UTRECHT UNIVERSITY

MASTER THESIS

Creativity Behind the Prompts: Automated Creativity Assessment in Prompting for Text-to-Image Models

Author:
Saif ABDOELRAZAK

Supervisor:
Dr. A.A. (Almila) AKDAG
Second Supervisor:
Dr. D.P. (Dong) NGUYEN

A thesis project submitted in fulfillment of the requirements for the degree of Artificial Intelligence

August 7, 2023

Utrecht
University

UTRECHT UNIVERSITY

Abstract

Faculty of Science
Graduate School of Natural Sciences

Artificial Intelligence

Creativity Behind the Prompts: Automated Creativity Assessment in Prompting for Text-to-Image Models

by Saif ABDOELRAZAK

This thesis delves into the intersection of artificial intelligence and creativity, specifically focusing on the application of text-to-image synthesis models. These models, gaining significant attention in recent years, hold the potential to redefine the boundaries of human imagination and challenge conventional notions of creativity. However, they also raise pertinent questions about originality, copyright, and the role of human input in the creative process.

The study investigates the use of prompt engineering to augment the creativity of the generated artworks. Various prompt modifiers, including artist names and aesthetic quality descriptors, are employed to guide the synthesis process. The results indicate that the strategic use of these modifiers significantly enhances the creativity of the generated images, providing a concrete strategy for both novice and experienced users of these models.

The research also explores the use of topic modeling methods, such as Gibbs Sampling Dirichlet Mixture Model (GSDMM) and BERTopic. However, several challenges, including computational constraints and limitations in the clustering methods used, are identified. Despite these challenges, the research offers valuable insights into the potential of text-to-image synthesis models and the role of prompt engineering in enhancing creativity. Future work aims to address these challenges and further explore the potential of these models in various creative domains.

Contents

List of Abbreviations

AI Artificial Intelligence

AUT Alternate Uses Task

BERT Bidirectional Encoder Representations from Transformers

CAT Consensual Assessment Technique

CBOW Continuous Bag of Words

cTFIDF Class-based Term Frequency-Inverse Document Frequency

DMM Dirichlet Mixture Model

GAN Generative Adversarial Network

GloVe Global Vectors for Word Representation

GSDMM Gibbs Sampling Dirichlet Mixture Model

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

LDA Latent Dirichlet Allocation

LSA Latent Semantic Analysis

LIWC Linguistic Inquiry and Word Count

MGP Movie Group Process

NLP Natural Language Processing

sBERT sentence-BERT

TFIDF Term Frequency-Inverse Document Frequency

TTCT Torrance Tests of Creative Thinking

UMAP Uniform Manifold Approximation and Projection

Chapter 1

Introduction

The field of Artificial Intelligence (AI) has made tremendous strides in recent years, with major advances being made in text-guided synthesis of images using deep generative models. Various tools such as Stable Diffusion, Midjourney, and Dall-E have been made public for everyone to use, without needing prior knowledge of the underlying technologies of these models. These AI models are able to generate a wide range of artistic styles and forms, including paintings and drawings, by analyzing prompts or instructions given to them by humans. The prompts used to generate art with these models can take many forms, including text descriptions, images, and even audio clips. In this thesis however, the focus will be on the use of text descriptions, also known as prompts, to generate artworks. The emergence of AI-art generating tools have sparked many discussions over the definition of art and creativity. Discussions concerning AI-generated art include questions about the ownership of the generated artefacts (Eshraghian, 2020), constitutional concerns over art and copyright (Ihalainen, 2018),questions about the role of human intention and originality in AI art generation (Weiner, 2018; Clarke, 2022), and questions about whether prompting is really creating art (McCormack et al., 2023). Recently, (Hertzmann, 2020) has argued that text-to-image models such as Stable Diffusion and Dall-E do not create art themselves, but that the artists and technologists who apply them as tools are the ones creating art.

In this study, the notion of creativity in the process of creating art with text-to-image models will be examined, as human creativity will be studied through the analysis of the prompts that humans use to generate art with the recently introduced AI-art generation algorithms. After all, it is the user who crafts the prompt, often meticulously, iterating over different versions of the prompt before arriving on the final prompt. In the ongoing discussions surrounding the artistry and creativity inherent in AI-generated art, the aim is to evaluate the prompts' role in gauging how this emerging AI technology triggers human creativity.

In this study, the primary objective is to examine the creativity inherent in prompts, by using the methodologies of Natural Language Processing (NLP). NLP, a specialized branch of AI, is dedicated to equipping computers with the capability to comprehend text and spoken words in a manner akin to human understanding. It combines computational linguistics with statistical models, machine learning, and deep learning methodologies, providing a robust framework for text analysis.

By utilizing NLP techniques, the analysis of text prompts in diffusion models provides insight into the factors contributing to satisfactory AI-generated art outcomes. This involves examining the types, structure, and formatting of the prompts. The objectives are thus: (1) to reveal the prompts used by people for art generation, offering an understanding of their preferences in the creative process, and (2) to determine the specific components of the prompts that influence overall creativity. However, there are several challenges to address. Firstly, a creativity assessment framework must be defined for the short-text prompt data, which varies in length and verbosity. For instance, comparing the creativity of "`a cat`" to "`photography of a Cat sitting on a box in a cyberpunk street, award-winning photo, saturated, colored, colors, 100mm, sharp, high res`" raises questions. Secondly, a suitable framework for text analysis needs to be established to critically evaluate prompt creativity. Lastly, the aim is to assess whether deep learning techniques can accurately predict prompt creativity, similar to existing methods for assessing creativity.

For the current research project, the following research question and subquestions will be used to guide the research:

RQ: *"How can NLP methods be used to assess the creativity of short-text prompt data in AI-art generation?"*

This research question will be guided by the following subquestions:

SQ1: *How does the level of detail and specificity in prompts impact the perceived creativity of the generated images?*

SQ2: *What role do elements like artist names and aesthetic quality modifiers play in enhancing or diminishing the perceived creativity of a prompt?*

Overall, the goal is to better understand the role of prompting in text-to-image art generation by using NLP methods, and to find out in what way practitioners use their creativity within prompts to get desired results.

Chapter 2

Literature

2.1 Creativity

Machine creativity is an emerging field that explores the capacity of artificial intelligence systems to generate novel and valuable outputs. It encompasses various domains, such as music composition, visual art, and storytelling, where machines are programmed to exhibit creative behavior. With advancements in deep learning and generative models, machines can now mimic human creativity by generating original and aesthetically pleasing compositions. By using large datasets and complex algorithms, these systems can learn patterns, explore new possibilities, and produce outputs that exceed traditional boundaries. One of such applications of machine creativity are deep learning text-to-image models like Dall-E, Midjourney, and Stable Diffusion. These models have become more popular and widespread within a short period of time. They have ignited many discussions on the internet over the originality and creativity of the generated artworks (Roose, 2022). Since these art-generator models have been trained on data which include the original artworks of various artists without giving consent to the inclusion of their art in the training data, it has also led to lawsuits (Edwards, 2023) and concerns over copyright and ownership (Strange, 2022). These text-to-image models hold immense potential for pushing the boundaries of human imagination and challenging conventional notions of what we consider to be creative. However, due to the nature of how these models work, it also raises questions about the nature of creativity, the role of human input, and the implications of machines producing creative works on society.

2.1.1 Defining Creativity

To investigate the creative aspects of the text-to-image generation process, it is important to establish a clear definition of creativity. Throughout the years, the concept of creativity has sparked considerable debate, resulting in multiple definitions by researchers and authors across different disciplines. This section provides an overview of various definitions of creativity from different fields and authors. The goal is to identify a widely accepted definition that aligns with the objectives of this thesis project.

Creativity is a multi-faceted concept with various definitions across different contexts and areas of research. According to Dictionary.com, creativity is defined as "the ability to transcend traditional ideas, rules, patterns, relationships, or the like, and to create meaningful new ideas, forms, methods, interpretations, etc.; originality, progressiveness, or imagination."

According to (Cropley, 2011), the definition of creativity has evolved over time, with the central idea of novelty remaining consistent across different definitions.

He distinguishes between two meanings of creativity: (1) production of novel products, in the sense that they have only recently come into existence, and (2) production of relevant and effective novelty, which he refers to as "sublime" creativity. The author also notes the distinction between "minor" creativity, which extends the known, and "major" creativity, which goes beyond the known and can lead to a paradigm shift.

In the context of AI-generated art, the introduction of these art generators has caused a paradigm shift and is analogous to the industrial revolution according to (Newton and Dhole, 2023). The act of producing a piece of art through the use of a prompt can be considered both "minor" and "major" creativity. The output produced by the algorithm is based on the data it has been trained on and extends what it knows, but is also novel for each instance of generation, therefore going beyond what is already known. Additionally, (Cropley, 2011) discusses the relationship between creativity and intelligence, stating that conventional intelligence relies on recognition, recall, and reapplication, while creativity requires the production of novelty through departure from facts and finding new solutions.

In (Runco and Jaeger, 2012), the authors revise the Standard Definition of creativity. They go over the history of creativity studies, rooted in the 1930s, 40s, and 50s, and make a correction to the Standard Definition of creativity, which is defined as follows: "Creativity requires both originality and effectiveness". Originality is required to be creative: if something is not unusual, novel, or unique, it is commonplace, mundane, or conventional. When something is not original, it is not creative. However, originality is not sufficient to deem something as creative. Ideas that are just original might also be useless, hence creativity also requires effectiveness, or usefulness according to the Standard Definition. (Feist, 2021) defines this second component of creativity not as effectiveness or usefulness, but as meaningfulness: “to be classified as creative, thought or behavior must also have meaning to other people”.

(Boden, 1990) states her definition of creativity as follows: "the ability to come up with ideas or artifacts that are new, surprising, and valuable". This definition aligns with (Cropley, 2011) and (Runco and Jaeger, 2012)’s central notions of novelty and value being key properties of creativity in different definitions of the concept. Boden’s definition has been adopted for the study of machine creativity, as discussed in (Franceschelli and Musolesi, 2022b) and (Franceschelli and Musolesi, 2022a), due to its consideration of the three crucial factors: value, novelty, and surprise. Value refers to the comparison of an object to others in its class based on factors such as utility, performance, or attractiveness, as stated in (Franceschelli and Musolesi, 2022b). Novelty represents the deviation of an artifact from previously established artifacts within its class. Lastly, surprise is defined as the degree of discrepancy between the real input and the observer’s expectation, according to (Berlyne, 1973). Considering the widespread adoption of Boden’s definition of creativity, this project will adopt the same definition to refer to creativity.

2.2 Assessing Creativity

Having defined creativity, the next step is to explore methods for assessing this concept. Creativity assessment is the process of evaluating and measuring an individual's creative thinking skills and abilities. It aims to gauge their capacity to generate unique ideas, think outside the box, and approach problems from innovative angles. Various methods are employed to assess creativity, including tests, questionnaires, and performance-based tasks that encourage divergent thinking and originality. These assessments typically consider factors such as fluency, flexibility, originality, and elaboration. Creativity assessment plays a crucial role in identifying and nurturing creative talent, guiding educational programs, and supporting innovation in various fields, such as arts, sciences, and business. This section will discuss diverse methods of creativity assessment and explore their application in the context of text.

2.2.1 Creativity Assessment Techniques

Torrance Tests of Creative Thinking

One way to assess creativity is through the use of the Torrance Tests of Creative Thinking (TTCT). The TTCT is a widely-accepted measure of creativity, pioneered by E. Paul Torrance, a prominent figure in the field of creativity assessment (Torrance, 1972). Extensive research, as demonstrated in a review by (Kim, 2006), has validated the TTCT as a reliable and valid tool for assessing creativity and divergent thinking skills. The TTCT consists of two parts: a verbal test and a figural test, each designed to evaluate creativity across different domains. In the verbal test, for instance, one subtest prompts the test taker to generate uncommon uses for ordinary objects, such as a brick or a newspaper. Another subtest requires the test taker to speculate on the causes or consequences of specific events, like envisioning what would happen when a cow jumps over the moon. The figural test assesses fluency, originality, elaboration, and abstractness through activities such as picture construction, picture completion, and lines tasks (Lubart et al., 2022).

The TTCT has gained popularity in studies related to child behavior and development due to its demonstrated reliability and consistency over time. Individuals who achieve high scores on the TTCT tend to exhibit higher levels of creativity throughout their lives, indicating that the test serves as a valuable indicator of long-term creative potential.

Consensual Assessment Technique

Another popular method of assessing creativity is Consensual Assessment Technique (CAT) (Amabile, 1982). The CAT method is specifically designed to examine the social and environmental influences on creativity. In this assessment method, experts in a particular field evaluate and rate the creativity of products within that field. These experts use their own criteria and standards to judge the level of creativity exhibited by each product, and their ratings are then averaged to obtain a final score.

The CAT addresses the challenge of defining an objective criterion for creativity by using the collective judgment of the domain experts. By combining the perspectives of multiple judges, it provides a nuanced understanding of creative achievements. Additionally, the CAT method has demonstrated its reliability, even when a relatively small number of judges are involved (Baer and McKool, 2009).

One of the advantages of the CAT method is its applicability across different cultures, making it a valuable tool for studying creativity in multicultural contexts. Research by (Hennessey et al., 2008) highlights the cross-cultural reliability of the CAT in assessing creativity, indicating that it produces consistent results across diverse cultural backgrounds. This characteristic of the CAT contributes to its versatility and enhances its usefulness in exploring creativity from a global perspective.

It is however important to acknowledge that the CAT method also has limitations. A major bottleneck lies in the need to acquire experts with sufficient domain expertise, which can be challenging. Finding individuals who possess the necessary knowledge and experience to evaluate creative products accurately is important for ensuring the validity of the assessments. Nevertheless, despite this bottleneck, the CAT method has demonstrated its reliability and effectiveness in measuring creativity in various fields and cultural contexts.

Alternate Uses Task

One more popular method of assessing creativity is the Alternate Uses Task (AUT) (Guilford, 1967). The AUT is a widely used psychological test designed to assess a person's divergent thinking abilities. Divergent thinking refers to the ability to generate multiple solutions, ideas, or responses to a given problem or task. Guilford, the author of the AUT, believed that divergent thinking was a critical component of creativity and innovation.

The AUT is a simple and straightforward test that can be administered individually or in a group setting. The test consists of presenting individuals with a common item and asking them to generate as many alternative uses for that item as they can within a specified time limit. For example, if the object is a brick, participants might generate alternative uses such as a paperweight, a doorstop or a bookend.

The primary goal of the AUT is to measure the fluency, flexibility, and originality of a person's thinking. Fluency refers to the total number of alternative uses generated by an individual. A higher number of responses suggests a greater ability to produce a large quantity of ideas. Flexibility refers to the different categories or types of alternative uses generated. A more diverse range of uses demonstrates cognitive flexibility and the ability to think outside the box. Originality is a measure of how unique or uncommon the generated alternative uses are. It reflects the individual's ability to produce new and innovative ideas. The AUT is often used in research settings to study various aspects of creativity and cognitive processes. Beyond research, the AUT has practical applications in fields such as education and personnel selection. In educational settings, for example, it can be used to assess and develop students' creative thinking skills. Benefits of using the AUT is that the test is easy to administer and can be adapted to different context. Drawbacks include that it may not be able to capture different aspects of creativity and it may not reflect real-world creative performance or potential.

2.2.2 Automating Creativity Assessment

Though subjective creativity scoring methods such as the CAT and AUT have been valuable in creativity research, these approaches are hampered by a significant limitation: the reliance on human raters. The involvement of human raters not only introduces subjectivity into the evaluation process but also incurs substantial labor costs. To address these limitations of subjective scoring and labor costs, research has been done to automate the creativity assessment process (Acar and Runco, 2014; Dumas, Organisciak, and Doherty, 2021). One tool for automating creativity assessment Latent Semantic Analysis (LSA) which quantifies the semantic distance between concepts within a semantic space, first introduced by (Landauer, Foltz, and Laham, 1998). For instance, the semantic distance between 'hammer' and 'nail' would be small, indicating high similarity, while the distance between 'hammer' and 'tissue' would be larger, indicating lower similarity. This aligns with the associative theory of creativity, which posits that creative thinking involves making connections between seemingly remote concepts.

However, while LSA and other tools like Linguistic Inquiry and Word Count (LIWC) provide valuable insights, they may not fully capture the complexity of creativity. For example, LIWC categorizes each word in a text into one of 80 predefined dictionaries and calculates the percentage of words in each category. While this can predict psychologically relevant outcomes, it may not be the most suitable method to assess creativity (Zedelius, Mills, and Schooler, 2018; Ahmed and Feist, 2021).

In contrast, the SemDis platform presented by (Beaty and Johnson, 2020) addresses many of these issues. SemDis automates creativity assessment using NLP by calculating the semantic distance between a response and a given item. The authors argue that semantic distance can capture the novelty of ideas generated in tasks such as the AUT, and compare different methods of combining semantic distances across words. They demonstrate that SemDis can provide reliable and valid scores of novelty that correlate with human ratings. SemDis works by calculating the semantic distance between the response and the given item, in either a creative or common condition. The creative condition requires out-of-the-box thinking for the responses, whereas the common condition requires uses for the item that are commonplace. For example, when given the item 'brick' and the response 'outdoor dog bed' within the creative condition, SemDis will return a semantic distance of 0.97994, meaning that the given response scores high on semantic distance, and is therefore 'more creative' than responses that return a lesser semantic distance score. SemDis can thus be used as a substitution of human raters of textual data. However, the authors warn that meaningless responses yield high semantic distance. Therefore, users are encouraged to carefully screen the data before calculating the semantic distance. For the prompt dataset, SemDis can prove to be useful tool for quickly analyzing the semantic distance between different prompts. However, since SemDis emulates the AUT, this also poses a challenge, since the prompt dataset is not structured nor created with the AUT in mind.

In conclusion, while tools like LSA and LIWC provide valuable insights into creativity, SemDis in particular stands out for its ability to automate creativity assessment and its correlation with human ratings. This makes SemDis a crucial tool for this study and for the field of creativity research.

2.3 Text-to-Image Algorithms

2.3.1 Art Generation

The field of deep learning has seen incredible progress in the synthesis of images guided by text since the introduction of Generative Adversarial Network (GAN)s in 2014 and Google's Deep Dream in 2015 (Goodfellow et al., 2014; Mordvintsev, Olah, and Tyka, 2015). However, it was the launch of OpenAI's CLIP in 2021 that really accelerated advancements in text-to-image generation (Radford et al., 2021).

CLIP, which is a model that pairs images with their corresponding visual categories, was trained on a great amount of images and text from the internet. This extensive training enabled CLIP to learn a broad range of visual concepts and associate them with their respective labels. This ability to link natural language with images has made CLIP a valuable tool in generative systems.

In the realm of text-to-image generation, CLIP's first significant application was in GAN-based image generation systems. When used as a discriminator in a generative deep learning architecture, CLIP can guide the generator to create digital images that closely match a given text prompt. The release of CLIP prompted AI enthusiasts to develop GAN+CLIP based systems specifically for generating digital art. One such system, "Big Sleep," combined a GAN called BigGAN with CLIP (Brock, Donahue, and Simonyan, 2018). This inspired the creation of an even more powerful system that paired VQGAN with CLIP, which became a standard technique for generating artworks in 2021 until it was surpassed by diffusion-based systems.

Diffusion models, which were first introduced in 2015, are a type of generative model that can generate new samples by simulating a random walk through a latent space (Sohl-Dickstein et al., 2015). The idea behind diffusion models is to start with a simple distribution, such as a Gaussian, and then gradually transform it into a complex distribution that represents the data. This transformation is achieved through a series of small, reversible steps, each of which slightly modifies the distribution. This process is similar to a particle undergoing Brownian motion, or diffusion, which is why these models are named diffusion models.

In the context of image generation, diffusion models can be used to generate new images by starting with a random noise image and then gradually transforming it into a realistic image through a series of diffusion steps. Each step is guided by a neural network, which is trained to predict the next step given the current state of the image.

The popularity of diffusion models for art generation began to rise in 2021 with the introduction of models like Dall-E and Stable Diffusion. These models improved upon the original diffusion models by introducing techniques to make the diffusion process more stable and efficient, allowing for the generation of higher-quality images.

It was specifically the release of OpenAI's DALL-E in 2021 that really brought diffusion models into the mainstream (Ramesh et al., 2021). DALL-E is a variant of GPT-3, a powerful language model, that was trained to generate images from textual descriptions. Unlike previous models, which used GANs in combination with CLIP. However, the specifics of DALL-E's architecture and training process are proprietary, meaning they are not publicly disclosed or available for use by the broader AI community. Additionally, in order to use Dall-E, users are required to pay a monthly fee.

Another popular model and a contender to OpenAI's Dall-E is Stable Diffusion (Mostaque, 2022). The defining feature of Stable Diffusion is its open-source nature. The model's architecture, training process, and codebase are all publicly available, which has democratized access to this advanced technology. This transparency encourages a collaborative approach to development, allowing researchers and developers worldwide to use, modify, and build upon the model. This has led to a lot of enhancements and applications, contributing to the model's popularity and its status as a mainstream tool in text-to-image generation. Futhermore, the open-source nature of Stable Diffusion has cultivated a vibrant community of AI enthusiasts and professionals. This community continually contributes to the model's development, pushing the boundaries of what's possible in the realm of text-to-image generation. As such, Stable Diffusion remains a highly relevant and influential model in the field, alongside DALL-E and Midjourney. In this study, the focus will be on art generated through Stable Diffusion.

2.3.2 Prompting

In AI art generation, prompting refers to the process of providing input to the AI model in order to generate a desired output. This input can take many forms, depending on the specific model and the type of art it is generating. For example, a model that generates images might be prompted with tags such as colors, shapes, or objects. A model that generates music might be prompted with genre, instrumentation, or tempo (Dhariwal et al., 2020).

In this study, the focus will be on textual prompts that are provided for art generation algorithms. Prompting allows the user to control the type of content generated by the AI model, thus enabling them to create more art that meets their creative vision. In (Rombach, Blattmann, and Ommer, 2022) prompting is described as a practice in which carefully selected and composed sentences are used to achieve a certain visual style in the synthesized image". Prompt engineers are individuals who are responsible for designing the input that is provided to AI art generation models. The goal of a prompt engineer is to create input that is both specific and engaging, in order to produce the best possible output from the model. This can be a challenging task, as the AI model may have a very different way of interpreting the input than a human would. Prompt engineers must carefully craft their input in order to effectively guide the model and produce high-quality art. An example of such a prompt could be "`a purple night sky over green grassy hills, oil painting, van Gogh`", which would generate the image as shown in 2.1. In some cases, prompt engineers may also be responsible for tuning the model itself in order to improve its performance and generate even better output. Since the AI-art explosion of 2022, various prompting guides and online prompt stores have spawned, signifying the importance of the prompt in the art generation pipeline.

