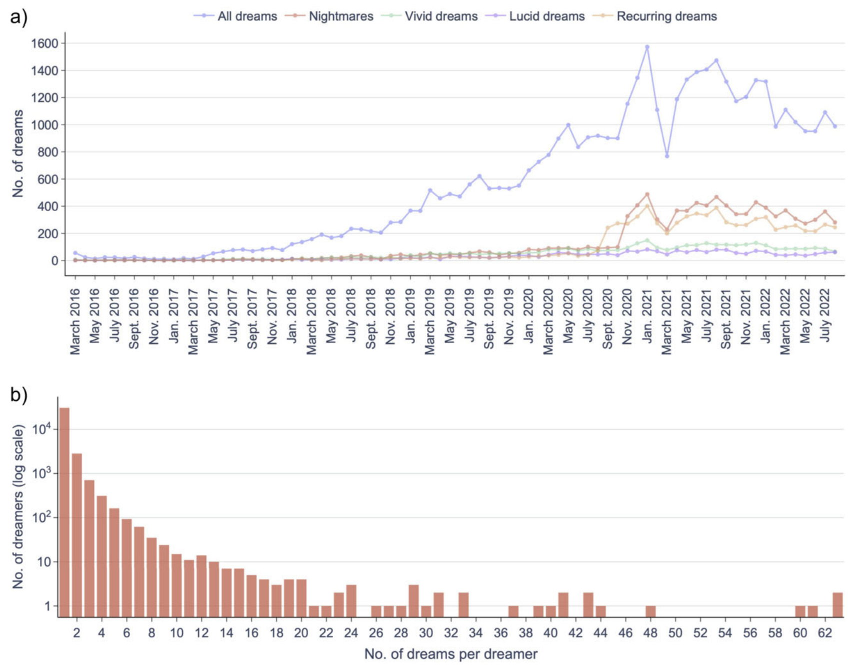

Figure 2 Reddit dream reports statistics: (a) number of reports of different types through time, and (b) number of reports per dreamer

Table 1 Dream reports statistics for the data from March 2016 to September 2022

Dream type	All dreams	Vivid dreams	Nightmares	Lucid dreams	Recurring dreams	Other dreams
# dream reports	44,213	3962	9823	2272	7597	24,673
# users	35,006	3819	8959	2107	7201	19,466
Mean # sentences	17 ± 18	25 ± 23	20 ± 20	27 ± 27	14 ± 16	16 ± 16
Mean # words	290 ± 285	434 ± 382	341 ± 323	465 ± 424	253 ± 256	263 ± 255
Mean # characters	1455 ± 1444	2192 ± 1951	1713 ± 1640	2343 ± 2152	1274 ± 1299	1319 ± 1289
Median # sentences	12	18	14	19	10	11
Median # words	205	325	252	338	178	186
Median # characters	1023	1636	1258	1697	895	930

that contained dream reports, we only kept the 44,213 unique posts tagged with dream-type flairs (Table 1). Different dream categories are not mutually exclusive, but they can overlap (Figure S1). Out of our 44,213 dream reports; 24,673 do not belong to any of the aforementioned categories (*other dreams* in Table 1), and the remaining ones belong to at least one. In our analysis, a *dream report* was the concatenation of the title and body of each of the Reddit posts.

2.2 Unsupervised mixed-method dream content discovery

The core part of our framework shown in Fig. 1 is the unsupervised mixed-method *topic modelling* approach and it involves: 2.1) *extracting topics* using an advanced NLP topic model, such as BERTopic [17]; 2.2) *grouping topics into themes* using a clustering method to group embedding representations of individual topics discovered in the previous step into themes, and 2.3) *filtering non-dream content and adjusting themes*, which is a mixed-method part of the proposed methodology, requiring human, ideally dream experts knowledge to filter out topics or themes that do not pertain to the actual dream

Table 2 Top 20 topics by the number of dream reports in which they are found. The number of dreamers who dreamt of these topics is also shown

Topic	# dream reports	# dreamers
lady, face, looked, head	13,020	11,356
dream, girl-dream, dreamt, have	6785	6337
lights, sun, pitch-black, wearing	4950	4488
mall, restaurant, eating, ice-cream	4660	4114
bus, driving, cars, train-station	4291	3966
ex, years, dating, talk	3786	3590
kitten, lion, birds, owl	3281	3060
dreams-mean, staring, talk, people-dream	3194	3018
beach, ocean, swim, river	3119	2900
doors, house, rooms, mansion	2984	2741
death, died-dream, going-die, death-dream	2825	2711
classroom, teachers, principal, student	2769	2565
dream-dad, dad-dream, dream-mom, mom-dream	2327	2197
mum, aunt, mom-came, called-mom	1893	1793
felt-real, dream-felt, real-dream, real-like	1840	1769
dream-ends, story, endings, recurring-theme	1677	1584
know-make, sure-dont, know-think, know-sure	1649	1580
school-dream, dream-school, college, dream-high	1632	1553
creature, bite, mouth, face	1621	1555
path, garden, hills, plants	1595	1494

content, as well as to adjust any topic to theme associations, if needed. For more details about any of the steps of our approach, please refer to the Methods section.

2.3 Taxonomy of reddit dream topics

Using the topic modeling techniques, we have identified 217 semantically-cohesive topics that emerged from the dream reports analyzed (details in Methods). We assigned to each dream report the list of unique topics that our model was able to extract from the report text. The distribution of number of dreams associated with a topic is broad (Figure S5), with only 38 topics being associated with at least 1000 dream reports. Table 2 shows the 20 most frequent topics.

To offer a more concise representation of dream topics, we automatically clustered the 217 fine-grained topics into 22 *themes* (see Methods), which we then manually parsed to assess their conformity to categories from the dream coding system by Hall and Van de Castle (*HVdC*). *HVdC* defines 12 categories and several subcategories of elements empirically relevant for quantitative dream analysis. All themes but one (a miscellaneous theme containing diverse topics) matched some *HVdC* category or subcategory (Table 3).

At a high level of abstraction, *HVdC* views the dream as *(i)* a cast of characters, *(ii)* interacting with each other, while being immersed in *(iii)* some background setting. These three aspects emerged in the most prominent themes extracted from Reddit dream reports. The largest theme, both in terms of the number of topics it contains (17) and the number of dream reports associated with it (20K), is *People and relationships*. The main topics included in this theme are *family* members and relationships, and *intimacy and romance* (see Table 3 and Table S1). Characters and interactions are also represented in themes concerning *Animals*, *Supernatural entities*, and *Religious and Spiritual*, which map directly to the two HVdC subcategories of Animals and Imaginary Characters. Interactions respectively of aggressive and friendly type are represented in the themes *Violence and Death* and *Life Events*. The second most frequent group of themes involves events or elements that are typical of well-characterized places, such as home or the workplace. Among them, the most prominent theme is *Indoor locations*, which includes 16 topics such as *house*, *hospital*, and *mall*. These themes map to different subcategories of the *HVdC* category of Settings.

The remaining themes correspond to the categories of Activities, Objects, Emotions, and Time in the *HVdC* framework. With respect to Activities, Reddit users more frequently reported their mental activities and perceptions, rather than physical actions, when recounting their dreams. Specifically, visual and auditory sensory experiences were often recounted. The most commonly described categories of Objects included body parts (often with gruesome details) and personal objects, with a particular emphasis on technological devices such as phones or elements from virtual worlds of computer games. Emotions were represented by a single theme, characterized by common formulas for describing emotions of any type, with negative emotions being more prominently featured. Finally, while one theme captured temporal scales, such references were infrequent, appearing in fewer than 1000 dreams.

Some of the *HVdC* categories did not map directly to any of the themes in our taxonomy. This is the case with *Misfortune/Fortune*, for example. We found that these types of events are usually reported as elements in dreams that were predominantly characterised by other themes, such as *Life events*, *People and relationships*, or *Violence and death*, for example.

The emerging themes in the dream reports do not exist in isolation; they often co-occur in the same reports and jointly construct their narratives. Figure 3 presents the backbone of the network of co-occurrence of these themes, where the connections between themes are weighted proportionally to the number of dreams in which they co-occur. *People and relationships* and *Indoor locations* are the most frequent and central themes, occurring often with many other ones. Feelings are mainly associated with actions and relationships rather than objects or settings; conversely, sensory elements are more associated with settings, especially outdoor locations. Some of the central topics, such as relationships and emotions, are highly valued in rating scales by dream researchers. However, other themes, such as indoor/outdoor settings, are frequently omitted in *HVdC* coding research.

Besides the prominence of both types of settings, their proportions differ somewhat. Our data show a greater predominance of indoor settings, while *HVdC* norms indicate lower levels of indoor settings [10]. This difference could reflect our data, which includes two years of the COVID pandemic, whereas the comparison *HVdC* studies do not.

Our findings challenge traditional dream analysis literature that found males mostly dream of male characters [18]. The two most recurring topics in our data concern the presence of female characters in dreams, despite the young male dominance in the Reddit user base [19]. A possible reason for this contradicting finding could have to do with *HVdC* ratings not being exactly equivalent to ours. In our method, the strength of association of a dream with topics concerning male characters and indoor settings is stronger the more mentions of male characters and indoor settings appear in the report. Conversely, in *HVdC*, a male character who is mentioned only once in the dream gets the same score as one who is referred to in every sentence of the report. Likewise, an indoor setting is scored in the same way regardless whether it is inferred from one mention in the dream or whether the dream account is largely a description of an indoor setting. In that sense, our methods allows for a richer characterisation and *quantification* of dream content.

