Digitizing Assessment of Creative Aptitude: A Human-Centred Design Approach

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

By

Nandita Bhanja Chaudhuri
(Roll No. 176105102)

Under the supervision of

Dr. Debayan Dhar
Department of Design, IIT Guwahati

Department of Design
Indian Institute of Technology Guwahati
Guwahati – 781039, India
2022

Department of Design
Indian Institute of Technology Guwahati
Guwahati, Assam-781039, India

Declaration

I hereby declare that the work contained in this thesis entitled "Digitizing Assessment of Creative Aptitude: A Human-Centred Design Approach" is my work and done under the guidance of Dr. Debayan Dhar, Assistant Professor at the Department of Design, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, it contains no materials previously published or written by another person or substantial properties of the material which has been accepted for the award of any other degree or diploma at Indian Institute of Technology Guwahati or any other educational institution, except where due acknowledgment is made in the thesis. Any contribution made to this research by others, with whom I have worked at Indian Institute of Technology Guwahati or elsewhere is explicitly acknowledged in the thesis. I declare that the intellectual content of this thesis represents my work and words. I have adequately cited and referred to the original work where others' ideas, work, and words have been included. I also declare that I have adhered to all principals of academic honesty and integrity and not misrepresented or fabricated or falsified any idea/ data/ fact/ source in my submission.

Place: IIT Guwahati

Date: 08.08.2022

Nandita Bhanja Chaudhuri

Roll Number: 176105102

Department of Design
Indian Institute of Technology Guwahati
Guwahati, Assam-781039, India

Certificate

This is to certify that the work contained in this thesis titled "Digitizing Assessment of Creative Aptitude: A Human-Centred Design Approach" submitted by Ms. Nandita Bhanja Chaudhuri to the Indian Institute of Technology Guwahati for the award of the degree of Doctor of Philosophy has been carried out under my supervision. This work has not been submitted elsewhere for the award of any other degree or diploma.

Place: IIT Guwahati
Date: 08.08.2022

Dr. Debayan Dhar
Assistant Professor
Department of Design
Indian Institute of Technology Guwahati
Guwahati, Assam-781039, India

Acknowledgment

I would like to thank *Almighty* for his abundant grace and giving me strength and ability in all phases of Ph.D. I take this precious opportunity to convey my gratitude and deep respect for all those who have supported and motivated me towards the successful completion of the Ph.D. thesis. It has been an extensive experience that kept me constantly engaged with this research entitled "Digitizing Assessment of Creative Aptitude: A Human-Centred Design Approach". It bestowed me an immense knowledge in human-centred design and computational creativity.

I would like to thank my supervisor Dr. Debayan Dhar for his immense motivation, guidance, and inspiring thoughts to complete my research. I am really grateful to Prof. Pradeep G. Yammiyavar, for his support and belief in me. I would also like to convey my gratitude to doctoral committee members, Dr. Sougata Karmakar, Dr. D. Udaya Kumar, and Dr. Chandan Karfa. They have periodically reviewed my work, provided significant insights, and supported me to complete the research successfully. I am obliged to all committee members for reviewing an earlier draft of this thesis and providing me valuable suggestions.

I am deeply indebted to Ministry of Human Resource and Development (MHRD) for the financial support during my PhD thesis. I convey my sincere thanks to Research and Development section, IIT Guwahati for providing me financial support in filling a patent that was part of my research. I would like to thank all faculties and lab members of usability engineering and human computer interaction lab in the Department of Design for providing necessary resources required during the research and supporting me. I would also like to thank other academic institutes like IIT Bombay, MIT Institute of Design, and VIT, from where I received required data that were significant to my research. I am also thankful to members of other academic institutes from where I received data through online medium and received direct or indirect support for completion of this thesis. Special thanks to all colleagues at IIT Guwahati for their voluntary participation in experiments. I am grateful to the non-teaching staffs of the department for their continuous support whenever required.

I am forever thankful to my parents, sister, husband, and brother-in-law (Mr. Swapan B.C., Mrs. Gayatri B.C., Mumu, Anindya, Prosenjit) for their unconditional love and support, without whom I could not have overcome the highs and lows during these years.

Nandita Bhanja Chaudhuri

Abstract of the thesis

Assessing creative aptitude is an inherent criterion in entrance examination of Design education. The assessment process of creative aptitude in Design entrance examination starts from two major perspectives- 1) Identifying characteristics of framing questions by Design pedagogues that test students' creative aptitude, 2) Evaluating students' creative responses (Aburas & Nurunnabi, 2019). Formulating creative question plays a significant role in Design entrance examinations that has the potential to instigate creative responses among students (Constantinou, 2021; Rashid & Qaisar, 2016). Pedagogues remain ever-inquisitive to know whether the questions framed by them are really creative enough to instigate creative responses. During this process of framing questions that triggers creativity in students, pedagogues solely rely on their individual experiences. To ensure least interference of individual biases among Design pedagogues that affects quality of questions while formulating them, this research study has been undertaken to inquire and investigate ways through which an optimized system of support for pedagogues can be designed. To address this situation, a computational design model is proposed that assesses the questions framed by the pedagogues and analyses it to check whether it would instigate creative responses from students.

While types of question plays a significant role in instigating creative responses, evaluating these responses is a major challenge. During assessment of these responses novelty is the most important factor Design pedagogues often look out for. It is a significant criteria in synthesizing creative responses (Jagtap, 2019; Liberati et al., 2018; Saaksjarvi & Goncalves, 2018; Sarkar & Chakrabarti, 2011). Assessing novelty in creative responses requires subjective evaluation, which is generally dependent on experts' knowledge, choice, and persuasion (Ma et al., 2017). Presently in Design entrance examinations of India, evaluation of novelty is conducted by pedagogues possessing expertise in assessing creative skills. During this evaluation process, they are confronted with multiple challenges such as individual stress, evaluation time, etc. which often lead to inconsistencies in the evaluation process and often reduce self-confidence among pedagogues (Gonzalez et al., 2017; Richards et al., 2017). Predominantly, creative responses are subjective and generally evaluated based on pedagogues' frame of reference, leading to inconsistency of evaluation across different pedagogues. In situations like this, there might be a reduction in trust of students in the evaluation procedure. To mitigate these challenges, multiple computational design models have been proposed in this thesis with an

objective to automate the process of evaluating novelty from creative responses exhibited through various patterns of responses such as descriptive, labelled image, and annotated image-based responses.

The research investigation reported here focuses on addressing challenges faced by the Design educators community and can be directly related as a contribution from the perspective of Design Praxiology as proposed by (Cross, 1999; Gasparski, 1979). The proposals made in this research work possesses an approach that intends to prepare Design education community specifically Design pedagogues to embrace changes in existing ways of framing creative questions and assessing students' responses. This study has addressed a Design problem, which is specifically based on addressing human-errors in the process of assessing creative responses. Furthermore, studies involved in this research would save time for evaluation, disseminate faster results, and reduce other logistics such as paper, storage spaces, etc., for education practitioners.

Highlights of the thesis

- *To empathize the experience of framing creative questions that tests creative aptitude and evaluating students' creative responses.*
- *To systematically identify features of creative questions that instigate creative responses among students by human-centred design approach.*
- *To propose and implement a computational design model to identify creative questions from a bunch of other non-creative questions.*
- *To systematically identify features of evaluating creative responses by human-centred design approach.*
- *To propose and implement computational design models for assessing creative responses.*
- *To validate all the proposed models by measuring the agreement among the outcome of the models and human-based assessment.*

Table of contents

List of figures

List of tables

Abbreviations used in the report

AICTE	All India Council for Technical Education
GATE	Graduate Aptitude Test in Engineering
CEED	Common Entrance Examination for Design
IIT	Indian Institute of Technology
IISc	Indian Institute of Science
NID	National Institute of Design
NIFT	National Institute of Fashion Technology
FTII JET	Film and Television Institute of India Joint Entrance Test
SRFTI	Satyajit Ray Film and Television Institute
NAT	Numerical Answer Type
MCQ	Multiple Choice Question
MSQ	Multiple Select Question
LMS	Learning Management Systems
H-novelty	Historical novelty
P-novelty	Psychological novelty
Bi-LSTM	Bi-directional Long Short Term Memory
NIST	National Institute of Standards and Technology
TREC	Text REtrieval Conference
Kernel PCA	Kernel Principal Component Analysis
SVM	Support Vector Machine
LDA	Latent Dirichlet Allocation
TF-IDF	Term Frequency-Inverse Document Frequency
DL	Deep Learning
MNIST	Modified National Institute of Standards and Technology
SVHN	Street View House Numbers
CIFAR	Canadian Institute For Advanced Research
MRI	Magnetic Resonance Imaging
COIL	Columbia Object Image Library
COCO	Common Objects in Context
Mask-RCNN	Mask-Region Convolution Neural Network
AQG	Automated Question Generation
RCNN-RoBERTa	Recurrent Convolution Neural Network-Robustly optimized BERT approach
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
BERT	Bidirectional Encoder Representations using Transformers
NLP	Natural Language Processing
RQ	Research Question
MSE	Mean Square Error
MAE	Mean Absolute Error
API	Application Programming Interface

CBOW	Continuous Bag of Words
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BIRCH	Then Balanced Iterative Reducing and Clustering using Hierarchies
OCR	Optical Character Recognition
VGG	Visual Geometry Group
RGB	Red Green Blue
SIFT	Scale Invariant Feature Transform
YOLO	You Only Look Once
MSCOCO	Microsoft Common Objects in Context
CNN	Convolutional Neural Network
GloVe	Global Vectors for Word Representation
QA	Question-Answering
CAT	Consensual Assessment Technique
RAT	Remote Associates Test
TTCT	Torrance Tests of Creative Thinking
DTR	Decision Tree Regressor
MOR	MultiOutput Regressor

Chapter 1: Introduction

Abstract

Assessing creative aptitude is a significant factor in Design education pedagogy. Question plays a major role in entrance examinations of Design education in triggering creative aptitude of students (Rashid & Qaisar, 2016). While formulating these types of questions, pedagogues remain ever-inquisitive to know whether the questions framed by them can really instigate creative responses. Currently pedagogues rely on their past experience to frame questions. This may lead to inconsistencies across questions among pedagogues intending to assess creative aptitude. The study conducted during this thesis work intends to first identify the challenges faced by pedagogues during question framing. With this intention an exhaustive state-of-the-art literature review was conducted to identify tools, techniques, and scientific contributions associated with question papers in pedagogy.

While types of question plays a significant role in instigating creative responses, evaluating these responses is a major challenge as well. Literature highlights that during evaluation of these creative responses, novelty is the most important factor Design pedagogues often look out for (Jagtap, 2019; Liberati et al., 2018; Saaksjarvi & Goncalves, 2018; Sarkar & Chakrabarti, 2011). Evaluation of novelty is subjective and depends on pedagogues' choice and persuasion. Presently, evaluation of novelty in Design education is manual, which is dependent on pen-and-paper-based techniques. Many a time, it is based on pedagogue's self-referential metrics, which might vary among multiple experts, thereby making the evaluation process inconsistent. During this evaluation process, examiners are confronted with major challenges such as errors in evaluation due to stipulated time of assessment, prolonged working hours, and repeated tasks on a large scale. Therefore, initially the focus of this research was to identify state-of-the-art techniques that were reported in literature addressing the concern of consistent evaluation of novelty of creative responses. Based on literature review findings further studies were conducted to address the research gaps to propose solutions to the identified Design problems.

Highlights

- *An overview of assessment of creative aptitude in formulating questions and evaluating their responses in entrance examinations of Design education.*

- *Detailed literature review related to creative questions that instigate creative responses, evaluating novelty that pedagogues look for in creative responses, and stress factors of pedagogues.*
- *Identifying research gap, research questions, aim, and objectives.*
- *Highlighting the structure of the thesis.*

1.1 Overview of Design entrance examinations in India

India reflects a very competitive scenario in higher education system. Higher education in India intends to give access, value, and quality with accountability at a moderate expense to all hopeful citizens with extreme straightforwardness to guarantee manageable monetary improvement of the country. It is accomplished through creation, utilization, and dispersal of information. The All India Council for Technical Education (AICTE) manages the activities of technical education and professional programmes in India (*All India Council for Technical Education- Approval Process Handbook 2021-22*, 2021). However, the conduct of entrance examination such as Graduate Aptitude Test in Engineering (GATE), Common Entrance Examination for Design (CEED), Undergraduate Common Entrance Examination for Design (UCEED), etc. is governed by old premiere institutes like the Indian Institute of Technology (IIT), and Indian Institute of Science (IISc) Bengaluru (Sohoni, 2016).

Students appear in examinations like UCEED, CEED, National Institute of Design (NID), National Institute of Fashion Technology (NIFT), Film and Television Institute of India Joint Entrance Test (FTII JET), Satyajit Ray Film and Television Institute (SRFTI), and other entrance tests for getting admission to bachelor's, master's, and Ph.D. programmes of creative schools. The type of entrance examinations conducted by these schools focus at assessing numerical, analytical, logical abilities, environment, social awareness, and multiple creative skills such as visualization, spatial ability, observation, design sensitivity, language, creativity, design thinking, problem identification, problem solving, and sketching. The question paper pattern in these examinations are broadly classified as objective and subjective questions. The objective part consists of questions related to numerical, logical, etc., whereas the subjective part possesses questions associated with sketching, form sensitivity, visual sensitivity, problem identification, design thinking, etc. The solutions to objective questions are straightforward and based on a strict set of options, whereas subjective responses are evaluated based on the relevance of ideas and comparison of uniqueness among responses for a given problem.

Subjective evaluation is dependent on individual persuasion and is relatively complex than objective evaluation (Chaudhuri et al., 2021c).

Question patterns in these nationalized Design examinations are structured into two categories viz., objective and subjective patterns of questions. Objective question patterns consist of Numerical Answer Type (NAT), Multiple Choice Question (MCQ), and Multiple Select Questions (MSQ). There are model solutions for these types of questions. On contrary, there are no model solutions for subjective questions (Bombay, 2021a, 2021b). The responses are mostly sketches and creative write-ups that are open-ended. The solution of objective type of questions is evaluated automatically by computer-based systems. On the other hand, subjective questions accept open-ended responses that are dependent on students' unique ideas. Open-ended solutions are not pre-defined and are presently evaluated based on novelty of a response by domain-specific experts.

1.2 Questions instigating creative responses

Unlike engineering and other disciplines that intends at assessing learning and recalling abilities of students through questions, entrance examination conducted by Design institutes focusses on framing questions that instigates creative responses. To frame such questions, deep critical thinking is required by pedagogues. While outlining these questions, teachers frequently self-assess, re-thinks, and re-defines their thought processes until they frame questions that has features to instigate responses that are truly creative in nature. During this cycle of formulating questions, they are often confronted with the challenge to identify whether the questions framed by them have characteristics that can trigger creative responses. Group discussions, brainstorming, and other similar methods in these circumstances are generally adopted and they provide significant insight, but they have their own impediments. Individual attributes and past experience referencing impacts often biases question formulation. Also, peer-reviewing a question often leads to divergence of ideas rather than convergence.

In the present scenario, domain-specific experts formulate creative questions that instigates creative responses from students of Design entrance examinations. Pedagogues utilize their knowledge and expertise to frame questions in order to capture students' creative aptitude. One way of verifying whether such questions really captures creativity of students is by assessing the students' responses to the questions. But, in situations like entrance examinations, which is a critical one, validity of a question regarding its ability to test creativity is crucial otherwise the objective of identifying appropriate candidates through these entrance examinations would fall flat and making these examinations redundant. It is therefore important to verify questions before such crucial examinations. During the process of framing these questions, apart from self-enquiry by pedagogues, another way of verifying a creativity-seeking question is through peer-reviewing, but it has its own demerits as well, as it involves biases of experts. Yet another way of verifying a question that can capture creativity from students is to find whether technology can support a situation like this. It is essential to find a technique devoid of human biases that can identify whether a question may instigate creative responses among students, i.e., questions precisely having features of creative questions. To propose measures for identifying creative questions, an exhaustive state of the art review was conducted to identify the features of creative questions or creative questioning techniques.

1.3 Creative aptitude evaluation in Design education

Responses of students to creative questions exhibits creativity, where novelty is an important factor that pedagogues look out for assessing their creative aptitude. Novelty is a significant factor of evaluation in preferably all sectors of society such as education, industry, agriculture, etc., that requires evolving over a period of time. It requires interchanging of ideas among function, structure, and behaviour of embodiments (Srinivasan & Chakrabarti, 2008). In dictionaries like Merriam-Webster, novelty is defined by multiple ways such as “something new or unusual”, or “the quality of being novel”. The same source provides various synonyms of novelty, such as “freshness” or “newness”, or “originality” (Merriam-Webster, 2021).

Literature highlights various definitions of novelty that plays a significant role in creativity assessment during entrance examinations conducted for Design educational institutes. For example, novelty in Design is defined as “a concept strictly related to unusualness or unexpectedness, using idea infrequency as a key measure”, or “originality of ideas” or “original and unexpected ideas” or “non-obviousness on the frequency of responses among design teams” or “something never seen before” (Fiorineschi & Rotini, 2021). Other studies defined novelty as “the quality of something being new and unusual for the student and at that level of education” (Demirkan & Afacan, 2012). In Learning Management Systems (LMS), novelty is defined in terms of its “infrequency and rarity”, and its “uniqueness and newness” (McDaniel et al., 2017). It is also defined as “something not previously experienced or one that deviates from everyday routine” (González-Cutre et al., 2016). Majority of these definitions provide a similar concept of novelty of an idea or solution or response being unique and newly introduced in a specified context.

Novelty is acknowledged as an important parameter for judging creative responses. However, literature highlights various degrees of novelty that are often evaluated in such responses. These are the degree of uniqueness in terms of historical novelty (H-novelty) and psychological novelty (P-novelty). H-novelty is considered the highest degree of novelty, where an invention is never experienced earlier. On contrary, P-novelty refers to the degree of improvement over its previous versions (Jagtap, 2019; Park et al., 2020; Sarkar & Chakrabarti, 2011). There is another classification of novelty depending on its degree of uniqueness with respect to time, namely, short-term novelty, long-term novelty, and complete novelty. Short-term novelty refers to a response that has not been encountered in the last few minutes. In contrast, solutions or responses that have not been experienced for some intermediate amount of time or a few days are termed as a long-term novelty. Further, a response that has never been experienced before is considered a complete novelty (Barto et al., 2013). Degree of novelty has also been classified as very high novelty, high novelty, medium novelty, and low novelty. A response possessing very high novelty has never been experienced before its creation, whereas conversion of state or input illustrates high novelty in a response. Medium novelty corresponds to transition of physical phenomena or effects, and difference among responses only in terms of physical parts or organs indicates low novelty (Ruiz-Pastor et al., 2021).

Creative aptitude in Design is evaluated based on unexpectedness of responses that is often considered as novel. Infrequent responses compared among cohort of students in an examination is considered to possess relatively higher novelty (Omari et al., 2016). It is analogous to personal or psychological novelty that has a relatively lesser degree of novelty, and responses of which are unique to a community, and might be an improvement of an earlier version of responses. The novelty score measured for a response depends on the frequency or the number of times a response is generated for a given problem in an examination. The more number of times a response or an idea is repeated, a relatively lower score is awarded to a response, and vice versa (Sluis-Thiescheffer et al., 2016).

In Design entrance examinations, creative aptitude is generally demonstrated through creative responses that are subjective in nature. Subjective responses illustrate designer’s experience, choice, opinion, and persuasion. In contrast, objective solutions are specific and pre-determined and can be matched for evaluation with any databases possessing similar responses (Florez & Castro-Lacouture, 2013). There are various types of responses that are identified in examinations of Design education, such as quantitative, analytical, and qualitative responses (Asunda & Hill, 2007). Generally, creative aptitude is associated with qualitative responses exhibited through sketch or image-based responses and descriptive responses (Oman et al., 2013). Various types of creative responses are illustrated in Figure 1.1.