In a three-month ethnographic study by (Oppenlaender, 2022), the authors provides a taxonomy of prompts. After researching online art generation communities, they found that the aesthetic qualities and subjective attractiveness of images can be improved by adding certain keywords and key phrases, referred to as prompt modifiers, to the text input. They found six categories of prompt modifiers:

- Subject terms (indicate the desired subject to the text-to-image system)

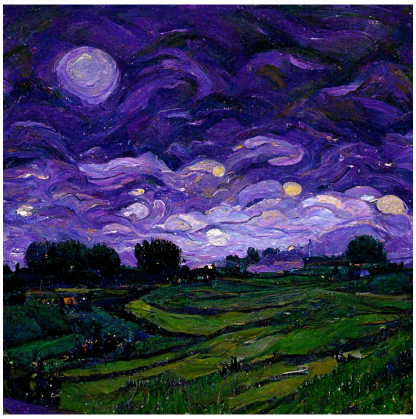


FIGURE 2.1: An image generated with the prompt "a purple night sky over green grassy hills, oil painting, van Gogh", using Mid-Journey

- Image prompts (provides the text-to-image system a target for synthesis of the image)
- Style modifiers (added to a prompt to achieve a certain style)
- Quality boosters (added to the prompt to increase aesthetic qualities)
- Repetitions (to potentially strengthen the associations formed by the generative system)
- Magic terms (to introduce randomness to the image that can lead to surprising results)

From the six modifiers the style and quality modifiers are the most commonly used. In (Liu and Chilton, 2022) , the authors also conducted a study exploring what prompt keywords and model hyperparameters can help produce coherent outputs. The authors use a series of five experiments to address key questions around prompt engineering. They test different phrasings of the prompt, different random initializations, vary the number of iterations for optimization, explore styles as a parameter of the prompt to understand the breadth of styles the system can reproduce, and explore subjects as a parameter of the prompt to understand how subject and styles interact with each other. They found significant differences between the quality of generations that fell into different categories of style as well as subject and style. In their study, the authors utilized the VQGAN+CLIP text-to-image generation model to conduct a series of experiments. The results of these experiments were distilled into a set of design guidelines intended to assist users in effectively utilizing art generation models.

(Pavlichenko, Zhdanov, and Ustalov, 2022) presents a human-in-the-loop approach to learning the most useful combination of prompt keywords for text-to-image models using a genetic algorithm. The authors propose a method for evaluating the quality of generations produced by different prompt templates and show that their approach can improve the aesthetic appeal of images depicting the same descriptions. Examples of prompt keywords are 'trending on ArtStation', or 'highly detailed'. They found that adding their prompt keywords increases the aesthetic quality of the produced images better than the most popular prompt keywords from the art generation community. The data and code used are publicly available, and is therefore useful for the research conducted in this thesis.

Lastly, more research on image prompts alongside text prompts for text-to-image models was conducted by (Qiao, Liu, and Chilton, 2022). The paper discusses how image prompts alongside text prompts can improve subject representation in AI generated art. The authors conducted an annotation experiment to quantify the effect of initial images on generations of abstract, concrete plural and concrete singular subjects. They found that initial images improved subject representation across all subject types with the most noticeable improvement in concrete singular subjects. Though image prompts are not within the scope of this thesis project, their research shows how multimodal prompting can provide better subject representation in the output.

2.4 NLP and Text Analysis

NLP is a subfield of AI that explores how computers can understand and manipulate natural language text or speech to perform useful tasks. NLP researchers aim to learn how humans understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform desired tasks (Goodfellow, Bengio, and Courville, 2016).

With the proliferation of online documents that contain short-text data, such as texts generated through social media platforms like Twitter, Instagram, and Facebook, the need for a special type of text analysis arises. Short-text data often contain slang, emojis, misspellings, abbreviations, and grammatical errors, which present challenges for traditional text analysis methods (Ahmed et al., 2022). As a result, short-text clustering has become an essential task for clustering various unlabeled texts into meaningful clusters.

2.4.1 Topic Modeling

Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. Traditional topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003), are probabilistic methods that analyze the words of the original texts to discover the themes that run through them. However, when it comes to short-text data, traditional topic modeling techniques may not perform as well due to the sparsity and high dimensionality of the data.

Gibbs Sampling Dirichlet Mixture Model (GSDMM)

Gibbs Sampling Dirichlet Mixture Model (GSDMM) is a Bayesian nonparametric method specifically designed for short-text corpora (Yin and Wang, 2014). It assumes a mixture of Dirichlet distributions and uses Gibbs sampling to estimate the model parameters. GSDMM has been shown to outperform Latent Dirichlet Allocation (LDA) in terms of topic coherence for short-text corpora, making it a suitable choice for such data.

2.4.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a model in the field of NLP. Introduced by researchers at Google AI Language in 2018, BERT has significantly influenced the landscape of NLP research and application (Devlin et al., 2018a).

BERT is based on the Transformer architecture, which uses an attention mechanism that weighs the influence of different input words on each output word (Vaswani et al., 2017). Unlike other models that processed words in an input sequence one at a time in order (either left-to-right or right-to-left), BERT takes into account the full context of a word by looking at the words that come before and after it. This bidirectional approach allows BERT to understand the context and semantic meaning of a word within a sentence more accurately.

BERT has been pre-trained on a large corpus of text, which includes the entire English Wikipedia (2.5 billion words) and BooksCorpus (800 million words). This pre-training step involves learning to predict missing words in a sentence and learning to predict whether a sentence follows another sentence. The pre-trained BERT model can then be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks, without requiring big task-specific architecture modifications.

BERTopic

BERTopic is a topic modeling technique that uses the power of transformers and class-based Term Frequency-Inverse Document Frequency (TFIDF) to create dense clusters representing topics (Grootendorst, 2022). It uses BERT embeddings, which capture the semantic meaning of words and their context within a document, to transform documents into embeddings. These embeddings are then clustered using a density-based algorithm to group similar documents together. Each cluster represents a potential topic. BERTopic has been shown to generate more coherent and distinct topics compared to traditional topic modeling techniques, and is capable of handling large text corpora and identifying fine-grained topics due to its use of dense vector representations.

In conclusion, both GSDMM and BERTopic provide efficient solutions for topic modeling of short-text data, offering more coherent and distinct topics than traditional methods. These advanced methods are particularly useful for later analysis of creativity in short-text data, as they allow for topics to be found within the prompt data. A more detailed description of the inner workings of GSDMM and BERTopic are provided in the Methods section 3.

Chapter 3

Methods

This chapter shows the methods used in this study. It begins by detailing the data used for the project, namely the DiffusionDB dataset, and how it was collected. The chapter then moves on to explain the data clustering methods used, which are the Gibbs Sampling Dirichlet Mixture Model (GSDMM) and BERTopic. The chapter concludes with an explanation of SemDis, an online platform used to compute semantic distance, which is used in this project to generate creativity scores for the prompt dataset.

3.1 Data

To evaluate the creativity of prompts used in generating art, a dataset of prompts is required for this project.

AI art is usually generated through the websites that provide these services, such as Dall-E (Ramesh et al., 2022) and Imagen (Saharia et al., 2022), or through the use of writing prompts on specific Discord servers, which is the case for Stable Diffusion (Rombach et al., 2022) and Midjourney. As these text-to-image have become more popular over the last few years, more and more data has been collected from these sources, one of which is the DiffusionDB dataset (Wang et al., 2022), which contains data collected from the Stable Diffusion Discord channel. For the purpose of this thesis project, the DiffusionDB dataset will be worked with, as it contains raw, authentic data from real users, with useful metadata that can be used for analysis.

3.1.1 The DiffusionDB dataset

For this research project, the DiffusionDB-2M dataset will be used, which is part of the DiffusionDB dataset (Wang et al., 2022). The DiffusionDB dataset contains over 14 million images and prompts generated with the Stable Diffusion model. The text in the dataset is mostly English, but also contains other languages such as Spanish, Chinese, and Russian. According to the dataset description, the data was collected by scraping user-generated images on the official Stable Diffusion Discord server. The Stable Diffusion data was chosen because it was at the time the only open-source large text-to-image generative model, and all generated images have a CC0 1.0 Universal Public Domain Dedication license that waives all copyright and allows uses for any purpose. Furthermore, the Stable Diffusion Discord server was chosen, because it is public, and it has strict rules against generating and sharing illegal, hateful, or images not suitable for work (NSFW). However, the dataset might still contain generated harmful images that were not detected by the NSFW filter in the server or removed by the server moderators. The dataset contains images generated with prompts by users from the Discord server. The server is open for everyone, meaning that the prompts are generated at any level of prompting expertise.

3.1.2 DiffusionDB Data Collection

The data in the DiffusionDB dataset is collected from the Stable Diffusion discord channel, using DiscordChatExporter (Tyrrrz, 2023). This tool saves chat messages in a HTML file format. For the DiffusionDB dataset, the creators of the database focused on channels in which Discord users can request a bot to run Stable Diffusion to generate images with the prompt provided by the user. The usernames of the Discord users have been anonymized by only including their SHA256 hashes in the dataset distribution. Since some images might have made it past Discord's NSFW filter, the creators of the dataset also applied their own NSFW classifier to predict NSFW images and prompts. The original DiffusionDB dataset contains over 14 million images and 1.8 million unique prompts. For prompt analysis, the creators also constured DiffusionDB2-M, which is a subset of DiffusionDB which includes 2 million images and about 1.5 million unique prompts. For this research project, the DiffusionDB2-M dataset will be used. Since we are interested in the prompts used, we will use the metadata file of DiffusionDB-2M, which excludes the images in the dataset. Each instance in the dataset represents a prompt-image pair prompted by a user in the Stable Diffusion Discord channel.

3.2 Data Clustering

The data requires to be clustered before SemDis can be applied to the data. This will be done by applying two methods of topic modelling to the data, namely GSDMM and BERTopic. After having found clusters with similar prompts, these clusters will then be labeled, which will serve at the 'item' to which the prompts within a cluster will be compared with in SemDis.

3.2.1 GSDMM

The GSDMM is a topic modelling method that can deal well with the scarce and high-dimensional issue of short-text data, making it an attractive approach for finding topics within the prompt dataset (Yin and Wang, 2014). GSDMM is a modified version of LDA that assumes that a document (such as a text, tweet, or in this case, a prompt) has just one topic. This is different from LDA, which assumes that a document can have more than one topic. The model combines the Dirichlet Multinomial distribution and Gibbs sampling to effectively group similar data points together. The core formulas used in the GSDMM algorithm include the updating rule of Gibbs sampling and the Dirichlet Multinomial distribution. GSDMM can automatically determine the amount of clusters, and has a quick convergence time, making it a better method than LDA when it comes to finding topics from short text data (Mazarura and De Waal, 2016).

The Movie Group Process

To understand the inner workings of GSDMM, the authors of the <paper> use an analogy known as the Movie Group Process (MGP) to explain how the algorithm works. Consider a situation in which a teacher wants to organize their students into film discussion groups. The teacher believes that students who have watched similar films will have more to discuss within the group. This situation is analogous to the GSDMM algorithm, grouping documents (the students) by shared words (the movies).

In order to find the right discussion groups, the teacher asks the students to write down the names of movies they have watched. The list of movies is constrained by the students' memory and a time limit, which results in a short list of movies, typically containing recently watched films.

The students (documents) are randomly assigned to K tables (potential clusters). The subsequent process of populating the tables (clusters) follows two rules:

1. A student prefers to choose a table with more students (documents in the same cluster tend to be more similar).

2. A student prefers to choose a table where the other students have watched similar movies (documents sharing more words are more likely to belong to the same cluster).

As this process continues iteratively, after each iteration, some tables (clusters) will attract more students and grow, while others may lose all their students and vanish. Eventually, only some tables remain, each representing a group where the students share similar movie interests.

The two above rules relate to goals of clustering, namely completeness and homogeneity. This means that completeness aims to for all members of a ground true group to be assigned to the same cluster, while homogeneity aims for a cluster to contain only members of a single ground true group. Rule 1 of the MGP results in a higher completeness, since it leads to popular tables to be more popular, which the authors of the paper refer to as 'the rich get richer' property. Rule 2 of the MGP results in a higher homogeneity, because it it leads the students in the same table to be more similar to each other. At the same time, these rules also represent the parameters within the GSDMM model, which are α and β. The parameter α influences the form of the probability distribution, or more specifically, the likelihood of a document being assigned to a cluster. In the MGP analogy, this means the likelihood of a student picking a table. The parameter β refers to the resemblance of one text to another. In the MGP analogy this is the likelihood that a student would join a table with similar movie interests.

In the context of GSDMM, these remaining tables represent the final clusters of documents, and the students at each table represent the documents that share common words. This process of iteratively shifting documents between potential clusters is similar to the Gibbs Sampling process used in GSDMM.

The DMM explained

To understand how GSDMM works, it is necessary to understand how a Dirichlet distribution works. A Dirichlet distribution is a Beta distribution over many dimensions, which in this case entails documents. A Beta distribution is a distribution of probabilities that represent the prior state likelihood of a document joining a cluster as well as the similarity of that document to the cluster. Two parameters, α and β influence the shape of this Beta distribution. The two rules of the MGP and the parameters of the model are the same:

- α: This parameter affects the shape of the distribution. This parameter represents the probability that a document will be grouped in a cluster. Recall the first rule of the MGP: *A student prefers to choose a table with more students (documents in the same cluster tend to be more similar)*. In other words, α represents the probability of a student choosing a table.

- β: This parameter also affects the shape of the distribution. Beta represents the similarity of words within a document compared to words in another document. Recall the second rule of the MGP: *A student prefers to choose a table where the other students have watched similar movies (documents sharing more words are more likely to belong to the same cluster)*. In other words, β represents the probability a student will join a table with similar movie interests. Note that when β is 0, the student will only choose tables with movies they have in common.

The Dirichlet Mixture Model (DMM) model then also uses the following parameters:

- ϕ: ϕ is the multinomial distribution of clusters over words such that $p(w|z = k) = \phi$, where w = words, z = cluster label, and k = the number of clusters.
- θ: θ is the multinomial distribution taking α into account, so that $p(d|z = k) = \theta$, where d = document, z = cluster label, and k = the number of clusters.

All of this is then combined into the following probability where document d is generated by a cluster k, with the assumption of Dirichlet priors, meaning that the priors follow the aforementioned Dirichlet distribution. This is all bundled in the following formula:

$$p(d|z = k) = \prod_{w \in d} p(w|z = k) \tag{3.1}$$

Now, the Gibbs Sampling part of GSDMM refers to the method of iterating over conditional distributions of variables, whose distribution over states converges to the true distribution after several iterations. In this case, that means that clusters are reassigned based on a conditional distribution after each iteration.

GSDMM implementation

To apply the GSDMM algorithm on the cleaned DiffusionDB dataset, the `GSDMM` python package will be used (rwalk, 2017). This python implementation is modeled after the original paper by (Yin and Wang, 2014), using the principles mentioned above. Initialization of a GSDMM model is shown in 3.1, with `K` being the max amount of clusters that will be found, `alpha` = α , `beta` = β , and `n_iters` being the amount of iterations of the algorithm.

LISTING 3.1: the GSDMM clustering algorithm

```
mgp = MovieGroupProcess(K=10, alpha=0.1, beta=0.1, n_iters=30)
vocab = set(x for doc in docs for x in doc)
n_terms = len(vocab)
y = mgp.fit(docs, n_terms)
```

3.2.2 BERTopic

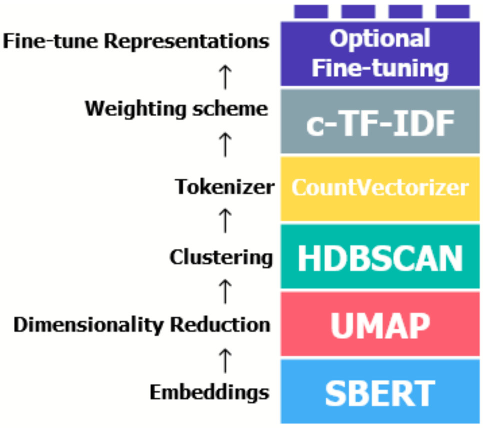


FIGURE 3.1: BERTopic visual overview. Taken from (Grootendorst, 2022)

BERTopic is an algorithm that leverages BERT and transformer embeddings from HuggingFace in order to generate topics using state-of-the-art embeddings (Grootendorst, 2022). BERT is a transformer-based language model developed by Google that is pre-trained on large amounts of unlabeled text data. It is designed to understand and generate human language, achieving state-of-the-art performance on various natural language processing tasks (Devlin et al., 2018b). BERTopic uses a class-based variation of TFIDF named Class-based Term Frequency-Inverse Document Frequency (cTFIDF) to extract coherent topic representations. The first stage of BERTopic involves using sentence-BERT (sBERT) to generate document embeddings. These high-dimensional vectors capture the semantic meaning of documents in a corpus of data. Afterwards, Uniform Manifold Approximation and Projection (UMAP) is applied. UMAP is a dimensionality reduction technique that aims to preserve the global structure of high-dimensional data while revealing its underlying patterns in a lower-dimensional space. Thus, by applying UMAP, it reduces the dimensionality of the sBERT embeddings, which reduces the computational load.

Then, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is applied to the UMAP embeddings to create a cluster of documents, with each cluster representing a topic. HDBSCAN is a density-based clustering algorithm that identifies clusters of varying densities in data and assigns data points to clusters based on their local density (McInnes and Healy, 2017). The documents within a cluster are semantically similar to each other. As a last step, BERTopic applies a cTFIDF vectorization to the documents within each cluster. This generates a list of terms for each topic, showing the importance of each term. This is determined by the frequency of a document within the topic and the infrequency of its appearance in other topics. Figure 3.1 visually lays out the stages in the BERTopic procedure.

BERTopic implementation

BERTopic is initialized as seen in 3.2. First, the data is transforme into a list of documents. Then the BERTopic model is initalized with the n-gram range set to (1, 1), indicating that only unigrams will be taken a look at. The model was fit to the data using the `fit_transform` method, which returns the predicted topics for each document. No specific sentence model is specified, meaning that BERTopic will work with the out-of-the-box pre-trained transformer model "all-MiniLM-L6-v2".

LISTING 3.2: BERTopic initialization

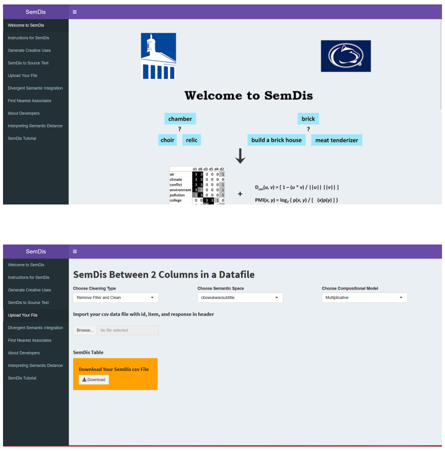
```
from bertopic import BERTopic

# Create documents
docs = dataset['prompt_tokenized'].values.tolist()

# Initialize BERTopic
topic_model = BERTopic(verbose=True, n_gram_range=(1, 1))
topics, _ = topic_model.fit_transform(docs)
```

3.3 SemDis

SemDis is an open online platform to efficiently compute semantic distance, which emulates the Alternative Uses Task (AUT) and has shown correlation with creativity measures(Beaty and Johnson, 2020), as previously described in 2.2.2. SemDis uses NLP to quantify the semantic relatedness of texts by using several semantic models, including Global Vectors for Word Representation (GloVe), an unsupervised learning algorithm for obtaining vector representations for words, and Continuous Bag of Words (CBOW), a word embedding model for predicting word from context.



Welcome to SemDis

$D_{cos}(u, v) = [\ 1 - (u * v) / ||u||\ ||v||\ ]$

$PMI(x, y) = \log_2 \{\ p(x, y) / [\ (x)p(y)\ ]\ \}$

SemDis Between 2 Columns in a Datafile

Choose Cleaning Type: Remove Filler and Clean

Choose Semantic Space: cbowukwacsubtitle

Choose Compositional Model: Multiplicative

Import your csv data file with id, item, and response in header

Browse... No file selected

SemDis Table

Download Your SemDis csv File

Download

FIGURE 3.2: The SemDis online homepage (top) and upload page (bottom)

SemDis , an accessible online platform, computes semantic distance resembling the AUT and which has shown correlation with creativity measures (Beaty and Johnson, 2020). SemDis leverages NLP techniques to quantify the semantic relatedness of texts through the application of various semantic models, including GloVe, an unsupervised learning algorithm generating vector representations for words, and CBOW, a word embedding model predicting words from context. Semantic distance is calculated by using the common variance from these semantic models, which forms a latent semantic distance variable. The semantic distance scores are derived from the semantic models and are used to predict human judgments of creativity on the AUT. The latent semantic distance variable has been found to reliably and strongly predict human creativity and novelty ratings across a range of creativity tasks, which is why SemDis will be used to generate creativity scores for the prompt dataset. In terms of the calculation of semantic distance, the online platform provides users the option to extract and download a latent factor score comprised of the five semantic models. For this study, the `cbowukwacsubtitle` semantic space will be used to calculate creativity scores. The `cbowukwacsubtitle` semantic space is particularly suited for this project for several reasons.

Firstly, The `cbowukwacsubtitle` model is built on a concatenation of the `ukwac` web crawling corpus and the subtitle corpus. The ukwac corpus contains about 2 billion words from web pages in the .uk domain, and the subtitle corpus contains about 385 million words from movie and TV subtitles. These sources provide a wide range of language use and contexts, making it a good fit for assessing the creativity of prompts, which can also cover a wide range of topics and styles.

Secondly, the `cbowukwacsubtitle` model has demonstrated good performance when it comed down to human judgments of relatedness, lexical decision speed, and free associations. This means that the semantic distances it calculates are likely to align well with human judgments of semantic relatedness, which is beneficial for assessing the creativity of prompts.

Additionally, the `cbowukwacsubtitle` model uses the CBOW architecture, which predicts a target word based on its context. This model is particularly suited for assessing creativity because it captures the semantic relationships between words based on their co-occurrence in the same contexts. This is in line with the assumption that words that share similar contexts also share meaning.

Finally, the `cbowukwacsubtitle` model uses a context window size of 12 words, 300 dimensions, and the most frequent 150,000 words. This means it captures a broad range of semantic relationships and can handle a large vocabulary, making it well-suited for assessing the creativity of a diverse set of prompts.

The SemDis online platform also provides an option between two compositional models to calculate the semantic distance, namely the additive model and the multiplicative model. These models are used to combine individual word vectors into a single compositional vector, which represents the semantic content of the entire response. The additive model simply adds together the vectors of each word in the response. This is a straightforward method, but it can sometimes lead to issues. For example, results from the SemDis paper showed that responses with a higher word count tend to receive lower semantic distance scores, which can be problematic if longer (and potentially more creative) responses are being penalized.

The multiplicative model, on the other hand, multiplies the vectors of each word in the response. This model has been found to perform better than the additive model in predicting human creativity ratings. It also circumvents the issue of word count affecting the semantic distance scores. With the multiplicative model, responses with more words (i.e., more elaborate responses, or in this case longer prompts) receive a boost in semantic distance, which aligns better with human creativity ratings that also tend to favor more elaborate responses. Therefore, when calculating the creativity scores during this study, the multiplicative model will be used.

SemDis requires a `.csv` datafile that has the following column names `id, item, response`. The `id` is a unique index number, `item` corresponds to the 'item' in the AUT, which in this case is the topic label of a group of prompts, and `response` corresponds to the 'response' within the AUT, which will be a prompt in the case of this project. In the next chapter, the necessary preprocessing steps will be shown that result in the required `.csv` format for SemDis.

Chapter 4

Experimentation and Results

In this chapter, the investigation of the research objectives is presented, employing topic modeling and topic labeling methodologies to find meaningful clusters within the dataset. The central aim is to derive creativity scores by using SemDis and to analyze the results. The procedure of experimentation unfolds in three steps:

1. Exploration and Preprocessing of the Dataset: This section starts by thoroughly exploring the initial dataset and carefully applying preprocessing steps to prepare the data for further analyses.

2. Comparative Analysis of Topic Modeling and Labeling Techniques: In this section, two distinct topic modeling methods and two automated topic labeling methods are applied to a subset of the dataset. The outcomes of these techniques are evaluated and compared.

3. Scaling Topic Modeling and Labeling to a Large Dataset: In this final section, the analysis of the data is expanded by using one topic modeling and one topic labeling method on a larger portion of the data. This comprehensive analysis allows for a deeper understanding of the underlying patterns and insights that emerge within the broader data context, which will aid in the understanding of how creativity is being represented in the prompts.

4.1 Exploration and Preprocessing of the Dataset

The dataset used in this study is the DiffusionDB dataset, as outlined in 3.1.1. Before utilizing the dataset with SemDis to derive creativity scores for the prompts, it is necessary to format, clean, and cluster the data. In this section, the data will be explored to understand its structure. Additionally, the data will undergo a cleaning and reformatting process to ensure its suitability for later stages of the study.