Table 3 Dream themes we identified and their example topics linked to the corresponded Hall/Van de Castle categories. The themes are grouped based on the categories to which they were associated (e.g., People and relationships, Animals, Supernatural entities, and Religious and spiritual themes were linked to categories within Characters and Imaginary Objects)

Theme	Top-3 topics under this theme	# topics	# dreams	# dreamers	Hall/Van de Castle cat.
People and relationships	lady, face, looked, head dream, girl-dream, dreamt, have ex, years, dating, talk	17	20,786	17,798	Characters; Social Interactions
Animals	kitten, lion, birds, owl spider, maggots, batman, thanos snakes, alligator, turtle, bite	13	5546	5048	Characters -> Animals
Supernatural entities	creature, bite, mouth, face shadows, shadowy-figure, dark-figure zombies, zombie-apocalypse, outbreak	11	3436	3173	Characters -> Imaginary
Religious and spiritual	demons, devil, monster, demonic church, cult, lot-people, dream-god demon, devil, demons, angel	7	2859	2697	Characters -> Imaginary; Objects -> Architecture
Indoor locations	mall, restaurant, eating, ice-cream bus, driving, cars, train-station doors, house, rooms, mansion	16	13,890	11,868	Settings -> Location -> Indoor
Outdoor locations	beach, ocean, swim, river path, garden, hills, plants cave, tunnels, underground, tower	6	5020	4539	Settings -> Location -> Outdoor
School	classroom, teachers, principal, student school-dream, dream-school, college tests, failed, exam, professor	3	3639	3352	Settings -> Location
Work	new-job, boss, jobs, shift	1	1077	1024	Settings -> Familiar
Weather, esp. storms	rain, tornado, hurricane, started-raining snow, snowing, cold, winter tornado, storms, april, category	3	1037	985	Settings
Mental reflection and interactions	know-make, sure-dont, know-think yeah, like-wtf, hell, like-fuck asked-doing, tell, help, begged	18	6808	6157	Activities -> Thinking
Sights and vision	lights, sun, pitch-black, wearing reflection, looked-mirror, looking-mirror blinded, vision-blurry, blur, recite	11	6185	5548	Activities -> Visual; Auditory
Movement and action	left, wanted-leave, time-leave, home continued-walking, continue, street run, started-running, run-like, sprinting	12	3457	3208	Activities -> Movement
Sounds and lack thereof	singing, songs, lyrics, stage voices, heard-voice, hear-voice footsteps, ghost, noises, ringing	7	3125	2892	Activities -> Auditory
Violence and death	death, died-dream, going-die, death pistol, shooting, shot-head, shotgun police, officer, officers, kidnapped	15	8203	7523	Social Interactions -> Aggression
Life events	birth, babies, pregnancy, newborn party, invited, having-party, brother giving-birth, dreamt, dream, twins	6	2230	2134	Social Interactions -> Friendliness
Personal objects	dressed, mask, naked, clothing phones, ringing, check-phone, battery pages, pen, ink, letters	17	4639	4135	Objects
Media and tech	theater, anime, movie-like, tv minecraft, games, vr, game-like game-dream, dream-playing, vr, games	6	2970	2677	Objects -> Communication
Human body, esp. teeth and blood	blood, skin, humanoid, arms teeth, falling, tongue, pain teeth, tooth, falling, gums	5	2163	2055	Objects -> Body Parts
Space	sun, earth, eclipse, phases space, space-ship, nasa, oxygen meteors, earth, asteroid, coming	3	448	438	Objects -> Nature
Feelings	felt-real, dream-felt, real-dream, real woke-crying, started-crying, crying feel-pain, painful, pain-dream, felt-pain	16	6351	5907	Emotions
Time, time travel and timelines	timeline, time-travel, time-skip, like time-travel, dream-world, universes noon, 00 am, early-morning, evening	3	777	739	Descriptive Elements -> Temporal Scale
Other topics	dream-ends, story, endings, recurring huge, inches, like-size, big-small pov, 3rd-person, person-perspective	21	5475	5018	-

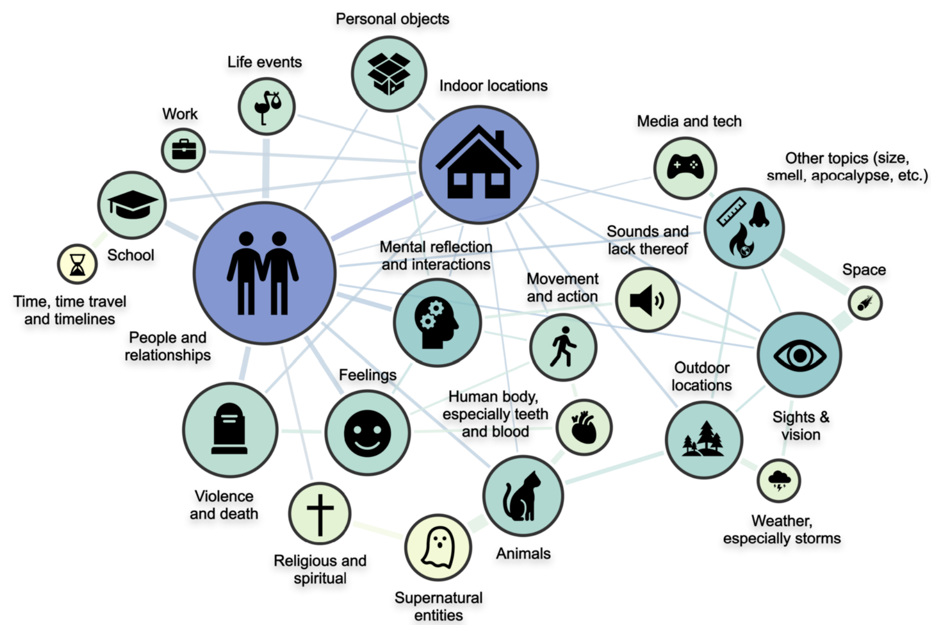

Figure 3 The taxonomy of dream themes that we automatically uncovered — *a map of dreams*. Nodes represent dream themes (i.e., groups of topics); node size corresponds to the frequency of the theme in our reports; node color represents the theme centrality in the network (the color scale goes from light yellow to green to dark blue, and the darker the color, the more central the topic); finally, edge thickness corresponds to the frequency of co-occurrence of the theme topics in the dream reports

2.4 Topics and themes by dream type

Using odds ratios (see Methods), we uncovered that certain topics (Table 4) and themes (Fig. 4) appeared more frequently in dreams of specific types.

2.4.1 Topics and themes in nightmares

Results in Table 4 revealed among the top topics specific for nightmares, keywords such as *shadows*, *rape*, *sexual-assault*, *scary*, *creepy*, *violence*, *demons*, *911*, and *blood*. In terms of themes, *Religious and spiritual* is the most prominent in nightmares, followed by *Feelings*, *Supernatural entities*, and *Violence and death*, while the least prominent ones were *Media and tech* and *Time, time travel and timelines*.

2.4.2 Topics and themes in vivid dreams

Results presented in Table 4 reveal among topics specific for vivid dreams, keywords such as *felt-real*, *religious*, *felt-right*, *felt-wrong*, *apocalyptic*, *felt-pain*, *baby birth and pregnancy*, *aliens*, *mirrors*, and *nuclear war*. This has translated into the most prominent themes being *Feelings*, followed by *Sights and vision* and *Life events*.

2.4.3 Topics and themes in lucid dreams

Results presented in Table 4 reveal among topics specific for lucid dreams, keywords such as *control*, *alternate-reality*, and *reflection and mirrors*, *felt-real*, *couldnt-speak*, *falling*, *heard-voices* and *demon*. When considering themes, *Mental reflection and interactions*, followed by *Movement and action*, *Sights and vision*, and *Feelings* were the most prominent.

Table 4 Odds-ratio (OR) analysis of individual topics in the four dream types. The OR scores are equal to the probability of the topic appearing in dreams labeled with a given dream type (e.g., nightmare), divided by the same probability calculated on dreams that are *not* labeled with that type. Scores higher/lower than one indicate that dreams of the given type feature the topic more/less often than other dreams. Unlike in Fig. 4, in which we reported OR across themes; in this table, we report OR across individual topics to finely characterize a certain dream type by a list of specific topics appearing in it