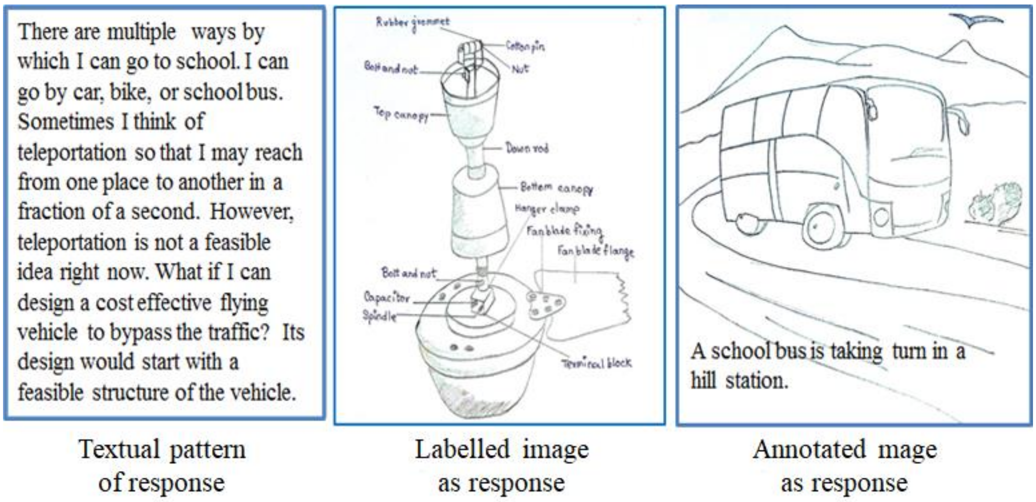


Figure 1.1: Typical example creative responses exhibiting descriptive, labelled images, and annotated image-based patterns

Creative responses in Design entrance examination are primarily received in qualitative form and can be categorized into- 1) only textual or descriptive response 2) image with labels and 3) image with annotations (Goldschmidt, 2009, 2014; Goldschmidt & Sever, 2011; Self, 2019; Welch et al., 2000). Creative responses that are text only generally consists of a write-up of a context, representation of a storyline or a narrative, multiple approaches to a design solution, materialistic specifications, behavioural specifications, functionalities, and many more. Segments of images or sketches marked with text are referred to as images with labels (Forbus et al., 2011; Kembhavi et al., 2016; Landa, 2018; Taborda et al., 2013). Annotated images are pattern of creative response that exhibits drawings or sketches agglomerated with textual descriptions (Song et al., 2017; Wang et al., 2018).

In Design, pedagogues prefer divergent and unique approaches to problem-solving. Each response for a given problem is compared with a cohort of responses, and novelty is measured. The assessment of novelty thus depend on variance of responses and relevance of a response to a problem (Omari et al., 2016; Wei et al., 2016). Variance of responses means creative ideas that are divergent, whereas relevance refers to the quality of being appropriate in a given context.

Presently, entrance examination system in Design education reflects two broad patterns of evaluation viz., objective and subjective. The objective part of evaluation is automated, and depends on computer-based comparison of students' solutions with electronically stored modelled solutions. While, subjective evaluation involves assessment of novelty that are still primarily dependent on manual pen-and-paper-based evaluation. This type of mass examination involves large number of experts conducting assessments on a large scale in a stipulated time. In situations like this, experts might have frustrations and fatigue due to repeated monotonous tasks conducted on a large scale, different evaluators may have different frames of reference for assessing the responses resulting in inconsistency in the evaluation process.

1.4 Role of Design Pedagogues in assessing creative aptitude

Pedagogues in Design education formulate questions that instigate creative responses from students. Pedagogues drill their thought processes to formulate questions that are creative in nature. They think, reflect, and review to ensure that the questions framed, precisely possess features that instigate creative responses from students. During this process of question formulation, pedagogues compare and contrast their ideas. They remain ever-inquisitive to know whether questions framed by them are truly the ones that can instigate creative responses. And in order to ensure this, many a time peer review approach is adopted by the pedagogues to assess the quality of the questions framed. But peer review has its own disadvantages as well. Individual characteristics and past experiences of individual experts/pedagogues can influence frame of reference, and as such each pedagogue may have their own interpretation of the quality and type of questions that can trigger creative responses. This difference in individual characteristics across each pedagogues might lead to more differences than convergences.

Unlike other fields of education, pedagogues evaluate creative responses in Design education subjectively. Creative responses do not have a strict set of answers and are open-ended in nature with their own set of constraints (Landa, 2018). Framing questions that trigger creative responses, pedagogues often refer to self-referential metrics based on their choice, knowledge, and expertise. Mass examinations in Design education require large scale subjective evaluation by pedagogues. During this process, pedagogues get highly stressed and fatigued due to repeated tasks. Difference in opinion across different evaluators arising out of individual frames of reference may lead to major inconsistencies in the evaluation process.

1.5 State of the art review

The literature review presented in this section is conducted from the perspective of identifying sources of pain points in assessment of creative aptitude. Assessment process of creative aptitude conducted during entrance examinations for Design educational institutes primarily has two main sections- 1) Identifying characteristics of questions that can instigate creative responses from students, and 2) evaluating students' creative responses. In examination of disciplines like science and engineering, exam questions focusses on testing student's recall and learning abilities (Park et al., 2016). While questions in Design triggers creative responses (Zolfaghari et al., 2011).

In order to clearly define the creative assessment approach, it is important to conduct an extensive and detailed review on the nature of creativity and how it is being studied historically. Creative aptitude illustrating novelty is a genre that stimulated scientists for ages (Amabile, 1988). The idea of creativity and novelty has made considerable progress from the time of early Greek logicians, who characterized it as a supernatural force offered by the Divine. During the classical period of ancient Greece, creative skill was considered as individualistic and had "an association with madness and frenzied inspiration". Several centuries later, scientists in this domain all the while attempting to comprehend the quintessence of innovativeness, tried to investigate what made a small population of people more inventive than others (Mehta & Dahl, 2019). In the field of psychology, the historical background of novelty can be categorized into two major eras. The first era was before 1950s, and the second era started after 1950 (Chaudhuri et al., 2020). Though literature reports a dearth of research associated with novelty before 1950, remarkable inventions in the copper age, early and middle bronze age, and early modern age marked the notion of creative aptitude.

Novelty is a significant factor of creative aptitude (Ranjan et al., 2018) that existed from the age of Platos and much before that, but creative skill in a formal and broader sense was probably presented for the first time around sixty-nine years back by J.P. Guilford, who was the president of the American Psychological Association. In the paper titled "Creativity" published in the year 1950, his explanations and thoughts of creative aptitude were more inclined towards psychological features (Runco & Plucker, 2001). He described creative skill in person as "The creative person has novel ideas" (Still & d'Inverno, 2016). Traditionally, creative aptitude was considered individualistic, but later, it was found to be either individualistic or collectivistic. The cycle of development of creative aptitude from individuals to groups was initiated with a self-propagating novelty in work and later advanced by an aggregate formation of numerous individuals instead of a solitary individual. The scientific, technological, and artistic evolution significantly impacted novelty in ideas. There are four stages of the perspective study of creative aptitude. The formal and initial stage was between 1950s and 1960s, when this genre was investigated to identify skilful individuals. The second stage was between 1970s and 1980s; during this period, it was studied from the perspective of a person's perception, cognition, and thought processes. The third stage was between 1980s and 1990s, studies reported during this decade were from the perspective of societal and cultural impact of novelty. The present stage, which started in 1990s and is continuing till now, highlights research works where the study of novelty is associated with multiple disciplines such as education, design, advanced technologies like artificial intelligence, robotics, etc. (Watts & Blessinger, 2016).

1.5.1 Features of questions triggering creative responses

Question is a medium by which pedagogues confirm students' learning and creativity (Baloche & Platt, 1993). Multiple categories of questions are framed by teachers with an objective to understand and capture students' learning (Aziza, 2018). Creative questioning is an art and science that invokes creativity in students (Zolfaghari et al., 2011). All questions do not trigger creative responses. Questions that have the potential to instigate creative responses from students can be identified based on the responses illustrating divergent unique ideas.

One of the important aspects that pedagogues generally considers while framing creative question is that it must trigger higher order critical and creative thinking among students while they try to respond to the questions. Basic questioning skills involve unambiguity, relevance, providing guidelines and reference points. Advanced questioning skills include additional characteristics such as switching the demands of cognitive level of students, maintaining question chronology, framing questions capable of monitoring and stimulating ideas of students, interactivity between student and examiner through questions (Halim et al., 2018).

Extensive studies have been reported on evaluating creativity from the perspective of process, individual, product, press, and creative systems (Said-Metwaly et al., 2017). Literature highlighted numerous features of questions, but very few articles have focussed on identifying characteristics of questions that has the potential to instigate creative responses from students. The review presented in this section mainly highlights the characteristics of questions framed for examinations associated with creative domains. Few of the articles were associated with descriptive questions and artificial question generation techniques. Though the responses of the descriptive questions were open-ended, but it primarily focussed on assessing a students' recall and past learning. It did not necessarily focussed on the students' ability to generate creative responses to the problem provided in the question (Kurdi et al., 2020).

Sometimes questioning strategy is itself questioned to check for its impact on students' learning. Various studies highlighted the types of questions that influenced students' learning (Ellis, 1993). Since the scope of this review focussed on identifying creative questioning techniques and its features, further literature explorations towards student learning were not conducted. A question can assess creative aptitude of students from responses. The features of responses highlighting characteristics of creative thinking are: flexibility of ideas, interpreting the problem perspective, variation in idea generation, idea screening, selection, and elaboration of an idea (Cropley, 2000). Literature highlighted multiple features of creative questions such as seeking views of respondents, seeking a well-explained solution, appreciating multiple responses, seeking imaginative answers, framing a branched question, and open-ended questions (Zolfaghari et al., 2011). However, there exists a substantial difference in the patterns and features of questions across different domains of study.

Socratic questioning strategy focussed mainly on optimization of critical thinking by multiple features like inspecting intensive concepts, fact-seeking questions, and hypothesis scrutinization. Socratic questions involved branched questions, seeking consequences based on an action, verifying questions, factual questions, and opinion-seeking questions. Experiments with Socratic questions were mostly associated to improve critical thinking of primary level children (Chew et al., 2019). The art of Socratic questioning is significant as it verifies a question from multiple perspectives such as, whether goal of a question is clear or not?, is it seeking appropriate information?, is it seeking important concept?, is it looking for consequence of an action?, is it verifying a context?, is it a fact-seeking question or subjective question?, is it seeking alternative of solutions?, is it clarifying all facts?, is it capable of targeting a context from multiple perspectives? (Paul & Elder, 2019). Socratic questioning is widely applicable in education and counselling. The factors seem relevant in questioning and also have an impact on critical thinking (Sahamid, 2016). However, it is essential to verify these factors in context of Design entrance examinations.

Majority of studies in literature elaborated multiple features of questions that lead to critical thinking. The questions were classified into three categories- why, what, and how. The keyword ‘why’ in questions indicates the inquisitive nature of the asker, ’what’ prompts in recalling certain knowledge, and the examiner focuses on the category ‘how’ to find out the process of certain things. Swenson (2016) laid more emphasis on the after-effect of using the classified question (Swenson, 2016). However, these categories of questions are often associated with objective questions and used in examinations where knowledge evaluation and recall are the primary concern (Park et al., 2016). Generally, objective questions and recalling abilities of students do not influence creativity.

Davis et al. (2002) reported multiple categories of questions for classroom assessment and projects such as ‘short answer’, ‘group task’, and ‘reflective essay’. Indicators of these type of questions were, seeking definitions of design processes, teamwork, communicative environment, evaluating procedures, relation among concepts, and write-up ability (Davis et al., 2002). However, these features are based on classroom environment, and the present context of the study (mass entrance examination for Design) is different. Unlike any mass examination, design and development of a project in the classroom or organization follow a different approach in conceptualization, ideation, development, deliverables, timeline, etc. Seeking definitions in questions of any education settings is a method to recall concepts and is uncorrelated with creativity. Questions seeking teamwork and communicative environment are unrelated to entrance examinations in Design as assessments are made individually.

Majority of the studies claimed that questions that seek opinion of students has the potential to capture students' creativity. They highlighted multiple features of opinion seeking questions. Opinion seeking questions have been considered a factor of creative questions as the answers depend on one’s choice and persuasion. Stoycheva (2010) highlighted significant features associated with creativity. These are open-ended, subjectivity, capable of accepting multiple answers, and ambiguity (Wan & McAuley, 2016). Literature also highlighted open-ended, subjectivity, and ambiguity as factors that are positively correlated with creativity (Stoycheva, 2010). But, ambiguity might be inappropriate in a mass examination as it restricts comprehension of a question. Mass examination requires extensive consistent and unambiguous content for inquiry and questioning techniques capable of providing the same interpretation across all respondents.

Demir & Sahin (2014) investigated assessment of open-ended questions by four factors, namely fluency, flexibility, originality, and scientific knowledge. Fluency indicates putting forward ideas that are part of any creative aptitude testing process, whereas flexibility refers to the number of ideas generated based on different domains. Seeking fluency in mass examination is possible where ideas are required to be presented for a given question. Originality was referred to seeking uncommon answers or it can be termed as novelty as well (Demir & Sahin, 2014). Majority of the articles that reported on fluency and originality had hardly presented any empirical data, hence those study results should be interpreted cautiously.

Daly et al. (2014) reported assessment criteria for creativity mainly associated with the context of a classroom setting of an engineering discipline. The features of creativity were outcomes of a qualitative research technique, specifically by triangulation of interviews, surveys, and analysis of course materials. Assessment criteria for all courses were studied, and analyses were conducted to identify features that were relevant and commonly used in classrooms. Idea generation process in classrooms included fluency as a factor, but there was no evidence of flexibility being applied to any course; originality was a significant feature for only one course whereas, elaboration and metaphorical thinking were not applied for any course. The features applied for courses were analysis, redefine, and evaluation for critical analysis of ideas. Synthesis was considered in a single academic course, and figuring relationships and solving ambiguity was not used for any course. For exploring ideas, only aesthetic sensitivity features were found in a single course, whereas there was no evidence of problem sensitivity, imagination, tolerance for ambiguity, intuition, and integration of dichotomies. Metacognition was applied for some of the courses (Daly et al., 2014). The features were well-tested in the classroom environment and required experimentation for mass examination settings. Competitive or nationalized exams require maintaining consistent interpretation of problems, time constraints to solve question papers, and questions depending on the level of test and educational background of a student (Wang, 2013).

Annamoradnejad et al. (2020) reported subjective attributes of a question associated with an online questioning platform. The features considered were- perceiving rationale of a problem, its communicative nature, its ability to invite short-length solution, its ability to seek facts, its ability to seek novelty in answer, its ability to generate interest, its ability to have multiple explanations, its ability to seek opinion, type of question, ability to seek comparison of generated solutions, seeking implications of an act, seeking definition, relation of question to an entity, seeking instructions, seeking processes, seeking explanation, spell-checking, and checking story-line (Annamoradnejad et al., 2020). These are the predictors of subjectivity in questions, but requires to be associated with particular contexts in educational settings such as classroom, project, etc.

Madabushi et al. (2018) reported question classification that involved capturing a syntactic map of questions revealing syntactic information. The study highlighted identifying question's headword in noun phrase, at the same time dealt with entity and phrase detection. Finally, a rule-base was used to map words at various locations in the syntactic map and to question categories using a hierarchical structure. But many a time, question misclassification occurs during usage of certain words that convey multiple meanings. For example, "what rank did you achieve in the examination?" and "what rank did she achieve in the military?", which showed usage of the word "rank" conveying different meanings according to its context. Therefore, rules were altered for such context in questions (Madabushi et al., 2018). A rule-base was utilized for categorizing questions. It is hard to identify creativity based on rule-based systems due to their complex nature and domain-specific characteristics.

Literature highlights the migration of rule-based techniques to advanced machine learning algorithms. Various algorithms included for this task were Long Short Term Memory (LSTM) network, support vector machine, and AdaBoost. Performance metric showed that LSTM outperformed as compared to other algorithms (Borg et al., 2021). Several studies focussed on identifying sarcasm and irony based on DL techniques. Recurrent Convolution Neural Network-Robustly optimized BERT approach (RCNN-RoBERTa) has been extensively used in literature studies as pre-trained models for Natural Language Processing (NLP) downstream tasks. An end-to-end model was created, where weight of pre-trained RoBERTa was combined with RCNN to acquire semantic and contextual details. Performance metric showed that this model outperformed when compared with other state-of-the-art approaches (Potamias et al., 2020).

Literature study highlighted seven state-of-the-art neural network architectures that provided a glimpse of performance of the models. Experiments were conducted with pooled Gated Recurrent Unit (GRU), LSTM, GRU with attention, 2D convolution with pooling, GRU with capsule, LSTM with capsule and attention, and Bidirectional Encoder Representations using Transformers (BERT) to classify multilingual hate speech and offensive language. Two evaluation metrics viz; macro-averaged F1 score and weighted F1 score were used for comparing all the models. BERT architecture provided the best F1 score for all the languages (Ranasinghe et al., 2019). The results were highly informative, but performance of the models is contingent on factors like size and type of dataset, context of usage, number of features, computational time and complexity, etc.

Measuring indicators of creative questions that instigate creative responses using a standardized procedure is essential to register the question as creative. However, majority of articles in literature attempted to measure creative aptitude by psychometric scale using subjective ratings and not by objectively scoring students' responses. Deep learning techniques are growing rapidly to solve important problems related to identifying characteristics of questions. Extensive literature studies reported lists of DL algorithms, their complexities, data repositories, performance metrics, etc., acted as a significant source of information for experimental design (Shinde & Shah, 2018; Zhu et al., 2017).

The insights of the literature review section are as follows:

i. The outcome of this review highlights that there is less focus on identifying creative questions that have the potential to instigate creative responses from students associated with Design entrance examinations.

ii. Majority of the articles reported in literature have highlighted multiple features associated with questions that seek critical thinking but did not highlight any feature of questions that instigate creative responses.

iii. Most of the studies reported in literature are associated with questioning skills and techniques concerning classroom and project environment and not with mass Design examinations.

iv. While there are multiple pedagogues involved in framing questions based on their experience and frames of reference, each pedagogues can visualize or define question that triggers creative responses in their own way. This may lead to lack of consistency across the systemic approach of an optimized assessment process for Design creativity.

v. Majority of the studies in literature highlighted the fact of a two-fold approach of problem-solving that initially captures features of domain of interest and then predicted desired outcome using DL techniques based on the requirement of context. There was hardly any study associated with digitized identification of creative questions that instigates creative responses from students.

1.5.2 Novelty in descriptive or textual content

While questions play a significant role in instigating creative responses, assessing these responses are also a major challenge. Novelty is an important factor of assessment that Design pedagogues often look out for (Sarkar & Chakrabarti, 2011). Literature highlights assessment of novelty in numerous contexts such as web content, scientific reports, etc. Nowadays, people are dependent on information on web, and innumerable volumes of documents are generated each day. The increasing volume of web documents leads to duplication and paraphrasing of documents, which raises the issue of genuineness of documents. There are investigations conducted to evaluate novelty of web documents that can be considered original documents.

Ghosal et al. (2018) proposed a novelty detection mechanism in which their model was trained with a synthetic dataset accumulated from various corpora. The textual dataset was converted into numerical form by Bi-directional Long Short Term Memory (Bi-LSTM). Vectors of the source document and target documents were compared to evaluate similarity by a cosine similarity function. A classification algorithm then categorized the input into a novel and not a novel document (Ghosal et al., 2018). However, novelty is an extremely critical concept, and it is hard to define it only in a binary expression. It is essential to consider a fuzzy concept that might express degrees of novelty between two extremities, i.e., novel and not novel. Moreover, evaluation of novelty need not based be on an absolute scoring; it is generally measured relative to other documents.

Several studies proposed models for evaluating novelty of documents. Classification algorithm was utilized to predict novelty from a given dataset. Graph-based representation of textual data led to the generation of feature sets (Gamon, 2020). The features were model-specific. The technique considered seem to be computationally expensive as adding data to the existing database resulted in updating connections of the graph, and as such remembrance of long relationships in the graph deemed difficult. Detecting novelty of scientific documents is highly essential as outcome of each research initiates a path for other related research. Dynich & Wang (2017) reported an algorithm to assess novelty of scientific research. They considered the following steps to evaluate novelty. Initially, a corpus was created comprising scientific documents, and those scientific data were pre-processed. Further, weights were assigned to words, and sentence-wise matching was performed in order to evaluate novelty (Dynich & Wang, 2017). However, assigning weights to words based on frequency failed to convey the meaning of documents in all possible cases. Usually, a document is considered novel in terms of originality and uniqueness. Unless themes or semantics of a document are considered, judging novelty of scientific documents seem redundant. It is also essential for scientific novelty detection algorithms to track the evolution of domain-specific work and inventions.

National Institute of Standards and Technology (NIST) made available Text REtrieval Conference (TREC) dataset for novelty detection. Multiple parameters were considered to detect novelty. TREC dataset was used to track records for evaluating novelty of documents, and categorizing sentences that are considered unique. Human judgments were used to find novelty in texts. Few of the decisions were based on machine learning techniques. The performance metric of detecting novelty was descent (Soborof & Harman, 2005). However, novelty was hardly looked into as factor that is expressed as a fuzzy concept with varying degrees such as low, moderate, and high; or can be expressed in the form of scores. Concept of novelty is subjective in nature and often depends on uniqueness related to domain of interest, target audience, intended usage of a document, etc. Therefore, selecting parameters and collecting data based on those parameters is essential to predict novelty, which might support measuring novelty in a given context.