4.1.1 Dataset Details

The original DiffusionDB dataset is comprised of a dataframe with dimensions of (2,000,000 × 13) and a total size of 26,000,000 data points. The data is collected from the Stable Diffusion Discord server between 2022-08-07 and 2022-08-19. Figure 4.1 shows the number of entries over this period of time. The dataset consists of 13 columns which are: `'image_name'`, `'prompt'`, `'part_id'`, `'seed'`, `'step'`, `'cfg'`, `'sampler'`, `'width'`, `'height'`, `'user_name'`, `'timestamp'`, `'image_nsfw'`, and `'prompt_nsfw'`.

The dataset holds a total of 2,000,000 prompts, with an average of approximately 197 prompts per user, and a median count of 54 prompts per user. The dataset contains a considerable number of duplicate prompts since users tend to

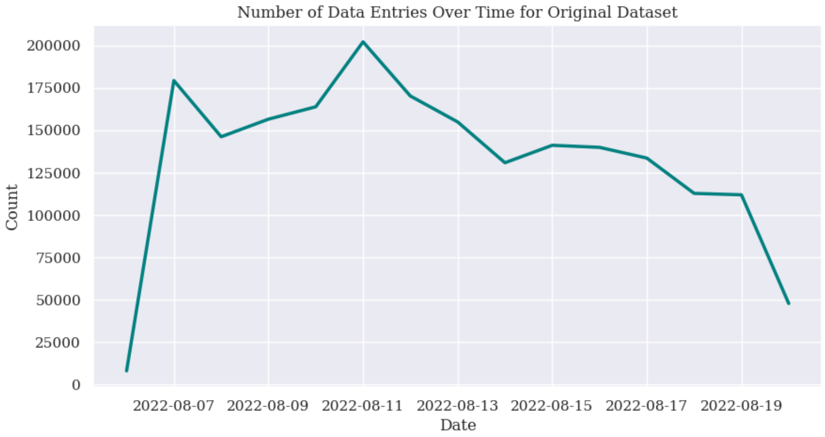


FIGURE 4.1: Entries Over Time in the Original Dataset

use the same prompt multiple times (in order to view different generated outcomes with the same prompt). Out of the total prompts, there are 1,528,514 unique entries.

The dataset includes data from 10,176 unique users. The user with the most prompts submitted 25,551 prompts. In total, there are 645,035 duplicate entries.

The average prompt length in the dataset is 24.56. In 4.2, the original dataset is compared with the resulting datasets after preprocessing. Table 4.1 shows the top 25 term pairs that occur in the same prompt. From this it can be seen that there is an overwhelming amount of artist names and prompt modifiers that are prevalent in the data, showcasing the need for removal in the next preprocessing steps of the data. The original dataset, though containing some duplicates, provides many unique prompts from a large number of distinct users, making this a qualified dataset for this project.

4.1.2 Dataset Preprocessing

The data for this study originates from the DiffusionDB, as previously mentioned. This dataset, available on HuggingFace's website, is loaded as a `.parquet` file before any preprocessing steps are applied to it. The initial data structure includes many unnecessary variables and duplicate instances, which are irrelevant for the further analysis of the creativity in the text. The aim of the preprocessing part is to reduce the original dataset to only the necessary instances that are relevant to this research, as well as to create two separate datasets: one dataset that includes artist names in the prompts, and another dataset that excludes artist names. This enables the evaluation of the influence of including artist names on the creativity of a prompt.

The first step in this preprocessing process reduces the dataframe to only three columns: `prompt`, `user_name`, and `timestamp`. After this step, other irrelevant columns are removed. This includes entries where `user_name` is listed as `deleted_account`, entries with an empty `user_name` field, and prompts that are empty. Cases with multiple `prompts` by a `user_name` are also removed. The first instance is left in the dataset. This is done because having duplicates of prompts prompted by one user is irrelevant, but the same prompt being prompted by different users gives a better idea of the behavior of the prompters in the dataset.

Top 25 Bigrams in the Original Dataset	
Bigram	**Frequency**
(’highly’, ’detailed’)	301,512
(’concept’, ’art’)	230,704
(’!’, ’!’)	224,221
(’greg’, ’rutkowski’)	217,787
(’trending’, ’artstation’)	205,858
(’sharp’, ’focus’)	187,858
(’8’, ’k’)	173,394
(’octane’, ’render’)	152,777
(’digital’, ’painting’)	150,502
(’unreal’, ’engine’)	107,588
(’digital’, ’art’)	102,773
(’4’, ’k’)	100,432
(’alphonse’, ’mucha’)	96,798
(’detailed’, ’digital’)	93,970
(’artstation’, ’concept’)	88,088
(’focus’, ’illustration’)	82,834
(’painting’, ’artstation’)	79,662
(’smooth’, ’sharp’)	75,450
(’cinematic’, ’lighting’)	72,704
(’artgerm’, ’greg’)	61,484
(’oil’, ’painting’)	59,885
(’art’, ’artgerm’)	59,334
(’intricate’, ’elegant’)	57,350
(’art’, ’smooth’)	56,982
(’illustration’, ’art’)	54,441
(’stylized’, ’character’)	50,891
(’anime’, ’art’)	50,780

TABLE 4.1: Top 25 Bigrams in the Original Dataset

After this part of the cleaning process, a new variable `prompt_length` is added to store the length of each prompt.

An extensive list of stopwords is created, which includes over 500 words that aid prompters in the generation of their art, also known as prompt modifiers. These terms do not add anything to the essence of a prompt; rather, they are added to aid with the aesthetics of the generated piece of art. Example prompt modifiers include: `’ambient’`, `’4k resolution’`, `’minimalism’`, `’steampunk’`. Prompters usually also include artist names as a prompt modifier, such as `’greg rutkowski’` and `’alphonse mucha’`. In order to analyze the influence of these artist names on the creativity of a prompt, we create two pre-processed datasets, one which includes the artist names and one without these names. The full list of stopwords and artist names can be found in C. The lists were initially compiled based on engineering guidelines from (Taylor, 2023), and subsequent additions were made manually.

After initializing the stopwords lists, these words are removed from the prompts and saved in a new column `prompt_full_tokens`, which aids in comparison between original and processed prompts. An additional `prompt_tokenized` column

	DiffusionDB	**Cleaned w/ artists**	**Cleaned no artists**
Dataframe shape	(2000000, 13)	(709975, 6)	(724775, 6)
Dataframe size	26000000	4259850	4348650
# of prompts	2000000	709975	724775
Mean prompts per user	196.53	69.81	71.26
Median prompts per user	54.0	31.0	32.0
# of unique prompts	1528514	696053	710919
Avg. prompt length	24.56	16.91	17.0
Median prompt length	22	13.0	13.0

TABLE 4.2: Dataset Comparisons

is created where the prompts undergo further processing, including tokenization and stemming by using methods from the `nltk` Python package (Loper and Bird, 2002). The preprocessing concludes with the removal of any rows where `prompt_full_tokens` is left empty after all modifications since this indicates that the original prompt contains nothing more than prompt modifiers for aesthetic reasons. As a final step, prompts that score high in similarity are removed. Highly similar prompts add noise to the clustering process in the later steps. This is done by partitioning the dataset per user and removing prompts with a cosine similarity higher than 0.7, if the prompt is prompted within a time window of 15 minutes of the original prompt.

4.2 Data Preprocessing Results

After having removed unnecessary columns, rows, duplicate entries, and similar entries from the dataset, the resulting two datasets have been greatly reduced in size. Table 4.2 shows a comparison between the original dataset and the two resulting datasets. As can be seen is that the amount of remaining prompts were reduced to 709975 and 724775 for the dataset including artists and the dataset excluding artists respectively. The reduction was mainly due to the removal of empty, duplicate, or highly similar prompts during the cleaning and filtering process. What can be noted is that the dataset with artist names included contains fewer prompts than the dataset without the artist names. This can be attributed to the fact that the prompts including artist names make prompts more similar and are therefore removed in the later stages of the preprocessing process.

Regarding the columns, the original dataset initially contained 13 columns. After preprocessing, a set of columns were added, which resulted in a total of 6 columns. The column names of the resulting datasets are `prompt`, `user_name`, `timestamp`, `prompt_length`, `prompt_full_tokens`, `prompt_tokenized`. Tables 4.3 and 4.4 show the top 25 bigrams for the resulting datasets after preprocessing. What can be seen is that it becomes more clear what the users are prompting after having removed the stopwords

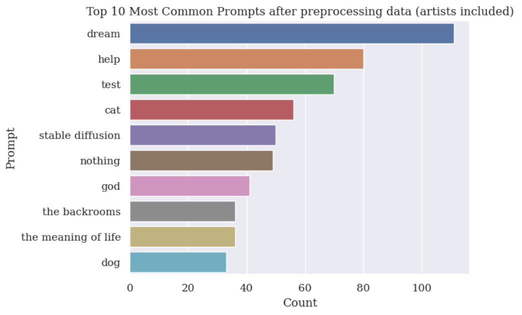


FIGURE 4.2: Top prompts in the cleaned dataset

from the original data, especially when the artist names are also removed. From the data including the artist names it becomes clear which artists’ styles are popular among prompters, signifying the importance of the aesthetic qualities that certain artists hold. Figure 4.2 shows the most common prompts in the pre-processed dataset. Note that duplicate prompts per user are removed and that these prompts are shared by multiple unique users.

Before delving into the next analyses and results, it’s important to understand the nature of the data under examination. The prompts used in this study exist in two forms: full prompts and cleaned prompts. Full prompts refer to the original, unaltered text descriptions used to generate artworks. However, for the purposes of this study, these prompts undergo a cleaning process to eliminate potential noise and redundancy, and make them more suitable for comparison and clustering in SemDis. This cleaned version of the prompts is what is referred to as ′cleaned prompts.

Bigrams with artist names	
Bigram	**Frequency**
(′greg′, ′rutkowski′)	41593
(′alphonse′, ′mucha′)	19296
(′artgerm′, ′greg′)	12810
(′rutkowski′, ′alphonse′)	10775
(′makoto′, ′shinkai′)	7953
(′ilya′, ′kuvshinov′)	7411
(′studio′, ′ghibli′)	6529
(′james′, ′jean′)	6077
(′smooth′, ′artgerm′)	6020
(′black′, ′white′)	5752
(′sci′, ′fi′)	5256
(′craig′, ′mullins′)	4404
(′donald′, ′trump′)	4265
(′movie′, ′still′)	3985
(′global′, ′illumination′)	3865
(′ruan′, ′jia′)	3749
(′tom′, ′bagshaw′)	3467
(′ross′, ′tran′)	3439
(′dan′, ′mumford′)	3201
(′krenz′, ′cushart′)	3166
(′golden′, ′hour′)	3125
(′norman′, ′rockwell′)	3093
(′joe′, ′biden′)	3048
(′lois′, ′baarle′)	3026
(′thomas′, ′kinkade′)	2908

TABLE 4.3: Bigrams with artist names

Bigrams no artist names	
Bigram	**Frequency**
(′black′, ′white′)	5838
(′sci′, ′fi′)	5427
(′donald′, ′trump′)	4282
(′movie′, ′still′)	4001
(′global′, ′illumination′)	3988
(′krenz′, ′cushart′)	3332
(′lois′, ′baarle′)	3199
(′golden′, ′hour′)	3194
(′joe′, ′biden′)	3084
(′emma′, ′watson′)	2861
(′gaston′, ′bussiere′)	2808
(′golden′, ′ratio′)	2787
(′character′, ′design′)	2719
(′walter′, ′white′)	2643
(′movie′, ′poster′)	2625
(′anime′, ′girl′)	2612
(′color′, ′scheme′)	2607
(′charlie′, ′bowater′)	2547
(′wayne′, ′barlowe′)	2478
(′victo′, ′ngai′)	2464
(′new′, ′york′)	2435
(′ilya′, ′rossdraws′)	2374
(′elon′, ′musk′)	2313
(′archdaily′, ′wallpaper′)	2312
(′comic′, ′book′)	2269

TABLE 4.4: Bigrams no artist names

4.3 Comparative Analysis of Topic Modeling Techniques

After data preprocessing, clustering is necessary before applying SemDis. Two topic modeling methods, namely GSDMM and BERTopic, will be used for clustering. Once clusters with similar prompts are identified, they will be labeled to serve as the 'item' value for comparison in SemDis.

4.3.1 Data Clustering

A subset of the two datasets will be used, in order to find the right parameter values. Using the same procedure as in (Udupa et al., 2022), different values for `alpha`, `beta` will be used to measure topic coherence and the amount of clusters generated with GSDMM. Similarly, different values for the `n_gram_range` will be used to evaluate the amount of topics found and the coherence within the topics. The effects of the `alpha` and `beta` values will be evaluated on both the artist and no artist datasets. GSDMM will be applied on the tokenized prompt data. Will be done for the two datasets, different parameter values for the two datasets for BERTopic. For GSDMM, we will only use the dataset without artist names, because of time constraints.

4.3.2 Validation of Topic Modeling

To evaluate the topic modelling process, the `u_mass` measure of coherence will be used. The `u_mass` measure is a metric used to evaluate the quality of topic models, such as LDA and GSDMM, by assessing the interpretability of the topics generated. The `u_mass` coherence measure is readily available through the `gensim` python library. Coherence measures help determine how coherent or meaningful the topics are by examining the degree of semantic similarity between the words within a topic. The `u_mass` assesses the coherence of topics by quantifying the likelihood of observing specific word co-occurrences within documents. It measures how well the words within a topic tend to appear together in the same context, indicating their semantic relatedness and coherence. The coherence measure benefits from being topic-model agnostic, meaning that it will work for both GSDMM and BERTopic's clusters. Additionally, it has a quicker runtime compared to other coherence measures (Röder, Both, and Hinneburg, 2015). A close to zero `u_mass` score indicates higher topic coherence (Tijare and Rani, 2020).

The `u_mass` measure of coherence is based on word co-occurrence statistics gathered from the corpus being modeled and does not depend on an external reference corpus. This calculation is designed to measure the semantic coherence of a topic, with the idea that words belonging to a single concept will co-occur within a single document.

The `u_mass` coherence score for a topic is calculated as follows:

Let $V(t) = (v_1, ..., v_M)$ be a list of the M most probable words in topic t. The coherence score is then defined as the sum of the log probabilities of finding word pairs (v, v') in the same document, where v and v' are words in $V(t)$, and v' appears after v in the list. A smoothing count of 1 is included to avoid taking the logarithm of zero (Mimno et al., 2011). Here it is shown as a formula:

$$C_{UMass}(t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m, v_l) + 1}{D(v_l)} \tag{4.1}$$

In this formula, $C_{UMass}(t)$ represents the U-Mass coherence of topic t. $D(v_m, v_l)$ is the number of documents in which words v_m and v_l co-occur, and $D(v_l)$ is the number of documents in which word v_l appears. The outer sum goes from the second most probable word to the Mth most probable word in the topic, and the inner sum goes from the first most probable word to the $(m-1)$th most probable word. The $+1$ in the numerator is the smoothing count to avoid taking the logarithm of zero.

4.4 Topic Modeling Results

4.4.1 Data Clustering - GSDMM Results

In this section Gibbs Sampling Dirichlet Mixture Model (GSDMM)'s performance will be examined using different α, β and `n_iters` values. Samples of size `50000` are taken from both datasets. Parameter `k` is set at `10` across all evaluations.

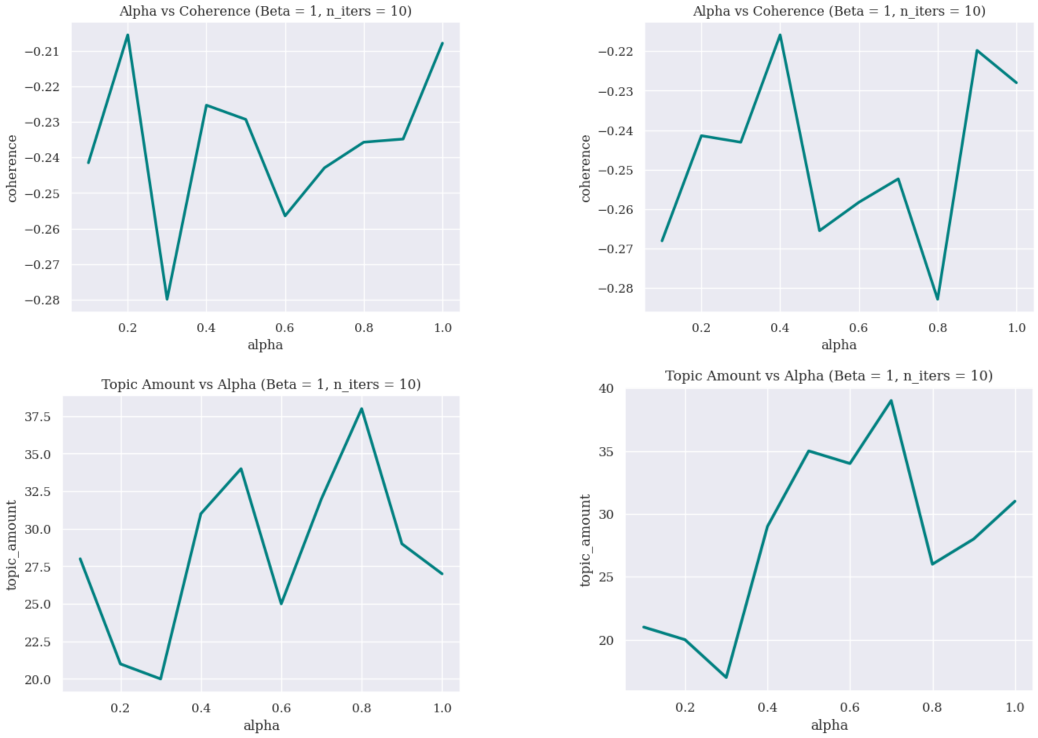


FIGURE 4.3: Comparison of Alpha against Coherence and Topics without artists (left) and with artists (right)

Effect of α

Different values of α are evaluated with the other parameters set at β `= 1` and `n_iters = 10`. The amount of clusters found and the coherence for the different values of α can be seen in 4.3. The α parameter in GSDMM topic modeling has an impact on coherence and the number of topics. Lower α values (close to 0) tend to result in higher coherence scores, indicating more meaningful topics. They emphasize document-topic sparsity, resulting in distinct and focused topics. Higher α values (closer to 1) tend to produce a larger number of topics by promoting topic-document sparsity, allowing documents to contain a mixture of more topics.

On this project's data, what can be seen is that the coherence scores stay between -0.30 and -0.20 for both datasets, with only slight variability. Higher values for α do seem to produce more topics, but not as consistently as expected.

Effect of β

Different values of β are evaluated with the other parameters set at α `= 1` and `n_iters = 10`. The amount of clusters found and the coherence for the different values of β can be seen in 4.4. Higher values of β tend to result in a lower number of clusters produced by GSDMM, as the MGP explains that students select tables with common movie interests, resulting in less clusters. For both datasets, this also seems to be the case, as lower values for β produce more topics, and higher values produce less. What can also be seen is that higher scores for β produce significantly higher coherence scores.

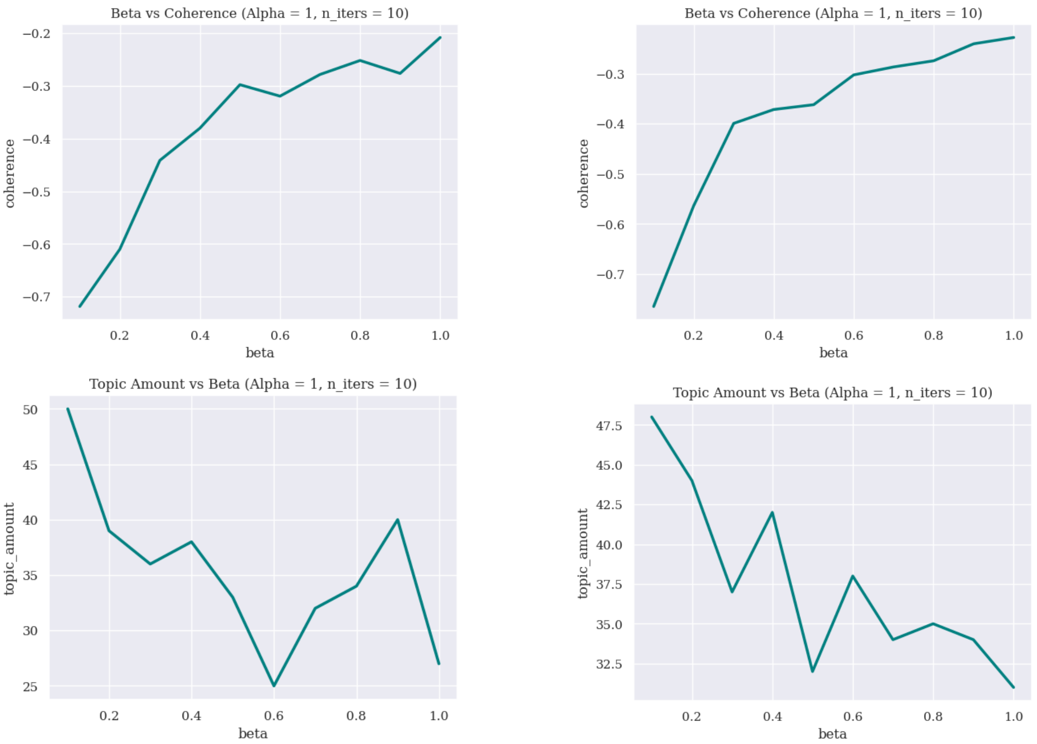


FIGURE 4.4: Comparison of Beta against Coherence and Topics without artists (left) and with artists (right)

4.4.2 Data Clustering - BERTopic Results

In this section BERTopic's performance will be examined using different `n_gram` values. Samples of size `5000` are taken from both datasets. The N-gram range is set from (1, 1) to (6, 6).

Effect of the N-Gram range

An N-Gram is a chain of n elements from a text or speech, with n representing the number of words in the sequence. For the sentence "I love eating ice cream," the n-grams are as follows:

Unigrams	Bigrams	Trigrams
"I"	"I love"	"I love eating"
"love"	"love eating"	"love eating ice"
"eating"	"eating ice"	"eating ice cream"
"ice"	"ice cream"	
"cream"		

TABLE 4.5: The sentence 'I love eating ice cream' in uni-, bi-, and trigrams

Using an appropriate N-gram range is beneficial to the quality of classifications by capturing the intricate composition of words. When multiple words are considered, the complexity of language is better understood. BERTopic, being a neural topic model that leverages semantic meaning, has better performance when the N-gram Range is sufficiently broad. This enables a more effective capturing of contextual information. To evaluate the effect of the N-gram range on the coherence of the topic modelling process, the N-gram range is varied between (1,1) and (6,6). In Figure 4.5 the results of varying between different N-gram ranges can be seen for the two different datasets. In general, it can be seen that from N-gram range (2,2) onwards, the coherence scores significantly improve and get closer to zero, which signifies a better topic coherence. What can be seen is that coherence scores for N-grams of type (x, x) where the minimum and maximum of the N-gram are equal, the coherence scores increase significantly, which suggest overfitting. What can also be seen is that as the minimum N-gram range increases, the coherence score increases. N-grams beyond $(3, 3)$ are producing coherence scores close to zero, meaning that the topic model is capturing highly coherent and contextually rich topics. In Figure 4.6 the number of topics within the dataset at different N-gram ranges are shown. The amount of clusters found range between 70 and 85 for both datasets regardless of the N-gram range. An examination of the Pearson correlation coefficients was carried out for both datasets. The Pearson correlation coefficient is a measure of the linear correlation between two variables. It has a value between +1 and -1, where +1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. The calculated coefficients yielded values of 0.115 and -0.201, respectively. Moreover, the corresponding p-values were determined to be 0.619 and 0.382. These findings suggest that there is no significant correlation between the N-gram range and the number of clusters in either dataset.