Topics	OR
Nightmares	
shadows, shadowy-figure, dark-figure, entity	3.986
rape, sexually-assaulted, abuse, nightmares	3.184
wasnt-scary, scary-just, frightening, scariest	3.017
really-creepy, weird-creepy, creepiest, sounds	2.880
tw, warning, nsfw, sexual-assault	2.799
violent, disturbing, dreams-past, weird-dreams	2.476
demons, devil, monster, demonic	2.455
escape, escaped, way-escape, managed-escape	2.421
911, dial, emergency, ambulance	2.405
dolls, porcelain, voodoo, haunted	2.267
panic, panicking, panicked, started-panic	2.26
footsteps, ghost, noises, ringing	2.235
feel-pain, painful, pain-dream, felt-pain	2.207
blood, like-blood, covered-blood, splatter	2.204
paralyzed, able, speak, couldnt-speak	2.109
lungs, couldnt-breathe, like-couldnt, heart	2.012
serial-killer, dream-killing, killed-dream, killers	1.956
blinded, vision-blurry, blur, recite	1.889
demon, devil, demons, angel	1.886
feel-right, felt-wrong, wasnt-right, feel	1.771
Lucid dreams	
control, couldnt-control, control-body, hopeless	4.119
universes, alternate-reality, different-dimension, travel	2.82
reflection, looked-mirror, looking-mirror, mirrors	2.702
felt-real, dream-felt, real-dream, real-like	2.196
paralyzed, able, speak, couldnt-speak	2.052
falling, hit-ground, fall-ground, let-fall	1.979
know, okay, yeah, know-going	1.855
voices, heard-voice, hear-voice, voice-head	1.627
dont-want, needed, doing-dont, remember-doing	1.535
demon, devil, demons, angel	1.355
flying, airport, helicopter, planes	1.355
Vivid dreams	
felt-real, dream-felt, real-dream, real-like	2.104
religious, atheist, catholic, supernatural	1.722
feel-right, felt-wrong, wasnt-right, feel	1.649
apocalyptic, dream-end, world-dreams, earth	1.614
feel-pain, painful, pain-dream, felt-pain	1.500
know-knew, knew-didnt, knew-dont, known	1.473
birth, babies, pregnancy, newborn	1.456
aliens, invasion, grey, race	1.445
path, garden, hills, plants	1.435
reflection, looked-mirror, looking-mirror, mirrors	1.419
giving-birth, dreamt, dream, twins	1.356
nuke, war, fires, missile	1.313
Recurring dreams	
cheating, having-dreams, dream-boyfriend, relationship	6.349
ex-boyfriend, dreams, having-dreams, relationship	4.021
teeth, tooth, falling, gums	3.760
school-dream, dream-school, college, dream-high	3.084
house-dream, dream-house, apartment, childhood-home	2.966
ex, years, dating, talk	2.462
dreams-mean, staring, talk, people-dream	2.087
dream-ends, story, endings, recurring-theme	2.047
dream, girl-dream, dreamt, have	1.974
apocalyptic, dream-end, world-dreams, earth	1.912
time-travel, dream-world, universes, dream-time	1.857
crush, falling-love, fell-love, fall-love	1.799
teeth, falling, tongue, pain	1.788
rape, sexually-assaulted, abuse, nightmares	1.679
paralyzed, able, speak, couldnt-speak	1.658
flying, airport, helicopter, planes	1.625
sex-dream, sexual, wet-dream, dream-sex	1.624
serial-killer, dream-killing, killed-dream, killers	1.556
room-starts, walking, family, house-starts	1.517
doors, house, rooms, mansion	1.465

2.4.4 Topics and themes in recurring dreams

Results in Table 4 revealed among the top topics specific for recurring dreams, keywords such as *cheating*, *ex*, *teeth*, *school*, *house and apartment*, *time-travel*, and *sex-dreams*. Looking at the themes, we found *School*, *Human body, especially teeth and blood*, and *Other topics (size, smell, apocalypse, etc.)* being those that characterize recurring dreams.

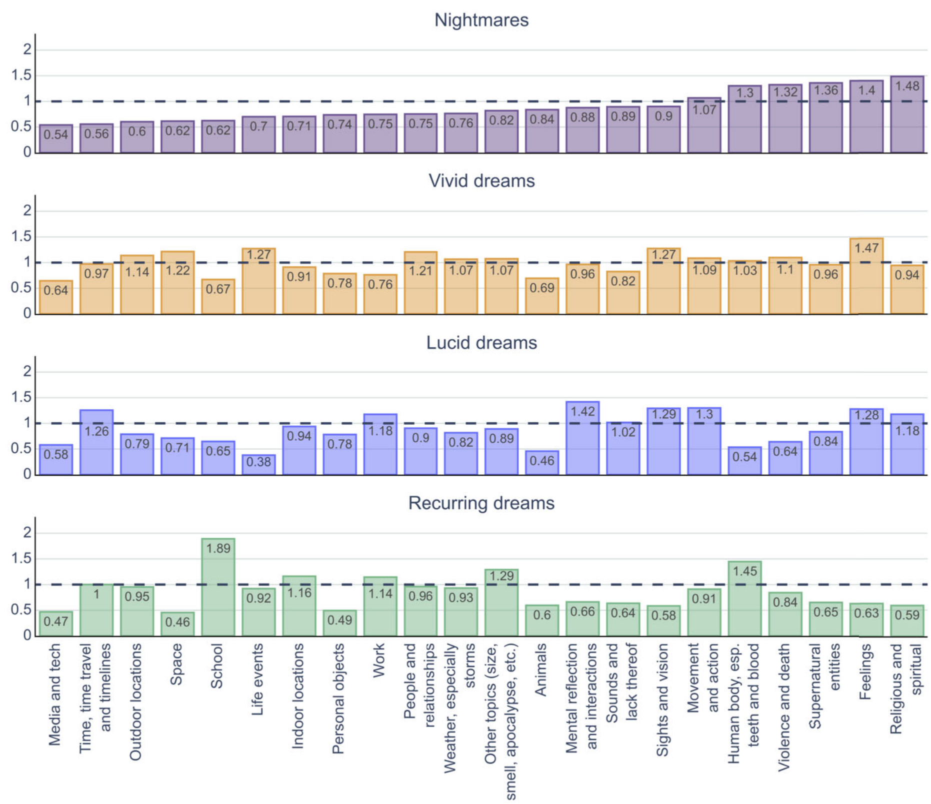

Figure 4 Odds-ratio (OR) analysis of themes in the four dream types. The OR scores (y-axis) are equal to the probability of the theme (x-axis) appearing in dreams labeled with a given dream type (e.g., nightmares), divided by the same probability calculated on dreams that are *not* labeled with that type. Scores higher/lower than one indicate that dreams of the given type feature the theme more/less often than other dreams

2.5 Topics and themes through time

Finally, we studied the evolution of topics and themes over time. We focused on the period with at least 300 monthly dream reports, namely from January 2019 to September 2022 (Fig. 2 (a)). We found that soon after COVID-19 started, there was a gradual collective shift in the content of dreams from `r/Dreams` (Fig. 5).

At the very beginning of the COVID-19 outbreak (February-March 2020), and even more so after the first peak of recorded deaths (April 2020), people gradually dreamt less of *Sight and vision*, *Outdoor locations*, *Movement and action*, and *Mental reflections and interactions*, while they dreamt more of *Religious and spiritual* figures, *Indoor locations*, and *Human body, especially teeth and blood*. An example dream from this period talks about *"spitting teeth and cornea onto the palm."* We found a sharp decrease in the frequency of *Life events* topics in February 2020, while in March 2020 there we recorded a peak of mentions of *Human body, especially teeth and blood*, which are predominantly found in nightmares, as our previous analysis showed. These trends continued throughout the time of the COVID-19 second death peak (January 2021), from when we also detect a stark decrease in dreams of other *People and relationships*, *Feelings*, and, for a while, of *Animals* and *Work*.

Finally, the start of the war in Ukraine (February 2022), is associated with a strong transition on the content of people's dreams towards violent topics. We found a sharp increase in

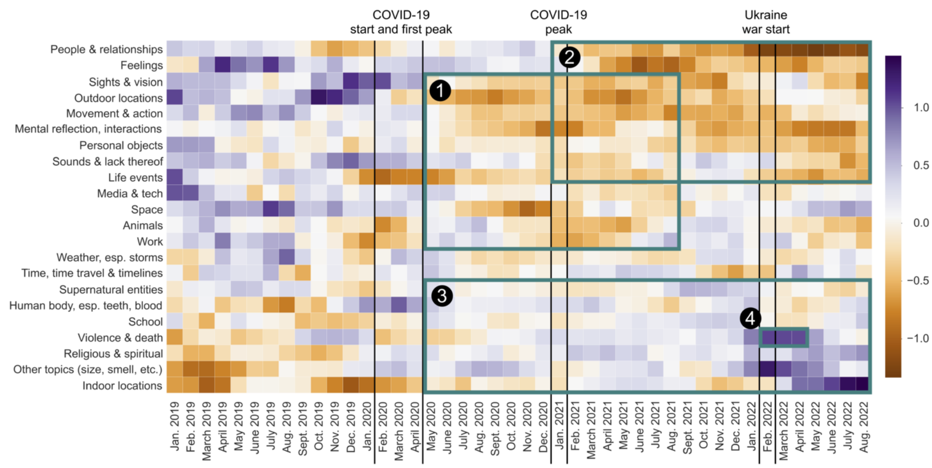

Figure 5 Temporal evolution of themes. The value of each cell is proportional to the average importance of a theme t across all the dream reports posted on a given month m ($I_{t,m}$ in Formula (4)). The importance of a theme in a dream is defined as the fraction of sentences in the dream report that belong to that theme ($I_{t,d}$ in Formula (2)). Values are standardized within each theme ($z_{I_{t,m}}$ in Formula (5)) so that positive (negative) values encode values of importance that are higher (lower) than the global average for that theme. Last, we smoothed values over a window length of 5 months to reduce noise ($\tilde{z}_{I_{t,m}}$ in Formula (6)). During period 1) themes of *Sight and vision*, *Outdoor locations*, *Movement and action*, *Mental reflections*, *Life Events*, *Space*, and *Work* plummeted, while in period 2) also the themes of *People and relationships* and *Feelings* dropped. During period 3) we see an increase in the themes such as *Supernatural entities*, *Human body, especially teeth and blood*, *Religious and spiritual*, *Indoor Locations*, and *Violence and death*; 4) the last two themes (*Violence & death* and *Other*) peaked around the time of the start of the war in Ukraine

topics from the themes of *Violence and death*, as well *Other topics (size, smell, apocalypse, etc.)*. Example dreams from the former group talk about *"being a murderer,"*,*"tooth falling out,"* and *"getting shot in the head."* Example dreams from the second group talked about *"temple spirit monster,"* and *"being attacked by an opossum."* The changes in response to external events were also evident at the level of individual topics, such as those about soldiers and nuclear war, which both peaked after the war in Ukraine started (Fig. 6). Another theme that also started increasing in frequency after the war start is *Indoor locations*.