Prediction of novelty from real-world data is highly challenging as it might comprise a substantial volume of noise and extraneous variables. Kernel Principal Component Analysis (Kernel PCA) algorithm was proposed in various literature studies to investigate the significant components from hand-written digits and cancer datasets, which are considered real-world records. The training dataset was first converted into its corresponding infinite-dimensional feature space. In that space, the kernel PCA captures the principal components of the data distribution. The subspace which was substantial, was used to measure the associated squared distance as a measure of novelty of the datasets. The outcome illustrated evidence of low classification error (Hoffmann, 2007). However, the detection of novelty is based on several parameters according to any given context. In evaluation procedure, the parameters of novelty are dependent on experts' point of reference for assessment. Utilizing such parameters might make novelty evaluation much closer to the way humans assess creative responses.

Literature highlights studies associated with evaluation of novelty in real-time scenarios, which is considered as a series of related data. Hayashi & Ohsawa (2015) reported strategies to evaluate novelty based on role and knowledge of a scenario. Novelty was evaluated based on the degree of problem-solving and combining knowledge of a solution. The study also highlighted relative distance between words and roles. Other factors considered were feasibility and usefulness of a solution (Hayashi & Ohsawa, 2015). However, other factors might also influence novelty in such a situation, such as narration, involvement of participants, persistence, divergent thinking, and level of social interaction (Jordanous, 2012). Measuring novelty based on narration is a significant aspect of a role, as it illustrates the uniqueness and originality of a character.

Tu & Seng (2012) reported that researchers often urged to choose themes or proposals that were novel in nature to initiate their research. It was essential to consider themes of research that were unique. A technique was reported that captured original themes in terms of published volume that depends on time, volume, and frequency. The model demonstrated life span of an emerging topic. An algorithm was proposed that was based on the selection of a topic based on volume and frequency of publications (Tu & Seng, 2012). However, one of the significant parameters to evaluate a novel theme is to identify its uniqueness. Therefore, an investigation is required to identify unique themes by comparing it with other themes within the research area. Novelty is also associated with the newness of a research area, the contribution to the domain, depth of work, etc. Moreover, novelty cannot be justified by binarized labels, i.e., novel and non-novel theme. There are varying degrees of novelty, depending on the contribution to a domain of interest.

Novelty detection of news was proposed in several literature studies. Large volumes of news are published each day, but some of the themes gets duplicated and do not possess novelty. Generally, news were introduced to annotators who judged the story based on novelty. Three popular novelty detection methods are as follows: (1) cosine similarity, (2) language model-based method, and (3) cover-coefficient based approach. In most cases, the outcomes of language model-based method outperformed (Aksoy et al., 2012). The cover-coefficient process calculates the probability of one document containing another document. However, there might be chances where two documents might completely possess different words, but they are representations of the same theme. Therefore, recognizing semantics of text is highly essential in order to find the uniqueness of a document.

Cutumisu & Guo (2019) reported Topic search on essays using Topic modelling technique. Manual evaluation of a large number of essays is impractical and inconsistent. Data pre-processing was conducted followed by Latent Dirichlet Allocation (LDA) in order to extract themes of an essay (Cutumisu & Guo, 2019). However, LDA doesn't reflect correlation between topics. An essay might use some keywords related to a theme, but as a whole document, it might mean something different and semantically might belong to a different set of themes or genre. Therefore, categorization of themes based on semantics of a document is essential. Moreover, to assign scores to essays, a score generator algorithm needs to be defined with a set of parameters that might be considered as true predictors of evaluating novelty in essays.

Humans detect novel ideas by multiple parameters, which is complex in nature, and when detection of the same needs to be performed by an automated engine, then the complexity reaches its paramount. Kim & Horii (2015) reported a method to detect novelty of ideas based on two significant parameters: (1) superficial similarity and (2) structural similarity. The first parameter showcases similarities between two ideas based on their attributes. The second parameter demonstrates similarities between ideas based on their relationship between objects. The ideas with low superficial similarity and high structural similarity was categorized as novel (Kim & Horii, 2015). However, novelty of ideas is also considered to be context specific. An idea might be considered novel if it solves a given problem and is relevant to a context.

Relevance in context of an idea is essential in order to consider an idea as novel. Because an idea might be novel, but if it doesn't solve a given problem, it is not considered useful. Feasibility of ideas in real-world might be another significant factor that can classify an idea as novel.

Uniqueness of ideas is essential in safeguarding intellectual properties. One of the significant parameters of an idea, product, or response to be established as a patent is novelty. Literature studies highlights a similarity detection method to identify patents that are unique and original in nature. Keywords and concept matching were considered to find similarity among patents. Similarity of textual information was measured based on Term Frequency-Inverse Document Frequency (TF-IDF) (Albitar et al., 2014), which was dependent on the frequency of keywords. Therefore, it hardly focussed on semantic analysis, i.e., no comparisons were made among patents based on their meaning and context. Cluster of patents was demonstrated based on their domain (Kasravi & Risov, 2009). However, in order to investigate domain of patents, semantic comprehensibility of theme is essential. Subsequently, mapping with its associated domain is essential based on context and contribution towards an area.

Literature highlights utilizing data mining techniques in finding novelty in textual documents. Often rules between words describe relationship between words in sentences. Generally, data mining techniques extract rules between words. The more distance between the antecedent and consequent of a rule, the higher is the degree of novelty (Basu et al., 2001). However, rules between sentences might have a higher impact than rules between words as framing rules between sentences might require analyzing the semantics of sentences that can support measuring the novelty of a document. Moreover, only depending on the distance between antecedent and consequent of rules may not be sufficient in all cases as highly irrelevant words might show a higher distance and render the outcome as novel. Therefore, this type of case needs to be addressed where both the relevance of words and sentences is accounted to classify a document novel (Chaudhuri et al., 2020).

The insights of this section of literature review are as follows:

i. Novelty is measured quantitatively across various fields, but there are hardly any studies that highlight digitized evaluation of novelty in creative responses illustrating descriptive creative aptitude in Design education.

ii. Majority of the studies have demonstrated measuring novelty based on its parameters. But, hardly any study illustrated the procedure of identifying and acquiring these parameters that can be used in digitized novelty evaluation of descriptive creative aptitude in Design education.

iii. Parameters of novelty are domain-specific, and hardly any studies focus on features of novelty associated with evaluating descriptive creative aptitude of students in Design education.

iv. Majority of the articles have shown evaluating novelty of textual documents and their associated models or architectures. But, there were hardly any models or architectures related to digitized evaluation of novelty of students illustrating descriptive creative aptitude in Design education.

1.5.3 Novelty in images or sketches

Over the past few years, there has been an excessive investigation in measuring novelty from images or sketches. Novelty is ubiquitous in all manifestations; along these assessment lines, they are broadly focused in literature through various viewpoints (Markou & Singh, 2003a, 2003b; Miljković, 2010; Pimentel et al., 2014). Literature featured in this segment is limited to assessing novelty from images or sketches. Evaluation of novelty in images refers to newness in image-based pattern of creative responses. Newness is evaluated based on comparing and contrasting responses in a cohort intended for a specific task or problem. Studies reported evaluation of anomalies in scanned electron microscope images. Boracchi et al. (2014) reported novelty in images that was identified by a patch-based approach where the model learned to detect patches based on training set. A model was built using sparse representation where mutually detecting reconstruction error and sparsity optimized performance of identifying anomalies outperformed as compared to other models (Boracchi et al., 2014). However, cases are subject to investigation when patch size varies. Moreover, evaluating novelty in image-based pattern of creative responses in Design education is dependent on features that might be different from other fields of study.

Some of the techniques of novelty evaluation illustrated in literature were based on the "out-of-distribution" data. Uniqueness of data and relevance to a given context were considered novel. Precisely, data that intersects with training space could be appended in a knowledge base for learning. Studies using these concepts illustrate images in the form of embeddings and neural networks. Deep Learning (DL) encompasses images through simultaneously interconnected nodes to distribute weights. A loss in this type of network adjusts embeddings to maintain a distance in embedding space between similar and dissimilar class labels (Chopra et al., 2005; Guillaumin et al., 2009; Hadsell et al., 2006). However, there might be circumstances that require "out-of-distribution" data during training of network for any unexplored context. Investigating feature of novelty assessment for a new context is essential to decently train a network for accurate prediction.

Masana et al. (2018) recommended a modified loss function by introducing a rule that assigns zero to an image when it is an "out-of-distribution" data, otherwise one. They utilized various datasets such as Tsinghua, Modified National Institute of Standards and Technology (MNIST), Street View House Numbers (SVHN), Canadian Institute For Advanced Research (CIFAR-10) for capturing "out-of-distribution" data and further training the models. Their modified function considers at least a single image as "out-of-distribution," and further computes a distance with other similar classes (Masana et al., 2018). However, datasets were too disparate in contrast with solutions related to Design assessments. The investigation intends to optimize loss function to accommodate "out-of-distribution" data, but results were hardly utilized for scoring of novelty of images.

Salehi et al. (2020) reported a novelty detection method for medical images. Numerous datasets were utilized such as medical data of haemorrhage and Magnetic Resonance Imaging (MRI), MNIST, Fashion-MNIST, CIFAR-10, and Columbia Object Image Library (COIL-100) in U-Net model (Network appears similar to letter U) with Deep Convolution Generative Adversarial Network implemented in subsequent networks (Salehi et al., 2020). Many a time, images are marked with labels or text descriptions embedded in them. There is a dearth of specifications related with handling variations of image datasets, such as images with labels embedded in them or images with annotations. The investigation highlighted optimizing efficiency and ablation studies; however, the results were domain-specific and associated with clinical studies. This further triggers the difference in features and results of novelty evaluation in Design education with any other fields.

Majority of the studies in literature reported novelty detection of medical images, where rare clinical conditions were identified and investigated. Clinical images are inclined to vulnerabilities, which might be because of seriousness of sickness, natural elements, and so forth, where irregularity and novelty detection is essential (Reinhold et al., 2020). Therefore, it is highly challenging to determine novelty in medical imagery associated with any disease or abnormality prediction where unsupervised and transfer learning techniques are essential. It is difficult to acquire labelled datasets for various new and infrequent diseases. Moreover, prediction of novel images in clinical domain requires higher accuracy (Szymkowski et al., 2020) as it deals with mainly real-time data generated from patients. But, clinical datasets and prediction of novelty in disease is different from evaluation of novelty in image-based creative responses. Features of novelty of clinical data and students' responses are different and responses mostly requires scoring mechanism.

Many studies were focused on novelty detection of planetary images (Bonnici et al., 2010; Kerner et al., 2019, 2020; Sintini & Kunze, 2020; Stefanuk et al., 2020). Multiple computational algorithms were compared such as autoencoders, generative adversarial networks, reed xiaoli detectors, and principal component analysis and further attempted to investigate the optimized method. However, there exists a difference in pattern of images between planetary objects and creative responses in Design, which restricted using any of those datasets or learning outcomes in future investigation. This genre of studies is hardly associated with novelty evaluation process in academics; therefore, features considered in those studies are distinct from Design education.

Wachs et al. (2018) reported a statistical learning technique for assessing novelty in graphic design shared over a network. Novelty was evaluated by contrasting features of current images with ones that were created earlier. As images were shared over network, user demographics and multiple network features such as number of followers, numbers of user unfollowed, number of duplicate outgoing connections, distance between nodes, and interpersonal ties were considered. Multiple image features were considered in this study, such as compositional features comprising aesthetics and colour and inception features involving contents of an image (Wachs et al., 2018). However, network contents are quite dissimilar from solutions in examination as novelty of an answer in mass examination isn't judged by the number of visits, hits, or followers. Generally, competitive examination focuses on problem solving (Amini et al., 2019); therefore, relevance between problem and its response is a fundamental aspect that pedagogues often look out for.

Few articles from literature focussed on proposing novelty detection method for a heterogeneous dataset consisting of a combination of text and image. Amarbayasgalan et al. (2018) highlighted a mechanism for measuring novelty using Common Objects in Context (COCO) dataset presented in a temporal window, which involved any time span of the network. Mask-Region Convolution Neural Network (Mask-RCNN) was used to combine text and image being predicted by machine learning models. Text and images are converted into their numerical vector representations by utilizing autoencoders (Amarbayasgalan et al., 2018). Further, novelty is evaluated using unsupervised algorithms (Amorim et al., 2019). The methodology was proposed for social media, which might not require any scoring function for novelty assessment.

Literature studies suggested novelty as a conceptual shift in design. A conceptual shift of a sketch involving a concept can be related to another concept. It is a significant procedure in Design education as it is engaged in evaluating a design from multiple perspectives and identify uniqueness in a response. Karmimi et al. (2019) proposed two parameters to evaluate novelty- 1) visual similarity and 2) conceptual similarity. Visual similarity was measured using QuickDraw dataset, where vector representations of visual sketches were computed. Further, feature vectors were clustered based on their visual similarity. Google News corpus was used to evaluate conceptual similarity. Word embeddings were created, and subsequently cosine similarity function was used to measure distance between concepts (Karimi et al., 2019). Assessment of creative responses in Design in a general context is different from examination on a large scale. Pedagogues initially intend to search for a relevant response (Abacha & Demner-Fushman, 2019; Esposito et al., 2020) compare it with other responses to evaluate novelty. Relevance is highly prioritized in examination because a response might possess aesthetic attributes, but it must meet the requirements of the question.

Literature highlights novelty detection by applying various classification algorithms (Désir et al., 2013; Schölkopf et al., 2001); however, it is challenging to acquire novel dataset as it is associated with newness of a concept. Bodesheim et al. (2015) proposed a shift of concept for every novel response. Similar to human evaluation strategy, irrelevant responses were discarded before scoring novelty. Further, each sample's local learning model was framed using the K-nearest neighbour algorithm. A response was considered novel if it was at a distance from its nearest neighbours (Bodesheim et al., 2015). However, datasets considered in the article were purely image-based solutions, and there were no case studies related to a combination of sketch and textual descriptions. Further, the study described evaluation of novelty for a generalized context; however, for a specific context such as evaluation for mass examination in Design based educational assessment, it is required to define context-specific parameters for evaluation.

Linder et al. (2013) highlighted procedures to assess novelty from Googles' list items of images. Initially, a sample of 3579 images was acquired, and search queries for sample were created algorithmically. Queries were utilized to perform search of images and search results were captured. A function was derived that can measure novelty by inversing the number of search results generated by an image (Linder et al., 2013). This mechanism was coherent but hard to correlate with dimensions of examination. Though search techniques on the web might be applied in a few cases, such as design registration or design exhibition, evaluation associated with mass examination requires comparing and contrasting ideas within a cohort of responses. Further, assessment in Design based entrance examination is associated with creative responses that are an amalgamation of visual and textual content. The study did not highlight possible categories of image-based pattern of creative responses. Moreover, the function defined was restricted to image-based responses over web and was not adapted to solution found in other media.

The insights of this section of literature review are as follows:

i. Novelty is measured quantitatively across various fields, but hardly any study highlighted digitized evaluation of novelty in responses illustrating image-based creative aptitude.

ii. Majority of the studies have demonstrated measuring novelty based on various parameters, however, the procedure of identifying and extracting these parameters were not categorically specified and detailed.

iii. Parameters of evaluating novelty are domain-specific, but hardly any studies focussed on features of novelty specifically associated with image-based creative aptitude of students specifically for Design or creative specialization domain.

1.5.4 Stress associated with pedagogy-related tasks

Any unpleasant stimulus that triggers a natural reaction in human body and mental state is known as stress (Yaribeygi et al., 2017). Empirical studies highlight multiple stress factors in pedagogues such as "indisciplined students", "time pressure", "less motivated students", "lack of supervisory support", "conflict with colleagues", "value conflict", and "student diversity". Pedagogues often get stressed by behavioural problems of students. They are highly persuaded by multiple activities such as prolonged working hours, meetings, preparation for classes, evaluation, and low motivation in students. A dearth of trust and acclaim from institutions, differences in opinion, and values is also a stress generator source. A large difference in students' abilities is termed "student diversity", where a student or a group performs extraordinarily well; on other hand, other students or groups not performing well may too cause stress. These factors are responsible for emotional stress, emotional exhaustion, and motivates one to quit a job, reduces self-efficacy, and most drastically reduces engagement in work (Skaalvik & Skaalvik, 2016). Lack of engagement leads to a reduction in efficiency, consistency and thereby increases errors in work.

Stress generators in pedagogues can be categorized from three main sources- 1) administrative work and pressure, 2) interpersonal differences, and 3) hectic tasks associated with classroom. These factors reduce self-confidence in pedagogues (Naghieh et al., 2015). This research mainly focuses on hectic tasks associated with examinations that involve prolonged and repeated evaluation of creative questions and their responses that have potential in capturing creative aptitude of students (Boyle et al., 1995). Stress generated due to this task minimizes self-confidence, efficiency, and consistency in pedagogues. They are filled with self-doubt and remain ever-inquisitive to know whether evaluations done on a large scale are correct and consistent. Any error in this evaluation process would reduce trust in the academic system (Tan, 2017).

Mass examination requires a fast evaluation of answers as it is essential to declare results within a stipulated timeline for students aspiring admission to Design schools. Checking subjective responses and identifying creative questions is time-consuming, and conducting it faster on a large scale affect its credibility (Jobanputra, 2019; Mihaylova et al., 2018, 2019). In situations like this, pedagogues often raise self-doubt, which is also a factor of stress generator. Many a time, they recheck their evaluation in order to verify mistakes in evaluation. Time optimization is significant in these processes; therefore, re-evaluation on a large-scale is hard to implement and infeasible in entrance examinations.

Consequences of stress in pedagogues have an effect at multiple levels in one's life, such as personal, interpersonal, and organizational. Effect of stress at a personal level lead to self-isolation, inadequacy, and anxiety. On an interpersonal level, it lead to a difference in opinion between students, colleagues, and parents. It leads to disagreement in role clarification and disempowering practices and policies (Prilleltensky et al., 2016). It has a deep impact on physical, mental, and emotional wellbeing of pedagogues (Skaalvik & Skaalvik, 2017).

The insights of this section of literature review are as follows:

i. There is a dearth of literature addressing stress of pedagogues in assessing creative aptitude in Design education.

ii. Literatures hardly reported any human-centred design and technology based intervention to address stress generated due to hectic creative aptitude assessment process of pedagogues specifically during subjective assessments of creative aptitude.

The existence of stress due to pedagogical-related activities and their substantial evidence is reported in literature. This review can be considered as the reason why digitization techniques are required in pedagogic activities. Predominantly, this state of the art review supported in identifying the questions that has the potential to instigate creative responses among students. It also aided in understanding the role of novelty in evaluating creative aptitude, and the patterns of responses in evaluating novelty in Design entrance examination, which is further framed as the objectives (Objective 1, 4, and 5) of this thesis in Section 1.12.

1.6 Research gap

Research gap has been identified from two major perspectives viz., 1) identifying features of creative question that instigate creative responses from students, and 2) identifying parameters for evaluating creative responses. Exhaustive literature review highlighted that there are abundant articles associated with subjective and objective evaluation techniques. Numerous literature has been identified in the domain of evaluating novelty in products and automation of assessment in engineering. Factors of stress has been also highlighted in literature studies.

But, there is a dearth of literature regarding features of creative questions and parameters for assessing their creative responses in entrance examinations conducted by Design institutes.

The outcome of the literature review highlighted that there is a dearth of studies that ever investigated ways pedagogues frame creative questions for assessing students' creative aptitude. Also, hardly any research papers have ever reported on ways to distinguish between a creative question from a bunch of non-creative questions. These gaps lead to research questions (RQ-1 and RQ-2) presented in Figure 1.4.

Presently, evaluating novelty of creative responses is manual and dependent on pen-and-paper-based techniques. Literature highlights stress in pedagogues due to academic tasks, administrative tasks, and interpersonal differences. It can also be interpreted thus, that during evaluation of large scale creative aptitude pedagogues may also get stressed as currently the assessment is manual in nature. Further, during such assessments pedagogues adopt subjective evaluation based on their individual referential metrics, this may lead to bias, and inconsistency due to involvement of multiple evaluators in the evaluation process. These gaps lead to RQ-3 and RQ-4, as illustrated in Figure 1.2.