4.5 Comparative Analysis of Labeling Techniques

To use SemDis for generating results, the found topics need to be appropriately labeled. These labels serve as the 'items' in the SemDis method, which enables the calculation of semantic distances between an `item` and a `response`. In this context, the responses correspond to the prompts within the dataset, while the items will be represented by the topic labels themselves. Manual topic labeling is a time-intensive process, as it requires observing documents from each cluster and assigning suitable labels. To aid this process, two automatic cluster labeling techniques will be employed on sample datasets, and the resulting SemDis scores will be compared for evaluation.

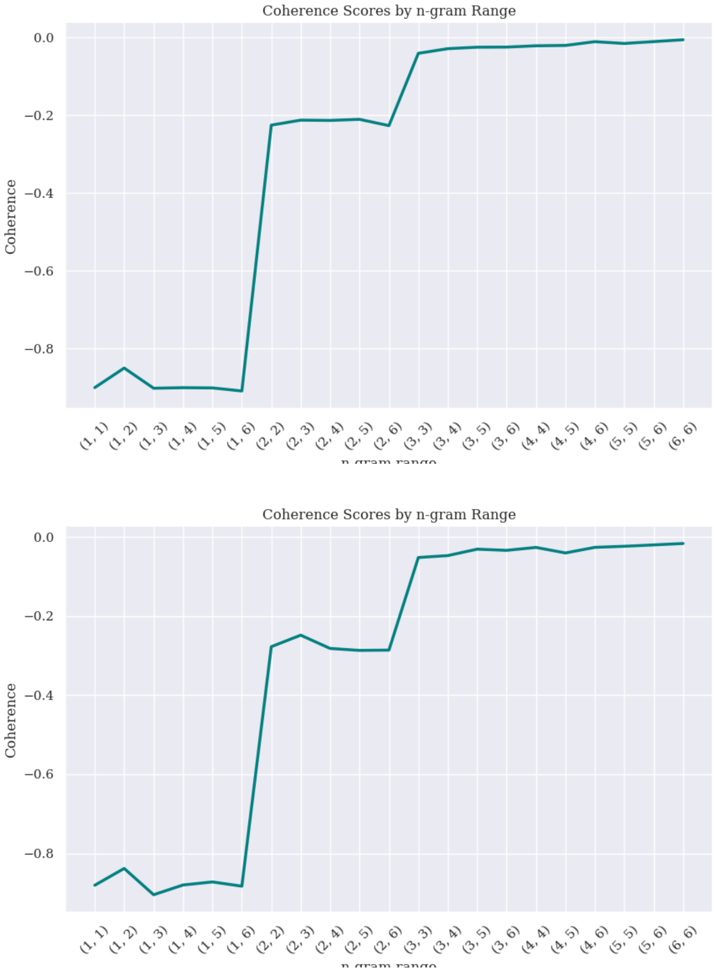


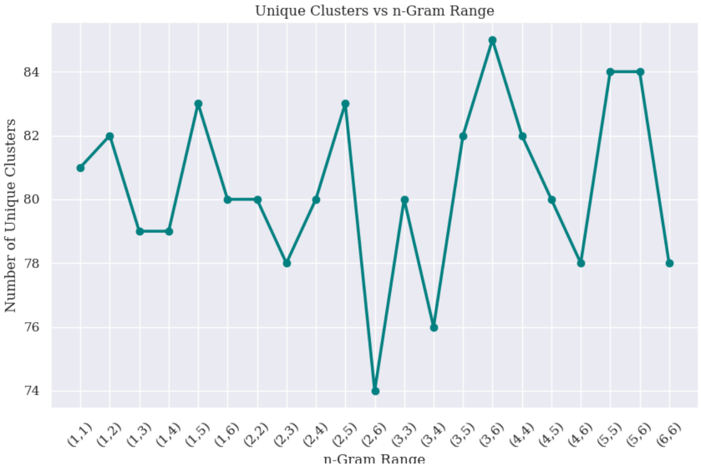
FIGURE 4.5: Comparison of n-gram ranges against Coherence without artists (top) and with artists (bottom)

4.5.1 `top_n` Topic Labeling

With the `top_n` topic labeling technique, the top frequently used terms within a topic category will be observed and used as the label for the topic. The `n` will be ranged between 1 and 5 terms, meaning that the topics will be labeled with 1 to 5 terms.

4.5.2 OpenAI Topic Labeling

In the OpenAI cluster labeling method, a sample of high-frequency words from each cluster is presented to the OpenAI API. Leveraging GPT-3.5's language modeling capabilities, this method generates comprehensive and meaningful cluster labels. This is done by feeding the top 10 frequently used terms in a cluster to the OpenAI API, and prompting it with the following task: `"I am given a set of related concepts: {topic_words}. Summarize these concepts in {max_words} words."`. The `max_words` will be ranged from 1 to 5 words, meaning that the cluster labels will be labeled with 1 to 5 words.



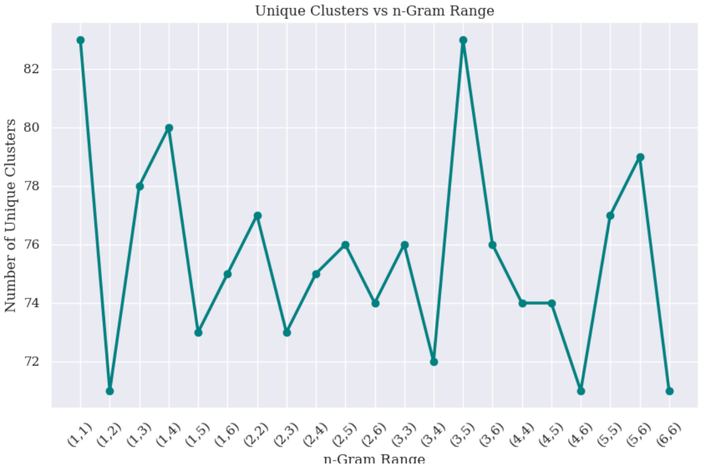


FIGURE 4.6: Comparison of n-gram ranges against number of topics without artists (top) and with artists (bottom)

4.5.3 Topic Labeling Setup

A sample of 3000 entries is taken from both datasets, with and without artist names and clustered with both GSDMM and BERTopic. For GSDMM, the following parameters are used: $\alpha, \beta = 0.5$, `n_iters = 10`, and `num_topics = 300`. For BERTopic, an `n_gram` range of (1,1) is used, since this range will result in single terms that can be used for the labels. The `top_n` and OpenAI cluster labeling techniques are applied on datasets for both topic modelling techniques, resulting in a total of 20 files on which SemDis will be applied to.

4.6 Topic Labeling Results

Upon labeling the clusters using the `top_n` and OpenAI methods, a substantial difference between the labels can immediately be observed. Table 4.6 presents the generated labels for a representative cluster. The `top_n` labels appear less descriptive, whereas the OpenAI labels more effectively encapsulate the subject matter of the topic. he `top_n` labeling method generates labels using the most frequent terms within the cluster. As a result, the variability between label lengths is minimal. However, for labels generated by the OpenAI method, increased label length

Label Length	top_n	OpenAI
1	woman	Gender
2	woman hair	Feminine Hair
3	woman hair wear	Feminine Stylish Hair
4	woman hair wear dress	Fashion Beauty Style Access
5	woman hair wear dress face	Fashion Runway Stylish Photography

TABLE 4.6: Generated labels for the prompt '`fashion young woman model jewel watch black woman white dress hasselblad medium format`'

results in a clearer representation of the subject matter. The topic of the example above contains 122 documents. Below, five of them are listed:

```
female character study cute young plus size todd solondz tan skin
clear faces screenwriter introvert outsider geek disturbed emotional
character sheet fine design kim jung gi pixar body head turnaround
front view back view

solarpunkthemed lolita outfit fabric pictures windmills solar panels
tall ecofriendly green buildings covered leaves lolita dress themed
lush green ecofriendly future city angelic pretty

woman holding inspired head shoulders white cute pixar character

female knight cat head wild nature armor design wayne barlowe blonde
hair symmetry sci fi dark perfect composition hd sense awe

red woman dark glasses eyes show blue makeup eyes single eyelashes
color eyebrows red red lipstick pink tongue shadow neck head nails
color holding small man brought mouth man wearing black suit black
tie pink shirt quiff dark glasses arms raised flat design solid colors
```

What can be seen is that the `top_n` labeling method directly uses terms from the prompts in the cluster labels, whereas the OpenAI generated labels uses terms that are not present directly in the documents. The prompts above include terms that are related to women, style, clothing, and beauty, which is encapsulated by both labeling techniques. However, OpenAI tends to overgeneralize the topic when the label length is set to 1, as seen in the 'Gender' label, which might not capture the full context of the cluster. However, as the topic labels extend beyond a length of 1, they tend to provide a much more comprehensive representation of the topic.

4.6.1 SemDis Results after Labeling

After having labeled the clusters, the data has been put through the SemDis algorithm, generating creativity scores for the prompts within the clusters. Figures 4.7

and 4.8 show the average SemDis scores for the prompts, grouped by topic label length. What can be seen immediately is that the mean scores are incredibly high, averaging between 0.95 and 1.00 for both clustering label methods, topic modelling techniques, and label lengths. However, the SemDis scores range higher for the OpenAI labeling method, which is expected since the OpenAI labeling method generates a label which is less likely to be term that is present in the prompts, thus creating more semantic distance between the the item and the response. The `top_n` labeling method uses terms that are directly derived from the prompts, resulting in less semantic distance. To gain more insight into the scores, the scores are normalized using min-max normalization.

Min-max normalization, also known as feature scaling, is a data normalization technique used to transform numerical data to a common scale. It rescales the values of a dataset so that they fall within a specific range, typically between 0 and 1.

The process of min-max normalization involves subtracting the minimum value from each data point and dividing it by the range (the difference between the maximum and minimum values). Mathematically, the formula for min-max normalization is as follows:

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4.2}$$

where x represents the original data point, and $x_{\text{normalized}}$ is the normalized value.

By applying min-max normalization to the SemDis scores, the values can be rescaled to the range between 0 and 1, which allows for better comparison and interpretation of the scores. Figures 4.10 and 4.9 show the average average normalized SemDis scores for the prompts, grouped by topic label length.

From the original SemDis scores, we see high variability for the scores in the BERTopic dataset, with `top_n` cluster labeling. However, the scores are still high, ranging between 0.95 and 1. OpenAI's approach to generating cluster labels shows a more uniform distribution of scores, regardless of the label length, with the SemDis scores uniformly high, varying between 0.97 and 1.The OpenAI methodology, which generates new cluster labels not necessarily present in the prompts, logically contributes to these higher SemDis scores due to the increased semantic distance between the item and the response. This pattern is replicated within the GSDMM dataset, as it can be seen that the OpenAI generated cluster labels score higher than the `top_n` cluster labels, reinforcing the assertion that OpenAI's approach generates more semantically distinct labels, thus enhancing the validity of the clustering process.

Upon applying normalization to the scores across each label length dataset, a significant change in the distribution of these scores can be seen. Now, they fall within a consistent range of 0.3 to 0.7, regardless of the dataset, labeling methods, and label lengths involved. The process of normalization serves an important function in making the data more comprehensible.

Upon examination of the results derived from the various labeling methods and label lengths, the decision has been made to proceed with the OpenAI labeling method for several reasons. Firstly, the scores of the OpenAI generated labels score higher than the `top_n` method on average, implying a more varied scoring of the creativity within the prompts. As previously stated, this is expected, since the generated labels allow SemDis to compare the prompts to a label (item) that is not (frequently) present in the prompts. Secondly, using the generated labels allows for a better encapsulation of the subject matter within the topics, making the topics more interpretable at first glance. Lastly, employing the OpenAI generated labels aligns more coherently with the principles of the AUT, what the SemDis procedure is originally based on. In the AUT, respondents are required to conceive as many creative uses as possible for a given item, such as a 'brick', with possible answers ranging from 'a step' to 'a diving aid'. The OpenAI labeling approach mimics this inventive generation process more closely, further solidifying its suitability for this application.

In summary, these points collectively establish a strong case for the use of the OpenAI labeling method. This method showcases its effectiveness by consistently generating high SemDis scores, enhancing data interpretation, and faithfully upholding the core ideas within the AUT, thus demonstrating its superiority over the alternative labeling method.

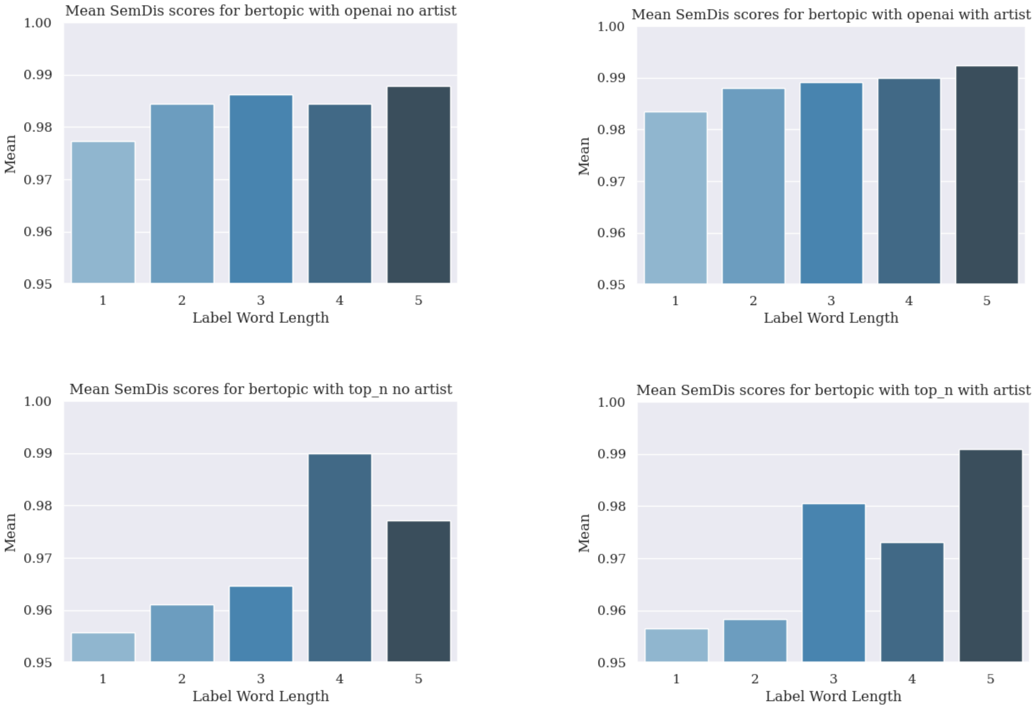


FIGURE 4.7: Effect of label word length (generated with openAI and top_n) on average SemDis scores after BERTopic clustering

4.7 Scaling Topic Modeling and Labeling to a Large Dataset

After having evaluated the two clustering methods, GSDMM and BERTopic, and two different methods of cluster labeling, `top_n` and OpenAI, we proceed to move forward and generate scores for a bigger piece of the dataset. We will use BERTopic to generate clusters, and OpenAI to generate cluster labels. This is done for several reasons. Firstly, BERTopic generates more and more equally distributed clusters. What can be seen after multiple uses of GSDMM using different parameters, is that GSDMM tends to overpopulate one specific cluster and minimally populate all other clusters. A suspected reason for this is the probabilistic nature of GSDMM. Within one prompt, several concepts and themes can be found, which can make the probabilistic clustering process that GSDMM uses more likely to group these prompts all together, and making small clusters with prompts that

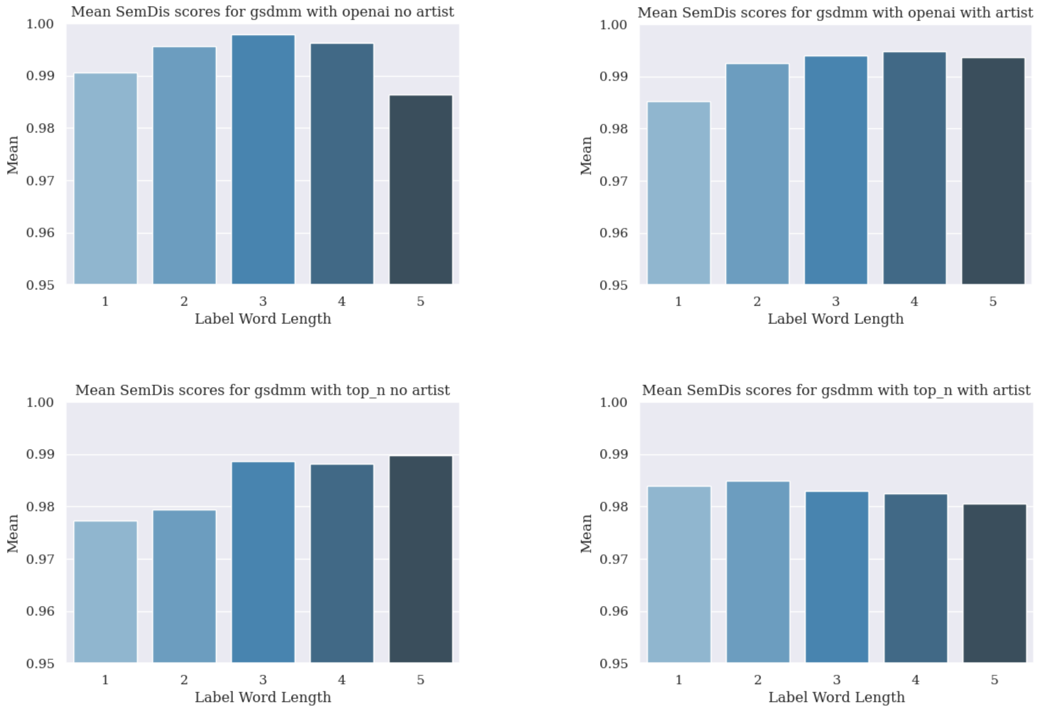


FIGURE 4.8: Effect of label word length (generated with openAI and top_n) on average SemDis scores after GSDMM clustering

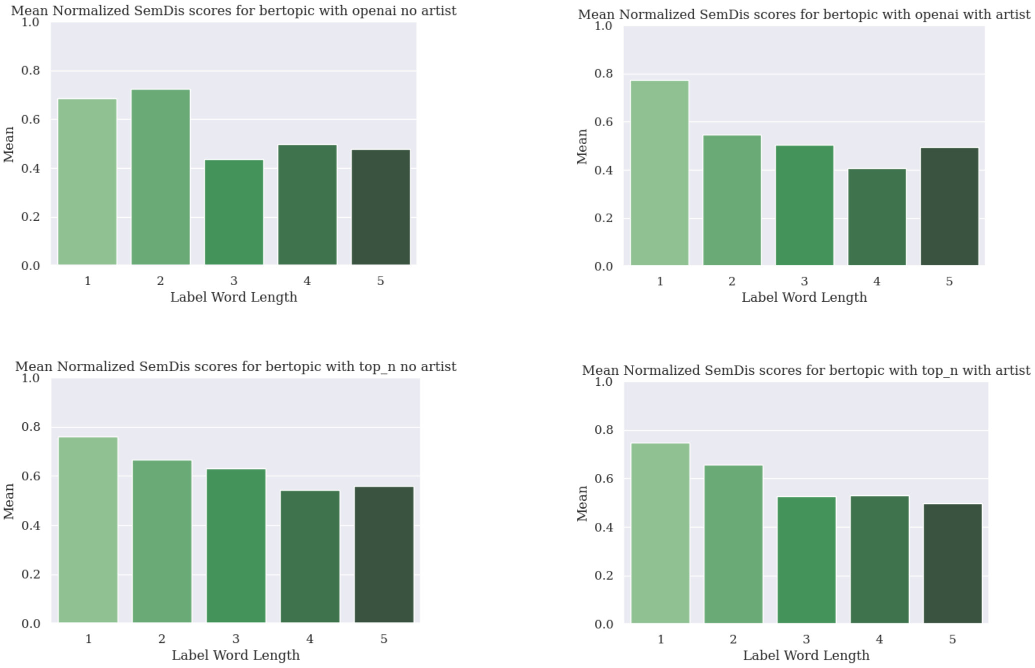


FIGURE 4.9: Effect of label word length (generated with openAI and top_n) on average normalized SemDis scores after BERTopic clustering

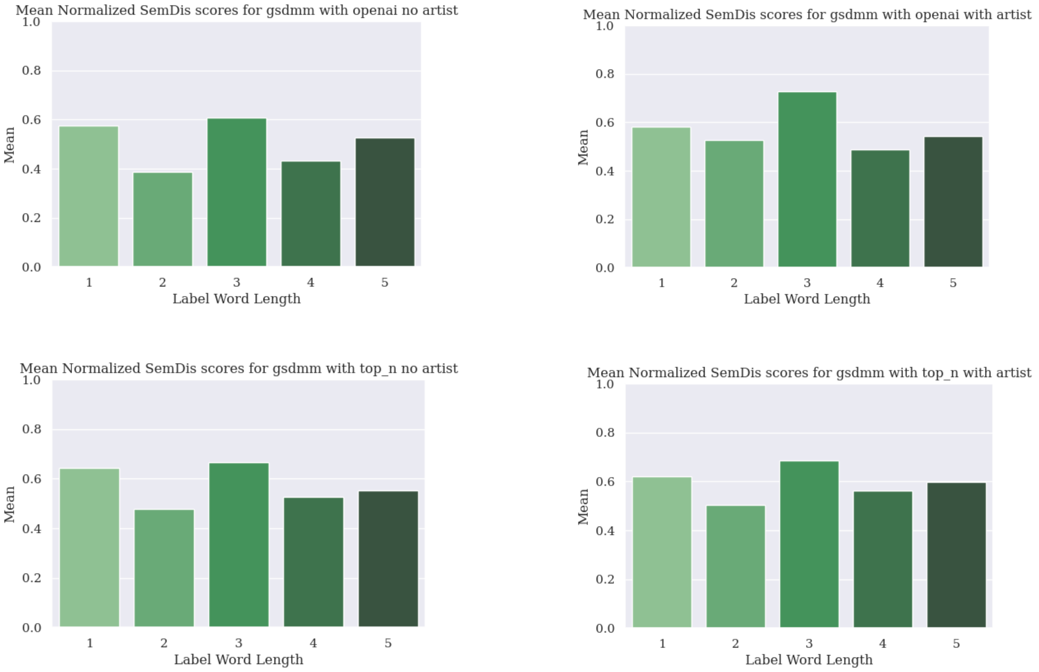


FIGURE 4.10: Effect of label word length (generated with openAI and top_n) on average normalized SemDis scores after GSDMM clustering

are actually similar. This is different compared to BERTopic, which leverages the pretrained BERT model to capture complex semantic relationships between words and sentences. In contrast, GSDMM is a probabilistic model that primarily relies on word frequency and distribution, lacking the same level of semantic understanding as BERTopic. BERTopic's use of contextual embeddings enables it to capture the meaning and context of words, leading to more accurate clustering results. Secondly, the BERTopic topic modelling process is significantly quicker than GSDMM. Where BERTopic can take up to a few hours on a dataset with 700,000 rows, GSDMM can take up to several days.

BERTopic was applied to a preprocessed dataset comprising a subset of 50,000 rows. This subset was taken, because SemDis only allows for files up to 40,000 rows. After removing the rows with prompts that were not assigned to a cluster, it is expected that the resulting dataset would be smaller than 40,000 rows. Two versions of the dataset were used, one with artist names and one without. The `n_gram` range chosen for feature extraction was (3,5), as it produced a considerable number of clusters while maintaining a favorable coherence score, as observed in experiment 1. To ensure the presence of substantial clusters, the `min_cluster_size` parameter in BERTopic was set to 50.