3 Discussion

The current study advances the field of dream science by implementing a new methodology to study the content of dreams in a more objective and ecological manner and also showcasing how this method can generate new insights into dreams.

3.1 Reddit as a source of dreams

Prior dream research has relied heavily on traditional laboratory, survey, and diary methods. Laboratory studies benefit from monitoring participants with PSG and waking them directly from REM sleep [20], which is known to increase dream recall drastically [21]. However, these dreams are not representative of natural dream content, as they are highly influenced by the laboratory setting [13]. Survey studies avoid this contextual bias on dream content, but often ask of a "recent" dream, which might be days or weeks prior to the survey response, thus suffering from memory distortion. Lastly, diary studies track dream content in a participant sample through morning diaries [22]. While these studies benefit from having dream content collected from ecological settings and with less retrospective bias, these studies are often limited by small sample sizes due to the high burden

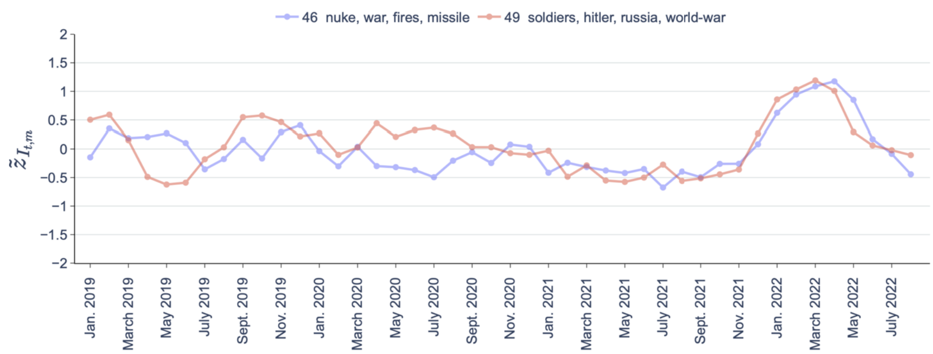

Figure 6 Temporal evolution of two selected topics: *soldiers, hitler, russia, world-war* and *nuke, war, fires, missile*. Smoothed z-scores $\tilde{z}_{I_{t,m}}$ (*y*-axis) of a topic's importance at a given month (*x*-axis) are shown for the period from January 2019 to September 2022. The values $\tilde{z}_{I_{t,m}}$ are proportional to the average importance of the topic *t* across all the dream reports posted on a given month *m* ($I_{t,m}$ in Formula (4)). The importance of a topic in a dream is defined as the fraction of sentences in the dream report that belong to that topic ($I_{t,d}$ in Formula (2)). Values are standardized within each topic ($z_{I_{t,m}}$ in Formula (5)) so that positive (negative) values encode values of importance that are higher (lower) than the global average for that topic. Last, we smoothed values over a window length of 5 months to reduce noise ($\tilde{z}_{I_{t,m}}$ in Formula (6)). Both topics peaked around the time when the conflict in Ukraine started

of participation. In the current study, we extracted morning dream reports from social media, thus capturing dreams in an ecological setting and also at a much larger scale. While other studies have investigated dedicated online dream forums [23]. Reddit is one of the most popular social media sites and its usage continues to grow at higher rates than specialized forums, suggesting that the current approach might continue to be a source of population dream content for scientific analysis.

3.2 Unsupervised generation of dream themes

In using this ecological data source to generate common dream themes, our results complement previous studies using more traditional survey methods. Though it is widely accepted that dream content varies based on individual personality and cultural differences, previous research suggests there might also be thematic "universals" that appear in a disproportionately high amount of dreams. Universal dream themes are typically quantified using surveys with predetermined thematic content developed by the researcher [24, 25], which are biased towards existing knowledge of dreams. In the current study, our unsupervised approach to developing common dream themes confirms some previously developed themes while also offering more specificity within them. For example previous survey studies using the Typical Dreams Questionnaire [24] or Dream Motif Scale [25] have identified common dream themes of failure, paranoia, snakes/insects and animal symbolism, alien life, fighting, and sex. Our results identified similar themes, while offering a finer-grained view with the subtopics that formed each theme. For example, we observed a popular theme of animals, and animal subtopics included what might be positive animal dreams (kitten, birds) and negative animal dreams (spider, maggots; snake, bite). It is difficult to compare the ranking of our dream themes with prior work, since prior work is heavily-dependent on the method of data collection [26]. A notable advance of our approach is the ability to "map out" the relationship between dream themes presence. The co-occurrence of common themes has not been studied extensively before, and future work comparing the co-occurrence of waking and dreaming themes might help to uncover what is truly unique about dream content.

3.3 Phenomenology of dream types

Dreams are highly varied experiences, and have long been grouped into subclasses or types of dreams (e.g., nightmares, lucid dreams). However, these dreams are often defined by a single feature (e.g., nightmares as intensely negative dreams) and the phenomenological variety within each subtype is not well understood. The present results offer new insights into the consistent-yet-variable content of nightmares, lucid dreams, vivid dreams, and recurring dreams, some of which have important clinical implications.

Nightmares are defined as intensely negative dreams, sometimes with a secondary requirement that they result in a direct awakening [27], and have immense clinical relevance in post-traumatic stress disorder (PTSD) [28] and other psychiatric diagnoses [29]. Our observation of keywords relating to sexual assault highlight prior observations of uniquely episodic event replay in PTSD patients [30], including survivors of sexual assault [31]. The additional observation of increased themes of Feeling and Violence suggest that the episodic replays are highly emotional recreations of violent events, and the inclusion of many escape-related words highlights the helplessness felt by many recurrent nightmare sufferers [32]. For nightmare sufferers, the negative affect during dreams [33] or the dream recall during the day might increase negative symptoms of other co-morbid diagnoses (e.g., anxiety) [34]. Lastly, the heightened presence of supernatural entities in nightmares might relate to the common state of sleep paralysis, an under-studied and cross-cultural phenomenon that occurs during sleep-wake transitions and frequently involves a feeling of helplessness amidst a hallucinated "demon" or otherwise frightening figure [35].

Vivid dreams are highly realistic (or similarly, well-remembered) dreams. Our unsupervised approach suggests that vivid dreams are not only realistic (e.g., keywords felt-real, real-like), but also that these dreams often contain major life events, strong emotions, and supernatural/religious experiences. Vivid dreams included births and pregnancies, missiles and war, apocalyptic events, and alien invasions. This dream subtype overlaps almost directly with a class of dreams referred to as "big dreams" [36]. Big dreams occur rarely, but when they do, they are highly meaningful experiences that make a significant and long-lasting impression on waking life. Thus, our analysis of vivid dreams might be representative of these dreams, given that they consisted of major life events and religious experiences that likely influenced future thinking. Interestingly, the presence of an Alien invasion theme in vivid/realistic dreams suggests that prior reports of UFO abductions might result from cases of dream-reality confusion [37], where a dreamt abduction is misinterpreted as a memory from waking life [38].

Lucid dreams are defined as those that include awareness of the dream while still dreaming [39], sometimes with a secondary requirement of having control over the dream [40]. The present unsupervised analysis was consistent with these defining features, with common themes and keywords in lucid dreams such as mental reflection and control. Additionally, our results confirm more recent preliminary findings about lucid dream content, such as frequent episodes of flying [41] and an overlap with sleep paralysis [42]. Though lucid dreams are generally regarded as more positive in valence than non-lucid dreams [43–45], there are more recent reports of extremely negative lucid dreams, or lucid nightmares [46–48]. Our results offer a cohesive explanation for these differential findings, in that we observed a general heightened realness and emotion in lucid dreams (themes of Feeling and Sights and visions and keywords of felt-real) without attachment to positive or negative valence. Our recent findings, focused on a different subreddit (r/LucidDreaming) suggest that positively-valenced lucid dreams are more likely to occur when dream control is involved [46], and the current results highlight the importance of focusing future clinical applications of lucid dreaming on the dream control rather than simply awareness of the dream (see [49] for a review of the clinical efficacy of lucid dreams to treat nightmares).

A recurrent dream is one that is experienced repeatedly, and these have been estimated to occur at least once in roughly 75% of the population [50, 51]. It is likely that the many themes and keywords we observed in recurrent dreams are related to waking anxieties or worries. The limited amount of prior work on recurrent dream phenomenology suggests that recurrent dreams are primarily related to waking anxieties [52] and other negative content [53], and also increase in frequency during periods of stress [51, 54]. Our analysis extends these findings by observing more specific anxieties in recurrent dream content. The top two themes were related to relationships, particularly negative aspects of relationships (i.e., cheating, ex-partners), and other common themes were even still related to relationships (e.g., sex, dating, crush). Other recurrent dream themes were explicitly negative, and at times ultraviolent; themes regarding serial killers, paralysis, sexual assault, and the apocalypse suggest that recurrent dreams are far more negative than positive or neutral. We suspect that many of these recurrent dreams would also be classified as nightmares (see statistics in Figure S1), and thus our analysis might help future work in the prediction/monitoring of nightmares via the inclusion of recurrent themes.