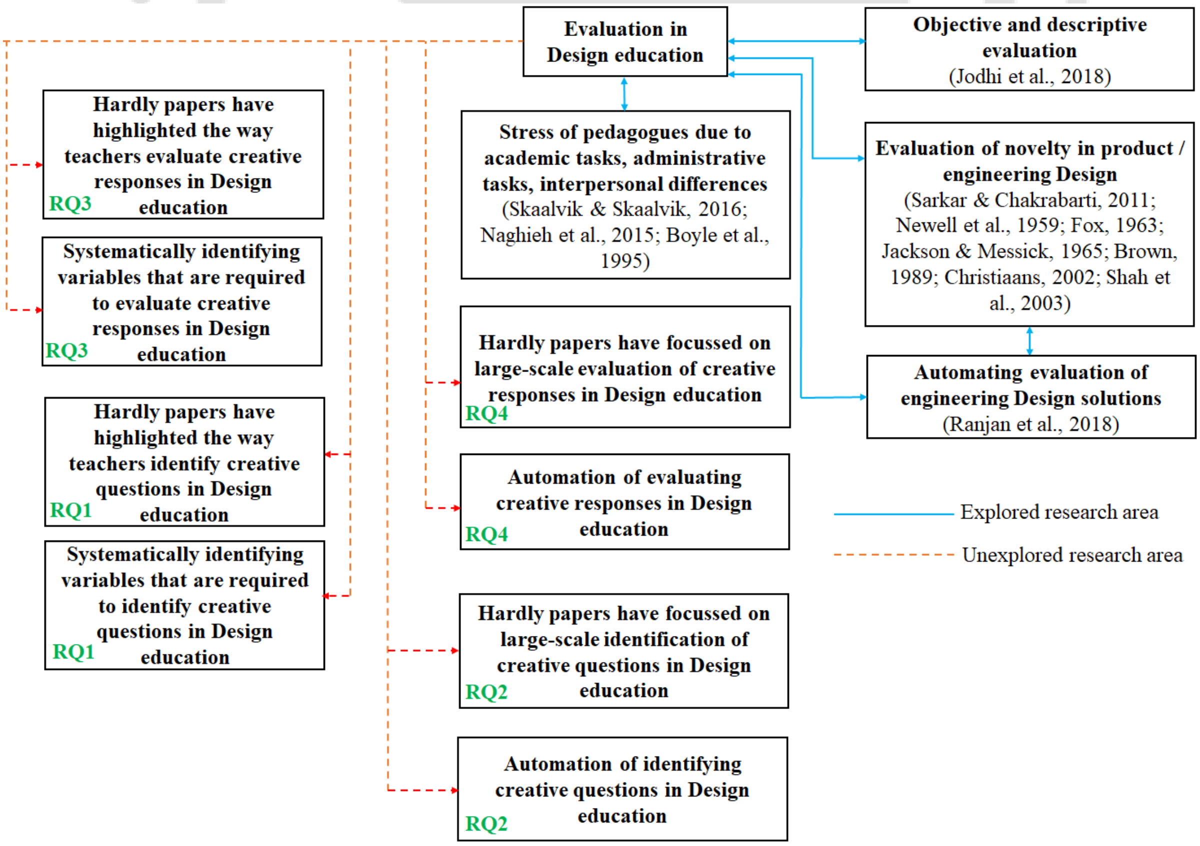


Figure 1.2: Research gap

1.7 Motivation of the research study

Ever-increasing population of students participating in entrance examinations of Design educational institutes requires formulation of creative questions that instigate creative responses from students and assessing their responses on a large scale. Presently, the assessment of creative aptitude in Design entrance examination is subjective, and depends on pedagogues' referential metric (Park et al., 2016). Conducting subjective evaluation on a large scale might lead to inconsistency in the evaluation process due to individual biases based on their past experience and monotony of the repeated task. Inconsistency in the evaluation process has a drastic drawback to the education system. They are as follows- 1) Ambiguity in the selection process of students aspiring admission to Design schools leading to frustration in students, 2) Self-doubt of pedagogues in the process of identifying creative questions and assessing their creative responses. These driving factors lead to focus on the following research questions presented in the subsequent section.

1.8 Research questions (RQ)

RQ1: What are the features of a question that triggers creative responses?

RQ2: How to identify a creative question from a set of non-creative questions?

RQ3: While assessing creative responses of students, what are the factors that Design educators consider for assessing novelty of the responses?

RQ4: Can the existing process of subjective manual evaluation of creative aptitude be automated?

1.9 Aim

To digitize the creative aptitude assessment process in Design education with an intention to offer support to Design pedagogues and optimize the assessment process.

1.10 Objectives

Objective 1: To identify questions that has the potential to instigate creative responses among students.

Objective 2: To identify variables of questions that has the potential to instigate creative responses among students.

Objective 3: To design a digitized system to identify creative questions that has the potential to instigate creative responses among students.

Objective 4: To examine the role of novelty in assessment of creative aptitude.

Objective 5: To examine types of responses in evaluating novelty in creative aptitude.

Objective 6: To identify the factors of novelty in creative aptitude evaluation.

Objective 7: To design a digitized system for novelty assessment in creative aptitude.

1.11 Focus of research

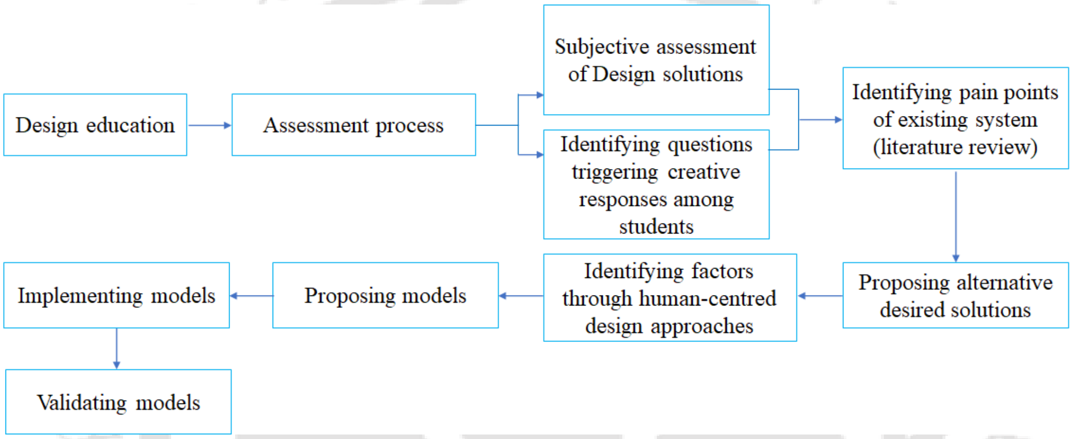


Figure 1.3: Focus of research in this thesis

The primary focus of this thesis lies in Design discipline, wherein the research is narrowed down to highlight a research gap. This thesis mainly focusses on assessment process of creative aptitude in the context of Design entrance examinations conducted by various institute in India. Two categories of assessment is considered here, viz., subjective assessment of question formulation that instigates creative responses from students, and subjective assessment of the responses to these type of questions. The research study further identifies the pain points associated with these existing type of assessments. Alternative design solutions have been proposed to address inconsistences involved in assessment due to repeated tasks on a large scale. Factors of subjective assessment are identified using human-centred design approaches that encourages human involvement in capturing the parameters of assessment by survey, interview, etc. Models are proposed based on these factors. Further, the design of models are implemented using numerous algorithmic techniques. Finally, the implemented models are validated by comparing the outcomes of the models with human-based evaluations. The focus of this thesis is illustrated in Figure 1.3.

1.12 Expected outcome(s)

i. Identifying dimensions of questions that have the potential to instigate creative responses from students.

ii. Supporting pedagogues in identifying creative questions that possesses feature for instigating creative responses.

iii. Identifying dimensions of novelty in evaluating creative aptitude in Design education.

iv. Supporting pedagogues in evaluating creative responses illustrating creative aptitude on a large scale by digitization.

1.13 Structure of the thesis

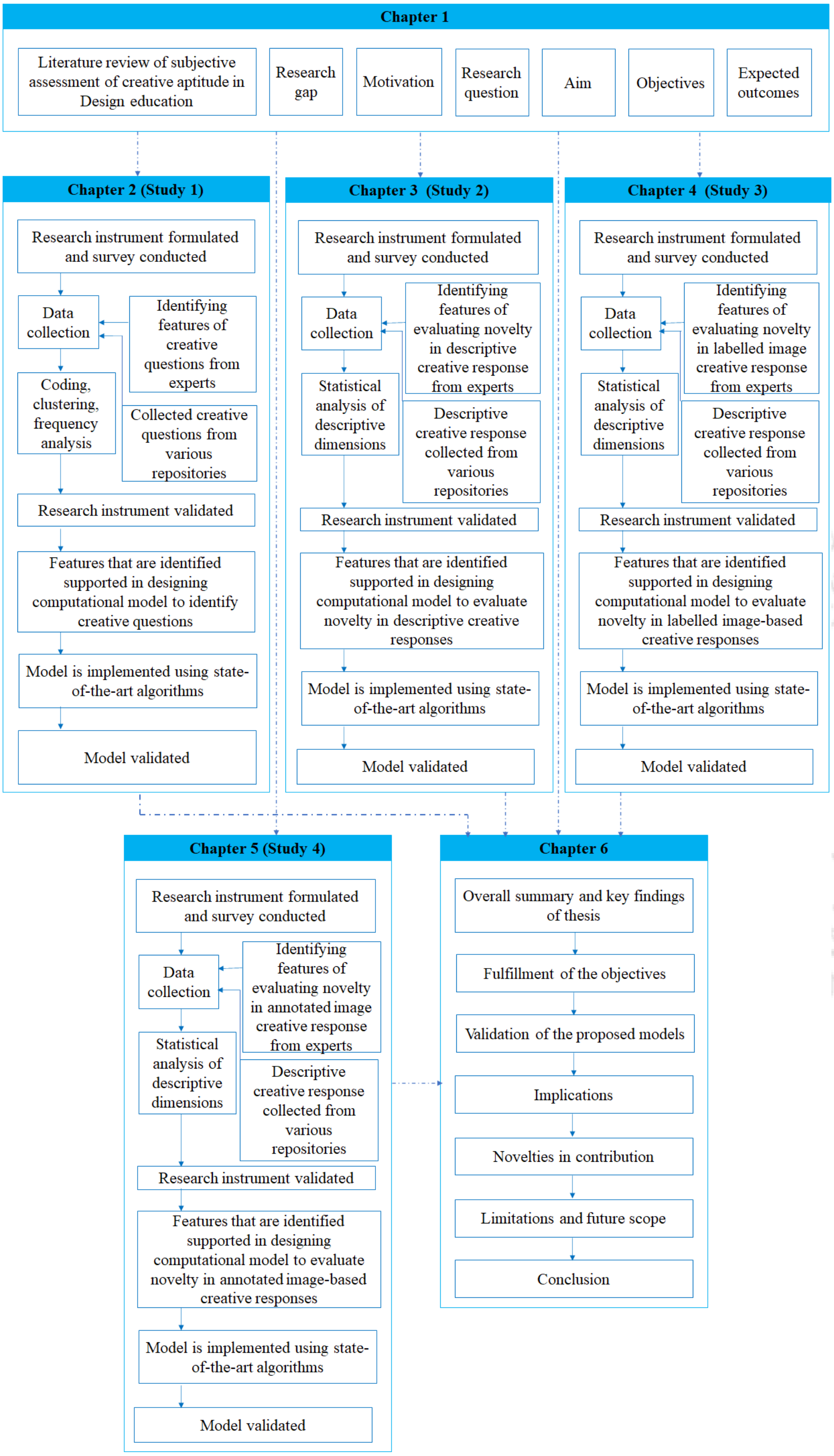


Figure 1.4: Workflow of this thesis

Table 1.1: Research questions, objectives organized in various chapters and publications

Chapter No.	Research Questions	Objectives	Publications
I,II	**RQ1:** What are the features of a question that triggers creative responses?	**Objective 1:** To identify questions that has the potential to instigate creative responses among students. **Objective 2:** To identify variables of questions that has the potential to instigate creative responses among students.	**Chaudhuri, N. B.,** Dhar, D., & Yammiyavar, P. G. (2021). Do Design Entrance Exams in India Really Test Creative Aptitude? An Analytical Study of Design Tests Conforming Creativity Benchmarks. In Design for Tomorrow—Volume 2 (pp. 371-383). Springer, Singapore. [Scopus indexed] ***[Identifying question format of Design tests in India]***
II	**RQ2:** How to identify a creative question from a set of non-creative questions?	**Objective 3:** To design a digitized system to identify creative questions that has the potential to instigate creative responses among students.	**Chaudhuri, N. B.,** Dhar, D., & Yammiyavar, P. G. (2021). A human-centred deep learning approach facilitating design pedagogues to frame creative questions. *Neural Computing and Applications*, 1-28. **https://doi.org/10.1007/s00521-021-06511-8** [SCIE/Scopus indexed, IF: 5.606] ***[Identifying factors and automated identification of creative questions in Design education]***
I,III,IV,V	**RQ3:** While assessing creative responses of students, what are the factors that Design educators consider for assessing novelty of the responses?	**Objective 4:** To examine the role of novelty in assessment of creative aptitude. **Objective 5:** To examine types of responses in evaluating novelty in creative aptitude.	**Chaudhuri, N. B.,** Dhar, D., & Yammiyavar, P. G. (2020). A computational model for subjective evaluation of novelty in descriptive aptitude. *International Journal of Technology and Design Education*, 1-38. **https://doi.org/10.1007/s10798-020-09638-2** [SCIE/Scopus indexed, IF: 2.177] ***[Identifying factors and automated evaluation of descriptive pattern of creative response]***
III,IV,V	**RQ4:** Can the existing process of subjective manual evaluation of creative aptitude be automated?	**Objective 6:** To identify the factors of novelty in creative aptitude evaluation. **Objective 7:** To design a digitized system for novelty assessment in creative aptitude. ***N.B.:*** *These objectives are repeated for three types of creative responses*	**Chaudhuri, N. B.,** Dhar, D., & Yammiyavar, P. G. (2021). A System For Evaluating Novelty Of Creative Write-Up. **Application no. 202131006753** [Patent filed in India] **Chaudhuri, N. B.,** Dhar, D., & Yammiyavar, P. G. (2021). Automating assessment of Design exams: A case study of novelty evaluation. *Expert Systems With Applications*, ISSN 0957-4174, **https://doi.org/10.1016/j.eswa.2021.116108** [SCIE/Scopus indexed, IF: 6.954]

	[Identifying factors and automated evaluation of labelled image and annotated image-based pattern of creative response]
	Chaudhuri, N. B., & Dhar, D.(2021). Designing deep-network based novelty assessment model in Design education (In Review) ***[Proposing assessment model]***
	Chaudhuri, N. B., & Dhar, D.(2021). Digitizing creativity evaluation in Design education: A systematic literature review (In Review) ***[Literature review of digitized assessment systems]***

1.14 Summary of chapters

This thesis is divided into six chapters is as follows:

Chapter 1: Introduction (present chapter)

The first chapter highlights the overview of Design entrance examinations in India, question patterns, types of question that instigates creative responses from students, patterns of creative responses, categories of novelty, understanding and evaluating novelty in mass examinations of Design institutes. State of the art literature review highlighted methodologies, algorithms related to descriptive and image-based content assessment, and evaluation of question formulation. Subsequently, insight is drawn from literature review and research gap, research questions, aim, objectives, expected outcomes, and structure of this thesis is presented.

Chapter 2: Identifying parameters of creative questions and designing a digitized system to support design pedagogues for framing questions

The second chapter reports the study conducted to identify the parameters of questions that have the potential to instigate creative responses from students. A systematic mixed-method technique is reported that have been used to capture the features of creative questions from experts. Further, design to automate the identification of creative questions is illustrated. The design is transformed into computational models by implementing it using various algorithmic techniques. Finally, inter-rater reliability among the model and examiners was measured, and the outcomes were reported to show the subjective agreement among them.

Chapter 3: Identifying parameters to assess novelty of descriptive creative responses and digitizing its evaluation process

The third chapter describes the comprehensive investigation of identifying dimensions or parameters of evaluating novelty from descriptive creative response. This identification process involves studies with human experts possessing expertise in evaluating creative aptitude. The descriptive statistic of experts involved in the study is presented in this chapter. Further, design of the computational model to automatically evaluate descriptive creative response is illustrated. Subsequently, the model is implemented by using various algorithms. Finally, the model is validated by comparing the outcome with human-based assessment.

Chapter 4: Identifying parameters to assess novelty of labelled image-based creative responses and digitizing its evaluation process

The fourth chapter highlighted the investigation of identifying the features of evaluating novelty from labelled image-based creative responses. This identification process involves studies with human experts possessing expertise in evaluating creative aptitude. The descriptive statistic of pedagogues involved in the study is illustrated in this chapter. Further, design of the computational model to automatically evaluate labelled image-based creative response is shown. Subsequently, the model is implemented by using various computational procedures. Finally, the model is validated by comparing the outcome with human-based assessment.

Chapter 5: Identifying parameters to assess novelty of annotated image-based creative responses and digitizing its evaluation process

The fifth chapter describes the comprehensive investigation of identifying dimensions or parameters of evaluating novelty from annotated image-based creative response. This identification process involves studies with human experts possessing expertise in evaluating creative aptitude in design education. The descriptive statistic of experts involved in the study is presented in this chapter. Further, design of the computational model to automatically evaluate annotated image-based creative response is illustrated. Subsequently, the model is implemented by using various computational methods. Finally, the model is validated by comparing the outcome with human-based assessment.

Chapter 6: Discussion and conclusion

The sixth chapter presents a comprehensive discussion of the overall summary and key findings of the thesis. The fulfillment of the objectives are mapped with each of the chapters. A detailed description of the validation of each model is presented. Implications of the research are derived from the perspective of design process and human-centred approach. Further, contribution to knowledge-base, methods, design, and design education are highlighted. Finally, limitations, future scope, and overall conclusion of the thesis are presented.

Chapter 2: Identifying parameters of creative questions and designing a digitized system to support design pedagogues for framing questions

Abstract

Creative question triggers students' creativity and pedagogues attempts to capture it by the creative questioning technique. Creative questions are a major component of examination in Design education for testing creative aptitude. Examiners drill their thought processes and grapple with their ideas to frame questions that are creative in nature, which is capable of capturing creative aptitude in students. While framing creative questions, examiners often self-evaluate, compare, and contrast their ideas before finally phrasing the question. During this process, they remain ever-inquisitive to know whether questions framed by them are really creative; to be more precise, do the questions framed by them really capture features of creative questions? Peer review in these situations provides meaningful insights into construction of these questions. However, peer review has its own demerits. Individual characteristics and past experiences influence frame of reference of all individuals. And as such peer-reviewing a question paper might lead to more differences than convergences. The investigation presented in this thesis is exactly geared towards this issue. Our objective is to explore whether technology can support examiners in situations like these. Nowadays, increasing number of Deep Learning (DL) techniques are widely applied in Question-Answering (QA) platforms to assess success of a question. DL is often used to recognize features of questions. However, it has been hardly used to identify creative features in questions that attempt to trigger creative response from students. The study presented here, investigates features of creative questions through mixed-method research techniques. A model is proposed based on DL algorithms that can find out inherent creativity factors in questions and identify whether a question is creative. This process of identifying creative questions triggers decision-making of examiners by which they update their questions based on the outcome of the DL-based system. This model is implemented using Bidirectional Encoder Representations using Transformers (BERT) and Long Short Term Memory (LSTM) method for identifying creativity in questions and their performance is compared. Results highlight that BERT overrules LSTM mechanism, thereby optimizing the trust in BERT algorithms. A comparative study between the outcome of the model and examiner’s opinion of categorizing creative questions are mapped, thereby further building trust in the model. A major contribution of this research is to capture creative features in a question and categorize whether a question is creative in design education. This model

highlights human-machine collaboration and promotes examiners' decision-making process to frame effective questions. It attempts to reduce uncertainty of examiners and assists in quick decisions to include creativity features in their questions by providing feedback on whether a question is creative.