4.8 Final Results

After running BERTopic on the two datasets, we find a coherence scores of -0.0084 and -0.0028 for the datasets with and without artists respectively indicating that the clusters may overlap. By using the BERTopic method, we have generated 88 distinct clusters. After removing the prompts that were assigned to cluster '-1', which is the topic number for prompts that BERTopic was not able to assign to a cluster. This resulted in datasets with 19710 rows without artists, and 19445 rows with artists.

Next up in the process is the labeling of the topics, before the data is inserted in SemDis. In Experiment 2, we see that for both the datasets with and without artists, a label length of 3 has lower normalized SemDis scores compared to the other label lengths. This is favourable, since this allows for a more diverse range of scores in the data. Additionally, a label length of 3 allows for a better encapsulation of the subject matter within the topics.

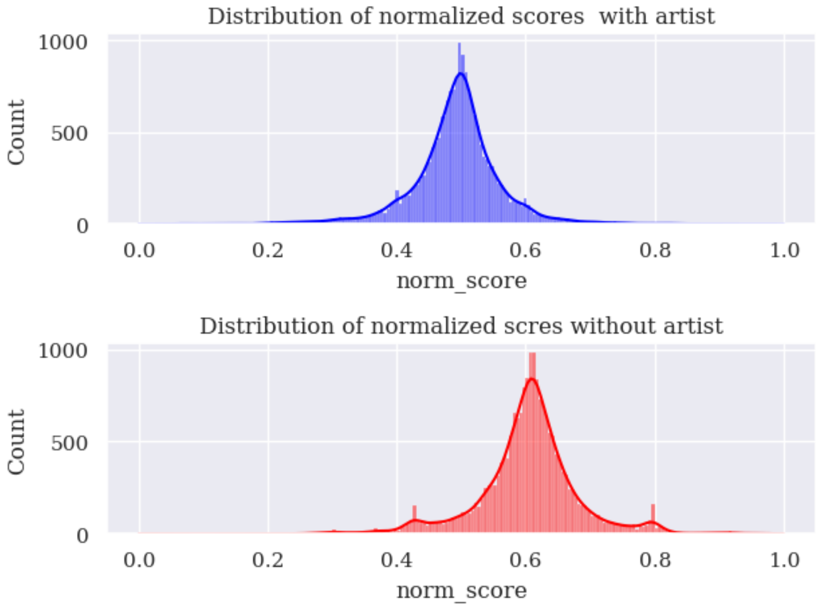


FIGURE 4.11: Distribution of the normalized scores in the with artist and without artist data sets

Table 4.7 presents the initial fifteen topics and the number of associated documents within the clusters, after the generation of the cluster labels. In the dataset including artist names, it is expected to to find clusters centered around these names, represented by their respective labels.

After reviewing Table 4.7, this assumption holds true. For instance, artist names such as *Rutkowski* and *Mucha* are part of the topic labels, implying that these topics are organized based on the artist names present in the prompts.

For a complete overview of the generated topics and labels, refer to Appendix C. This list strengthens the observation that the artist names in the dataset significantly impact how the topics are clustered.

After putting the data through SemDis, rows with a NaN value for the score were removed. The next step in the process is to feed the resulting clusters with their cluster labels into SemDis. After generating the SemDis scores, the scores are normalized. The resulting dataset descriptions can be seen in Table 4.2. As the results of the analysis are explored in the subsequent sections, it should be noted that these findings are based on the cleaned prompts. The difference between the results ’with artists’ and ’without artists’ lies in these cleaned prompts. In the ’with artists’ case, the artist names are included in the cleaned prompts, while in the ’without artists’ case, they are not. This distinction is important for understanding the differences in topic comparison and clustering in SemDis.

no_artist		with_artist	
Topic	#	Topic	#
Fashion, Beauty, Art	2275	Manga, Art, Adventure	1595
Science Fiction Hybrid	718	Beauty, Art, Shade	1147
Tabbi Cat Fur	412	Food, Bowl, Candy	887
Art, Politics, Fear	333	Art, Rutkowski, Mucha	861
Appropriate Icons	315	Golden Ratio, Compositing, Faces	806
Armor, Knights, Dragons	259	Tabbi, White, Grey	478
Electronic, Dance, Music	233	Art, Women, Portraits	433
Aging, jewelry, symbolism.	223	Colour, Warm, Artist	431
Early photography technology.	203	Color, Monochromatic, Soft	410
Extraterrestrial, Flying, Dragons	196	Cyberpunk, Steampunk, Syn	373
Nature, beauty, light.	190	Comic, Joshua, Albuquerque	366
Awards, Horror, Gloom	175	Art, Geometry, Surrealism	332
Geometric Shapes.	162	Artificial Intelligence, Storytelling, Post	331
Urban, Destruction, Desolation	151	Automobiles, Technology, Suzuki	295
Heavy, Grain, Cinema	148	Medieval, Armor, Magic	282

TABLE 4.7: Top 15 topics for the no_artist and with_artist Datasets

4.8.1 Statistical Assessment of Creativity Scores

ANOVA

To find whether there is a correlation between the users and the creativity scores, and the topics and the creativity scores, an ANOVA test can be conducted. Before an ANOVA test can be conducted, the data has to meet the three assumptions of an ANOVA test, which are:

- Independence of observations: The bservations should be independent of each other.
- Normality: the residuals of the model should approximately follow a normal distribution.
- Equality of variances: also called homoscedasticity, this assumes that the variance of data in groups should be the same.

The datasets are not suitable for conducting an ANOVA test due to several reasons. Firstly, the extremely low p-values obtained from the Shapiro-Wilk test (0.0 for both datasets) indicate that the data does not follow a normal distribution, which is a fundamental assumption for ANOVA. Additionally, the p-values from the Levene's test (1.17×10^{-195} for the dataset with artist names and 6.64×10^{-140} for the dataset without artist names) reveal that the variances among the groups are significantly different. This violation contradicts the assumption of homogeneity of variances required for ANOVA. Consequently, due to these violations, it might be more appropriate to employ alternative non-parametric tests or data transformations for analyzing this dataset.

Kruskal-Wallis H Test

The Kruskal-Wallis H test is a non-parametric statistical test used to determine if there are statistically significant differences between two or more groups of an

	with_artist	no_artist
Statistic	Value	Value
Dataset length	19074	19445
Amount of different user_names	672	666
Amount of different topics	86	88
Average SemDis score	0.9881	0.9879
Median SemDis score	0.9942	0.9938
Average Norm_score	0.4927	0.6039
Median Norm_score	0.4962	0.6075
Mean number of prompts per topic	221.79	220.97
Median number of prompts per topic	143.0	142.5

TABLE 4.8: Comparative Statistics for 'with_artist' and 'no_artist' Datasets

independent variable on a continuous or ordinal dependent variable (McKight and Najab, 2010). It is especially applicable when the data does not meet the assumptions of parametric tests like ANOVA, which is the case.

The Kruskal-Wallis H test essentially ranks all the data from all groups together, then analyses the ranks to determine whether the data from each group tends to have higher or lower ranks than would be expected under the null hypothesis. The null hypothesis of the Kruskal-Wallis H test is that the medians of all groups are equal, meaning that there is no difference between the groups.

Within the context of the two created datasets, this means that the username or topics are the independent variables, which are categorical variables representing different users or topics, and the normalized SemDis scores as the dependent variable, a continuous variable representing the scores assigned to the users. The test will determine if there's a statistically significant difference in scores among the different users or topics.

To apply the test, the data will be grouped usernames or topics and rank all the 'score' values from lowest to highest, ignoring which user or topic they belong to. The test then calculates a test statistic based on these ranks, with a larger value indicating greater divergence between the groups.

The result of the test is a p-value, which tells the probability of observing the given data, assuming the null hypothesis is true. If the p-value is less than 0.05, the null hypothesis is rejected and it can be concluded that the usernames or topics do indeed make a difference to the 'score', meaning that the medians of the scores across different users are not equal.

The Kruskal-Wallis H test was thus used to determine if there were differences in the normalized scores across various user groups and topic categories. This test revealed significant differences in both cases. Specifically, a p-value of approximately 4.92×10^{-7} was found for user groups in the dataset with artists., and 0.00633 in the dataset without artists, indicating a significant difference in the scores associated with different users. For the topic categories, the p-value was approximately 1.54×10^{-43} in the dataset with artists, and 1.82×10^{-25} in the dataset without artists. These results point to a significant variation in scores across various topics. Also, it can be inferred that both the user who initiated the prompt and the topic category of the prompt do influence the creativity score the prompt received.

4.8.2 Topics

Figure 4.12 provides a comparison between the highest and lowest scoring topics. Notably, the higher scoring topics exhibit relatively low variability. For further insight, an examination of some prompts from both the highest and lowest scoring topics is conducted.

It is hypothesized that prompts associated with the highest scoring topics likely have a greater semantic distance from the cluster label, resulting in a higher creativity score. This suggests the inclusion of diverse elements not directly related to the words in the topic label. In contrast, prompts from lower scoring topics are anticipated to incorporate terms that closely align with the topic label.

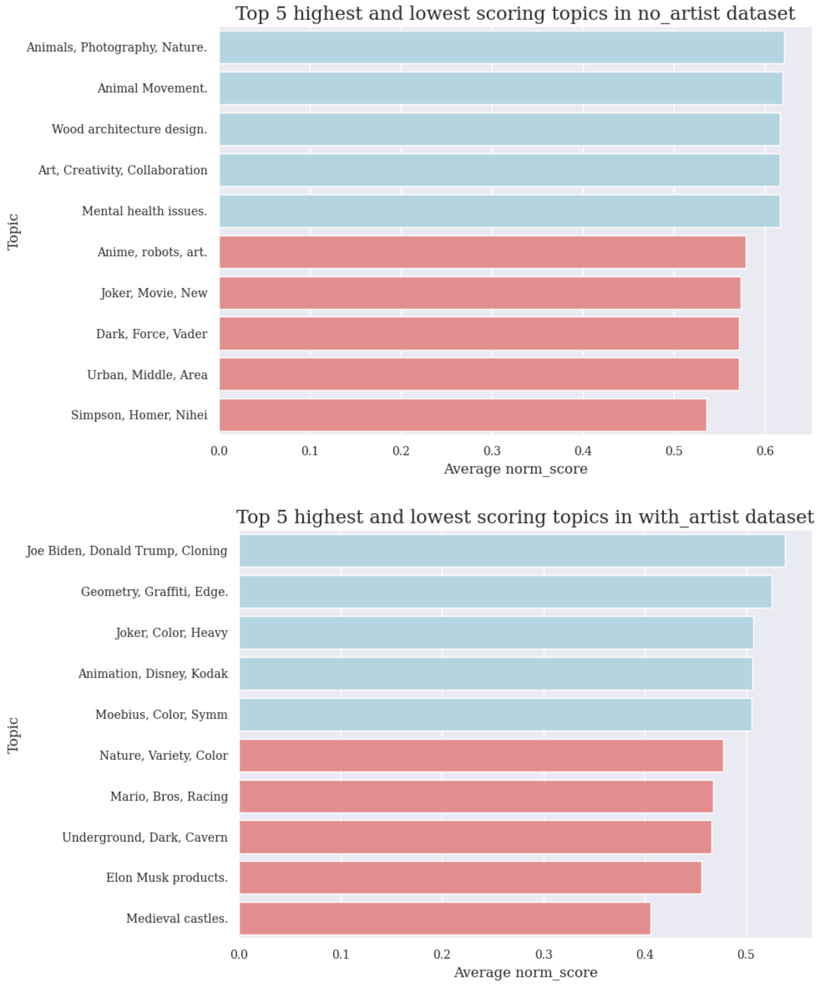


FIGURE 4.12: Highest and lowest scoring topics for datasets without artists (top) and dataset with artists (bottom)

Table 4.9 presents a subset of prompts from the *'Animals, Photography, Nature'* topic derived from the no_artist dataset, while Table 4.10 offers a selection of prompts from the *'Medieval Castles'* topic from the with_artists dataset. As anticipated, the scores for the *'Medieval Castles'* topic are generally lower, attributable to the frequent repetition of terms that match the topic label. Specifically, the term 'medieval' appears in all sample prompts, and 'castle' features in two. Conversely, the typically higher scoring *'Animals, Photography, Nature'* topic includes prompts that relate to the topic but refrain from using the label terms. A prime example is the highest scoring prompt: `a frog bumblebee, a hybrid animal that is half - frog half - bumblebee, cute animal photography`, which mentions 'animal' and 'photography' yet incorporates additional terms not present in the label.

Prompt	Username	Norm_Score
a dog and a racoon fusing via dance, anime art style, cgi render	00bc(...)	0.6056
an illustration of a jaunty ferret with a monocle	0199(...)	0.5686
a portrait of a jaunty ferret with a cane by edward gorey	0199(...)	0.6973
a frog bumblebee, a hybrid animal that is half - frog half - bumblebee, cute animal photography	0240(...)	0.7606
a fox squirrel, a hybrid animal that is half - fox half - squirrel, cute animal photography	0240(...)	0.7541

TABLE 4.9: Topic: Animals, Photography, Nature - Average Norm Score: 0.6215

Prompt	Username	Norm_Score
mideval fantasy castle in an epic forrest clearing in style of Jhonen Vasquez, dawn, photorealistic, detailed, trending on artstation	016e(...)	0.5163
ancient medieval art depicting a shoggoth entering into this reality	0195(...)	0.5411
medieval depictions of lovecraftian sights	0195(...)	0.4473
grand fantasy castle by martin deschambault	01cc(...)	0.5286
medieval prince, origami	0240(...)	0.3628

TABLE 4.10: Topic: Medieval Castles - Average Norm Score: 0.4057

Building upon the understanding of topic-specific prompts, another key factor comes into play: the unique individual preferences of users in the prompting process. A question: Do users tend to gravitate towards and prompt only in specific topics? Data reveals that users, on average, prompt across 10.43 different topics. However, outliers exist, such as user '0f0c127(...),' who has demonstrated a wider interest by prompting across 52 distinct topics.

A next area of investigation is the potential correlation between a user's inclination to prompt across multiple topics and their creativity scores. For this purpose, a correlation analysis is performed.In the analysis, a Pearson correlation test was used to determine the relationship between the number of topics users prompted and their respective creativity scores in both the 'with_artist' and 'no_artist' datasets. The correlation coefficients yielded are 0.054 and 0.028 respectively, which indicates a marginal positive linear relationship.

This suggests a slight trend of increased creativity scores as users prompt across a wider array of topics. However, given the weak correlation, the influence can be considered minor. Consequently, it is concluded that the diversity of topics a user prompts in, has a negligible impact on their creativity scores. IIt can thus be stated that a user's proliferation in topics does not have a significant effect on their creativity score.

4.8.3 Users

Figure 4.13 shows a comparison between the highest and lowest scoring users in the two data sets. Similar to how it is expected that prompts resemble the topic label when inspecting the prompts within the dataset, we expect that higher scoring users in the dataset exhibit prompts that are semantically distant to the topic labels that they prompt in. Lower scoring users are expected to prompt highly similar to the topic label or prompt with short texts. displays a User '05f(...) has displayed the highest average scores across all users in both datasets. Looking into their prompts, it shows that this user has only prompted once, with the following prompt `'Darth Vader walking his dog through a snowy forest, Thomas Kinkade painting'` in the topic 'Dark, Force, Vader' for the no_artist data, and the topic 'Star Wars, Vader, Movies' in the with_artist dataset. This prompt was rated with a normalized score of 0.75. This prompt exhibits a high degree of novelty, as it combines elements from disparate domains (Star Wars and Thomas Kinkade's art style) in an unexpected way. This aligns with the definition of creativity as the production of novel and relevant ideas, and also maintains a sufficient semantic distance from the topic labels, which grants it a relatively high creativity score.

A test for correlation between the amount of prompts a user has to their name and the creativity scores can be done. Again, a Pearson correlation test was used to determine the relationship between the number of prompts users have prompted and their respective creativity scores in both 'with_artist' and 'no_artist' datasets. The correlation coefficients were found to be 0.0123 and -0.0132 for 'with_artist' and 'no_artist' respectively, which indicates a weak positive linear relationship in the dataset with artist names included, and a faint negative linear relationship for the data excluding the artist names. Both correlation scores are very weak, implying the effect is again, not strong and negligible. Again, the number of prompts assigned to a user does not have a significant effect on a user's creativity score.

4.8.4 Prompts

An examination of the highest and lowest scoring prompts in both datasets - one with artists and one without - is conducted. It's expected that the top-scoring prompts will use words that diverge significantly from the topic label, indicating a larger semantic distance. This divergence is what contributes to a higher score, as it demonstrates more creativity and originality.

On the other hand, prompts that score lower are likely to adhere closely to the topic labels, demonstrating a smaller semantic distance. While these prompts may be relevant, they lack novelty, which is why they receive lower scores.

What can be seen from the higher scoring prompts is that they demonstrate a significant semantic distance from the topic labels. These prompts contain imaginative and unique ideas that diverge from the conventional expectations associated with the given topics. Take for example prompts like "`highly detailed full body portrait of a zombie swimming underwater in a zombie-apocalypse, in a swimming pool, style of plants vs zombies`" and "`boa constrictor and Komodo dragon mutant animal`" These prompts incorporate unexpected combinations and elements, indicating higher creativity and originality.

On the other hand, it can be seen from the lower scoring prompts that they adhere moreclosely to the topic labels. They lack novelty and creativity, presenting more conventional and expected ideas. For instance, prompts such as "wlop

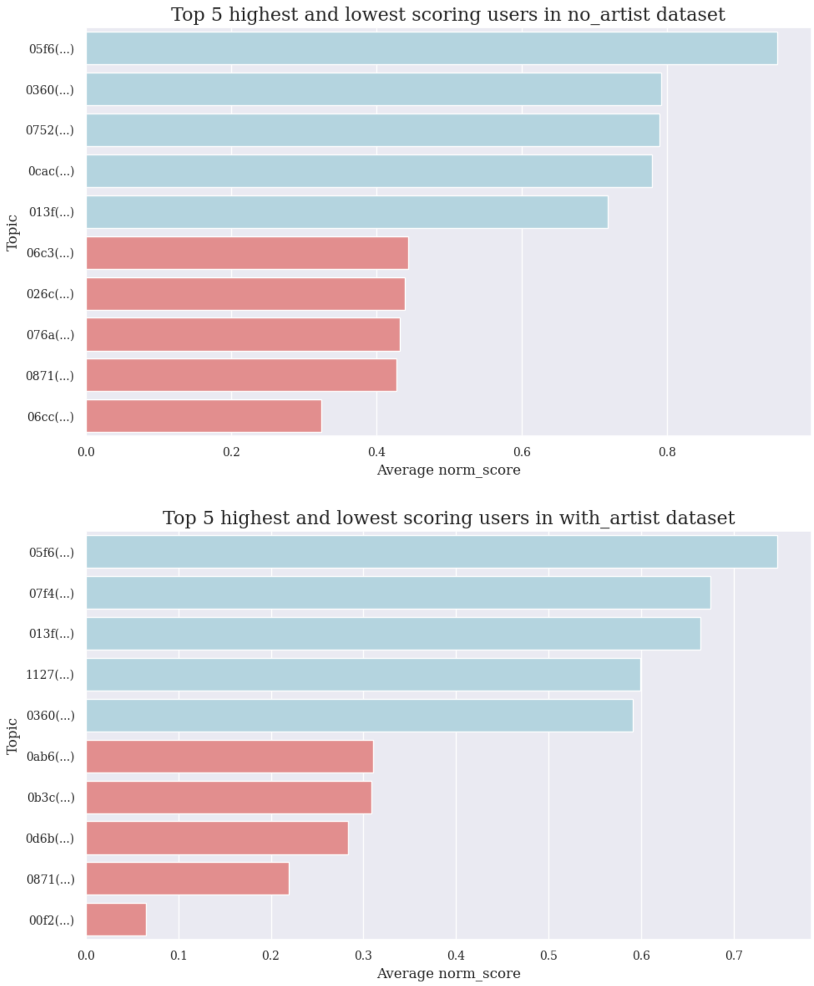


FIGURE 4.13: Highest and lowest scoring users for datasets without artists (top) and dataset with artists (bottom)

masterpiece, Emma Watson as Hermione portrait, detailed, high quality, high resolution" and "Homer Simpson as the Discobolus" demonstrate adherence to the topic labels and receive lower scores due to their lack of uniqueness. The semantic distance is calculated on cleaned prompts, where prompt modifiers are removed, leaving only the essential terms used to calculate semantic distance from the topic label. Because of this, it is logical for the scores to be lower, as the remaining terms in the prompts overlap with those in the label.

A test for correlation between a prompt length and the creativity scores can be done. Again, a Pearson correlation test was used to determine the relationship between the number of prompts users have prompted and their respective creativity scores in both 'with_artist' and 'no_artist' datasets. The correlation coefficients were found to be 0.06 for both datasets, indicating a positive but weak correlation between the length of a prompt and the creativity scores. This suggests that while longer prompts may see a slight increase in creativity scores, the relationship is not strong. In other words, while there's a hint of a trend suggesting that longer prompts could lead to slightly higher creativity scores, the correlation is too weak to conclude that prompt length significantly impacts creativity scores. Thus, the length of a prompt seems to have minimal influence on its creativity score.

Tables 4.11, 4.12, 4.13, and 4.14 show the highest and lowest scoring prompts for both datasets.

Prompt	norm_score	Topic
highly detailed full body portrait of a zombie swimming underwater in a zombie - apocalypse, in a swimming pool, style of plants vs zombies	1.00	Zombies, Swimming, Underwater
highly detailed full body portrait of a zombie swimming underwater in a zombie - apocalypse, in a swimming pool, style of plants vs zombies, octane render	1.00	Zombies, Swimming, Underwater
boa constrictor and Komodo dragon mutant animal	0.98	Animals, Photography, Nature.
an extremely detailed Darth vader standing portrait in front of a highly detailed landscape of a big and structured city inpired by Star Wars, portal to outer space, digital art, 8k, disney render, disney	0.97	Dark, Force, Vader
an ultra detailed Darth vader standing portrait in front of a highly detailed landscape of a big and structured city inpired by Star Wars, portal to outer space, 4k digital art, octane render, trending on artstation, digital art, 8k, disney render, disney	0.97	Dark, Force, Vader
an ultra detailed Darth vader standing portrait in front of a highly detailed landscape of a big and structured city inpired by Star Wars, portal to outer space, 4k digital art, octane render, trending on artstation, digital art	0.97	Dark, Force, Vader
pixar darth vader giving two thumbs up, render, 3d modelling,	0.97	Dark, Force, Vader
close - up portrait of isabelledeltore with tattoos wearing a balaclava with colourful intricate psychodelic patterns, by edgar maxence and caravaggio and michael whelan and delacroix style, artistic, intricate drawing, light brazen, realistic fantasy, extremely detailed and beautiful aesthetic face, establishing shot, 8 k resolution, dramatic lighting	0.97	Tattoo, Design, Cate
darth vader made of wool and felt, soft toy, textures, puppet, wool, knitted	0.97	Dark, Force, Vader
hawk and crocodile mutant hybrid, bipedal, realistic picture taken in zoo	0.97	Animals, Photography, Nature.