3.4 Impact of major events on collective dreams

Previous research has revealed that major personal and cultural events might influence dream content. For example, an increase in nightmares was observed after the terrorist attacks of September 11th, 2001 [55] and during the COVID-19 pandemic [56, 57]. Dreams during the COVID-19 pandemic have also been shown to include pandemic-related content. The present results expand on these prior findings by offering a finer-grained view into the topical and temporal impact of major events on dreams. Rather than a categorical increase in nightmares or pandemic-related content, our analysis allowed us to identify specific sub-themes of pandemic-related topics and how they change over time. During the COVID-19 pandemic, population dreams transitioned from outdoor to indoor locations and decreased in social interactions, as did our waking experiences during the pandemic. Interestingly, these two effects had qualitatively different time courses, where the location change of dreams was longer-lasting and more persistent than social changes. This might reflect the mass adoption of technical communications (e.g., Zoom gatherings) that occurred while people were still mostly indoors. These effects are consistent with prior work showing a continuity between wake and dream content [7, 58] and future work might evaluate how subtopics contribute uniquely to this continuity [58]. These results also contribute to hypothesized dream functions, such as the drop in social content contradicting the Social Simulation Theory that predicts a compensatory effect of social activity in dreams [59].

While dreams during the COVID-19 pandemic have been investigated at great length [8, 56, 57, 60], much less work has been dedicated to observing the influence of the Russo-Ukrainian war on sleep and dream patterns. We observed a population increase in negative war-related topics (e.g., violence, death) after the start of the war, which is notable because our social media sample is not expected to be mainly those people who were directly exposed to the war. The association between the war and population dreams could be a result of international media exposure (see also [55], highlighting a cognitive continuity between dreams and wake, where it is not daily activities per se, but the thought processes and internal imagery that predicts dream content). The negative content of population dreams during the war has important implications, given recent findings that negative dreams, including specifically dreams of death, are predictive of next-day negative affect [33] and nightmares have extensive mental health implications [29].

3.5 Limitations

These user characteristics could also influence the self-reported dream types we used as labels (flairs). However, our manual review of a sample of dreams showed that users' labels generally align with expert expectations, and the distinct topics identified across dream types suggest users typically label their dreams accurately. Furthermore, unlike the clear associations observed with COVID as a major event, those related to the Ukraine war may be incidental and would require future research to confirm any effects as the conflict continued to evolve. Finally, the users of Reddit are not a representative sample of the general population; they are known to be more male, young, educated, and urbanites [19].

The second limitation is about the elements from dream reports not reflecting the dream content. Given the free-form social media format, the `r/Dreams` users would not always describe only their dream report, but sometimes they would include some contextual information, for example recounting how they felt when they were woken up from the dream, or how the people they dreamt of are related to them in real life. For example, a Reddit user might drop in their dream report a sentence like *"It was about 5 am when I woke up from this dream..."* Such contextual information is not a direct description of the dream content, yet it often helps to qualify it and it is therefore useful to our analysis. Dream researchers analysing a small number of dreams could read through each report and manually remove such instances to focus on dream content only. Given the automatic topic extraction method that we employed, such a data cleaning step was not possible. Some common categories of contextual information that were unrelated to the dream experience (e.g., the author explaining when they last time met the person they dreamt of, or what time they woke up from the dream) emerged as independent topics in our topic analysis and we removed them (see Methods, Clustering Topics).

The third limitation of our work is shared with other content analysis methods [2]. We inevitably lose some dream information that could not be captured by the topics. This also means that our approach cannot represent subtle aspects of the individuality of each dreamer.

3.6 Implications

Our work has three main implications:

3.6.1 Developing an unsupervised dream content analysis method

The development of dream scoring systems has historically favored certain aspects of dream reports over others, based on assumptions about what are the most emotionally and socially important dimensions in dream analysis, rather than considering all potential topics or types of human experiences. Previous attempts to apply machine learning to dream content analysis have either replicated existing scoring systems [7] or focused on specific types of dream content (e.g., symptoms [8]). In contrast, our study used a deep learning approach that prioritizes no particular topic. By analyzing dream content from an exclusively linguistic standpoint, we were able to discover and quantify themes that have not been previously considered.

3.6.2 Uncovering the first ecological taxonomy of dream topics

Our results demonstrate that many of the themes we discovered align with those captured by existing scoring systems, particularly the Hall and Van de Castle scale, as it includes topics such as interpersonal relationships, emotions, and friendly and violent interactions. However, our approach also revealed differences in the frequency of certain topics, such as a significant category of weather-related topics that have not been previously considered in any dream scoring system. The significance of weather as a conversational topic varies, being mentioned as a safe subject for small talk or as the subject of jokes regarding dull conversations. Nevertheless, considering the evolutionary perspective, weather played a crucial role during the era of limited shelters and the absence of temperature controls, which shaped human instincts. Hence, weather likely holds a deep-rooted importance for our well-being and survival. To sum up, our findings support the continued use of traditional scoring systems in clinical psychology research and psychotherapy, where there is a strong rationale for focusing on the more emotional and social content of dreams. At the same time, our results suggest that AI tools can provide a more detailed and nuanced understanding of dream content, and may be mature enough to support dream analysts in their work.

3.6.3 Collaborating with AI in scientific discovery

The AI's ability to categorize dream content in ways unknown to human researchers is reminiscent of the scenario when chess and go-playing programs began surpassing human players. These programs did not merely excel at the strategies employed by humans, instead, they developed unique strategies that had never been observed by humans before [61]. It was assumed that humans approached both games based on evolutionary instincts developed for social interactions, while the AI adopted a more objective perspective, solely focusing on the game rules without any assumptions influenced by human endeavors. Similarly, AI comprehends texts based solely on their intrinsic content, without filtering them through instinctual categories primarily designed for interactions during waking states.

4 Materials and methods

4.1 Data pre-processing

For the purpose of topical analysis, we employed the `en_core_web_sm` model from *Spacy* [62] to segment dream reports into individual sentences. The corpus consisted of 44,213 dream reports, for a total of 761,619 sentences. On average, each dream report contains 17 sentences and 290 words. Notably, recurring dreams tended to be shorter, with 14 sentences and 253 words, while lucid dreams were longer, with 27 sentences and 465 words (Table 1). It is worth mentioning that the distribution of dream reports was not concentrated among a few individuals; the majority of users shared a single dream report, whereas only a small group of frequent users contributed more than 30 dream reports (Fig. 2 (b)).

We discarded the top frequent 10,000 sentences with the fewest characters. These sentences typically contained non-dream related content, such as greetings to the reader (hello, hi, etc.), sentences consisting solely of special characters, and similar instances. This step of removing a significant number of sentences that were not relevant to the dream report itself ensured that our subsequent labeling process for non-dream topics was sufficient to preserve predominantly dream topics.

4.2 Topic modelling

Our topic modelling procedure consisted of the following five core steps: i) extracting topics, ii) grouping topics into themes, iii) building dream topics taxonomy, iv) finding topics and themes by dream type, and v) finding topics and themes through time. In addition to the core steps, our procedure includes filtering and validation steps to ensure high-quality topics and themes, prior to constructing the dream topics taxonomy.

4.2.1 Extracting topics

BERTopic [17] is largely based on a neural model that is designed to identify latent topics within document collections. Unlike conventional topic modeling techniques such as Latent Dirichlet Allocation, BERTopic leverages semantic information by utilizing embeddings as an initial step to cluster documents into semantically cohesive topics. Each discovered topic in BERTopic is described by a list of 10 *topic words*, which are the most distinctive words associated with that particular topic. The topics are numbered based on their frequency rank, indicating their prevalence within the corpus.

In its default configuration, BERTopic assigns a single topic to each document. However, our manual inspection of the dream reports revealed that most dreams cannot be adequately characterized by a single topic. Instead, they often encompass multiple topics, such as the dream's location, the people involved, and the emotions experienced by the dreamer. Additionally, dreams are known for combining various elements from waking experiences, resulting in a sense of bizarreness. To address this, our initial alternative was to modify BERTopic to associate a distribution of topics with each dream report. We tested this approach by allowing up to ten topics per report. Subsequently, we applied a threshold to the probabilities associated with each topic to identify the relevant topics for each report. However, we encountered two issues with this method. Firstly, due to the substantial variability in the length of dream reports, we often missed relevant topics in longer dreams. Secondly, even with varying the threshold, over 55% of dreams ended up being associated with no topic, resulting into a considerable loss of data. To mitigate this issue, we opted to consider individual dream sentences as input documents for BERTopic. Applying BERTopic at sentence-level is a practice recommended by the BERTopic authors on the official website of the tool. Such a solution enabled us to associate over 88% of dream reports with at least one topic.

To establish robust topic representations, we configured the hyperparameters of BERTopic. We set the minimum frequency threshold (`min_df`) to 10, ensuring that a word appears in at least 10 sentences before it is considered for inclusion in the topic representation. This criterion helped to ensure that topics were formed based on words with a reasonable level of occurrence within the dream reports. Additionally, we employed the Maximal Marginal Relevance (MMR) algorithm with a diversity parameter set to 0.4. The MMR algorithm was utilized to enhance the diversity of the topic words, preventing a dream's topic representation only from near-synonyms. By incorporating this diversity measure, the resulting topics encompassed a wider range of relevant terms, capturing distinct aspects and avoiding redundant or similar descriptions within the topic representation.

Our model successfully extracted a total of 288 topics which included both dreaming and non-dreaming subjects. An exhaustive manual inspection of the topic words demonstrated a significant degree of semantic coherence across the majority of topics: for most topics, the top four topic words provided sufficient information to understand the essence of the topic. To ensure topic specificity, we excluded all instances of the terms "dream" and "dreams" from the list of topic words associated with each topic. Furthermore, we aggregated topics related to multiple sentences within the same dream, consolidating them into a comprehensive list. This approach allowed us to capture the overarching themes and content associated with individual dreams more effectively.