Highlights

- *Human-centred design approach to identify parameters of questions that has the potential to instigate creative responses from students.*
- *Proposing a computational design model for assessing creative questions.*
- *Implementing the model using various tools and algorithmic techniques.*
- *Validating the model by measuring the inter-rater agreement.*

2.1 Introduction

India is rich in art, culture, and design, but structured and systematic design education was imparted to only a few of the schools until 2004. After this period, design schools in India increased at an exponential rate, and reports highlight more than seventy design institutes in the year 2016 (Gemmell & Vyas, 2016; A. Sharma & Kirloskar, 2008). Design schools attempt to test creative skills of students; however, each design school has its own strategy of testing creativity. It is essential to identify the features of questions that have the potential to instigate creative responses from students. Creativity has various forms, such as big-creativity associated with immense novel inventions like the invention of a steam engine; pro-creativity tends to possess a relatively lesser degree of creativity than big-creativity. Pro-creativity measures the novelty of outcomes of professionals in their domain of expertise. Little-creativity involves a lesser degree of creativity than pro-creativity and is associated with contributions in day-to-day life activities, whereas mini-creativity possesses the least degree of creativity, and it is evaluated within a particular community, or any person considers an embodiment as creative (Kaufman & Sternberg, 2010). Creativity is also classified as H-creativity or Historical creativity and P-creativity or Psychological creativity or Personal creativity. H-creativity is involved in predominant creativity from the perspective of uniqueness in the entire human race, whereas an idea or a response is P-creative, if it is creative with respect to the mind of the person or the community concerned (Gigerenzer, 1994). Creativity in an examination is mostly associated with mini creativity or personal or psychological creativity, where creativity is evaluated relative to other responses (Csikszentmihalyi & Wolfe, 2014).

Extensive methods of creativity tests are reported in the literature. Many creativity tests lack reliability and validity in recognizing creativity, while others are not meant to be used in the context of examination. But in examination context, experts may decide to choose a combination of these techniques to extract creativity from different perspectives. Literature highlights some of the creativity tests like Consensual Assessment Technique (CAT) that evaluate products, art, theory, or artifacts based on expert's opinion (Pritzker & Runco, 2011). Remote Associates Test (RAT) tests creativity based on divergent thinking. It examines the degree of unrelated ideas combined to form a coherent whole of an idea. One of the significant tests reported in literature is the Torrance Tests of Creative Thinking (TTCT), which is similar to the tests that are conducted for selecting students in design schools. It scrutinizes creativity from verbal and figural perspectives. The verbal part is checked for one's creativity by analyzing words with which they frame narrations. In contrast, the figural part is tested based on the usage of visual elements, completeness of the art, and degree of modifiability of visual elements (Kaufman et al., 2012).

Some tests are related to examining creativity of children, like Wallach and Kogan's method of creativity testing, which desires to test creativity and intelligence of fifth-standard students (Silvia, 2008). Similarly, Getzel and Jackson's study reveals the fact of testing sixth-grade gifted creative students (Getzels & Jackson, 1962). These categories of tests are associated with testing intelligence and divergent thinking of children. However, creativity of adults is different from creativity of children. An artifact created by children may seem creative, but the same developed by an adult might not appear creative. However, each of these tests is dependent on examiners' choice and persuasion.

A highly consistent selection process is essential in order to qualify students on a large scale entrance examinations conducted in India and other countries. This study highlights Indian context due to the large population of students participating in entrance examinations. Most of the tests associated with Design education have objective and subjective question structure. The objective part consists of questions related to numerical answer type, multiple-choice, and multiple select questions, whereas the subjective part contains questions associated with sketching, form sensitivity, visual sensitivity, and problem identification that attempts to capture creativity of students (Bombay, 2021b, 2021a). The solutions to objective questions are straightforward and based on a strict set of options, whereas creative responses are marked by the usefulness of ideas and comparison of novelty relative to other responses (Sarkar & Chakrabarti, 2011). Subjectivity evaluation is based on individual persuasion and is relatively complex than the objective evaluation.

Literature highlights multiple types of creativity testing for different contexts such as creativity tests of young and adults, testing of products, different question patterns of design tests in classroom, etc. A lacuna has been investigated from the evidence of literature that there is less focus on the ways of capturing creativity in the design entrance exams. There is also a lack of assurance in literature, whether optimized creativity is captured based on all its factors and requirements of the design schools. Further, hardly studies focused on proposing a standardized format of testing creativity for entrance examinations in design schools. This leads to the following questions: What factors in questions can instigate creative responses among students? Further, do Design tests in India confirms the systematic identification of creative questions from a bunch of other non-creative questions?

Creative question is a significant component in Design education that attempts to trigger creativity in students. Questioning is a medium by which pedagogues confirm learning and recalling ability of students (Baloche & Platt, 1993). It triggers creative and critical thinking in students. There are multiple types of questions by which teachers try to capture learning and knowledge from students (Aziza, 2018). Creative questioning is an art and science that invokes creativity in students (Zolfaghari et al., 2011). All questions are not creative, and a way of assessing quality of a question is the divergent creative responses acquired from students. However, it is debatable that quality of a student determines the degree of creativity in a response. Further, in a mass examination, assessment of quality of questions is essential prior to it is delivered to students.

This research is intended towards the direction of identifying creative features of questions that instigate creative responses. This study is not focused on identifying features of creative questions through peer-review techniques, observations, or individual subjective assessment. The aim of this study is to find out features of creative questions from experts in this field by mixed-method research techniques and further digitize identification of creative questions to achieve a consistent identification process. This study informs examiners whether a question framed by them is creative or not, which further assists pedagogues in decision-making of reformulating questions, as illustrated in Figure 2.1.

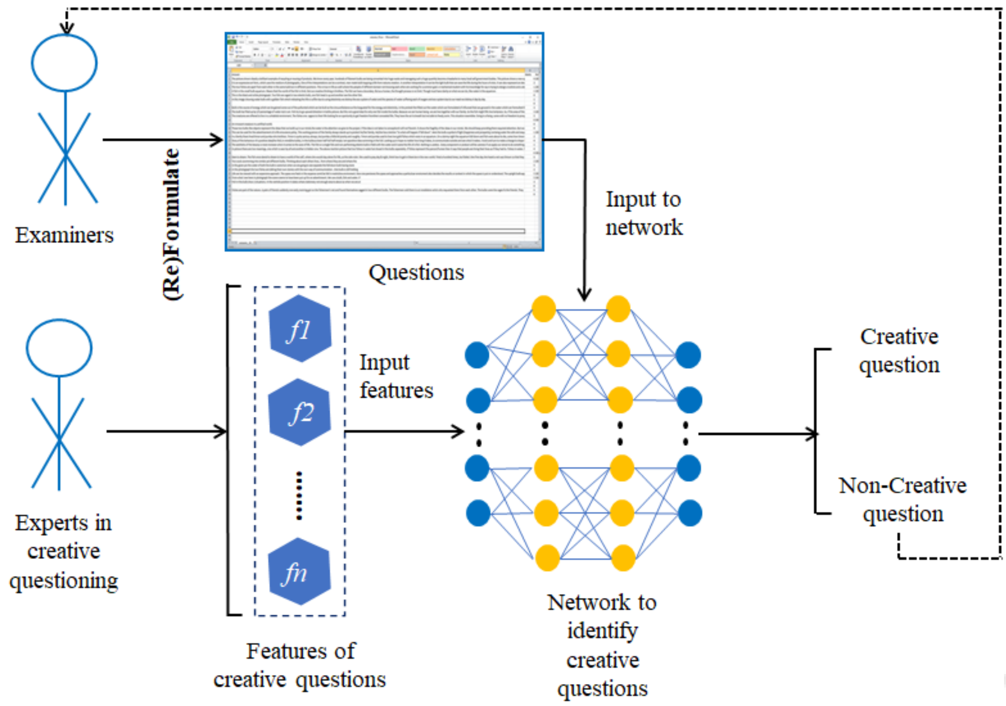


Figure 2.1: Overview of the model

Features of creative questioning (f1, f2,…, fn) were acquired from pedagogues associated with Design education experienced in formulating creative questions. A network was built using Deep Learning (DL) techniques capable of identifying creativity based on those features. Questions framed by examiners serve as input to the network, categorizing it as creative and non-creative. A question that turns out to be non-creative assists in decision-making of examiners that reformulating a question is essential to turn it into a creative one. However, this research is restricted to the decision of whether a reformulation of a question is required or not. Human-engagement is essential in this context to reframe any question. This study does not focus on assisting examiners on how a reformulation can be done in order to make a question creative.

The investigation in this chapter attempts to address the research gaps highlighted and reported in the state-of-the-art literature review presented in subsection 1.5.1, subsequently corresponding research questions and objectives reported in section 1.9 and 1.12. The research questions and the objectives are stated below again for reference.

RQ1: *What are the features of a question that triggers creative responses?*

RQ2: *How to identify a creative question from a set of non-creative questions?*

Objective 1*: To identify questions that has the potential to instigate creative responses among students.*

Objective 2: *To identify variables of questions that has the potential to instigate creative responses among students.*

Objective 3: *To design a digitized system to identify creative questions that has the potential to instigate creative responses among students.*

2.2 Method

This study focuses on two research questions. The first question pointed out identifying features that are essential to recognize a creative question that instigate creative responses. To accomplish the solution for this question, an unstructured interview was planned, followed by mixed-method analysis to identify features of creative questions. The second research question highlighted identifying a creative question from a bunch of other non-creative questions. In order to accomplish this goal, creative and non-creative questions were collected from multiple repositories to test DL models to identify creative questions. The implication of this system is that it assists examination paper setters in deciding whether reformulation of a question is required in order to register it as creative.

The investigation reported here highlights features of creative questions and ways to identify them using mixed-method research techniques (Östlund et al., 2011). Here, mixed-method research encompasses qualitative data gathered using a semi-structured interview technique to find out features of creative questions and quantitative techniques to identify creativity in questions. Here, open coding was applied to generate codes, and frequency analysis of codes assisted in the removal of insignificant codes (Hoddy, 2019; Rack et al., 2018). The frequent codes that affirms creativity were considered as features of creative questions and are further utilized for identifying them using DL algorithms. The outcome of the algorithms would assist examiners in optimizing decision-making in reformulating creative questions if required. The overall architecture followed in this study is illustrated in Figure 2.2.

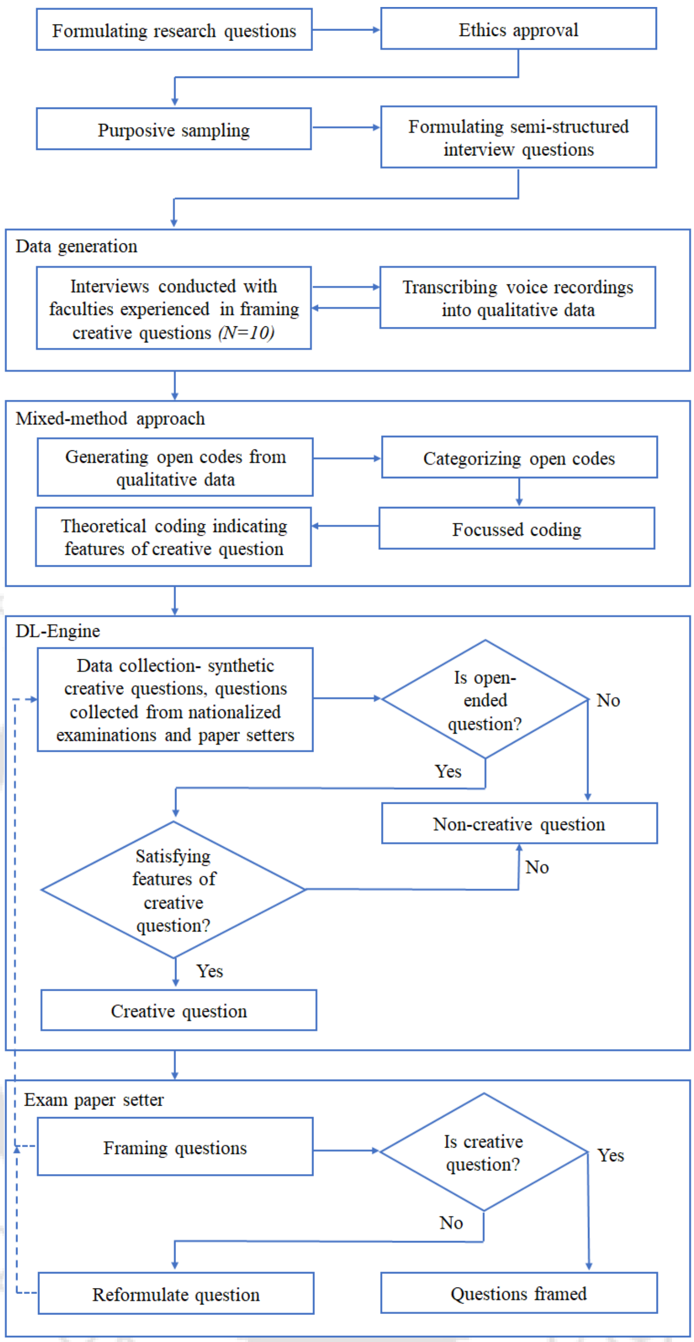


Figure 2.2: Architecture of proposed model to recognize creative questions and to optimize decision-making of exam paper setters

Domain-specific data considered in majority of general classification and prediction problems in literature are usually acquired from existing data repositories. Further, algorithmic approaches using those data are utilized to classify and predict the required tasks (Ray, 2019). However, in typical research problems like this, where creativity in questioning of Design education needs to be identified, requires mixed-method approach involving both human and algorithmic intervention. The typical problems that are unconventional, lacks existing documentation, and require understanding its characteristics from its legitimate source. Literature highlights similar types of approaches, where qualitative studies involving humans were conducted to gather data or collect theme, and quantitative and machine learning approaches are considered for building intelligent systems (Alshanqiti et al., 2020; Nappi, 2017).

Basias & Pollalis (2018) proposed a research methodology, where they suggested executing a research problem in following steps. Firstly, identifying research questions and briefing the objectives were essential. Secondly, research methods or techniques were identified based on qualitative, quantitative, or mixed-method approaches. Finally, results were highlighted based on outcomes of the research methods. The methodology chosen for this study was influenced by the research methodology framework of Basias and Pollalis that attempted problem-solving by both qualitative and quantitative approaches. The methodology that was followed is- 1) Initially identifying research questions for this study was significant for exploring the problem in a focused way and planning the methods in executing them. 2) Further, techniques for problem-solving were essential involving mixed-method research and algorithmic approaches to execute this critical task. 3) Finally, highlighting results by providing questions as input and categories of question as output was essential to identify questions that instigate creative responses (Basias & Pollalis, 2018).

Earlier Recurrent Neural Networks (RNN) were widely accepted in research problems, which highlighted a cyclic architecture that enabled it to update the current state based on its previous states. However, when the sentences were too long, and the output was dependent on input present at initial stages, Vanishing Gradient problem hindered the long-term dependencies. During Back Propagation, weights get updated with a chain rule, which transforms the weight into a very small value. Again, when derivative of that weight is considered, then it becomes an even smaller and negligible value. LSTM can tackle the problem of long-term dependencies by introducing gates, which can add or remove information to the cell state (Wang & Jiang, 2015; Yu et al., 2019). Literature highlighted overwhelming results of LSTM in Natural Language Processing (NLP) tasks and usage of LSTM in the field associated with creativity sector (Anantrasirichai & Bull, 2021; Shedko, 2018), thereby influenced this study to utilize and experiment with this advanced state of the art technique (G. Rao et al., 2018).

BERT is a significant innovation supporting NLP tasks. BERT is computationally effective as, unlike ELMo and ULMFiT that require specific architecture for additional features for human perception of the end task, BERT's fine-tuning approach requires no specific architecture. BERT is a pre-trained model that possesses language understanding, and further fine-tuning enables to perform NLP tasks (Devlin et al., 2018; Xu et al., 2019). Literature focusses on the overwhelming results of BERT in comparison with other state-of-the-art algorithms for intended NLP tasks (Garcia-Esteban, 2017). This pre-trained model was also found associated with creative fields (Hossain et al., 2020). Therefore, this model is utilized and experimented within the present methodology to identify creative questions, and its performance is evaluated. The architecture highlighted the plan of identifying a creative question and optimizing decision-making process of examination paper setters, who were earlier dependent on their beliefs and frames of reference.

Initially, research questions were formulated and further research plans were devised. Institute human ethics committee approval was received prior to this study. Qualitative data was captured using semi-structured interview approach. Later these qualitative data were transformed into multiple stages of codes, and their frequencies were measured. The affirmative codes were later considered as features of creative questions. A DL-engine was developed to identify creative questions based on these features. Questions framed by examiners are provided as input to DL-engine, which categorized whether a question was creative. In case of questions lacking creativity, it suggests reframing the question. The major contribution of this study was to conceive, design, and develop an intelligent system that recognizes a creative question and supports pedagogues in decision-making while formulating a question paper to test creative aptitude. In addition to this, baseline comparison was conducted among algorithms to compare and identify which among them outperformed in identifying creative questions. Further, a comparative study was also conducted among examiners and the proposed model to identify the degree of inter-rater reliability.

2.2.1 Questionnaire preparation and interview

Semi-instructed interview questions were prepared with reference to the features acquired from literature review. The list of questions used in the semi-structured interview is provided in Appendix A. The objective of this questionnaire was to identify the factors usually Design pedagogues refer to while framing questions that instigate creative responses from students.

The objective of the study conducted with Design pedagogues was to identify the factors to ascertain whether the factors that literature review has suggested are the ones that they also consider in practice. State of the art review was conducted to examine the findings from literature that contributed to the identification of factors for framing questions that triggers creative responses as shown in Table 2.1.

The factors listed above were acquired from the state-of-the-art literature are mostly associated with seeking critical thinking and problem-solving abilities from students. Design pedagogues were shown these factors and were asked to identify the ones which they consider for framing creative questions in educational settings. One of the reasons for using this technique was due to limited number of studies that investigated the factors of questions that instigate creative responses; therefore, open-ended questions were used for identifying the experience and point of reference for framing this type of question by Design pedagogues. Secondly, this technique enabled collecting first-hand data from Design pedagogues who are experienced in formulating such questions for Design entrance examinations.

Semi-structured interview questions were framed with reference to the features acquired from literature review. The questions were confirmed by expert review and face validity. Purposive sampling or expert sampling was used to find out subjects having expertise in framing creative questions. Ten experts *(N=10)* from different renowned design institutes were chosen, possessing experience in the range of *10-30* years of framing creative questions for nationalized examinations. The pattern of data acquired after the fifth subject tends to be repeated; therefore, the interview segment was concluded after the tenth recording of the participant. The recordings of the telephonic interview provided a rich source of information highlighting the important features of creative questions.

Table 2.1: Factors for framing creative questions (questionnaire given in Appendix A) that are associated with literature articles

Factors	State-of-the-art literature
Asker intent understanding, conversational, expect short answer, fact seeking, has commonly accepted answer, interestingness others, interestingness self, multi intent, not really a question, opinion seeking, type choice, type compare, type consequence, type definition, type entity, type instructions, type procedure, type reason explanation, type spelling, well written	Annamoradnejad et al., 2020
Seeking views of respondents, seeking a well-explained solution, appreciating multiple responses, seeking imaginative answers, framing a branched question, open-ended questions	Zolfaghari et al. 2011
Inspecting intensive concepts, fact-seeking questions, hypothesis scrutinization, branched questions, seeking consequences based on an action, verifying questions, factual questions, opinion-seeking questions	Chew et al., 2019
Whether goal of a question is clear or not, is it seeking important concept, is it looking for consequence of an action, is it verifying a context, is it a fact-seeking question or subjective question, is it seeking alternative of solutions, is it clarifying all facts, is it capable of targeting a context from multiple perspectives	Paul and Elder, 2019
Open-ended, subjectivity, capable of accepting multiple answers, ambiguity	Wan and McAuley, 2016
Seeking fluency, seeking flexibility, seeking originality, and scientific knowledge	Demir and Sahin, 2014

2.2.2 Mixed-method analysis

Mixed-method analysis comprises both qualitative and quantitative analysis. First-hand data collected from semi-structured interviews were in the form of voice recordings. The qualitative analysis was conducted in multiple stages. In stage one, voice recordings are transcribed. In stage two, the transcriptions were converted into open codes based on their semantics acquired from a sentence or paragraph (Charmaz, 2006). In stage three, open codes were clustered or grouped based on their semantic similarity (Sbaraini et al., 2011; Wiesche et al., 2017). In stage four, each cluster was labelled to get concise names. These names were considered as features for further processing. A representation of coding stages is illustrated in Table 2.2. Further, quantitative analysis, specifically frequency analysis was conducted to identify the repetition of codes in a cluster to identify their significance. The step-by-step approach of mixed-method analysis is depicted in Figure 2.3.