TABLE 4.11: Top 10 Highest Scoring Prompts in no artist dataset

Prompt	norm_score	Topic
wlop masterpiece, emma watson as hermione portrait, detailed, high quality, high resolution	0.00	Emma Watson Hermione
photorealistic homer simpson	0.00	Simpson, Homer, Nihei
Homer Simpson as the Discobolus	0.00	Simpson, Homer, Nihei
tlingit art of homer simpson	0.00	Simpson, Homer, Nihei
homer simpson	0.00	Simpson, Homer, Nihei
a renaissance style portrait painting of Bart Simpson	0.13	Simpson, Homer, Nihei
a close - up photo - real delicate ceramic porcelain sculpture of sophie turner with flowing hair detailed on an intricate white surface by victo ngai and takato yamamoto, micro detail, subsurface scattering, translucent, thin porcelain, octane renderer, white color, physically based rendering, japanese pottery, trending on cgsociety	0.16	Artistic, Sculpting, Design
man eating potato chips, Daguerreotype	0.18	Eating, Photography, Sweets
human nervous system with spinal skeleton surrounded with geometric luminous polished glass cube, carbo, oxyd and glass materials, anatomical sculpture	0.21	Artistic, Sculpting, Design
darth vader enjoying a day at that beach with lightsaber umbrella, by rossdraws and annie leibovitz, artstation	0.22	Dark, Force, Vader

TABLE 4.12: Top 10 Lowest Scoring Prompts in no artist dataset

Prompt	norm_score	Topic
scenic view of a Star Trek Enterprise spaceship flying in a starry outer space, realistic epic image	1.00	Space Exploration, Sci-Fi, Flying
photograph of a rusty generational spaceship landing on a new hopefully habitable planet.	0.99	Space Exploration, Sci-Fi, Flying
an enormous alien space ship flying towards earth. menacing, dangerous. dark. depressing. insane intricate detail. sophisticated compostion. feature shot. high fidelity. ultra hd. 8 k. gold. yellow	0.98	Space Exploration, Sci-Fi, Flying
a space ship landing on a foreign planet, concept art, highly detailed, ultra realistic	0.95	Space Exploration, Sci-Fi, Flying
doom eternal eating an hot dog, mutant in a shape of a burger, tubes fused with the body, painted by stanley lau, painted by greg rutkowski, painted by stanley, artgerm, masterpiece, digital art, trending on arts	0.94	Color, Monochromatic, Soft
vector style the abstract painting of an image of a lady artistic flat illustration, cyber punk minimal figure art, soft colors mono chromatic, art in the style of Bryen Frost	0.94	Color, Monochromatic, Soft
mise en abyme!!!, joe biden paints a painting of joe biden who is in the frame painting a frame of joe biden, droste effect!!!, mise en abyme!!!	0.94	Color, Monochromatic, Soft
a boat sitting on top of a lush green forest, an art deco painting by Romare Bearden, behance, synthetism, ukiyo-e, matte drawing, picasso	0.94	Color, Monochromatic, Soft
fractals by picasso, by dan mumford, moebius, sss, barclay shaw, karol bak, jean baptiste monge, high quality, high resolution, smooth 8 k octane rendered with volumetric cinematic dramatic light rutkowsky	0.92	Color, Monochromatic, Soft
A cyberpunk street riot painted by Claude Monet	0.92	Color, Monochromatic, Soft

TABLE 4.13: Top 10 Highest Scoring Prompts in with artist dataset

Prompt	norm_score	Topic
spaceship inspired by star trek orbiting a planet, 4 k, nasa photography, hd	0.00	Space Exploration, Sci-Fi, Flying
new space giant planets and dwarf planets, full body high detail, high modernization, high stylization, divine render	0.00	Space Exploration, Sci-Fi, Flying
Mythological creature holding planet Earth in space	0.01	Space Exploration, Sci-Fi, Flying
alien space ship flying in space, dramatic,, matte painting, super resolution, beautiful render, octane render, concept art by wlop	0.03	Space Exploration, Sci-Fi, Flying
alien space ship flying in space, matte painting, super resolution, beautiful render, octane render, concept art by wlop	0.03	Space Exploration, Sci-Fi, Flying
vintage, minimalism, figurative, beautiful, colorful, cyber punk, minimal, figure art, soft colors, mono chromatic, black color on white background, casey baugh, 1 1 4 5	0.06	Color, Monochromatic, Soft
vintage, minimalism, figurative, beautiful, colorful, cyber punk, minimal, figure art, soft colors, mono chromatic, black color on white background, casey baugh, 3 3 7 4 2 1 6 5 8 1	0.06	Color, Monochromatic, Soft
vintage, minimalism, figurative, beautiful, colorful, cyber punk, minimal, figure art, soft colors, mono chromatic, black color on white background, casey baugh, 1 1 4 3	0.06	Color, Monochromatic, Soft
vintage, minimalism, figurative, beautiful, colorful, cyber punk, minimal, figure art, soft colors, mono chromatic, black color on white background, casey baugh, 2 2 2 1 1 4 3	0.06	Color, Monochromatic, Soft
vintage, minimalism, figurative, beautiful, colorful, cyber punk, minimal, figure art, soft colors, mono chromatic, black color on white background, casey baugh,	0.06	Color, Monochromatic, Soft

TABLE 4.14: Top 10 Lowest Scoring Prompts in with artist dataset

Chapter 5

Discussion and Conclusion

This study aimed to investigate the creativity of the prompt dataset, designed by (Wang et al., 2022), based on the automated version of the Alternate Uses Task (AUT), titled SemDis, created by (Beaty and Johnson, 2020). The goal was to find and apply a creativity assessment tool on prompts, in order to investigate the creativity of short-text prompt data. The results from this study contribute to a better understanding of the design processes behind the generation of AI art, a rapidly evolving field that lies at the intersection of technology and art.

5.1 Findings and Interpretation

The main research question, *How can Natural Language Processing (NLP) methods be used to assess the creativity of short-text prompt data in AI-art generation?*, was constructed to assess the current capabilities and limitations of methods used for creativity analysis. The study used various NLP clustering techniques to ultimately examine the text prompts used in diffusion models to generate art, providing insights into the elements that contribute to a satisfactory final outcome.

The exploration and preprocessing of the dataset, which prepares the prompts for further analysis, was found to be an important step in the study. This process not only removed potential noise from the data but also highlighted the difference in results when artist names were included or excluded from the cleaned prompts. The resulting datasets provided a base for the subsequent analyses done in this project.

The primary research question of this study was guided by two subquestions. The first subquestion, *How does the level of detail and complexity in prompts impact the perceived creativity of the generated images?*, was formulated to analyze the perceived creativity, as indicated by the creativity score, in relation to the complexity and intricacy of a prompt. The results suggest that prompts that are more detailed and specific, often involving complex and imaginative scenarios, are perceived as more creative.

This observation can be attributed to the bigger semantic distance between the item and the prompts to which they were compared, indicating a more creative use of language. This finding aligns with the idea that divergent thinking contributes to higher perceived creativity (Syahrin, Suwingyo, PRIYATNI, et al., 2019).

These results are consistent with the definition of creativity as the generation of ideas that are both novel and surprising (Runco and Jaeger, 2012). This suggests that the ability to generate detailed and specific prompts, which inherently require a higher level of divergent thinking, may be a crucial aspect of creativity in the context of text-to-image models.

Therefore, it can be said that prompts exhibiting higher complexity and divergence can lead to the generation of more creative ideas, and subsequently, more creative art. This finding has significant implications for the field of AI art generation, as it highlights the importance of the quality of prompts in influencing the creativity of the generated art.

The second subquestion, *What role do elements like artist names and aesthetic quality modifiers play in enhancing or diminishing the perceived creativity of a prompt?*, was formulated in response to the prompting guidelines that are widely distributed on the internet to guide new prompters in their art creation journey (Oppenlaender, 2022; Pavlichenko, Zhdanov, and Ustalov, 2022; Taylor, 2023). These guidelines often recommend the use of prompt modifiers, which are intended to enhance the outcome of the text-to-image process. These modifiers typically include aesthetic quality descriptors and artist names.

In the course of the study, it was observed that prompters frequently incorporate artist names and style modifiers into their prompts. These elements make up a significant portion of the terms found in the prompts, indicating their perceived importance in the art creation process. The inclusion of these elements can provide a more detailed and specific context for the AI model, potentially enabling it to generate images that are aesthetically more to the prompter's liking.

To analyse the influence of artist names on the perceived creativity of a prompt, two pre-processed datasets were created: one that includes artist names and one that excludes them. The results revealed that the inclusion or exclusion of artist names influenced the labeling of the topics found in the dataset. Interestingly, the dataset that included artist names had a slightly lower average perceived creativity score compared to the dataset without artist names. However, this difference was not statistically significant. This suggests that while artist names and style modifiers are frequently used in prompts, their impact on the perceived creativity of a prompt may not be as substantial as initially thought.

The skill of prompting has gained significant importance, not only in the realm of AI-art generation but also in other areas where tools like ChatGPT are used. The process of prompting requires the formulation of our ideas in a manner that yields the desired results. Therefore, understanding how individuals interact with these tools and construct their prompts is crucial.

This is particularly important when considering the enhancement of diffusion models through improved prompt understanding, as demonstrated in (Lian et al., 2023). The study illustrates that the quality of the generated image is not solely dependent on the AI model's capabilities, but also on the quality and clarity of the prompts provided. This observation aligns with the focus of our study on the role of prompts in AI-art generation and the creativity behind them.

Moreover, the findings of this study suggest that regardless of the perceived creativity in the text itself, the quality of prompts helps users achieve the generated images they desire. This underscores the importance of prompt construction in the process of AI art generation. It suggests that the careful selection and use of prompt modifiers, including artist names and aesthetic quality descriptors, can be a powerful tool for enhancing the creativity of the generated art. This insight could be valuable for both new and experienced prompters, as it provides a concrete strategy for enhancing the creativity of their prompts and the art they generate.

5.2 Limitations

This research encountered several challenges that should be discussed.

Firstly, the GSDMM method, which was initially anticipated to be a good fit for the data, fell short of expectations. The method proved to be time-consuming and failed to generate evenly distributed clusters, thereby limiting its effectiveness in this context.

Secondly, computational constraints restricted the application of BERTopic to the entire dataset. As a result, the study had to resort to using a subset of the data for BERTopic application. This limitation could potentially skew the results and does not provide a comprehensive view of the entire dataset. Additionally, the file size limit of SemDis posed another challenge, making it impossible to generate creativity scores for larger file sizes.

Thirdly, the study applied SemDis to short-text data, which is not the typical data structure SemDis is designed to handle. This deviation from the intended use of SemDis could potentially affect the validity of the scores. Furthermore, the absence of human ratings on the creativity of these prompts means the study lacks a benchmark against which to validate the results of the SemDis algorithm.

Lastly, the nature of the data prevented the execution of an ANOVA test, necessitating the use of a nonparametric validation technique, specifically the Kruskal Wallis H Test. While this test is a valid statistical method, it may not provide the same depth of insight as an ANOVA test, thus limiting the study's findings.

Despite these limitations, the study offers valuable insights into the application of AI in creativity assessment. However, these constraints should be taken into account when interpreting the results and planning future research.

5.3 Future Work

This study's findings suggest several intriguing future research directions.

Firstly, the study successfully employed SemDis, an automated creativity assessment tool, on a prompt dataset. However, this usage doesn't fully align with the AUT, suggesting that the results might not capture all aspects of human creativity assessment. Future research could enhance this by integrating a human element into the creativity assessment process. Human judges could rate the creativity of prompts, providing a validity measure for SemDis's automated results. This would offer a better understanding of the prompts' creativity and a robust validation of SemDis results. Additionally, human judges could help explore the relationship between prompts and AI-generated art, potentially revealing how to optimize prompts for enhanced creativity and aesthetic appeal.

Secondly, the study primarily focused on the creativity inherent in the prompt text data, generating a creativity score for these prompts. Future research could extend this by evaluating the AI-generated artworks themselves. Assessing whether the creativity scores of prompts correlate with the perceived creativity of the generated artworks would provide a holistic view of the creative process. Human judges could evaluate the creativity of the generated artworks, providing a subjective measure of creativity and insights into audience perception. Alternatively, using an already rated dataset, such as the dataset assembled in (Pressman, Crowson, and Contributors, 2022), could offer a more objective assessment of creativity.

Thirdly, the rise of text-to-image models and AI prompt marketplaces presents an opportunity to integrate creativity score mechanisms. Implementing a system that quantifies the potential creativity of the to-be-generated art could provide users with insights into their prompts' creative possibilities before the art generation process. This could save time and computational resources, fostering a more informed and creatively rich process. A neural network model, trained on a large dataset of prompts, could predict a given prompt's creativity score. Incorporating creativity scores into AI prompt marketplaces could revolutionize the user experience, making AI art generation more transparent, interactive, and creativity-focused.

5.4 Conclusion

In conclusion, this thesis has explored the intersection of creativity and artificial intelligence, specifically focusing on how prompts can be used to generate artworks through text-guided synthesis of images using deep generative models. The research has delved into the intricacies behind the prompting process which is a fundamental part of creating art with text-to-image models.

The study has also examined the use of different topic modeling methods, which were applied to understand and analyze data, which in turn helps understanding the prompts. The findings from this research have shed light on how creativity is represented in the prompts and how it can be objectively measured by text-mining models.

However, it's important to acknowledge that this field is still rapidly evolving, and there are many aspects of creativity and AI that are yet to be explored. Future work could delve deeper into the understanding of how different prompt templates can improve the aesthetic appeal of images and how the use of different topic modeling and labeling techniques can provide further insights into the data.

Overall, this research marks a significant contribution to the expanding body of knowledge on AI and creativity, offering valuable insights of relevance to both academics and practitioners in the field. In a period where the concepts of art and creativity are undergoing an evolution, the need to understand the creative process behind AI art generation intensifies. This study addresses that need, laying a foundation for future discourse and investigations of art and creativity within the AI context.

Appendix A

Ethics and Privacy Scan

Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?

- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?

- No

H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?

- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)

- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)

- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?

- No

H6. Does your project give rise to a realistic risk of harm to the researchers?

- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?

- No

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?

- No

H9. Is there a realistic risk of other types of negative externalities?

- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information. Please see the guidance on the ICS Ethics and Privacy website on what happens on submission.

Z0. Which is your main department?

- Other, namely::
  Natural Sciences

Z1. Your full name:
Saif Abdoelrazak

Z2. Your email address:
s.abdoelrazak2@students.uu.nl

Z3. In what context will you conduct this research?

- As a student for my master thesis, supervised by::
  Almila Akdag

Z5. Master programme for which you are doing the thesis

- Artificial Intelligence

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):
coordinator-ai-master@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):
d.p.nguyen@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:
Creativity Behind the Prompts: Automated Creativity Assessment in Prompting for Text-to-Image Models

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):
In this thesis project I will assess the creativity of short-text prompt data that is used to generate art with text-to-image algorithms. The data that is being used is lifted from the Stable Diffusion Discord channel. To be more specific, the open source database named DiffusionDB will be used. To analyze the creativity, I will apply GSDMM and Bertopic to find categories in the dataset, to prepare it for a SemDis analysis, which will emulate the Alternative Uses Task, essentially creating a creativity score for the prompts.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

- Not applicable

Scoring

- Privacy: 0
- Ethics: 0

Appendix B

List of stopwords

Alphonse Mucha
Mark Rothko
René Magritte
Henri Matisse
Leonardo da Vinci
Gustave Doré
James Gurney
Bill Watterson
Funko Pop
Toshi Yoshida
Caravaggio
Galen
Frank Gehry
Simon Stålenhag
Norman Rockwell
Edward Hopper
H.R. Giger
Fellini
Robert Hargreaves
Samuel Daniell
Pablo Picasso
Georges Braque
Frida Kahlo
Johannes Vermeer
Georges Seurat
Ralph McQuarrie
Francisco Goya
Claude Monet
Lucian Freud
Diego Rivera
Frank Frazetta
Pieter Brueghel
Saul Leiter
Banksy
Dr. Seuss
Lisa Frank
Brueghel the Elder
Keith Haring
Ivan Shishkin
Piet Mondrian
Andy Warhol
Dean Cornwell
Vincent van Gogh
Thomas Cole
Zdzislaw Beksinski
Pierre Auguste Renoir
Edvard Munch
Yoji Shinkawa
Hatsune Miku
John Constable
Annie Leibovitz
Dorothea Lange
Hanna-Barbera
Axel Schaefer
Thomas Kinkade
Duffer Brothers
Tim Burton
Albert Bierstadt
Jackson Pollock
Roy Lichtenstein
Craig Mullins
Michelangelo
Greg Rutkowski
Claude Lorrain
Hayao Miyazaki
Katsushika Hokusai
Utagawa Kuniyoshi
Hieronymus Bosch
Salvador Dalí
Wes Anderson
Terry Richardson
Beatrix Potter
Keith Haring
Basquiat
Arcimboldo
M.C. Escher
Robert McCall

TABLE B.1: List of Artist Names

Hudson	River	School	Photograph
Trending	On	ArtStation	Control
The	Soul	From	Behind
Surrealism	Painting	Beautiful	Futuristic
Wide	Angle	Dadaism	Oil
Vivid	Utopian	Fisheye	Lens
Cubism	Line	Drawing	Professional
Dystopian	Telephoto	Fauvism	Charcoal
Drawing	Extremely	Detailed	Blade Runner
Wide	Angle	Renaissance	Macro
35mm	Photograph	Stunning	Cinematic
Panoramic	Impressionist	Album	Art
Cover	Wondrous	Fantasy	Bokeh
Baroque	One-line	Drawing	Fantastic
Elegant	Lens Flare	Rococo	Crayon
Drawing	Contest	Winner	No Mods
Hard Lighting	Romanticism	Pastel	Drawing
Postprocessing	Magnificent	Landscape	Realism
Watercolor	Detailed	Retrofuturistic	Drone
Pointillism	Pencil	and Watercolor Drawing	Trending On r/art
Awesome On	Canvas	Symbolism	Vector Art
8K	Transhumanist	Close Up	Neoclassicism
Stained Glass Window	4K Resolution	From Below	Art Nouveau
Van Gogh	Cartoon	VFX	Wormhole
From Above	Expressionism	3D Render	Rendered In UE
Eclectic	Through a Periscope	Constructivism	Da Vinci
Line Art	Octane Render	Epic	Through a Porthole
Futurism	Pixel Art	Digital Art	Tasteful Framed
Suprematism	Product Photography	Photo	Minimal
Minimalistic	Mixed Media	Monochrome	Moody
Mysterious	Mystical	Natural Lighting	Naturalistic
Night Scene	Nostalgic	Noir	Octane
Oil	Oil Painting	Organic	Origami
Panoramic	Pastel	Pastel Colors	Photography
Photorealistic	Photo	Pixel Art	Playful
Pointillism	Polished	Polychromatic	Portrait
Pop Art	Psychedelic	Realistic	Realistic
Retro	Romantic	Romantic	Rule of Thirds
Rural	Sci-Fi	Sensual	Serene
Serene	Saturated	Single Color	Sketchy
Soft Lighting	Square	Stark	Static
Style	Stylized	Subtle	Surreal
Surrealistic	Symmetrical	Textured	Textured
Tenebrism	Tranquil	Trending	Trending On
"Trompe l'oeil"	Ultra	Unreal	Unreal Engine
Urban	Utopian	Vanishing Point	Vertical
Vibrant	Vibrant	Vintage	Vivid
Watercolor	Whimsical	Whimsical	Wide Shot
Whispering			

TABLE B.2: List of Extra Prompt Modifiers

Appendix C

Data

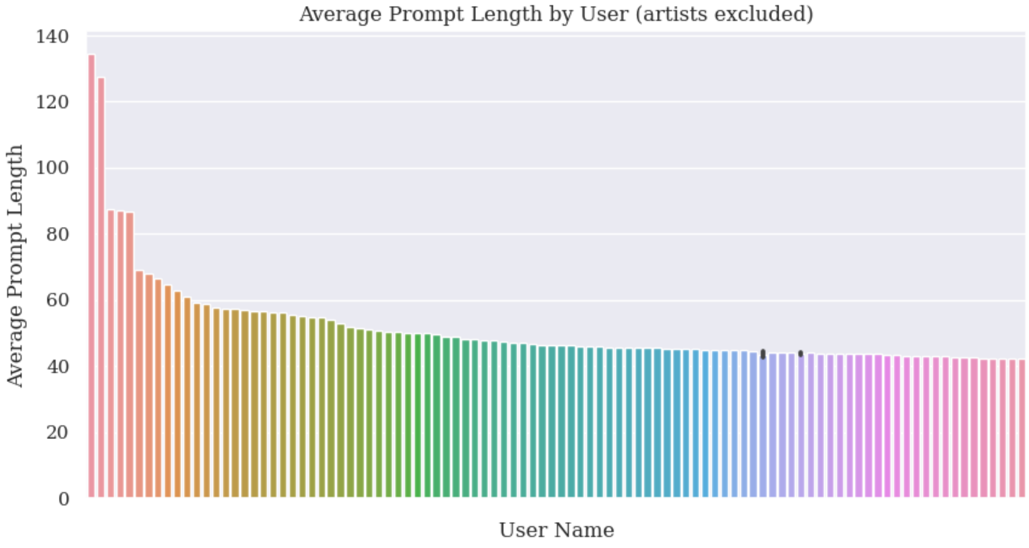


FIGURE C.1: Average Prompt Length (No Artist)

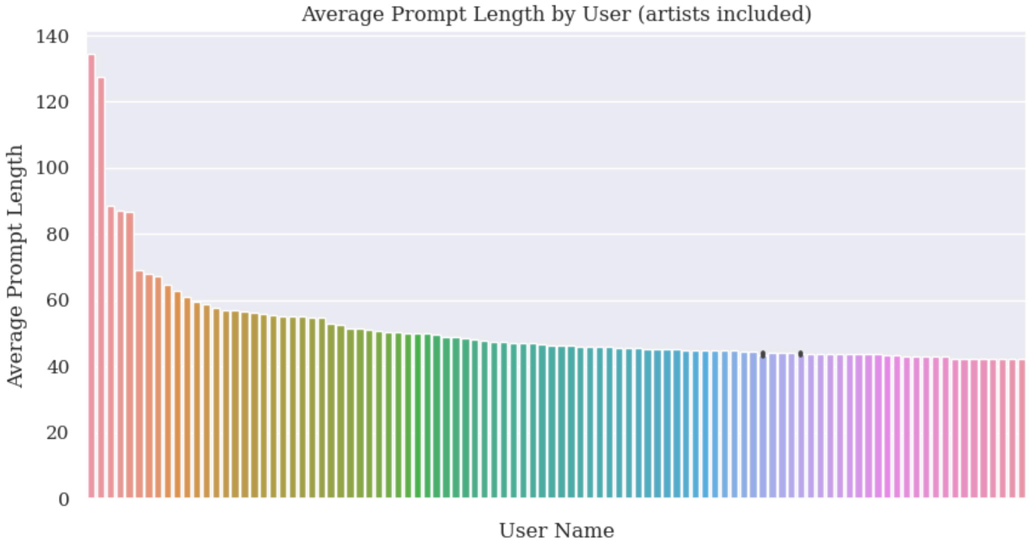


FIGURE C.2: Average Prompt Length (With Artist)

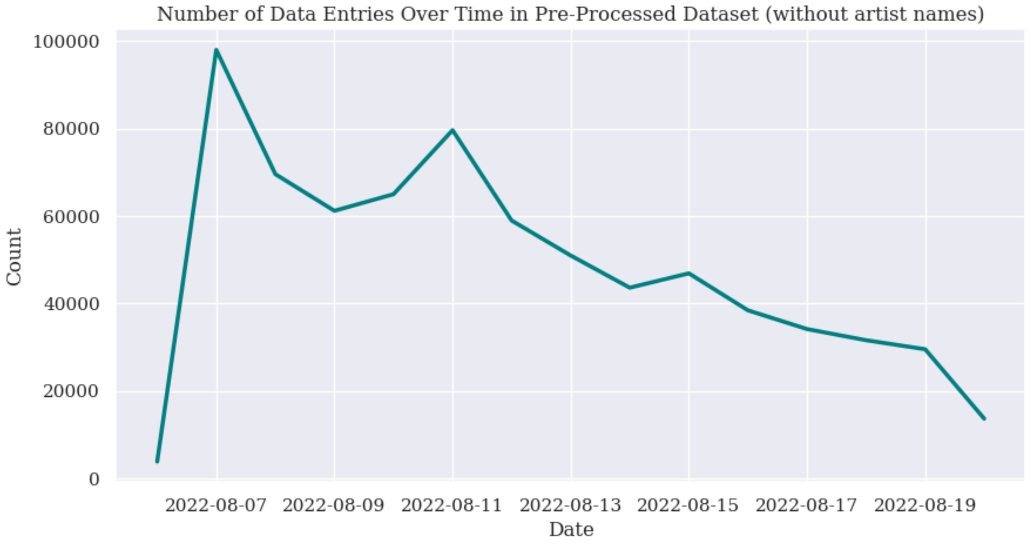


FIGURE C.3: Entries Over Time (Cleaned Dataset, No Artists)