4.2.2 Validating unsupervised topic modeling

To make sure that the clusters are semantically cohesive and well-separated, we compare the distribution of semantic distances between sentences within the same cluster and across clusters. We computed intra-cluster distance by calculating the mean Euclidean distance between L2-normalized [63] sentence embeddings pairs within the same clusters. For each cluster, we select a random sample of 4500 pairs, and we extract the distance of their embeddings calculated with the *all-mpnet-base-v2* model.

We computed inter-cluster distance as follows. First, we obtain the topic centroids. Given the centroid of a target topic X, we identify the 5% of clusters (n=14) whose centroid is closest to it — these are neighboring topics that are semantically closest to the target centroid. We then calculate the Euclidean distance between the embeddings of 322 pairs of sentences (s_i^X, s_i^Y), where s_i^X is a sentences taken at random from the target topic and s_i^Y is a sentence taken at random from a neighboring topic. After repeating this calculation for all neighboring topics, we obtained 4508 semantic distances, which we average to estimate the average distance between topic X and its neighboring topics.

Last, we plot the distribution of inter-cluster distance and intra-cluster distance for all topics. Figure 7 shows that intra-cluster distance is significantly smaller than intra-cluster distance, indicating that our topic are topically cohesive and well-separated between one another.

4.2.3 Grouping topics into themes

To provide a concise overview of the 288 identified topics, we used clustering techniques to group them into broader, yet semantically-coherent themes. To achieve this, we used the Sentence Transformer model *all-mpnet-base-v2* [14, 64, 65] to project each topic word into a 768-dimensional embedding space. For each topic t, every word w^t within that topic was assigned a probability (p_w^t) by BERTopic, indicating its contribution to the overall representation of the topic. To compute an embedding for a topic ($\vec{t}$), we calculated the weighted sum of the embeddings of its topic words, where each word's embedding was weighted by its normalized probability (p_w^t):

$$\vec{t} = \sum_{w^t} \text{emb}(w^t) \cdot p_w^t, \tag{1}$$

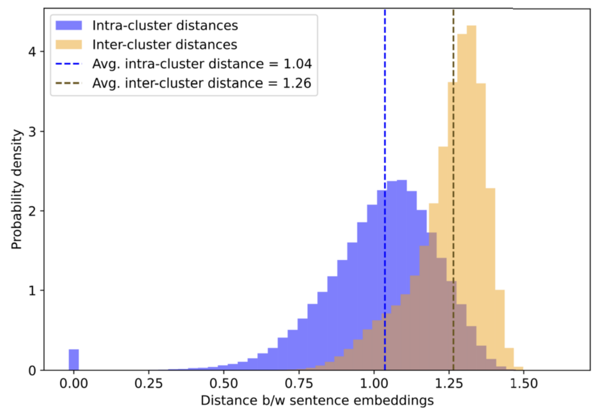

Figure 7 Distribution of intra & inter cluster distances for extracted topics. Here, each cluster corresponds to a collection of sentences associated with that topic. Kolmogorov-Smirnov (KS) test, comparing the two distributions, yielded a KS statistic of 0.575, ($p < .001$)

where *emb* is the sentence transformer model that maps the topic words into embeddings, and w^t are all the words associated with topic t.

To facilitate semantic clustering of the topics, we standardized their embeddings and reduced their dimensionality from 768 to 10 using Uniform Manifold Approximation and Projection (UMAP) [66] method for dimensionality reduction. We then explored two clustering algorithms, K-Means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [67], applied to the reduced topic embeddings.

Through manual inspection, we observed that the K-Means clustering method effectively generated semantically coherent clusters, with most topics within each cluster sharing a cohesive theme. We also observed that K-Means with 20 clusters produced the result with the best quality. Further details on the analysis can be found in Section S2.1.

4.2.4 Filtering non-dream content and adjusting themes

Our manual inspection of the themes and Reddit dream content more generally revealed that while most of it is dream content descriptions, there were some social media artifacts, as well as non-dream content explanations from the users. However, we also saw that BERTopic's remarkable semantic capabilities proved highly effective in distinguishing the actual dream recollections from other content. Leveraging this capability, we employed a filtering process to exclude non-dream themes and topics from our analysis.

Below, we first discuss how we have performed the filtering, and then follow with a more general guideline for each step derived from our learning, that can support other dream researchers working on similar tasks.

Our manual filtering steps

Step 1: Discard Themes with Non-Dream Content. We began by inspecting the 10 topic words associated with the topics within each theme. This was usually sufficient to classify the theme. In rare cases where it was not, we also reviewed three representative dream sentences generated by BERTopic for each topic, as well as 20 randomly sampled sentences from associated dreams. Themes such as *Dream types*, *Dreaming-waking interface*, and *Social media artifacts* consisted entirely of topics that could be easily identified as non-dream content. Consequently, we discarded these clusters entirely and proceeded with the remaining themes.

Guideline: Classify themes into one of the five categories listed below and retain only dream-related content.

Step 2: Split Composite Themes into Smaller Ones. Some themes contained topics that could be divided into smaller, semantically cohesive groups based on suggestions from dream experts. For example, the theme *Outdoor environment and space* was split into three distinct subthemes: *Outdoor locations*, *Space*, and *Weather*.

Guideline: This step may require expertise in dream research.

Step 3: Discard Topics with Non-Dream Content. We manually inspected all 288 topics and their top corresponding dream sentences to identify and separate non-dream content. For example, in the theme *Time, Time Travel, and Timelines*, we removed topics like *'5 am currently clock checked phone'* and *'date 2018 June'*, which referred to timestamps of real events. Similarly, in the theme *Mental Reflection and Interactions*, we excluded topics such as *'interpret, does really mean'* and *'dont remember details'*. This process ensured that topics predominantly reflected either dream content or waking content, allowing us to filter out non-dream-related topics.

Guideline: Focus on removing topics that likely reflect external explanations of the dream context or discussions about its interpretation.

Step 4: Minor Reassignment of Topics Between Themes. Nearly all topics were appropriately assigned to their respective themes, with only a few exceptions. For example, the topic *'lady, face, looked, head'* was initially placed in a cluster later termed *Other topics*, which served as a catch-all for topics that did not fit into a semantically coherent cluster. However, given its content, we reassigned it to the theme *People and relationships*.

Guideline: This step may not always be necessary but can benefit from the input of dream researchers.

We involved four annotators for the aforementioned steps: two with expertise in Natural Language Processing, and two senior dream researchers. Their diverse backgrounds ensured both computational rigor and alignment with traditional dream literature. Agreement among annotators was high across all steps: complete agreement in discarding non-dream themes initially, over 95% consensus in identifying non-dream topics, and over 90% agreement in splitting composite themes and reassigning topics. Disagreements were resolved through consensus discussions, prioritizing insights from the dream researchers.

These four steps resulted in a final set of 217 dream-related topics, clustered into 22 distinct dream themes.

Five Theme Categories.

1. Dream-related content – elements such as dream locations, animals, and people.
2. Dream types – including nightmares and recurring dreams.
3. Dreaming-waking interface – experiences like waking up crying or feeling confused, as well as being awakened by an alarm.
4. Waking phenomena – aspects such as mental health issues and the date and time of the dream.
5. Social media artifacts – elements like expressions of gratitude to the reader or requests for dream interpretation.

4.2.5 Validation of our filtering approach

To validate the accuracy and robustness of our filtering approach, we examined how the results would change if sentences were labeled prior to topic and theme clustering, while preserving the full dream context. Specifically, we randomly sampled 1000 sentences associated with dreaming topics and 1000 sentences linked to non-dream topics, as identified by our method. Using GPT-4o, chosen for its superior performance in complex labeling tasks [68, 69], we provided the LLM with individual sentences and their corresponding full dream reports, instructing it to classify each sentence as either dream content or non-dream content (e.g., meta-discussions, waking experiences). Results showed that 85.5% of sentences associated with dreaming topics and 81.3% of those linked to non-dreaming topics were labeled consistently with our method. Manual inspection of mislabeled sentences revealed that such cases often contained both dreaming and waking elements (e.g., *"I haven't visited my friend in 10 years, so I was surprised when I found myself dreaming about going out to dinner with her"*). In these instances, ChatGPT typically classified the sentence as non-dream content, whereas our method correctly mapped the dreaming portion (e.g., *"dreaming about going out to dinner with her"*) to a dream topic.

To assess whether this small proportion of mislabeled sentences significantly affected the overall distribution of dreaming topics, we investigated whether the relative size of topics (i.e., the proportion of dream sentences linked to each topic) remained stable. We sampled 10,000 sentences spanning 8005 dreams from reports associated with 217 dreaming topics and filtered out non-dream sentences using GPT-4o in the context of the full dream reports. For each topic, we calculated the number of dream sentences in the cleaned sample (Y_{topic}) and compared it to the original counts assigned by our method (X_{topic}). A high correlation between X_{topic} and Y_{topic} ($r = .977$, $p < .001$) confirmed that noisy associations had minimal impact on the relative proportions of dream sentences across topics. This finding validates our taxonomy as a reliable representation of the relative occurrence of topics and themes in people's dreams.

4.3 Building dream topics taxonomy (co-occurrence network)

To create a taxonomy of dream topics, we employed a network-based approach that explores the interplay between dream themes and their constituent topics within dream reports. To facilitate this analysis, we conducted an initial assessment of the frequency distributions of both topics and themes across the entire corpus of dream reports.