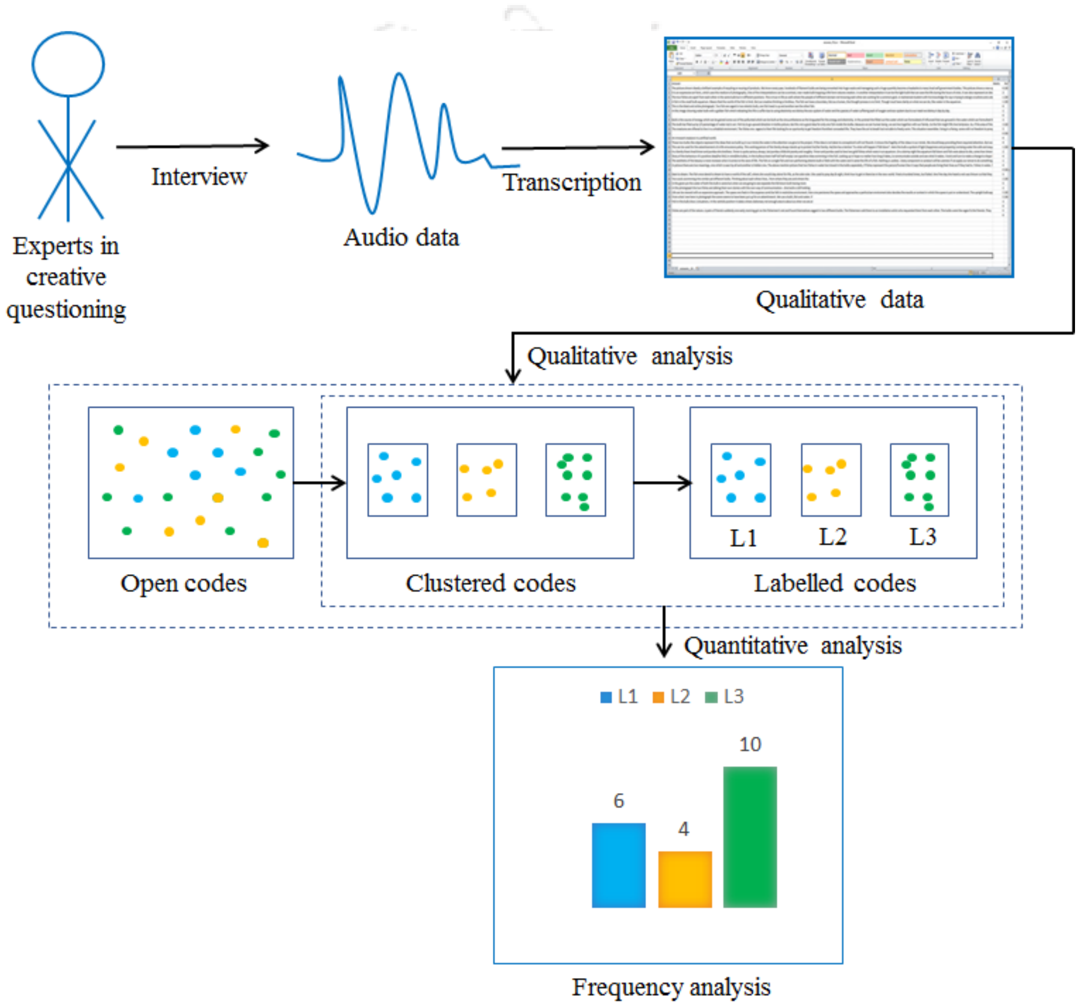


Figure 2.3: Mixed-method analysis

Table 2.2: Representative code formulation

Raw data	Open code	Cluster of code	Labelled code
Q. What are qualities/ ingredients a question must have to be classified as creative?			
[Ah…Like I mentioned on creative question one must kind of open-ended i.e. answer is open-ended. There will be no fixed answer to it.][i] [Ok; second, it can be interpreted in a multiple way.] [ii] [And also, sometimes it has some hidden aspect to find out.] [iii] [Sometimes through visualization.] [iv] [Sometime some context is given and how do they react; so how do they imagine and imagine themselves also. Its kind of a role-play. How can they imagine themselves in a situation. And also, how do they react to a situation.] [v]	[Open-ended][i] [Multiple interpretation][ii] [Hidden aspect][iii] [Visualization][iv] [Imaginative][v]	[Open-ended, Open-ended question] [i,viii]->[a]	Open-ended [a]
Q. What are qualities/ ingredients a question must have to be classified as creative?			
[Ok, one of the thing that I can think about at the top of my mind is that creative question may not have a specific answer, right. So, that means each and every person would respond to a creative question technically would think on their own way and we can always see who stands out the most. So might notice or gone through these design aptitude test or CEED or NID has their own exams, right.] [vi] [One of the things you might have noticed in example like in the CEED creative questions is that few criteria are	[Subjective response][vi]	[Multiple interpretation, Hidden aspect, Visualization, Imaginative, Subjective response, Subjective questions] [ii,iii,iv,v,vi,ix]->[b]	Subjective [b]

given, right. One is uniqueness, etc; so if you look at the problem solving questions you know that; I mean your end goal is solve the problem and your evaluation is based on whether you try to figure out some kind of solution to the problem.] [vii]	[Seeking novelty][vii]	[Seeking novelty] [vii]->[c]	Novelty [c]
[Creativity questions at least associated with these exams; it is seen as something open] [viii] [and subjective. There are no definite answers but some of the expectations of the answers are based on Google response; so the questions are made in such a way that are are some model answers, right. So this is how it could be. It can have a generic model answer but it will never have a model answer. And it might be surprising for the people who is evaluating. Ya somebody can think like that usually very open.] [ix]	[Open-ended][viii] and [subjective questions][ix]		

2.2.3 Feature extraction

There were *18* features found after the mixed-method analysis such as ‘open-ended questions’, ‘subjective questions, ‘application-oriented questions’, ‘intent understandability of questions’, ‘communicative questions’, ‘factual questions’, ‘procedure-seeking questions’, ‘seeking uncommon answers’, ‘other interested in question’, ‘verify creative questions’, ‘opinion-seeking questions’, ‘comparison of alternative of solution’, ‘seeking consequences of actions’, ‘seeking well explained solutions’, ‘question interpretation’, ‘narration of questions’, ‘qualities of examiner’, and ‘degree of creativity’. The feature ‘open-ended questions’ was considered as the basic criteria of a questioning to be registered as creative by experts. Therefore in this proposed model, a question was initially tested for open-ended or close-ended, then it would be further investigated for the presence of other features. The feature ‘qualities of examiner’ was not considered as a feature. It requires multiple psychological tests to assess qualities of examiners (Stronge, 2018). ‘Degree of creativity’ was also not considered though some of the experts mentioned in their interview as they were mostly unsure of the fact of measurement of degree of creativity, and it is totally subjective.

A question that was categorized as open-ended, were further tested for the presence of features such as ‘subjective questions, ‘application-oriented questions’, ‘intent understandable of questions’, ‘communicative questions’, ‘factual questions’, ‘procedure-seeking questions’, ‘seeking uncommon answers’, ‘other interested in question’, ‘verify creative questions’, ‘opinion-seeking questions’, ‘comparison of alternative of solution’, ‘seeking consequences of actions’, ‘seeking well-explained solutions’, ‘question interpretation’, and ‘narration of questions’. Kaggle’s Google quest (*Google QUEST Q&A Labeling | Kaggle*, 2020) data has *31* features from which ‘question identification’ is for reference to a specific question, and nine features associated with solutions were excluded. The remaining *21* features in Google’s data and *14* features found from mixed-method analysis were similar except ‘subjective questions’. Therefore additionally, ‘subjective questions’ to test subjectivity in a question and ‘question_polarity’ were included to perceive sentiments illustrated in questions.

Few additional features were included in this study such as ‘question_definition’, ‘question_entity’, ‘question_spelling’, ‘question_expect_short_answer’, ‘question_interest_self’, and ‘question_choice_type’. It was essential to identify whether question was seeking definition or not as questions associated with this attribute focus on recall of knowledge and not creativity. ‘question_entity’ verifies whether a question was associated with a particular object or product or not. Checking spellings was a basic criterion to verify questions for examination. ‘question_expect_short_answer’ was included for verifying length of a solution. ‘question_interest_self’ verified whether a question framed by an examiner found them interesting himself/herself. ‘question_choice_type’ was essential for verifying whether a question was subjective or objective.

2.2.4 Data preparation

Questions from multiple sources were collected to train and test the network for identifying creative questions. Firstly, data was collected from Amazon Question repository, which provided a significant resource of open-ended and close-ended questions (McAuley, 2018; McAuley & Yang, 2016; M. Wan & McAuley, 2016). Data was scraped from the website of the Common Entrance Exam from Design (CEED) (Bombay, 2021a). The questions of CEED were manually labelled as open-ended or close-ended. Google crowdsource team also possess a repository of creative questions possessing subjective features. Data was collected from Kaggle's Google quest (*Google QUEST Q&A Labeling | Kaggle*, 2020) and manually labelled as open-ended or close-ended questions. The questions from Amazon Question repository, IIT Bombay CEED website, and Kaggle's Google Quest contains both close-ended or yes/no and open-ended questions associated with constraints. Questions from all the repositories attempts in triggering creativity and does not focuses on recalling or learning abilities of individuals. The break up information of each question set are as follows: i) Amazon Question repository- *6079* records, ii) IIT Bombay CEED website repositories- *1818* records, and iii) Kaggle's Google Quest- *4261* records A total of *12,158* data was collected from all the sources.

2.2.5 Identifying creative questions

Qualitative analysis revealed that the open-ended nature of a question was highly essential and considered the base criteria to declare a question as creative. It indicates that if a question is open-ended, then only it is computationally effective to test the presence of other creativity features. Therefore, a pipeline mechanism was designed that filters questions that are open-ended to be declared as creative. Questions that are predicted as open-ended are further processed to identify the other features of creative inquiry by pre-trained transformers. Data integrated from multiple sources served as input to the transformers for identifying creative questions.

LSTM was implemented initially to recognize the open-ended nature of creative questions. LSTM was considered analogous to human perception mechanism as it considers previous words to understand any document. The network comprised of loops that permits information to be preserved and carried forward (Skansi, 2018; Y. Wang et al., 2016). The input to the model was the questions acquired after the data collection stage. Each input was converted into 300-dimensional vectors (Yilmaz & Toklu, 2020) followed by layers of Recurrent Neural Networks (RNNs) feedback connections that represent the internal state depending on weight of the input. Finally, the SoftMax layer provided the probabilities of each input that belong to a particular class as illustrated in Figure 2.4. The train and test dataset were divided into *80%* and *20%*, respectively. The validation accuracy is *81.006%*, with ten epochs.

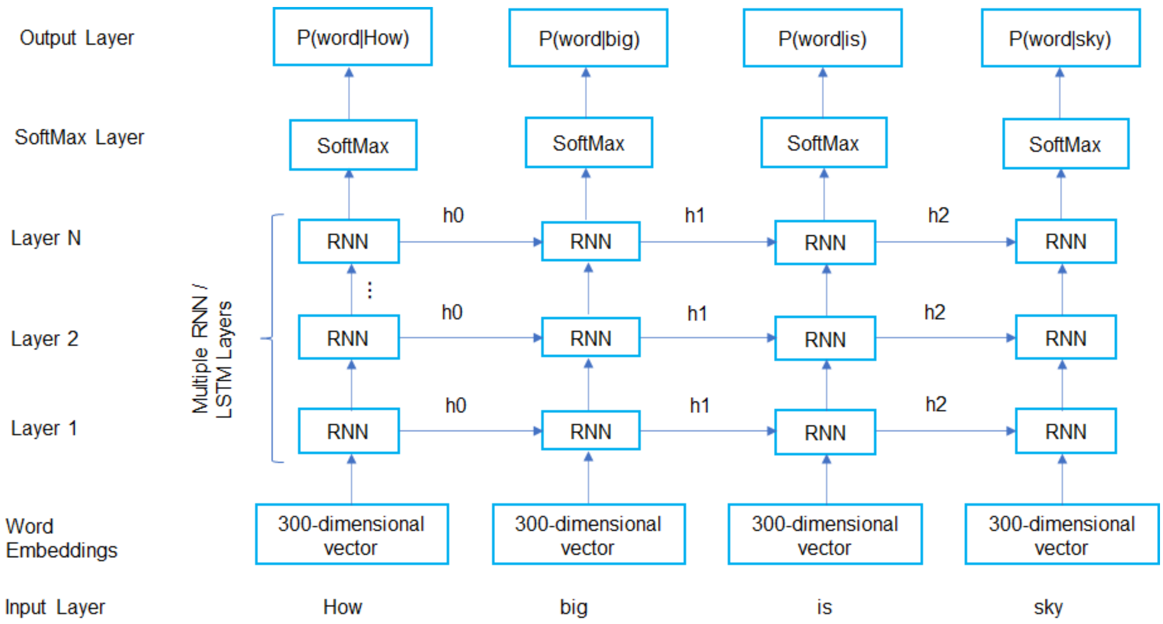


Figure 2.4: LSTM architecture

To further optimize the accuracy, BERT model was implemented for classifying the open-ended nature of questions (Clark et al., 2019). This model had multiple Transformer layers, feed-forward networks, and attention heads. The initial [CLS] token in input represented classification and [SEP] indicated demarcation of sentences. The input was carried to layers of encoders through a feed-forward neural net, applying self-attention to each layer. In the output layer, the first position was considered to receive classification, as shown in Figure 2.5. In this study, binary classification was conducted, which categorized open-ended questions from yes/no questions, i.e., questions that can be answered either by yes or no. The batch size used for this model was *6*, training epochs were *5* with a learning rate of *2e-05*. The loss of this model was *0.05*, with a validation accuracy of *99%*.

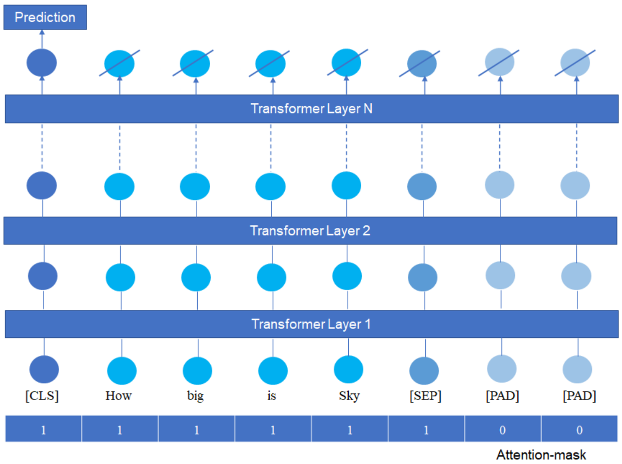


Figure 2.5: BERT architecture

The filtered output of BERT, which were open-ended questions in nature, were further processed for identifying other features of creative questions. Features of creative questions were investigated by mixed-method approach. Correlation between the features were found out to identify any overlapping characteristics of features. The heatmap below represents that there is a high correlation to the variable itself; otherwise, there is a weak correlation among variables, as shown in Figure 2.6. The colour codes range from black, corresponding strong correlation to white corresponding very weak correlation.

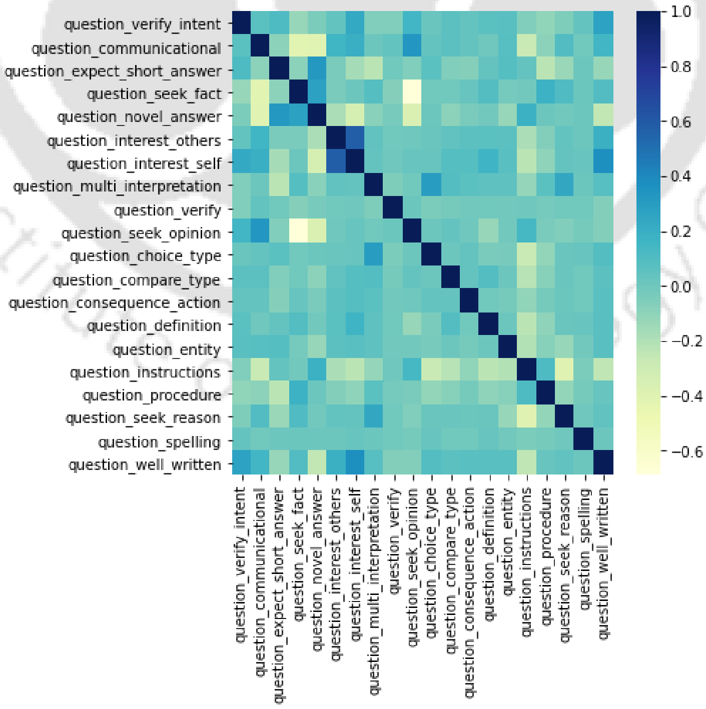


Figure 2.6: Correlation among features

Features that were common in Google quest data and outcome of mixed-method analysis were included in this study. Few of the additional features included from Google quest data were ‘question_definition’, ‘question_entity’, ‘question_spelling’, ‘question_expect_short _answer’, ‘question_interest_self’, and ‘question_choice_type’. Interviewees shared a common opinion of the features as found in Google quest. Further, BERT pre-trained model was implemented where features of it were mapped with the features extracted by qualitative analysis. This model identified the following features-‘question_verify_intent’, ’question_ communicational’, ’question_expect_short_answer’,’question_seek_fact’,’question_novel_ answer’,’question_interest_others’,’question_interest_self’,’question_multi_interpretation’,’q uestion_verify’,’question_seek_opinion’,’question_choice_type’,’question_compare_type’,’q uestion_consequence_action’,’question_definition’,’question_entity’,’question_instructions’,’ question_procedure’, ’question_seek_reason’, ’question_spelling’, ’question_well_written’ ‘question_ subjectivity’, and ‘question_polarity’ and classify it as creative. This model attempted to solve the research question of identifying a creative question from a bunch of other questions. The interpretation of the features is shown in Table 2.3.

2.3 Results and discussion

2.3.1 Mixed-method research results

The mixed-method research technique, specifically by multiple stages of coding and frequency analysis, generated features for utilization by DL-model. Expert’s opinion about creative questions that instigate creative responses from students was acquired by semi-structured interview. Qualitative data were acquired from *10 (N=10)* experts who possess experience in formulating creative questions for mass examinations. During interview, subjects were interrogated to identify features of creative questions. Data captured from them were transcribed. The transcriptions were then converted into open codes and further categorized based on their semantic similarity. Each semantically similar cluster was assigned a unique and relevant label that matches its intended purpose.

Table 2.3: Interpretation of features

Features	Interpretation
'question_verify_intent'	Whether rational of a question needs to be verified
'question_communicational'	Whether a question is conversational
'question_expect_short_answer'	Whether a question expects short solution
'question_seek_fact'	Whether a question expects factual solution
'question_novel_answer'	Whether a question expects novelty in solution
'question_interest_others'	Whether a question feels interesting to others
'question_interest_self'	Whether a questions feels interesting to self
'question_multi_interpretation'	Whether a question propagates multiple interpretation across all respondents
'question_verify'	Whether a question can really be reported as a creative question
'question_seek_opinion'	Whether a question expects opinion from respondents
'question_choice_type'	Whether a question is objective or subjective
'question_compare_type'	Whether a question expects alternative of solutions
'question_consequence_action'	Whether a question expects consequences of any particular action(s)
'question_definition'	Whether a question expects recall of information
'question_entity'	Whether a question is associated with any particular object or product
'question_instructions'	Whether a question expects instruction
'question_procedure'	Whether a question expects procedural solution
'question_seek_reason'	Whether a question expects well-explained solution
'question_spelling'	Whether a question seeks spell-check
'question_well_written'	Whether a question is narrated well
'question_subjectivity'	Whether a question is based on one's opinion, experience, and preference
'question_polarity'	Whether a question is negative, positive, or neutral in terms of interpretation

The frequency analysis depicted in Figure 2.7 illustrates the repetition of features by multiple experts. The x-axis represents cluster labels, and the y-axis shows the number of times the open codes were repeated against each cluster. The colour codes represent the number of subjects from whom the qualitative data was acquired. Firstly, variant terms of open-ended questions were repeated by ten subjects. The frequency of the specification of variants of open-ended questions by the subjects are as follows: *13* times: subject1, *3* times: subject2, *7* times: subject3, *9* times: subject4, *8* times: subject5, *2* times: subject6, *2* times: subject7, *5* times: subject8, *3* times: subject9, and *2* times: subject10, respectively. Similarly, this pattern is followed for other features of creative questions that are illustrated in the graph. Since the frequency is generated from qualitative data, so inclusion of features to the model is dependent not only on frequency but also on affirmative response to add the particular feature. For example, 'degree of creativity' has got a very low frequency, and most of the experts didn't knew or were unsure of exactly how to define it in the context of a question. So this feature was not included for two reasons, viz., low frequency and ambiguous response received from experts during interview. Similarly, 'communicative question', which represents any form of communication in a question showed low frequency. However, it was observed that question papers associated with nationalized examinations were conversational (Baron, 2018; Hargie, 2006; Holmes et al., 2017). Therefore, this feature was considered as input to the model. The frequency of 'qualities of examiner' was relatively moderate; however, this study was specifically focused on attributes of question, the characteristics of examiners were not studied. It would require multiple psychological tests to be included and further relate to creative questions. All other features have relatively moderate to higher frequency and contextually eligible to be included in the study.