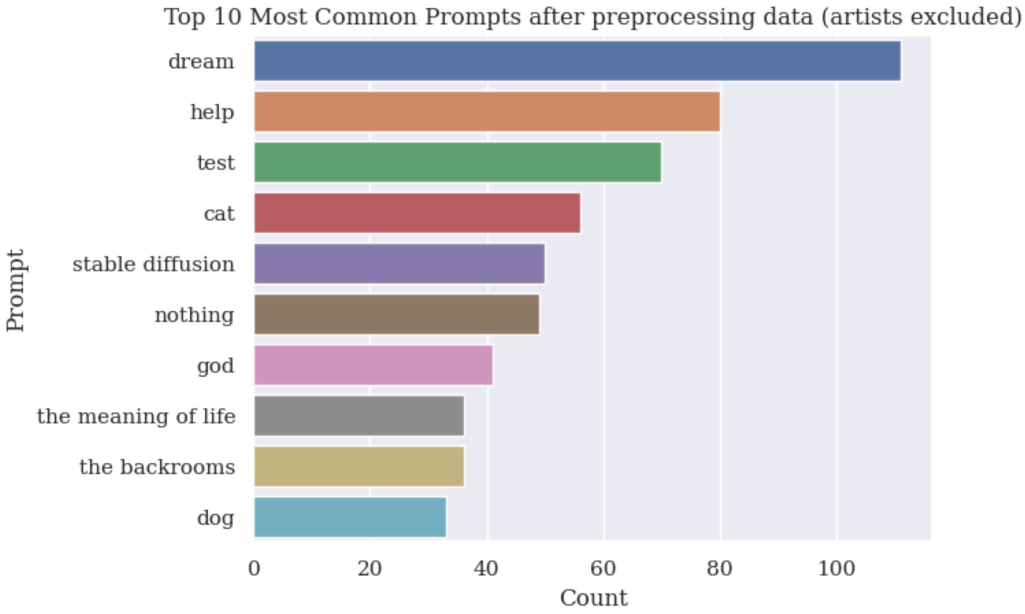


FIGURE C.4: Top 10 Common Prompts (Cleaned Dataset, No Artists)

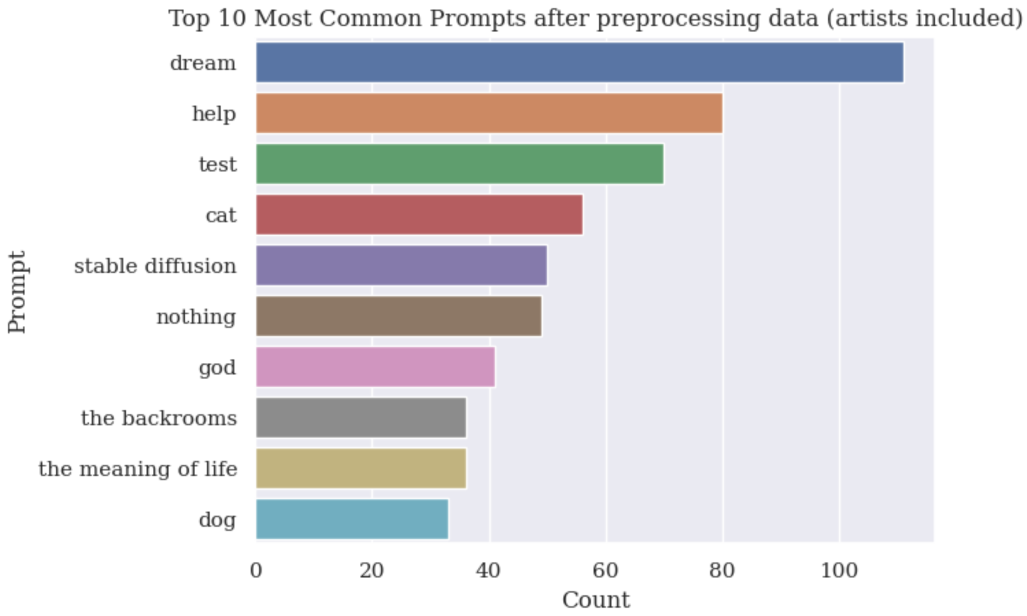


FIGURE C.5: Top 10 Common Prompts (Cleaned Dataset, With Artist)

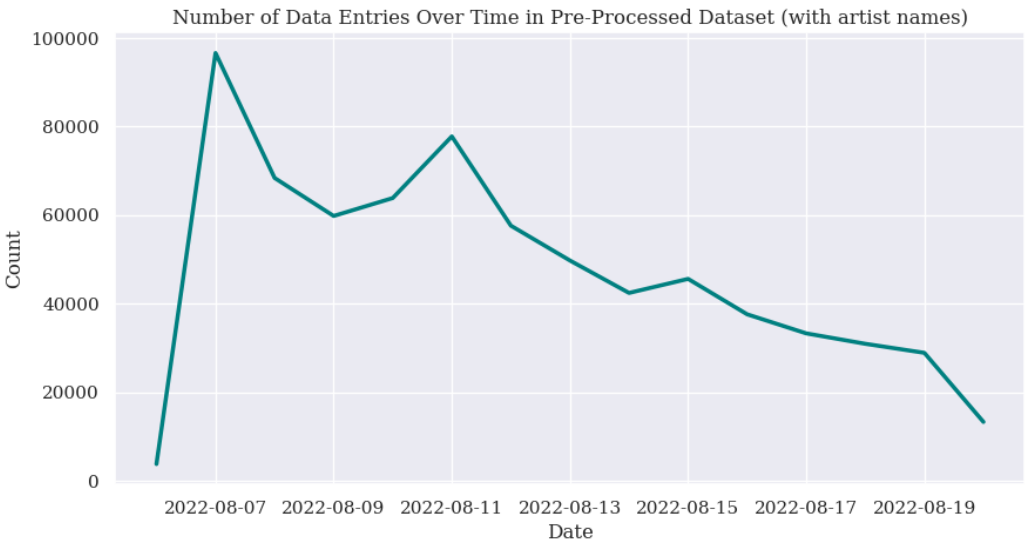


FIGURE C.6: Entries Over Time (Cleaned Dataset, With Artist)

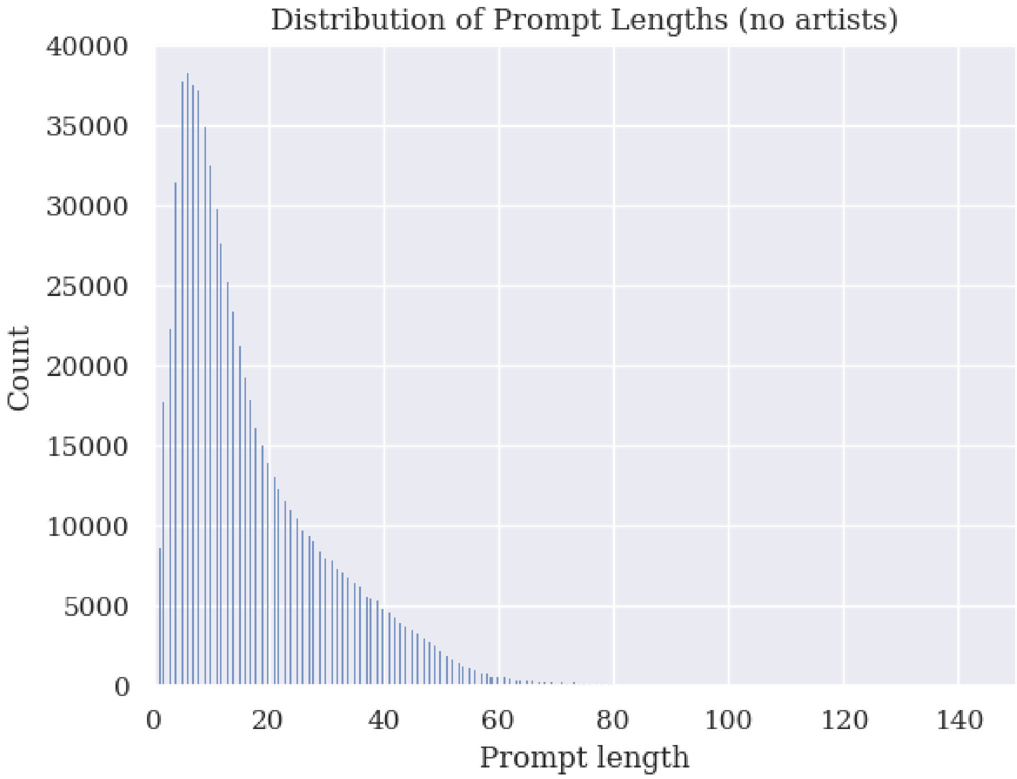


FIGURE C.7: Distribution of Prompt Lengths (No Artist)

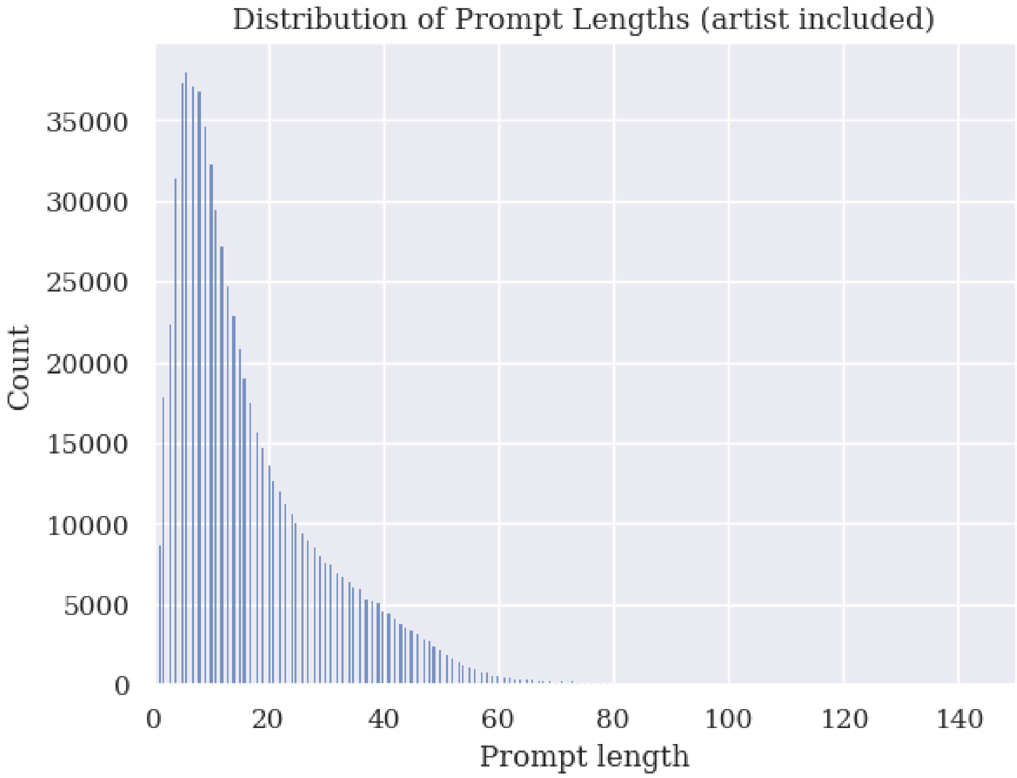


FIGURE C.8: Distribution of Prompt Lengths (With Artist)

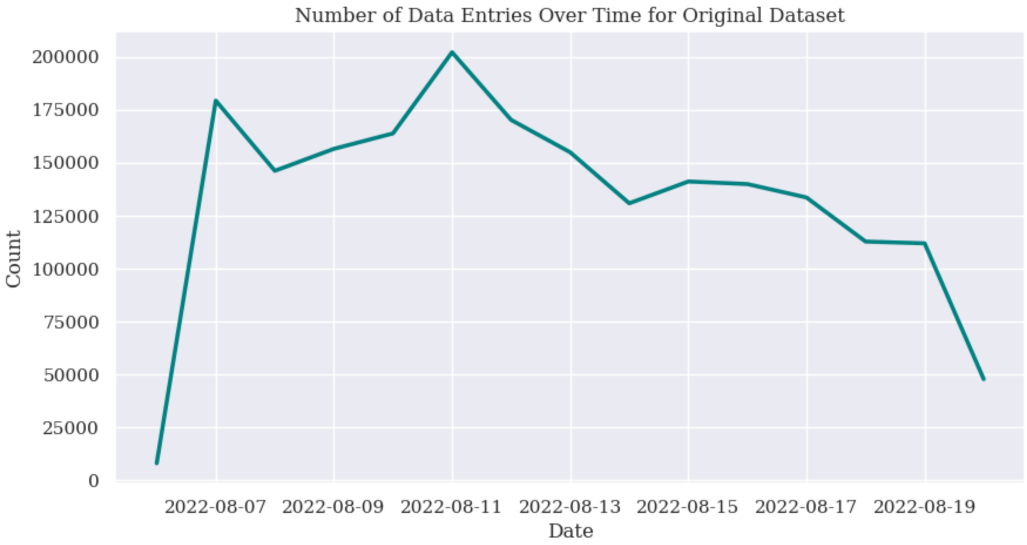


FIGURE C.9: Entries Over Time (Original Dataset)

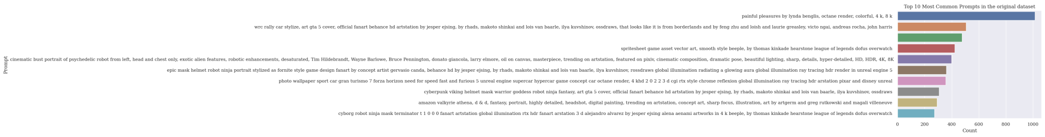


FIGURE C.10: Top 10 Prompts (Original Dataset)

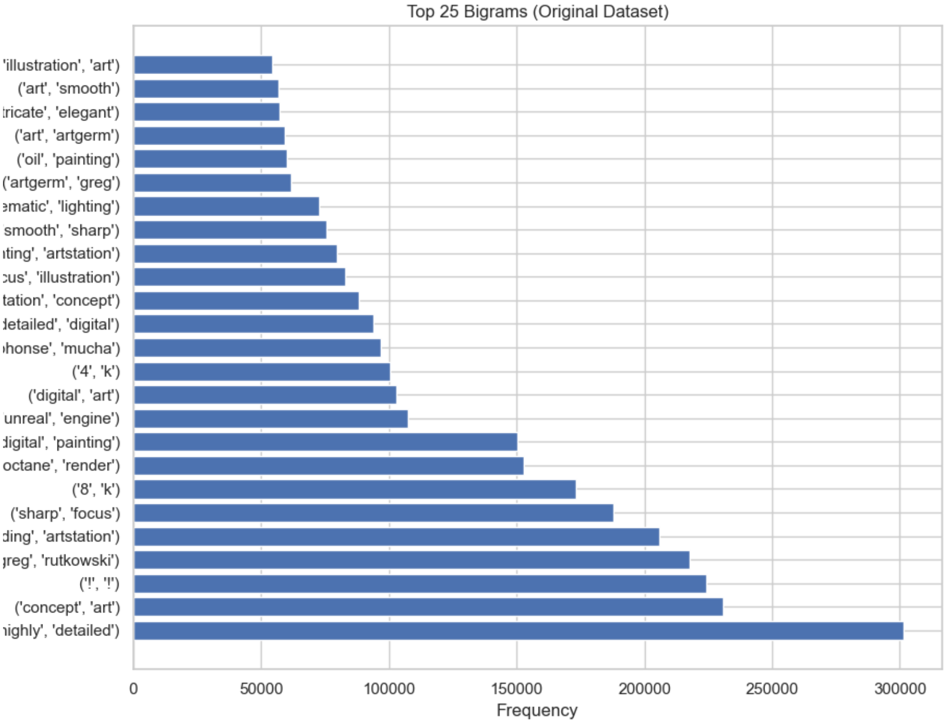


FIGURE C.11: Top 25 Bigrams (Original Dataset)

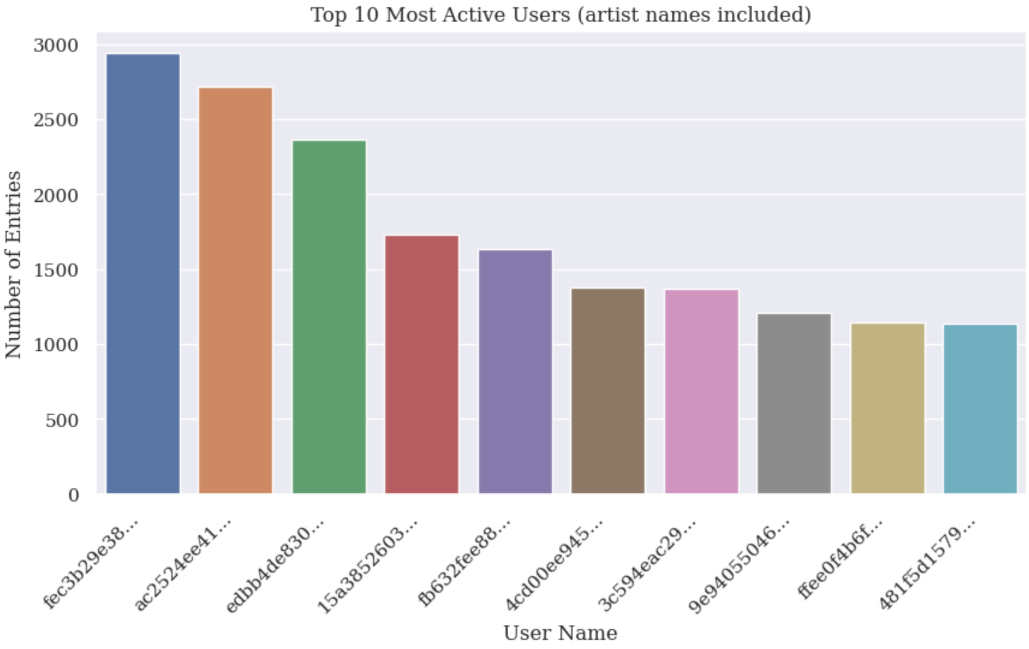


FIGURE C.12: Top 10 Active Users (No Artists)

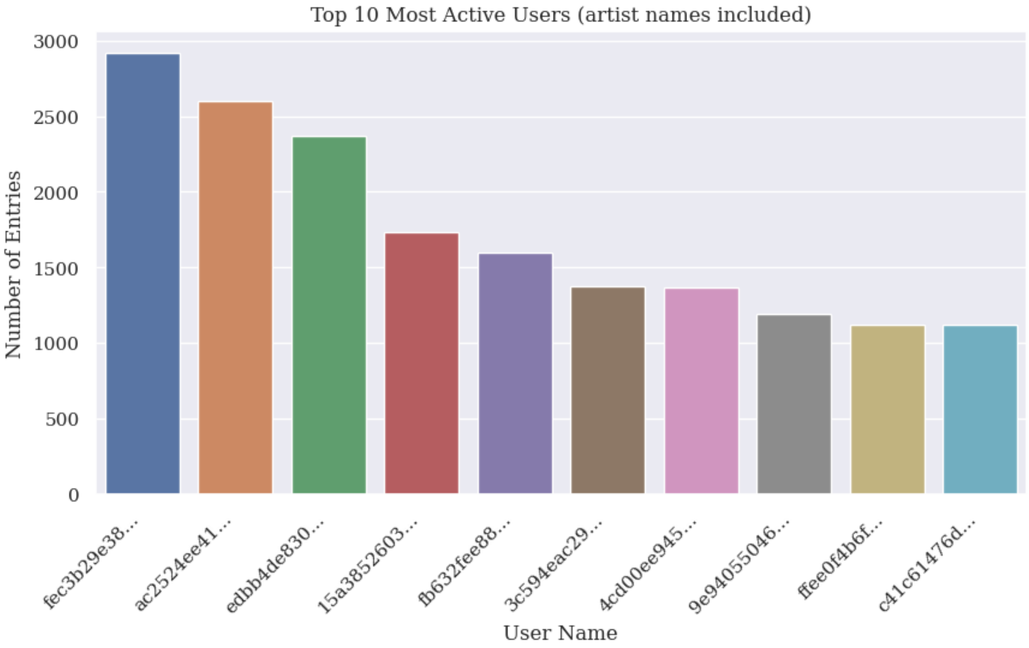


FIGURE C.13: Top 10 Active Users (With Artist)

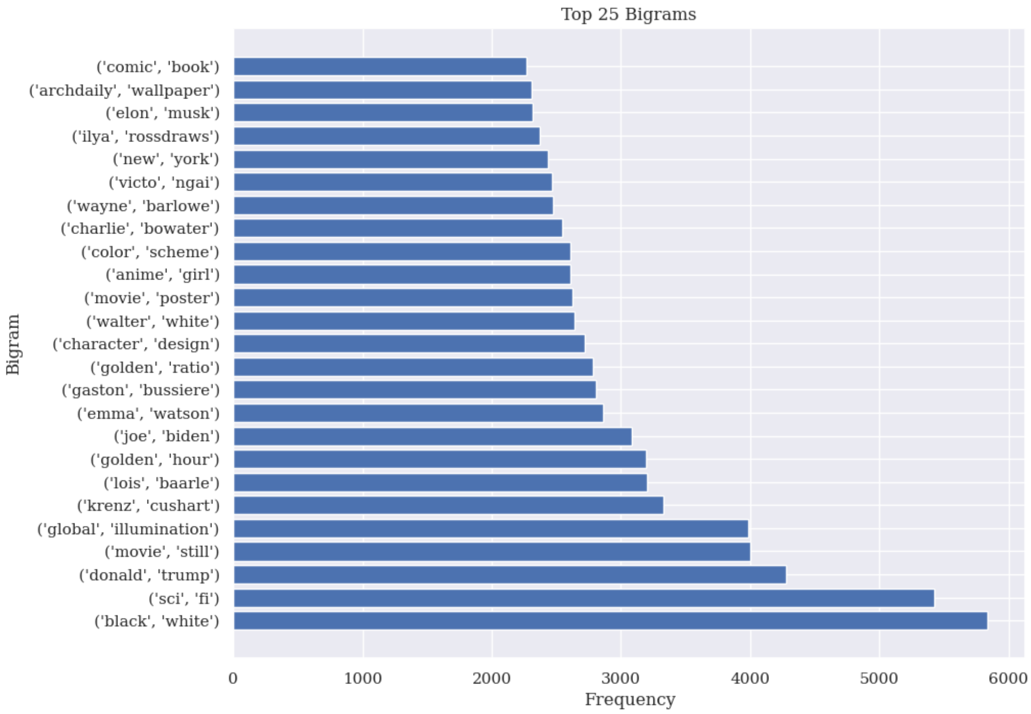


FIGURE C.14: Top 25 Bigrams (No Artist Names)