For topics, we linked each one to a dream report (/dreamer) if the topic was found at least once in the report (in all dream reports of that dreamer), and counted the number of dreams (/dreamers) associated with each topic. In Table 2, we present the topic name, top 4 topic words, number of dreams, and number of dreamers for the top 20 most frequent topics. Full statistics (e.g., the 10 topic words and number of sentences associated with each topic) for these and the rest of the topics, can be found in the Supplementary Information File (see Data Availability section).

Similarly, we linked a dream report (dreamer) to a theme, if any of the theme's constituent topics was found at least once in the report (in all dream reports of that dreamer). Additionally, we computed the number of topics in each theme associated with corresponding dream reports (dreamers). In Table 3, we present the theme name, top 3 topics associated with it, the total number of topics in it, and number of dreams, and number of dreamers associated with it. Additionally, we linked to a corresponding *HVdC* category each theme for which such a link is found. The full list of topics belonging to each theme can be found in Table S1.

Having these frequencies at hand, we built the co-occurrence network of dream themes as follows. Each node in the network represents a theme, and pairs of themes were connected by an edge if they co-occurred in a dream report. The edges were weighted by the number of dream reports in which such a co-occurrence was found. This undirected network had a single connected component with 22 nodes and edge weights ranging from 13 to 7643 (Mean = 762.31 $\pm$ 928.54 and median = 482.0).

For the purpose of visualization, we used backboning to sparsify the network by preserving the most important edges. We used noise-corrected backboning [70]—a technique that relies on a statistical null-model to identify and prune non-salient edges—with a backboning threshold of 3.8 (which reduced the network from 231 to 46 edges). We used Gephi [71] to visualize this network (see Fig. 3). We scaled the size of the nodes according to the number of dreams associated with each theme.

4.4 Finding topics and themes by dream type (odds-ratio analysis)

In addition to discovering common topics across all dreams, we studied whether specific types of dreams (i.e., nightmares, lucid, vivid, and recurring dreams) are characterised by particular topics and themes. Odds-ratio metric allowed us to do so as it compares the odds of a topic occurring in the specific type of dreams (e.g., recurring) to the odds of the same topic occurring in the rest of dreams. We first assigned the dream reports to the corresponding dream types by searching for relevant keywords ('nightmar', 'recurring' or 're-occurring', 'lucid', 'vivid') in Reddit post title and body or, if the dream type had a corresponding flair (which were present for nightmares and recurring dreams only). If either of these conditions were satisfied, we assigned the dream to the experimental subset; else to the control subset.

First, we defined a topic or theme t's importance in a dream d as:

$$I_{t,d} = \frac{\text{\# sentences mentioning topic t in d}}{\text{\# total sentences in d}} \tag{2}$$

We then computed the odds ratio for all topics and themes across the 4 dream types as follows:

$$\begin{aligned} \text{Odds Ratio (DT, t)} &= \frac{\text{Odds of DT association with t}}{\text{Odds of the rest of dreams association with t}} \\ &= \frac{\frac{\sum_{d \in DT} I_{t,d}}{\text{\# dreams in DT that do not contain t}}}{\frac{\sum_{d \notin DT} I_{t,d}}{\text{\# dreams that are not in DT that do not contain t}}} \end{aligned} \tag{3}$$

where DT is a dream type, t is a topic or theme.

4.5 Finding topics and themes through time

Analyzing the temporal dynamics of dream topics and themes holds particular significance, especially in light of major events such as the COVID-19 pandemic, known to have a substantial impact on collective dreaming patterns [8, 72, 73].

For our analysis, we focused on a monthly timescale, leveraging data spanning from January 2019 to August 2022, encompassing a total of 44 months. The inclusion criterion for each month required a minimum of 300 dream reports, ensuring robust statistical representation. Throughout this period, February 2019 recorded the lowest number of dreams ($n = 366$), whereas January 2021 exhibited the highest dream count ($n = 1573$). The average number of dreams per month was 925 ± 332, with a median of 935 dreams per month.

We used topic/ theme importance in a dream ($I_{t,d}$) introduced in Equation (2) to calculate topic/ theme importance at a given point in time m (i.e., month) as follows:

$$I_{t,m} = \frac{\sum_{(\text{d posted at time } m)} I_{t,d}}{\#\text{ dreams posted at time } m} \tag{4}$$

We tracked z-scores of topic/ theme importances $I_{t,m}$ through time, to understand the relative change of a topic/ theme w.r.t. itself:

$$z_{I_{t,m}} = \frac{I_{t,m} - \mu_{I_{t,m}}}{\sigma_{I_{t,m}}} \tag{5}$$

To additionally improve the quality of signals, we used the centered average with a window length of 5 for smoothing the temporal plots:

$$\widetilde{z}_{I_{t,m}} = \sum_{t=(t-2)}^{(t+2)} z_{I_{t,m}}. \tag{6}$$

Abbreviations
AI, Artificial Intelligence; API, Application Programming Interface; BERT, Bidirectional Encoder Representations from Transformers; HVdC, Hall and Van de Castle; NLP, Natural Language Processing; PTSD, Post Traumatic Stress Disorder; SI, Supplementary Information; UFO, Unidentified Flying Objects; MMR, Maximal Marginal Relevance; UMAP, Uniform Manifold Approximation and Projection; DBSCAN, Density-Based Spatial Clustering of Applications with Noise.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00554-w.

Additional file 1. (PDF 1.1 MB)

Acknowledgements
L.M.A acknowledges the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions
AD conceptualized, designed, and conducted the experiments, prepared figures, and contributed to writing, reviewing, and editing the manuscript. SŠ and LMA conceptualized and designed the experiments, prepared figures, supervised the project, and contributed to writing, reviewing, and editing the manuscript. RM, DB, and DQ supervised the project and contributed to writing, reviewing, and editing. All authors approved the final manuscript.

Funding information
Nokia Bell Labs provided support in the form of salaries for authors [SŠ, DQ], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Data availability
The datasets generated during the current study are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.23618064.v2.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Heritage Institute of Technology, Chowbaga Road, Anandapur, Kolkata, 700107, West Bengal, India. [2]Nokia Bell Labs, Broers Building, Cambridge, CB3 0FA, United Kingdom. [3]Uehiro Oxford Institute, University of Oxford, St Ebbe's Street, Oxford, OX1 1PT, United Kingdom. [4]IT University of Copenhagen, Rued Langgaads Vej, 7, Copenhagen, 2300, Denmark. [5]Pioneer Centre for AI, Øster Voldgade 3, Copenhagen, 1350, Denmark. [6]Northwestern University, 633 Clark St, Evanston, 60208, Illinois, United States. [7]Harvard University, 352 Harvard Street, Cambridge, 02138, Massachusetts, United States. [8]Politecnico di Torino, Corso Duca degli Abruzzi, 24, Turin, 10129, Piedmont, Italy.

Received: 8 June 2024 Accepted: 29 April 2025 Published online: 23 May 2025

References
1. Mota-Rolim SA, Bulkeley K, Campanelli S, Lobão-Soares B, De Araujo DB, Ribeiro S (2020) The dream of god: how do religion and science see lucid dreaming and other conscious states during sleep? Front Psychol 11:555731
2. Schredl M (2010) Characteristics and contents of dreams. Int Rev Neurobiol 92:135–154
3. Schredl M (2010) Dream content analysis: basic principles. Int J Dream Res 3(1):65–73
4. Payne JD (2010) Memory consolidation, the diurnal rhythm of cortisol, and the nature of dreams: a new hypothesis. Int Rev Neurobiol 92:101–134
5. Wamsley EJ, Tucker M, Payne JD, Benavides JA, Stickgold R (2010) Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. Curr Biol 20(9):850–855
6. Beck AT, Ward CH (1961) Dreams of depressed patients: characteristic themes in manifest content. Arch Gen Psychiatry 5(5):462–467
7. Fogli A, Maria Aiello L, Quercia D (2020) Our dreams, our selves: automatic analysis of dream reports. R Soc Open Sci 7(8):192080
8. Šćepanović S, Aiello LM, Barrett D, Quercia D (2022) Epidemic dreams: dreaming about health during the covid-19 pandemic. R Soc Open Sci 9(1):211080
9. Winget C, Kramer M (1979) Dimensions of dreams. University Presses of Florida, Gainesville
10. Hall CS, Castle RL (1966) The content analysis of dreams. The century psychology series. Appleton-Century-Crofts, New York, p 320
11. Elce V, Handjaras G, Bernardi G (2021) The language of dreams: application of linguistics-based approaches for the automated analysis of dream experiences. Clocks & Sleep 3(3):495–514
12. Colace C (2003) Dream bizarreness reconsidered. Sleep & Hypnosis 5:105–128
13. Picard-Deland C, Nielsen T, Carr M (2021) Dreaming of the sleep lab. PLoS ONE 16(10):0257738
14. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805
15. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J (2020) The pushshift reddit dataset. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 830–839
16. Keller JW, Brown G, Maier K, Steinfurth K, Hall S, Piotrowski C (1995) Use of dreams in therapy: a survey of clinicians in private practice. Psychol Rep 76(3):1288–1290
17. Grootendorst M (2022) Bertopic: neural topic modeling with a class-based tf-idf procedure. arXiv preprint. arXiv:2203.05794
18. Domhoff GW (2003) The scientific study of dreams: neural networks, cognitive development, and content analysis. American Psychological Association, Washington. https://doi.org/10.1037/10463-000
19. Hargittai E (2020) Potential biases in big data: omitted voices on social media. Soc Sci Comput Rev 38(1):10–24
20. Siclari F, LaRocque JJ, Postle BR, Tononi G (2013) Assessing sleep consciousness within subjects using a serial awakening paradigm. Front Psychol 4:542
21. Nemeth G (2022) The route to recall a dream: theoretical considerations and methodological implications. Psychol Res: 1–24
22. Schredl M (2002) Questionnaires and diaries as research instruments in dream research: methodological issues. Dreaming 12:17–26
23. Sanz C, Zamberlan F, Erowid E, Erowid F, Tagliazucchi E (2018) The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. Front Neurosci 12:7
24. Nielsen TA, Zadra AL, Simard V, Saucier S, Stenstrom P, Smith C, Kuiken D (2003) The typical dreams of Canadian university students. Dreaming 13(4):211
25. Yu Kai Ching C (2012) Pornography consumption and sexual behaviors as correlates of erotic dreams and nocturnal emissions. Dreaming
26. Mathes J, Schredl M, Göritz AS (2014) Frequency of typical dream themes in most recent dreams: an online study. Dreaming 24(1):57
27. Gieselmann A, Ait Aoudia M, Carr M, Germain A, Gorzka R, Holzinger B, Kleim B, Krakow B, Kunze AE, Lancee J, et al (2019) Aetiology and treatment of nightmare disorder: state of the art and future perspectives. J Sleep Res 28(4):12820
28. Campbell RL, Germain A (2016) Nightmares and posttraumatic stress disorder (ptsd). Curr Sleep Med Rep 2:74–80
29. Sheaves B, Rek S, Freeman D (2022) Nightmares and psychiatric symptoms: a systematic review of longitudinal, experimental, and clinical trial studies. Clin Psychol Rev: 102241
30. Phelps AJ, Forbes D, Creamer M (2008) Understanding posttraumatic nightmares: an empirical and conceptual review. Clin Psychol Rev 28(2):338–355
31. Krakow B, Schrader R, Tandberg D, Hollifield M, Koss MP, Yau CL, Cheng DT (2002) Nightmare frequency in sexual assault survivors with ptsd. J Anxiety Disord 16(2):175–190