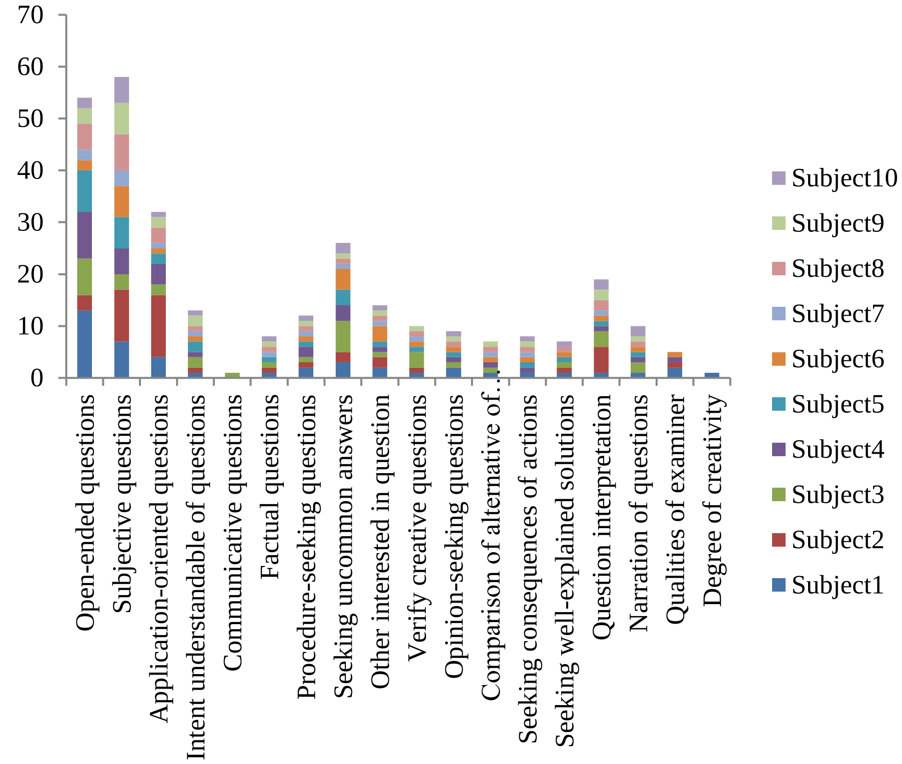


Figure 2.7: Frequency analysis of features evolved from coding

The major focus of this study was to identify creative questions from a bunch of non-creative questions. Creative question attempts to trigger creativity in students. The degree of creativity invoked in a response depends on the level of skill and creative thought process imbibed in students. Therefore, this study does not focus on student's response in order to identify creative question. It highlights factors acquired from experts using which they formulate creative questions to trigger students' creativity in mass examinations. Further, the model designed for this study would assist examination paper setters to verify whether the questions formulated by them are creative or not. Though examination paper setters in Design education possess experience in framing creative questions, but in some cases, it might lead to mistakes while formulating on a large scale. Self-bias is a major factor that might occur in formulating creative questions. To overcome these, this study would enable human-machine engagement to optimize decision-making of examination paper setters. This might lead to an iterative process, resulting in formulating creative questions as illustrated in Figure 2.8. Humans framing questions would serve as input to the model, which assists in predicting creative questions. If a question is predicted as creative, then it requires no further processing. A question predicted as not creative receives immediate decision of reformulating a question and thereby reducing human dilemma. This process iterates until a question is predicted as creative.

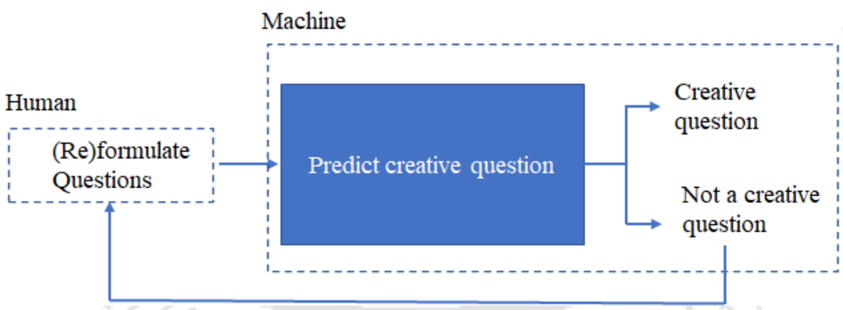


Figure 2.8: Human-machine engagement to formulate creative questions

There was another aspect of studying creative questions where one might be inquisitive about the degree of creativity. Literature highlights a study of degree of creativity of products (Sarkar & Chakrabarti, 2011). While acquiring qualitative data by semi-structured interviews, a question associated with degree of creativity of the questions in examination was asked. However, there was no appropriate response from experts in this field. The typical responses were *"Ahhhhh………I am not sure about the level of creativity."*, *"We do not assess the degree of creativity in a question. Sometimes a question can be relatively critical but again criticality* *is not creativity.”*, *“Sometimes I provide them with more constraints, but I am not sure about degree of creativity.”, “I don’t know the simplest of the things that we put it up and the problems of the questions and what for like a we do at the beginning of degree can be expressed like you feel it has very little creativity or you feel like if your personal goal in life is very high creative. So how can we afford that? I think it is very difficult to assess that. How much creative it is very subjective.”*, *“Aaaa….. I think its difficult to judge the level of creativity because I still feel it is difficult to put creativity into category; that’s my personal opinion.”* This outcome might have two interpretations- Firstly, paper setters in Design education do not formulate questions from the perspective of the degree of creativity, or secondly, it is a grey area that requires intensive research to investigate this problem.

2.3.2 Machine learning research outcomes

Theory from qualitative research evidence that the basic characteristic of a question to be creative was the open-ended nature of the question. The typical expert responses were *“Open-ended is the base criteria”*, *“Open-ended is the most significant parameter”*, *“Initial clarification to be a creative question”*, and *“Open-ended is considered as a common criteria”*. Open-ended questions do not limit the respondents with a limited set of options, nor do they bias in any form (Reja et al., 2003). Literature highlighted open-ended questions being positively correlated with creativity (Arsyad et al., 2017; Ghosh, 1993). It served as the basic ingredient in order to classify a question as creative.

A pipeline was designed that initially categorizes open-ended questions from close-ended types of questions. Questions that were categorized as open-ended were further investigated to find the probability of the presence of other creative features of questions. Initially, LSTM was used to classify open-ended questions (Zhang et al., 2019), which is often used in this type of problem. Given a question to the model, it classified it as open-ended question or yes/no question. Yes/no refers to the type of question which can be answered by either yes or no option as shown in Figure 2.9. However, literature highlights tackling similar types of problems using BERT algorithm (Sung et al., 2021). This classification problem was also experimented using BERT algorithm, which outperformed in this case, which further optimizes the trust in the model. The performance chart of both the models is shown in Table 2.4.

	question	type
0	can i take this off and clean it? dash lights ...	yes/no
1	I just started a diploma in software developme...	open-ended
2	do these fit a ATV Grizzly 550 ?	yes/no
3	Let $\mathcal{A}$ be a unital, commutative C*-...	yes/no
4	I've just added the following .htaccess rules ...	yes/no
...	...	...
12153	would it fit a nissan navara 2010 model	yes/no
12154	I once saw a sentence: \n\n\n I will go to a ...	open-ended
12155	With multiple monitors, I have so far been dra...	yes/no
12156	can you use it on shower glass?	yes/no
12157	is there a place to screw in the o2 sensor	yes/no

12158 rows × 2 columns

Figure 2.9: Classification of open-ended and yes/no questions

Table 2.4: Performance analysis of LSTM and BERT for classifying open-ended questions

	LSTM	**BERT**
Data size	12158	12158
Split size	70:30	70:30
Number of epochs	10	5
Learning rate	0.001	0.002
Loss	0.0850	0.0032
Validation accuracy	81.006%	99.999%

There are multiple features of creative questions that were investigated from qualitative analysis. The probability of presence of these features in questions was found out using pre-trained BERT model. In this experimentation, initially, there were *12,158* questions, and among them, *6,017* was classified as open-ended questions. Therefore, prediction of other creative features in questions was conducted on *6,017* data. This could be a disadvantage of this study where the dataset is considerably small, and the performance might optimize with an increase of dataset. The outcome of the prediction of BERT model is shown in Table 2.5, highlighting a negligible difference between the actual and predicted values. Due to unexplainable nature of artificial intelligence, the outcomes of the model was compared with gold standard data.

All features associated with creative questions were identified using theoretical coding, and similar features were present in Google quest data also. However, predictions of two features, ‘question_polarity’ and ‘question_subjectivity’ were not based on empirical data, and they were dependent on functions of sentiment analysis. The correlation between ‘question_polarity’ and ‘question_subjectivity’ with all other creativity features is illustrated in Figure 2.10. The heatmap demonstrates a strong correlation with black colour code and weak correlation with white colour code. Results show that there is mostly weak correlation between ‘question_polarity’ and ‘question_subjectivity’ with all other creativity features.

Table 2.5: Outcome of BERT model prediction

Features	**Predicted**	**Actual**	**Actual-Predicted**
‘question_verify_intent’	0.967752	1.000000	0.032248
‘question_communicational’	0.018866	0.000000	-0.018866
‘question_expect_short_answer’	0.528076	0.333333	-0.194743
‘question_seek_fact’	0.879446	1.000000	0.120554
‘question_novel_answer’	0.766153	1.000000	0.233847
‘question_interest_others’	0.711269	0.888889	0.177620
‘question_interest_self’	0.687939	0.888889	0.200950
‘question_multi_interpretation’	0.591925	1.000000	0.408075
‘question_verify’	0.001403	0.000000	-0.001403
‘question_seek_opinion’	0.452625	0.333333	-0.119292
‘question_choice_type’	0.426464	0.333333	-0.093131
‘question_compare_type’	0.008929	0.000000	-0.008929
‘question_consequence_action’	0.007309	0.000000	-0.007309
‘question_definition’	0.003906	0.000000	-0.003906
‘question_entity’	0.014018	0.000000	-0.014018
‘question_instructions’	0.831795	1.000000	0.168205
‘question_procedure’	0.219256	0.666667	0.447410
‘question_seek_reason’	0.153905	0.000000	-0.153905
‘question_spelling’	0.000685	0.000000	-0.000685
‘question_well_written’	0.956031	1.000000	0.043969
‘question_polarity’	-0.092857	NA	NA
‘question_subjectivity’	0.685714	NA	NA
Creative			

#NA indicates Not applicable

Generally, humans identify creativity in a question in current scenario (Mehta & Ingole, 2015), but algorithm computable dimensions need to be identified. Acquiring scores associated with each feature is essential to train and predict the outcome. In this context, scores associated with each feature were collected from Kaggle’s Google quest source. Crowdsourcing supported in acquiring continuous values in the range [0,1] for each feature. The predictions also were in the range of [0,1]. Each record had a question identifier that was positive integers. The remaining other features considered for this study contained scores in decimal form usually up to five decimal points. The nomenclature of the features was of the form: ‘question’_’characteristic_of_a_question’, which indicated that each feature contains an initial word as ‘question’ followed by an underscore, and further comprise the characteristic of the question. The sentiments of the questions were determined using TextBlob module in python to evaluate subjectivity and polarity. Psychologists and linguistic experts have labelled sentiments of words that were embedded in TextBlob module, which supported evaluating sentiments.

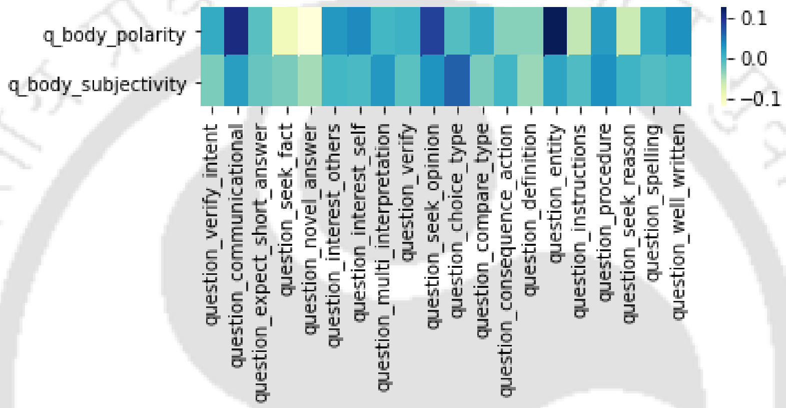


Figure 2.10: Correlation between polarity and subjectivity with other creativity features

The BERT model for predicting creativity features of questions was pre-trained in Google quest data comprising of questions associated with product design, art, culture, and technology. The data size was *6,017*, which was relatively less, but BERT tends to outperform without domain-specific and very less data (Garcia-Pablos et al., 2020). However, with an increase in size of the data, the performance of the model might be more descent. The model was executed on five epochs, and each epoch took around *5* minutes and *26* seconds. The performance of the model became better after the third epoch. The performance of the model is shown in Table 2.6.

Table 2.6: Performance of BERT model for predicting creativity

	BERT
Data size	6017
Split size	80:20
Number of epochs	5
Learning rate	0.00005
Loss	0.046
Accuracy	91.189%

2.3.3 Optimizing decision-making of examination paper setters

Creativity is a subset of cognitive ability and formulating a creative question is an extra-ordinary task as it requires special abilities to visualize things from a novel perspective (Ward & Kolomyts, 2010). Complex cognitive parameters such as fluency, flexibility, and ambiguity overlap with creativity making it really difficult to assess it (Almeida et al., 2008). Therefore, framing a creative question can be considered a highly complicated process. First-hand data was acquired during semi-structured interview session by inquiring the process of framing creative questions. Transcripts generated from interview highlighted that there was no algorithm or step-by-step procedure of framing creative questions. Some of the responses of experts in this domain were: *"A…..It is very difficult. Consciously what I follow is what I taught and I want to check whether they learned it or not. I also ask them to apply in a context. I do not ask direct question like what are the design principles? Instead I will give some situation where they have to find the principle; basically the application, ok. So I will show them the application and from there I will ask what design principles they can figure out."*, *"A…..It is a difficult question. I feel that there is no specific process that I can tell you."*, *"Question setter must be himself a very good educator and a creative person. That experience can only help in setting. They can refer books where there are previous examples for setting up the creative question."*

First-hand data highlights that there was no standardized procedure for formulating creative questions. So, creativity in question that instigates creative responses among students was totally dependent on expertise and experience of question paper setter. Examiners were always inquisitive whether questions framed by them were really creative, which could capture creativity in students. It is obvious that domain-specific experts formulate questions like this, but biases and human errors might occur due to tasks conducted on a large-scale. This system attempts to assist examination paper setters by providing a prompt decision of whether a question is creative or not. However, human intervention is essential to again reformulate the question. This enriches human-machine engagement, which optimizes experts' decision-making to formulate creative questions.

2.3.4 Performance evaluation with baselines

Table 2.7: Comparative study among models

Features	Actual	DTR	Actual-DTR	MOR	Actual-MOR	BERT	Actual-BERT
'question_verify_intent'	1	0.666667	0.333333	0.931785	0.068215	0.996752	0.003248
'question_communicational'	0.888889	1	-0.111111	0.698996	0.189893	0.870023	0.018866
'question_expect_short_answer'	0	0	0	0	0	0	0
'question_seek_fact'	1	1	0	1.072295	-0.072295	0.979632	0.020368
'question_novel_answer'	0.666667	1	-0.333333	0.699008	-0.032341	0.766153	-0.099486
'question_interest_others'	1	1	0	1.070356	-0.070356	1.000829	-0.000829
'question_interest_self'	0.444444	0.555556	-0.111112	0.596355	-0.151911	0.453769	-0.009325
'question_multi_interpretation'	0.555556	0.555556	0	0.595719	-0.040163	0.446782	0.108774
'question_verify'	0	1	-1	1.068452	-1.068452	0	0
'question_seek_opinion'	0	0	0	0	0	0	0
'question_choice_type'	0.333333	0	0.333333	0.596355	-0.263022	0.333012	0.000321
'question_compare_type'	1	0.333333	0.666667	0.359356	0.640644	1	0
'question_consequence_action'	0	0.333333	-0.333333	0	0	0	0
'question_definition'	0	0	0	0	0	0	0
'question_entity'	0	0	0	0.00386	-0.00386	0	0
'question_instructions'	0	0	0	0	0	0.000322	-0.000322
'question_procedure'	0	0.666667	-0.666667	0.71099	-0.71099	0	0
'question_seek_reason'	0	0.333333	-0.333333	0.355515	-0.355515	0	0
'question_spelling'	0.333333	0	0.333333	0.34933	-0.015997	0.333012	0.000321
'question_well_written'	0	0	0	0	0	0	0
Mean Absolute Error, MAE			*0.228*		*0.184*		*0.013*

The methodology presented in this study was tested with multiple baselines algorithms that intend to predict multiple dependent variables based on independent variables. The best-performing model was considered as the benchmark for this study. Initially, regression models were considered to predict multiple variables, but many of these models predicted a single value for each sample (Rao et al., 2017; Schmidt & Finan, 2018). Next, regression that allowed predicting more than two dependent variables based on multiple independent variables of both numerical and categorical types were identified from literature, and considered for predicting features of creative questions. Three models were considered viz, Decision Tree Regressor (DTR) (Wang et al., 2018), MultiOutput Regressor (MOR) using sklearn (Montiel et al., 2018), and BERT (Li et al., 2021), were experimented with labelled dataset of *6,017* instances, with *80:20* split. Differences among the predicted outcomes by the models and actual outcomes marked by humans rated in the range of [0,1] were evaluated as illustrated in Table 2.7.

The features obtained from qualitative analysis were provided to Design experts within the age group of *32-55* years and possessing *3-21* years of experience in formulating creative questions for Design entrance examinations. They rated each feature within the range [0, 1]. The positive decimal values closer to zero indicates relatively less composition of a particular feature in a question, whereas values closer to one indicate the relatively more composition of a particular feature in a given question. This decision was subjective and dependent on expertise of a pedagogue. The data acquired from humans was considered a gold standard.

A comparative study was conducted to compare the model with baselines to find the best model. Mean Absolute Error (MAE) (Wang & Lu, 2018) measured for the three models viz., DTR, MOR, and BERT were *0.228*, *0.184*, and *0.013*, respectively, which was evaluated using Equation 2.1, where n is the total number of samples, actual_rating is the ratings provided by humans, and model_rating is the predicted value of a model. A relatively low MAE of BERT indicates that its predicted value is closer to human judgement. A graph illustrated in Figure 2.11 indicates a relatively higher difference between DTR model's prediction and observation or actual rating as compared to MOR. BERT showed the least difference in its outcomes when compared with human ratings, and therefore considered as the best model to be implemented for the proposed methodology.

$$Mean\ Absolute\ Error, MAE = 1/\mathrm{n} \sum_{k=1}^{n} (actual\_rating_k - model\_rating_k) \qquad (2.1)$$

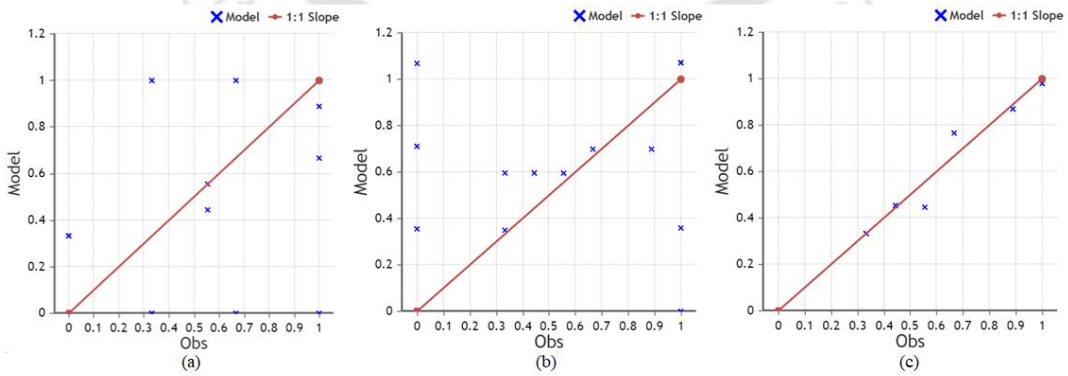


Figure 2.11: Error graph of Model's prediction versus actual human rating: (a) DTR, (b) MOR, and (c) BERT

Results highlight that BERT model outperformed in identifying features of creative questions, and its outcomes are relatively closer to human judgement when compared with other baseline algorithms, which increases trust in the model. The outcome of these models support in identifying whether questions framed by them really possess the features of a creative question or not. Since there exists a negligible error between BERT and human rating, therefore these results would influence pedagogues in decision-making of whether a question requires reformulation.