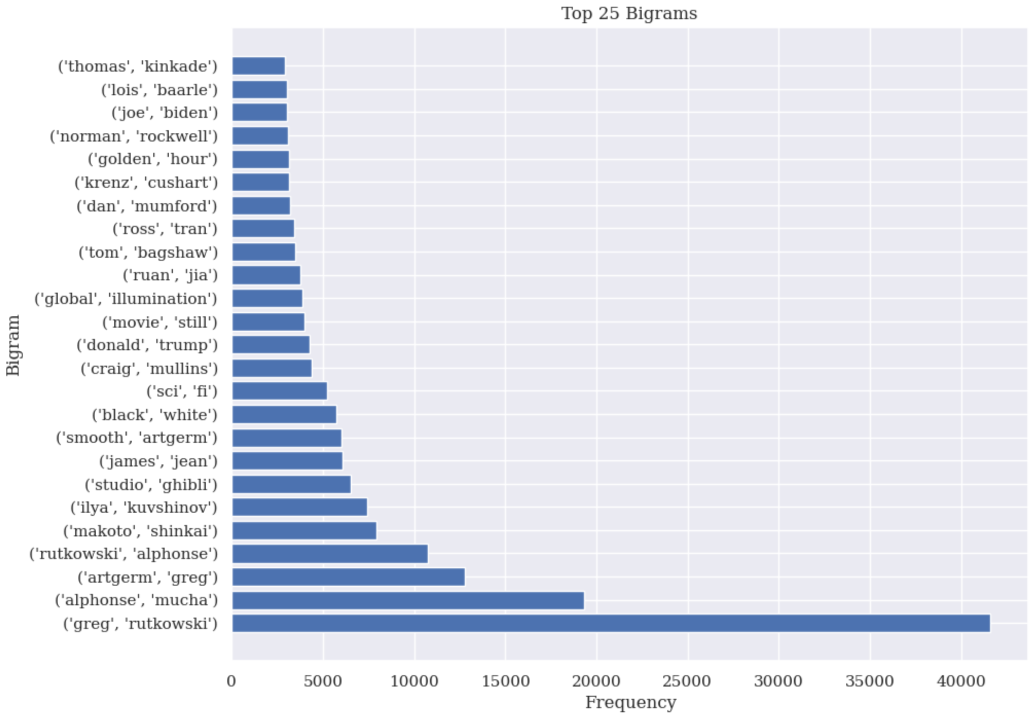


FIGURE C.15: Top 25 Bigrams (With Artist Names)

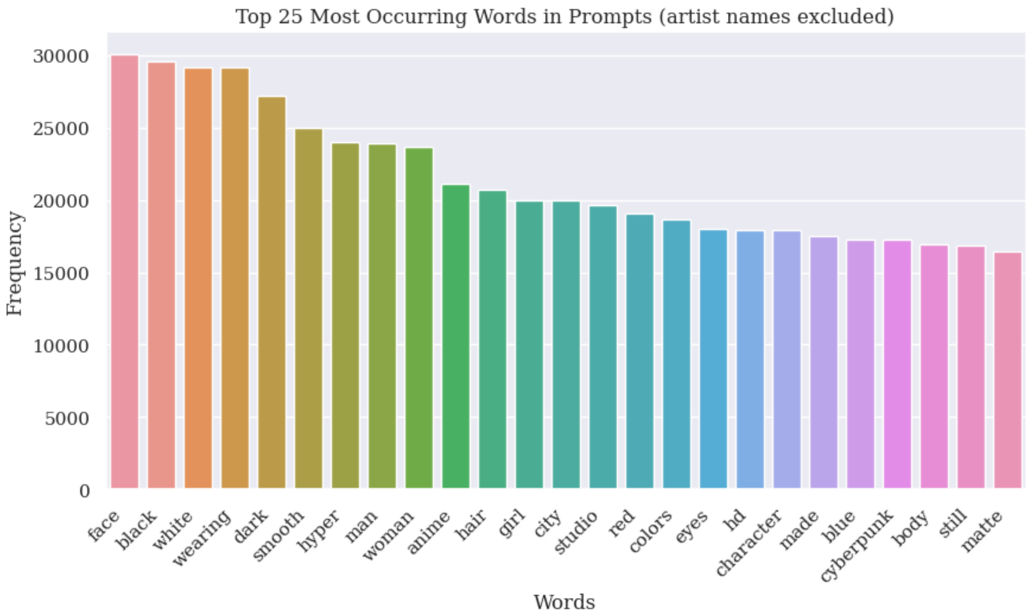


FIGURE C.16: Top 25 Occurring Words (With Artist)

TABLE C.1: SemDis results after BERTopic clustering and labeling

Label	**Length**	**Mean**	**Median**	**Std Dev**
top_n No Artist	1	0.9557	0.9777	0.1529
top_n No Artist	2	0.9610	0.9795	0.1736
top_n No Artist	3	0.9647	0.9807	0.1822
top_n No Artist	4	0.9900	0.9902	0.2134
top_n No Artist	5	0.9772	0.9845	0.1937
top_n With Artist	1	0.9565	0.9778	0.1652
top_n With Artist	2	0.9584	0.9772	0.1674
top_n With Artist	3	0.9805	0.9887	0.1740
top_n With Artist	4	0.9730	0.9858	0.2096
top_n With Artist	5	0.9909	0.9935	0.2232
openai No Artist	1	0.9773	0.9835	0.0817
openai No Artist	2	0.9845	0.9939	0.1006
openai No Artist	3	0.9863	0.9899	0.1111
openai No Artist	4	0.9845	0.9925	0.1047
openai No Artist	5	0.9878	0.9929	0.1264
openai With Artist	1	0.9834	0.9879	0.0848
openai With Artist	2	0.9881	0.9948	0.0941
openai With Artist	3	0.9891	0.9921	0.1124
openai With Artist	4	0.9899	0.9950	0.0944
openai With Artist	5	0.9923	0.9943	0.1394

TABLE C.2: SemDis Results after GSDMM clustering and labeling

Label	**Length**	**Mean**	**Median**	**Std Dev**
top_n No Artist	1	0.9774	0.9777	0.0944
top_n No Artist	2	0.9794	0.9818	0.1156
top_n No Artist	3	0.9887	0.9906	0.1304
top_n No Artist	4	0.9882	0.9926	0.1858
top_n No Artist	5	0.9898	0.9988	0.2313
top_n With Artist	1	0.9840	0.9859	0.1002
top_n With Artist	2	0.9849	0.9859	0.1035
top_n With Artist	3	0.9830	0.9858	0.1095
top_n With Artist	4	0.9825	0.9879	0.1414
top_n With Artist	5	0.9806	0.9885	0.1889
openai No Artist	1	0.9906	0.9912	0.0702
openai No Artist	2	0.9957	0.9966	0.0783
openai No Artist	3	0.9979	0.9972	0.1956
openai No Artist	4	0.9963	0.9970	0.1035
openai No Artist	5	0.9865	0.9952	0.1982
openai With Artist	1	0.9853	0.9904	0.0803
openai With Artist	2	0.9925	0.9944	0.0699
openai With Artist	3	0.9940	0.9976	0.0702
openai With Artist	4	0.9949	0.9975	0.1503
openai With Artist	5	0.9937	0.9969	0.2032

Topic	Frequency	Topic	Frequency	Topic	Frequency
Fashion, Beauty, Art	2275	Art, Comedy, Adventure	1548	Divine Proportionality	892
Science Fiction Hybrid	718	Jordan, Grimmer, Scene	516	Eating, Photography, Sweets	450
Tabbi Cat Fur	412	Space exploration speed.	377	High-tech dystopia	363
Art, Politics, Fear	333	Family, Jean, Peter.	330	Comic artistry.	317
Appropriate Icons	315	Storage, Backup, Imaging	297	Space, Travel, Luxury	285
Armor, Knights, Dragons	259	Paper, Worn, Yellow	254	Photography, Film, Color	253
Electronic, Dance, Music	233	Art, Creativity, Collaboration	230	Emma Watson Hermione	227
Aging, jewelry, symbolism.	223	Making, Experimenting, Creating	216	Fox, Fluffi, Tail	212
Early photography technology.	203	Donald Trump Presidency	200	Role-Playing Games	197
Extraterrestrial, Flying, Dragons	196	Science Fiction Movies	195	Vibrant, Luminous, Imag	194
Nature, beauty, light.	190	Starcraft, Mandi, Smooth	181	Artistic, Sculpting, Design	180
Awards, Horror, Gloom	175	Breaking Bad protagonist.	170	Joker, Movie, New	167
Geometric Shapes.	162	Fear, visuals, stories.	162	People, cars, racing.	157
Urban, Destruction, Desolation	151	Color coordination/harmony.	151	Modern, Super, Hyper	150
Heavy, Grain, Cinema	148	Dark, Force, Vader	146	Pirates, Smoking, Drinking	142
Dragons, Fire, Breath	138	Wood architecture design.	138	Futuristic, texture, gaming.	135
Art, Mysticism, Expression	133	Nature, Softness, Beauty	128	President, Vice President, Biden	127
Dungeons & Dragons	124	Atomic, Tropical, Relaxation	123	Medieval, Woman, Gaston	122
Platformer video game	120	Imagination, creativity, beauty.	120	Animal Movement.	117
Tattoo, Design, Cate	116	Love, Attraction, Confession	115	Exploring darkness caverns.	111
Sports, Fantasy, Basil.	110	Magic, Forest, Surround	110	Video game screenshots.	109
Collectibles, Toys, Decor	109	Female Vampire Look.	109	San Francisco landmark.	106
Simpson, Homer, Nihei	105	Innovator, Entrepreneur, Vision	105	Mental health issues.	105
Weather, Photography, Ritual	104	Comic book RPGs	98	Connectivity, Sharing, Networking	97
Wooden Doll Female	97	Patriotic symbolism.	96	Wrestling Iconography	96
Art, Women, Screaming	96	Interdimensional travel.	93	Zombies, Swimming, Underwater	93
Photography, lens, aperture	92	Double Exposure Tone	89	Creative, imaginative, Snoopi	86
Anime, robots, art.	86	Wildlife Surveillance Video	85	Animals, Photography, Nature.	85
Future, Believing, Beauty	84	Urban, Middle, Area	82	Roman, Mosaic, Byzantin	82
Plants, Water, Rocks	82				

TABLE C.3: Generated Topics for the no_artist Dataset

Topic	Frequency	Topic	Frequency	Topic	Frequency
Manga, Art, Adventure	1595	Food, Bowl, Candy	887	Beauty, Art, Shade	1147
Art, Rutkowski, Mucha	861	Golden Ratio, Compositing, Faces	806	Tabbi, White, Grey	478
Art, Women, Portraits	433	Colour, Warm, Artist	431	Color, Monochromatic, Soft	410
Cyberpunk, Steampunk, Syn	373	Comic, Joshua, Albuquer	366	Art, Geometry, Surrealism	332
Artificial Intelligence, Storytelling, Post	331	Automobiles, Technology, Suzuki	295	Medieval, Armor, Magic	282
Golf, Nature, Memory	278	Art, Sculpting, Ratio	275	Clive Barker horror.	275
Sea, Art, Fishing	271	Music, Collecting, Performance	267	Space Exploration, Sci-Fi, Flying	265
Anthropomorphic Foxes	254	Photography, Canon, Kodak	232	Dragons, clouds, monsters.	230
Sci-Fi, Action, Movies	215	Acting, Awards, Characters	213	Dogs, Sitting, Middle	212
Space, Cyberpunk, Emergence.	210	Art, Ink, Process	203	Fish, Swim, Land	198
Bright, Vibrant, Energy	190	Video Games, RPG, Playtime	187	Mathematics, Art, Animation	184
Labor, Survival, Renewal	174	Medieval castles.	172	Leonardo, Mona Lisa, Last	170
Donald Trump, Ride, Look	164	Joker, Color, Heavy	162	Business, Relationships, Attraction	161
Star Wars, Vader, Movies	157	Middle Eastern, Glass, Baseball.	151	Walter White protagonist	148
Wood, Architecture, Render	145	Ewa Czarniecka	144	Atey, Ghailan,	131
Animation, Disney, Kodak	126	Nature, Variety, Color	124	Fairytale, Nature, Fashion	121
Animals, Zoos, Hybrids	117	Moebius, Color, Symm	117	Circle, Age, Connection	117
Real, Life, World	114	Literature, Environment, Character.	114	Animation, Visuals, Bean.	114
Nature, cycle, moon.	112	Saul Goodman lawyer.	112	Elon Musk products.	111
Video Games, Screenshots, UI	110	Underground, Dark, Cavern	110	Geometry, Graffiti, Edge.	110
Football, Art, Algorithm	109	Mario, Bros, Racing	109	Urban, Artistic, Vibrant	108
Tarot, Magician, Ornat	108	Joe Biden, Donald Trump, Cloning	108	Sadness, Reflection, Morbid	108
Award, Rain, Polaroid	105	Video Surveillance, Recording, Quality.	105	Fantasy, Art, Clutter	105
Film, Grain, History	105	Tattoo, Design, Artger	103	Acting, Art, Beauty	103
Photography, Portraits, Aging	101	Color composition, aesthetics, tones.	99	American, Flag, Stand	96
High-Definition, Grainy, Night	96	Artists, Landscapes, Port	94	Collecting, Home Decor, Fun	94
Extinct, Mythical, Mythos	93	Swimming, Plants, Rocks	93	Modern, Hyper, Mega	92
Fantasy, portals, demons	92	Divine, Illumination, Nature	90	Bears, Eating, Lounging	86
Photography, Dinosaurs, Destruction	84	Tennis, Gogo, Basil	84	HD, Bodi, Headshot	81

TABLE C.4: Generated Topics for the `with_artist` Dataset

Bibliography

Acar, Selcuk and Mark A Runco (2014). “Assessing associative distance among ideas elicited by tests of divergent thinking”. In: *Creativity Research Journal* 26.2, pp. 229–238.

Ahmed, Majid Hameed et al. (2022). “Short Text Clustering Algorithms, Application and Challenges: A Survey”. In: DOI: 10.3390/app13010342. URL: https://doi.org/10.3390/app13010342.

Ahmed, Sana Tariq and Gregory J. Feist (Nov. 2021). “The Language of Creativity: Validating Linguistic Analysis to Assess Creative Scientists and Artists”. In: *Frontiers in Psychology* 12. ISSN: 16641078. DOI: 10.3389/fpsyg.2021.724083.

Amabile, Teresa M (1982). “Social psychology of creativity: A consensual assessment technique.” In: *Journal of personality and social psychology* 43.5, p. 997.

Baer, John and Sharon S McKool (2009). “Assessing creativity using the consensual assessment technique”. In: *Handbook of research on assessment technologies, methods, and applications in higher education*. IGI Global, pp. 65–77.

Beaty, Roger E and Dan R Johnson (2020). “Automating creativity assessment with SemDis: An open platform for computing semantic distance”. In: DOI: 10.3758/s13428-020-01453-w. URL: https://doi.org/10.3758/s13428-020-01453-w.

Berlyne, Daniel E (1973). “Aesthetics and psychobiology”. In: *Journal of Aesthetics and Art Criticism* 31.4.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). “Latent Dirichlet Allocation”. In: *J. Mach. Learn. Res.* 3.null, 993–1022. ISSN: 1532-4435.

Boden, Margaret A (1990). “The creative mind: myths and mechanisms Weidenfeld”. In: *Abacus Basic Books* 4, p. 344. URL: https://www.researchgate.net/publication/318494266_The_Creative_Mind_Myths_and_Mechanisms.

Brock, Andrew, Jeff Donahue, and Karen Simonyan (2018). “Large scale GAN training for high fidelity natural image synthesis”. In: *arXiv preprint arXiv:1809.11096*.

Clarke, Laurie (2022). “When AI can make art – what does it mean for creativity?” en-GB. In: *The Observer*. ISSN: 0029-7712. URL: https://www.theguardian.com/technology/2022/nov/12/when-ai-can-make-art-what-does-it-mean-for-creativity-dall-e-midjourney.

Cropley, A.J. (2011). “Definitions of Creativity”. In: *Encyclopedia of Creativity*, pp. 358–368. DOI: 10.1016/B978-0-12-375038-9.00066-2. URL: https://www.researchgate.net/publication/266031380_Definitions_of_Creativity.

Devlin, Jacob et al. (2018a). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

— (2018b). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. cite arxiv:1810.04805Comment: 13 pages. URL: http://arxiv.org/abs/1810.04805.

Dhariwal, Prafulla et al. (Apr. 2020). “Jukebox: A Generative Model for Music”. In: DOI: 10.48550/arxiv.2005.00341. URL: https://arxiv.org/abs/2005.00341v1.

Dumas, Denis, Peter Organisciak, and Michael Doherty (2021). "Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods." In: *Psychology of Aesthetics, Creativity, and the Arts* 15.4, p. 645.

Edwards, Benj (2023). *Artists file class-action lawsuit against AI Image Generator Companies*. URL: https://arstechnica.com/information-technology/2023/01/artists-file-class-action-lawsuit-against-ai-image-generator-companies/.

Eshraghian, Jason K (2020). "Human ownership of artificial creativity". In: *Nature Machine Intelligence* 2.3, pp. 157–160.

Feist, Gregory J (2021). "The creative personality: Current understandings and debates". In: *Creativity and Innovation*. Routledge, pp. 181–198.

Franceschelli, Giorgio and Mirco Musolesi (2022a). *Creativity and Machine Learning: A Survey; Creativity and Machine Learning: A Survey*. URL: www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

— (2022b). *DeepCreativity: Measuring Creativity with Deep Learning Techniques*.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Goodfellow, Ian et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems* 27.

Grootendorst, Maarten (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794*.

Guilford, Joy P (1967). "Creativity: Yesterday, today and tomorrow". In: *The Journal of Creative Behavior* 1.1, pp. 3–14.

Hennessey, Beth A et al. (2008). "A multi-cultural application of the consensual assessment technique". In: *The International Journal of Creativity & Problem Solving* 18.2, pp. 87–100.

Hertzmann, Aaron (2020). "Computers Do Not Make Art, People Do". In: *Commun. ACM* 63.5, 45–48. ISSN: 0001-0782. DOI: 10.1145/3347092. URL: https://doi.org/10.1145/3347092.

Ihalainen, Jani (Mar. 2018). "Computer creativity: artificial intelligence and copyright". In: *Journal of Intellectual Property Law Practice* 13.9, pp. 724–728. ISSN: 1747-1532. DOI: 10.1093/jiplp/jpy031. eprint: https://academic.oup.com/jiplp/article-pdf/13/9/724/25368526/jpy031.pdf. URL: https://doi.org/10.1093/jiplp/jpy031.

Kim, Kyung Hee (2006). *Can We Trust Creativity Tests? A Review of the Torrance Tests of Creative Thinking (TTCT)*.

Landauer, Thomas K, Peter W Foltz, and Darrell Laham (1998). "An introduction to latent semantic analysis". In: *Discourse processes* 25.2-3, pp. 259–284.

Lian, Long et al. (2023). "LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models". In: *arXiv preprint arXiv:2305.13655*.

Liu, Vivian and Lydia B Chilton (2022). "Design Guidelines for Prompt Engineering Text-to-Image Generative Models; Design Guidelines for Prompt Engineering Text-to-Image Generative Models". In: *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA* 1. DOI: 10.1145/3491102.3501825. URL: https://doi.org/10.1145/3491102.3501825.

Loper, Edward and Steven Bird (2002). *NLTK: The Natural Language Toolkit*. DOI: 10.48550/ARXIV.CS/0205028. URL: https://arxiv.org/abs/cs/0205028.

Lubart, Todd et al. (2022). "What do educators need to know about the Torrance Tests of Creative Thinking: A comprehensive review". In: DOI: 10.3389/fpsyg.2022.1000385.

Mazarura, Jocelyn and Alta De Waal (2016). "A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text". In: *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, pp. 1–6.

McCormack, Jon et al. (2023). "Is Writing Prompts Really Making Art?" In: *arXiv preprint arXiv:2301.13049*.

McInnes, Leland and John Healy (2017). "Accelerated hierarchical density based clustering". In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 33–42.

McKight, Patrick E and Julius Najab (2010). "Kruskal-wallis test". In: *The corsini encyclopedia of psychology*, pp. 1–1.

Mimno, David et al. (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272.

Mordvintsev, Alexander, Christopher Olah, and Mike Tyka (2015). "Inceptionism: Going deeper into neural networks". In.

Mostaque, Emad (2022). *Stable diffusion public release*. URL: https://stability.ai/blog/stable-diffusion-public-release.

Newton, Alexis and Kaustubh Dhole (2023). *Is AI Art Another Industrial Revolution in the Making?* URL: www.aaai.org.

Oppenlaender, Jonas (2022). *A Taxonomy of Prompt Modifiers for Text-To-Image Generation; A Taxonomy of Prompt Modifiers for Text-To-Image Generation*.

Pavlichenko, Nikita, Fedor Zhdanov, and Dmitry Ustalov (2022). *Best Prompts for Text-to-Image Models and How to Find Them*. DOI: 10.48550/ARXIV.2209.11711. URL: https://arxiv.org/abs/2209.11711.

Pressman, John David, Katherine Crowson, and Simulacra Captions Contributors (2022). *Simulacra Aesthetic Captions*. Tech. rep. Version 1.0. url https://github.com/JD-P/simulacra-aesthetic-captions. Stability AI.

Qiao, Han, Vivian Liu, and Lydia B Chilton (2022). "Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art; Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art". In: DOI: 10.1145/3527927.3532792. URL: https://doi.org/10.1145/3527927.3532792.

Radford, Alec et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR, pp. 8748–8763.

Ramesh, Aditya et al. (2021). "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR, pp. 8821–8831.

Ramesh, Aditya et al. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*. arXiv: 2204.06125 [cs.CV].

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.

Rombach, Robin, Andreas Blattmann, and Björn Ommer (2022). *Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models*. URL: https://github.com/CompVis/latent-diffusion..

Rombach, Robin et al. (2022). “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695.

Roose, Kevin (2022). *An a.i.-generated picture won an art prize. artists aren't happy.* URL: https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html.

Runco, Mark A and Garrett J Jaeger (2012). “The Standard Definition of Creativity”. In: *Creativity Research Journal* 24 (1), pp. 92–96. ISSN: 1532-6934. DOI: 10.1080/10400419.2012.650092. URL: https://www.tandfonline.com/action/journalInformation?journalCode=hcrj20.

rwalk (2017). *Rwalk/GSDMM: GSDMM: Short text clustering*. URL: https://github.com/rwalk/gsdmm.

Saharia, Chitwan et al. (2022). *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. arXiv: 2205.11487 [cs.CV].

Sohl-Dickstein, Jascha et al. (2015). “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International conference on machine learning*. PMLR, pp. 2256–2265.

Strange, Adario (2022). *Dall-e now allows anyone to cash in on AI art, but ownership gets complicated*. URL: https://qz.com/dall-e-now-allows-anyone-to-cash-in-on-ai-art-but-owne-1849597618.

Syahrin, Alfi, Heri Suwingyo, Endah Tri PRIYATNI, et al. (2019). “Creative thinking patterns in student‘s scientific works”. In: *Eurasian Journal of Educational Research* 19.81, pp. 21–36.

Taylor, Michael (2023). *Prompt engineering: From words to art and copy*. URL: https://www.saxifrage.xyz/post/prompt-engineering.

Tijare, Poonam and P Jhansi Rani (2020). “Exploring popular topic models”. In: *Journal of Physics: Conference Series*. Vol. 1706. 1. IOP Publishing, p. 012171.

Torrance, E Paul (1972). “Predictive validity of the torrance tests of creative thinking.” In: *The Journal of creative behavior*.

Tyrrrz (2023). *Tyrrrz/Discordchatexporter: Exports discord chat logs to a file*. URL: https://github.com/Tyrrrz/DiscordChatExporter.

Udupa, Abhinandan et al. (2022). “An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling”. In: *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*. IEEE, pp. 1–9.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Wang, Zijie J. et al. (Oct. 2022). “DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models”. In: URL: http://arxiv.org/abs/2210.14896.

Weiner, Ken (2018). *Can AI Create True Art?* en. URL: https://blogs.scientificamerican.com/observations/can-ai-create-true-art/.

Yin, Jianhua and Jianyong Wang (2014). “A dirichlet multinomial mixture model-based approach for short text clustering”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Zedelius, Claire M, Caitlin Mills, and Jonathan W Schooler (2018). “Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features”. In: DOI: 10.3758/s13428-018-1137-1. URL: https://doi.org/10.3758/s13428-018-1137-1.