32. Rousseau A, Belleville G (2018) The mechanisms of action underlying the efficacy of psychological nightmare treatments: a systematic review and thematic analysis of discussed hypotheses. Sleep Med Rev 39:122–133
33. Mallett R, Picard-Deland C, Pigeon W, Wary M, Grewal A, Blagrove M, Carr M (2021) The relationship between dreams and subsequent morning mood using self-reports and text analysis. Affect Sci: 1–6
34. Blagrove M, Farmer L, Williams E (2004) The relationship of nightmare frequency and nightmare distress to well-being. J Sleep Res 13(2):129–136
35. Denis D, French CC, Gregory AM (2018) A systematic review of variables associated with sleep paralysis. Sleep Med Rev 38:141–157
36. Bulkeley K, Hartmann E (2011) Big dreams: an analysis using central image intensity, content analysis, and word searches. Dreaming 21(3):157
37. Wamsley E, Donjacour CE, Scammell TE, Lammers GJ, Stickgold R (2014) Delusional confusion of dreaming and reality in narcolepsy. Sleep 37(2):419–422
38. Holden KJ, French CC (2002) Alien abduction experiences: some clues from neuropsychology and neuropsychiatry. Cogn Neuropsychiatr 7(3):163–178
39. Mallett R, Carr M, Freegard M, Konkoly K, Bradshaw C, Schredl M (2021) Exploring the range of reported dream lucidity. Philos Mind Sci 2:1–23
40. Windt JM, Voss U (2018) Spontaneous thought, insight, and control in lucid dreams. In: The Oxford handbook of spontaneous thought: mind-wandering, creativity, and dreaming, 385–410
41. Picard-Deland C, Pastor M, Solomonova E, Paquette T, Nielsen T (2020) Flying dreams stimulated by an immersive virtual reality task. Conscious Cogn 83:102958
42. Mainieri G, Maranci J-B, Champetier P, Leu-Semenescu S, Gales A, Dodet P, Arnulf I (2021) Are sleep paralysis and false awakenings different from rem sleep and from lucid rem sleep? A spectral eeg analysis. J Clin Sleep Med 17(4):719–727
43. Mallett R (2020) Partial memory reinstatement while (lucid) dreaming to change the dream environment. Conscious Cogn 83:102974
44. Schredl M, Fuchs C, Mallett R (2022) Differences between lucid and nonlucid dream reports: a within-subjects design. Dreaming
45. Voss U, Schermelleh-Engel K, Windt J, Frenzel C, Hobson A (2013) Measuring consciousness in dreams: the lucidity and consciousness in dreams scale. Conscious Cogn 22(1):8–21
46. Mallett R, Sowin L, Raider R, Konkoly KR, Paller KA (2022) Benefits and concerns of seeking and experiencing lucid dreams: benefits are tied to successful induction and dream control. Sleep Adv 3(1):027
47. Schredl M, Bulkeley K (2020) Lucid nightmares: an exploratory online study. Int J Dream Res: 215–219
48. Stumbrys T (2018) Lucid nightmares: a survey of their frequency, features, and factors in lucid dreamers. Dreaming 28(3):193
49. Ouchene R, El Habchi N, Demina A, Petit B, Trojak B (2023) The effectiveness of lucid dreaming therapy in patients with nightmares: a systematic review. L'Encéphale
50. Vallat R, Eskinazi M, Nicolas A, Ruby P (2018) Sleep and dream habits in a sample of French college students who report no sleep disorders. J Sleep Res 27(5):12659
51. Zadra A (1996) Recurrent dreams and their relation to life events and well-being. Trauma & Dreams: 231–247
52. Robbins PR, Tanck RH (1992) A comparison of recurrent dreams reported from childhood and recent recurrent dreams. Imagin Cogn Pers 11(3):259–262
53. Weinstein N, Campbell R, Vansteenkiste M (2018) Linking psychological need experiences to daily and recurring dreams. Motiv Emot 42:50–63
54. Duke T, Davidson J (2002) Ordinary and recurrent dream recall of active, past and non-recurrent dreamers during and after academic stress. Dreaming 12:185–197
55. Propper RE, Stickgold R, Keeley R, Christman SD (2007) Is television traumatic?: dreams, stress, and media exposure in the aftermath of September 11, 2001. Psychol Sci 18(4):334–340
56. Gorgoni M, Scarpelli S, Alfonsi V, De Gennaro L (2022) Dreaming during the covid-19 pandemic: a narrative review. Neurosci Biobehav Rev: 104710
57. Margherita G, Caffieri A (2022) An observatory on changes in dreaming during a pandemic: a living systematic review (part 1). J Sleep Res: 13742
58. Schredl M, et al (2003) Continuity between waking and dreaming: a proposal for a mathematical model. Sleep & Hypnosis 5:38–52
59. Tuominen J, Olkoniemi H, Revonsuo A, Valli K (2022) 'no man is an island': effects of social seclusion on social dream content and rem sleep. Br J Psychol 113(1):84–104
60. Solomonova E, Picard-Deland C, Rapoport IL, Pennestri M-H, Saad M, Kendzerska T, Veissiere SPL, Godbout R, Edwards JD, Quilty L, et al (2021) Stuck in a lockdown: dreams, bad dreams, nightmares, and their relationship to stress, depression and anxiety during the covid-19 pandemic. PLoS ONE 16(11):0259040
61. Shin M, Kim J, Opheusden B, Griffiths TL (2023) Superhuman artificial intelligence can improve human decision-making by increasing novelty. Proc Natl Acad Sci USA 120(12):2214840120
62. Honnibal M, Montani I, Van Landeghem S, Boyd A, et al (2020) Spacy: industrial-strength natural language processing in python
63. GitHub, UKPLab (2021) Sentence Transformers - Issue 1270. https://github.com/UKPLab/sentence-transformers/issues/1270. Accessed: 2025-04-07 (2021)
64. Song K, Tan X, Qin T, Lu J, Liu T-Y (2020) Mpnet: masked and permuted pre-training for language understanding. Adv Neural Inf Process Syst 33:16857–16867
65. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 3982–3992. https://aclanthology.org/D19-1410/
66. McInnes L, Healy J, Saul N, Großberger L (2018) Umap: uniform manifold approximation and projection. J Open Source Softw: 3(29)

67. Ester M, Kriegel H-P, Sander J, Xu X (1996) Density-based spatial clustering of applications with noise. In: Int. Conf. Knowledge Discovery and Data Mining, vol 240
68. Zhu Y, Zhang P, Haq E-U, Hui P, Tyson G (2023) Can chatgpt reproduce human-generated labels? A study of social computing tasks. arXiv preprint. arXiv:2304.10145
69. Byun C, Vasicek P, Seppi K (2023) Dispensing with humans in human-computer interaction research. In: Extended abstracts of the 2023 CHI conference on human factors in computing systems, pp 1–26
70. Coscia M, Neffke FM (2017) Network backboning with noisy data. In: 2017 IEEE 33rd international conference on data engineering (ICDE). IEEE Press, New York, pp 425–436
71. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the international AAAI conference on web and social media, vol 3, pp 361–362
72. Barrett D (2020) Dreams about covid-19 versus normative dreams: trends by gender. Dreaming 30(3):216
73. Schredl M, Bulkeley K (2020) Dreaming and the covid-19 pandemic: a survey in a us sample. Dreaming 30(3):189

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.