2.3.5 Validation of the proposed model

The outcome requires validation in order to be accepted by pedagogues. Therefore, data is collected from pedagogues, where they classified a question as creative or non-creative, which is considered as a golden standard. To evaluate the performance of the model, inter-rater reliability between subjective judges (gold standard data) and outcome of model was measured. In this study, results of inter-rater reliability showed the agreement between examiners and the outcome of proposed model. The evaluation procedure involved *21* subjective examiners *(N=21)* from different design institutes within the age group of *32-55* years *(M=37.76; S.D.=5.03)*, years of experience within *3-21* years *(M=9.04; S.D.=4.24)*, and *66.7%* were male, *28.6%* female and *4.8%* preferred not to say. Ten questions were provided to identify whether they were creative or non-creative questions, and thereby, the data collected from experts was of nominal dichotomous type (Zapf et al., 2016). The computation of Krippendorff's alpha reliability (Krippendorff, 2011) estimate between subjective examiners and the model at a nominal level of measurement was relatively high *(α = 0.96),* as shown in Table 2.8, with a total of *22* observers *(21 subjects and proposed model)*.

Table 2.8: Krippendorff's alpha reliability estimate between examiners and model

	Alpha	LL95%CI	UL95%CI	Units	Observrs	Pairs
Nominal	.9638	.9536	.9741	10.0000	22.0000	2310.0000

The relatively higher rate of alpha increased trust in the proposed model. This model intended to provide service in education sector where it can assist examiners in framing creative questions for instigating creative responses from students. Examiners are always inquisitive to know whether questions framed by them are really creative. It is obvious that examiners possessing experience are capable of formulating as well as identifying creative questions. Often, framing questions on a large-scale or self-bias are the factors that cause ambiguity in identifying whether a question is creative or not. Peer review, in some cases, seems beneficial, but it has its own demerits. It often gets biased by individual-experience, cultural, and psychological factors such as cognition, stress, and emotion of reviewers. It is essential for an examiner to eliminate self-bias and identify the nature of a question to capture creativity of students.

Majority of literature highlighted automating question formulation process and identifying the success of a question in capturing responses. But there is a dearth of literature that focuses on identifying creative questions that triggers creative responses. One major finding of this research is that it has the focal point of investigation in identifying creative questions capable of instigating creative responses from students to capture creative aptitude. Attempting to capture creative aptitude is different from other solutions that mainly comprise of definition or defining any processes. Creative questions expect open-ended responses that requires subjective assessment. There are multiple approaches to reach a solution to creative question in contrast to a strict set of a path in other types of questions.

The second contribution in this study is identifying the descriptors of creative questions capable of capturing creativity of students by a systematic human-centred design approach. Multiple studies in literature highlight various features of questions but without a prior method of acquisition of those features. However, this study focused on a systematic approach using a qualitative study to acquire features of creative questions. The qualitative data was acquired by a semi-structured interview. Further, the data was analyzed by mixed-method approach to investigate features that are essential to capture creative aptitude in students.

Another finding is that this research attempts in digitizing the process of identifying creative questions by a proposed model. State-of-the-art literature suggests automation of several facets of question-answering platforms covering the recognition of the success of domain-specific question, question generation, answer evaluation, etc. But there is less focus on digitizing the process of identifying creativity in questions capable of capturing creative aptitude. The digitizing processes involved in this study are acquisition of features by qualitative study, designing architecture of a proposed model, implementation of the entire architecture using DL models, and finally evaluating the performance of the proposed model by comparing the outcomes with human examiners.

The limitation of the current methodology is that it is extensively focussed on identifying creative questions specifically in Design education. It might require a different approach for identifying creative questions for any other domain. This methodology comprises an interviewing technique associated with capturing features of creative questions. Though this interviewing technique provide in-depth information of a phenomenon but triangulating the data acquisition process might provide more information. The disadvantage of an interviewing technique is that it is time consuming. The processing of interview data is also quite tedious as it requires transcription and further semantically analysing it. The sample size seems to be descent for identifying the features of creative question, but increasing the sample size might allow one to study the patterns of large volume of coded data. Further, there is a limited volume of data gathered from CEED question papers and Kaggle's Google quest data. The DL models might provide better performance by increasing the dataset. This methodology is restricted to provide question category. Pedagogues may decide to reformulate a question if required, based on the outcomes of this model. Reformulation needs to be done based on their own knowledge and persuasion.

This study is domain-specific and is associated with Design education. The features discussed in this research are not adaptable nor tested for any other domains. The model tested with questions of Design might provide unpredictable results for other domains. Another limitation of this study is the relatively small dataset. Increasing the dataset might further optimize the results and improve the accuracy of the model. The present research did not excavate the characteristics of degree of creativity in questions. There is a scope for future research where the fuzzy nature of creativity can be deeply mined, and more insights could be provided by researchers (Chaudhuri et al., 2021a).

2.4 Conclusion

This research intends to investigate creativity in questions, specifically in examination of Design education. Formulating creative questions is a highly complex task involving extensive cognitive dimensions. Therefore, questions for any examination are framed by domain experts. However, while framing questions, bias of an individual might lead to self-doubt. This study attempted in identifying feature of questions that has the potential of capturing creativity of students. Further, this research is directed towards automating the process of identifying creative questions. This study would support in optimizing decision-making of question paper setters who would be able to make prompt decisions based on the outcome of the proposed model. Therefore, this research assists in human-machine engagement, which supports decision-making and provides a verification of questions that has the potential of instigating creative responses from students.

Chapter 3: Identifying parameters to assess novelty of descriptive creative responses and digitizing its evaluation process

Abstract

Evaluating creative aptitude in Design education is subjective and generally depends on expert's referential metrics. Novelty is an important factor of assessment in creative responses that Design pedagogues look out for. Presently, practitioners in this field perform subjective evaluation of answers of prospective students, but many a time, humans are prone to errors when associated with repeated tasks on a large scale. Therefore, this chapter attempts to find parameters of evaluating novelty in descriptive pattern of creative responses and further support educators by automating the process of evaluating novelty using a proposed computational design model. Mixed-methods research is conducted based on structured questionnaires and analysis to investigate features of subjective evaluation of novelty practiced in Design entrance examinations. The survey resulted in features that closely resemble human evaluation strategies for evaluating novelty from descriptive pattern of creative responses. Further, a computational model is proposed, designed, implemented, and validated that evaluates novelty on a large scale. Scores are generated for each feature by unsupervised learning techniques, eventually calculating novelty scores by a scoring function. This model is validated by comparing the outcomes of the proposed model with expert evaluation, where the difference between them is found to be negligible. This model suggests unambiguous scores to responses, which might help in a consistent selection of students aspiring admission to Design schools. This study attempts to reduce pain points of educational practitioners by offering a voluntary automated technique for subjective evaluation and optimizing trustworthiness of students in examination process.

Highlights

- *Human-centred design approach to identify parameters of evaluating novelty in descriptive* creative responses *of entrance examinations.*
- *Proposing a computational design model for evaluating novelty in descriptive* creative *responses.*
- *Implementing the model using various tools and algorithmic techniques.*
- *Validating the model by comparing its outcome with human-based evaluation.*

3.1 Introduction

Education is a complex process involving enormous pedagogic activities. Examination process is a significant part of education system that intends to evaluate performance and learning of students. The evaluation procedure of an examination mainly comprises two evaluation categories 1) objective evaluation and 2) subjective evaluation (Thomas et al., 2008). Objective evaluation is associated with questions that have specific answers. Presently, majority of examinations are objective in nature that are automated and require mapping of questions with their corresponding answers. On the other hand, subjective evaluation, is highly complex as it comprises descriptive pattern of creative responses that are evaluated based on examiners choice, knowledge, and persuasion. Design education expects novelty in a creative response (Demirkan & Afacan, 2012), which indicates newness in a response. Unlike any other descriptive responses that can be automatically evaluated based on specific keywords, creative responses are an exception (Jodhi et al., 2018). Descriptive pattern of creative responses in Design tends to be open-ended in nature and evaluating such category of answers might be challenging as it is necessary to maintain consistency of grades across all evaluators when a single answer script is provided to several experts. There might be a difference in referential metrics among experts. Hence, an evaluation mechanism is required that can generate consistent grades for all students.

Inconsistency in evaluation might be due to prolonged working hours, repeated tasks, and evaluation over a stipulated timeline. These factors cause stress in pedagogues and often raise self-doubt in them. Therefore, it is essential to understand how pedagogues evaluate novelty in mass examinations of Design education. The understanding process involved identifying parameters of evaluation. Based on these parameters, an automated or digitized architecture for evaluating novelty in descriptive pattern of creative responses is proposed to provide a consistent subjective evaluation. The investigation in this chapter attempts to address the research gaps highlighted and reported in the state-of-the-art literature review presented in subsection 1.5.2, subsequently corresponding research questions and objectives reported in section 1.9 and 1.12. The research questions and the objectives are stated below again for reference.

RQ3: *While assessing creative responses of students, what are the factors that Design educators consider for assessing novelty of the responses?*

RQ4: *Can the existing process of subjective manual evaluation of creative aptitude be automated?*

Objective 6: *To identify the factors of novelty in creative aptitude evaluation.*

Objective 7: *To design a digitized system for novelty assessment in creative aptitude.*

The first research question (RQ3) attempts to identify the parameters that pedagogues refer to evaluating novelty in descriptive creative response in entrance examinations of Design education. Novelty is a significant feature for evaluating creative aptitude in Design education (Runco & Plucker, 2001). A scientific study is planned that intends at conducting a survey for identifying the features of evaluating novelty in creative aptitude in Design fraternity. Further, the second research question (RQ4) highlights whether the present manual system of evaluation can be automated. The need for a digitized or an automated system is raised for the difficulties of subjective evaluation on a large scale in mass examinations where pedagogues are confronted with multiple challenges of evaluation such as stipulated time, repeated task, maintaining similar referential metrics of evaluation among experts, etc. To address this research question, a model is proposed based on the factors of evaluating novelty that attempts in automating the process of evaluation. This model is further validated by confirming a negligible difference between the evaluation conducted by the proposed model and human experts.

The outcomes of this chapter are the parameters of evaluating novelty in descriptive pattern of creative responses exhibiting creative aptitudes such as grammatical mistakes, misspellings, relevance between question and a response, narration or coherence, and uniqueness of responses were identified using a mixed-method approach. Further, a model is proposed based on the identified parameters to evaluate novelty. Finally, the outcome of model is validated by comparing the results with a golden standard data. The overall model is illustrated in Figure 3.1.

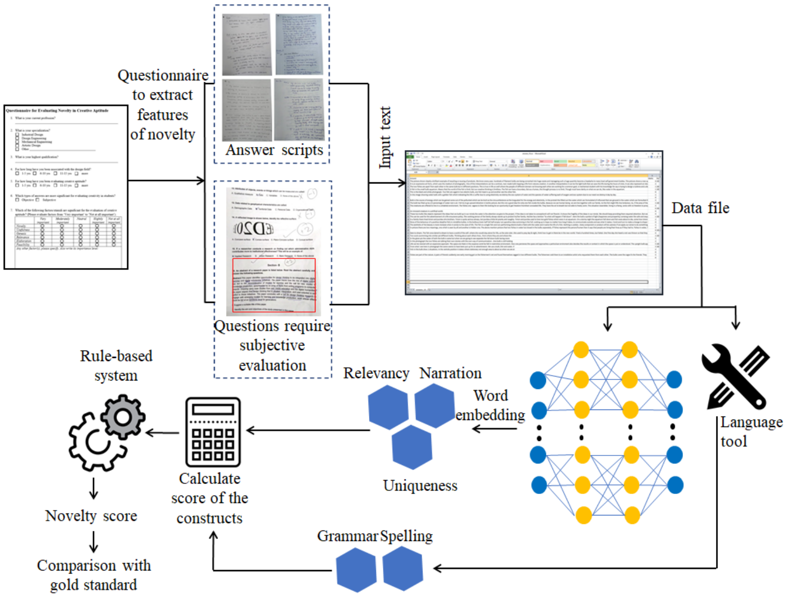


Figure 3.1: Overview of this model

3.2 Method

3.2.1 Questionnaire preparation and survey plan for identifying parameters

This experimental plan intends to evaluate novelty in creative aptitude exhibited in mass examinations of Design education. Novelty is a complex factor that constitutes several other subfactors. Literature highlights several subfactors involved in novelty, such as language processing, relevance between questions and their responses, narration link or coherence in responses, unique concept, etc. (Berbague et al., 2021; Camburn et al., 2020; Demirkan & Afacan, 2012; Schumann et al., 1996). A questionnaire was framed based on these features whose internal consistency was measured with Cronbach's alpha which was found to be *0.709* that belongs to an acceptable range (Sharma, 2016). The questionnaire used Likert-type scale ranging from 'very important'=*5*, 'slightly more important'=*4*, 'important'=*3*, 'slightly important'=*2*, and 'not at all important'=*1*. It possessed a provision of including additional features that Design practitioners considered significant apart from the existing features.

The questionnaire designed for finding the parameters to evaluate novelty in creative responses is illustrated in Appendix B. The questionnaires were written in the way (provided in appendix) to confirm whether the factors identified in literature review ascertain to the factors considered by Design pedagogues in practice. State of the art review was conducted to examine the findings from literature that contributed to the identification of factors that are referred for assessing products, solutions, ideas, etc. as illustrated in Table 3.1. A survey was conducted to collect user ratings using 5-point Likert-type scale (Demirkan & Afacan, 2012) to identify the factors preferred by experts in the evaluation process. The structure of the questionnaire is provided in Appendix C.

Table 3.1: Factors for evaluating creative aptitude (questionnaire given in Appendix B and D) that are associated with literature articles

Factors	State-of-the-art literature
Ideas of fluency	D'Souza, 2021; Benedek et al., 2016; Diedrich et al., 2015; Cheung et al., 2003; Kornish and Jones, 2021; Al-Zahrani, 2015; Mirabito and Goucher-Lambert, 2020; Dippo and Kudrowitz, 2013; Bayer-Hohenwarter, 2010; Almeida et al., 2008
Flexibility	D'Souza, 2021; Benedek et al., 2016; Diedrich et al., 2015; Cheung et al., 2003; Al-Zahrani, 2015; Dippo and Kudrowitz, 2013; Bayer-Hohenwarter, 2010; Chakrabarti, 2006; Almeida et al., 2008
Usefulness	Sarkar and Chakrabarti, 2011; Diedrich et al., 2015; Kornish and Jones, 2021; Takai et al., 2015; Bayer-Hohenwarter, 2010; Chakrabarti, 2006; McCarthy, 2018
Relevance	Camburn et al., 2020; D'Souza, 2021; Cheung et al., 2003
Uniqueness	Demirkan and Afacan, 2012; Diedrich et al., 2015; Cheung et al., 2003; Kornish and Jones, 2021; Al-Zahrani, 2015; Vargas Hernandez et al., 2012; Takai et al., 2015; Dippo and Kudrowitz, 2013; Bayer-Hohenwarter, 2010; McCarthy, 2018; Sarkar and Chakrabarti, 2011; Almeida et al., 2008
Clarity	Chaudhuri et al., 2020; Chakrabarti, 2006;
Choice of colours	Demirkan and Afacan, 2012; Chaudhuri et al., 2021b
Sketching ability	Takai et al., 2015; Garaigordobil, 2006; Demirkan and Afacan, 2012; Schumann et al., 1996; Almeida et al., 2008
Language processing	Benedek et al., 2016; Cheung et al., 2003
Narration/coherence	Pérez and Sharples, 2001; Wu, 2013; D'Souza, 2021; Demirkan and Afacan, 2012; Takai et al., 2015

The population size of this study is unknown, and there are fewer experts available in this domain. A confidence interval of *0.75* was considered in this study, which is commonly practiced in design, educational, and social researches with *0.05* marginal errors (Krejcie & Morgan, 1970). The sample size is calculated to be seventy-one *(N=71)* subjects. Subjects considered for this study are experts from reputed private and government Design schools, and Designers from industry are selected for this study. The average age of the subjects was between *39- 62* years *(M=47.92, S.D. =5.52)*, and *46%* were female *(N=33)*. The survey was conducted mostly on-site, i.e., in department of numerous Design schools and industries. However, *3%* of the data were acquired by online survey form due to unavailability of the experts. Pedagogues were inquired about parameters required to evaluate novelty in mass examination where students aspire admission to Design schools. The selection parameters were based on their opinion and a set of options provided in the questionnaire.

3.2.2 System architecture

A model is proposed based on the parameters captured from the survey viz., 1) grammatical mistake, 2) misspelling, 3) relevance between question and response, 4) narration or coherence, and 5) relative uniqueness of a response. The input to the architecture is the digitized questions and creative responses exhibiting creative aptitude, and the outcome is the novelty score. A score like this, which is based on a fixed set of parameters intends at generating consistent and unambiguous support for evaluating novelty in Design entrance examination.

Initially, language processing was conducted to identify grammatical mistakes and misspellings in students' responses. Language processing is significant in evaluating novelty as it supports experts in comprehending responses to their target audience. The processing is conducted on an online language processing tool. The Application Programming Interface (API) provides multiple types of semantic and syntactic errors. Next, relevance between a question and an response was verified in order to confirm that the response fits boundary of a question. The questions and responses were individually tokenized. The tokenized questions were fed into the doc2vec model (Devlin et al., 2018; Mikolov et al., 2013; Skansi, 2018). This model is responsible for converting textual data of questions into numerical vectors. The outcome of this model is *300* dimension words embedding of a given document. Similarly, tokenized creative responses were also fed into the doc2vec model. This model also provided *300* dimension word embedding. Both vectors were matched by using cosine similarity function (Skansi, 2018), which provided scores of similarity between a question and response in the range of -1 and 1, indicating lower relevance to higher relevance of descriptive creative responses.

Narration or coherence between sentences is also a pivotal construct to evaluate novelty. Coherence of a response indicates that the answer presents a proper description without diverging to irrelevant contexts. Coherence in a creative response is significant in evaluating novelty as it supports experts in comprehending novelty of responses. Coherence of responses was identified by Bidirectional Encoder Representations from Transformers (BERT), which predicts the probability of the next sentence being the next sentence. Explicitly, this model accepts a pair of a sentence and returned the tensor of two values. The first value, when passed through a normalization function, gave the probability of the next sentence being the next sentence. The second value, when passed through a normalization function, gave the probability of the next sentence not being the next sentence. If the probability of the first value was less than *0.75*, then it was considered to be a narration break (Le & Le, 2013). However, this threshold of considering breakage in narration can be altered depending on the type and difficulty level of an examination.

The relative uniqueness of a response determines its novelty. A clustering algorithm was implemented to evaluate uniqueness of responses, which groups semantically similar responses together in a cluster. The cluster containing fewer descriptive pattern of creative responses indicates that it is unique. However, to get a concise set of responses, three cases were considered and experimented before applying clustering algorithms. They were 1) responses were summarized by abstractive text summarization algorithm (Vodolazova et al., 2013), 2) responses were summarized by extractive text summarization algorithm (Kubat, 2017), and 3) responses were not summarized. The length of responses was considerably high; hence text summarization intend to find central theme of responses. However, many text summarization models require a large volume of data to get trained (Kubat, 2017). Since there was a dearth of dataset to train these models, summarized themes obtained from test dataset were irrelevant to a given context. The digitized responses were directly used for clustering to find infrequent responses. Initially, k-Means clustering algorithm was deployed to find groups of semantically similar responses, but it is subjected to drawbacks of initialization of clusters. Then Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering algorithm was used for clustering the answers, which provided a stable set of clusters (Zhang et al., 1997). A demonstrative rule-base system was framed to decide threshold for acceptance of responses, number of errors in responses, and scores. This rule-base was formed by brainstorming with experts possessing a minimum of ten years of experience in evaluating novelty in creative responses illustrating creative aptitude in Design education. However, these rules cannot be standardized, and it is entirely dependent on the type and level of an examination. Type of an examination indicates that the test may be institutional, nationalized, etc., whereas level of examination refers to the degree of difficulty of test, which may be easy, moderate, difficult, very difficult, etc. Moreover, pedagogues might change threshold value of parameters depending on their stringent or lenient behaviour towards students. Finally, summative assessment of all the features was performed to evaluate novelty score. The details of each of the processes involved in this model are illustrated in Figure 3.2. This model is validated using Mean Squared Error (MSE) and Mean Absolute Error (MAE), where the errors were found to be negligible.

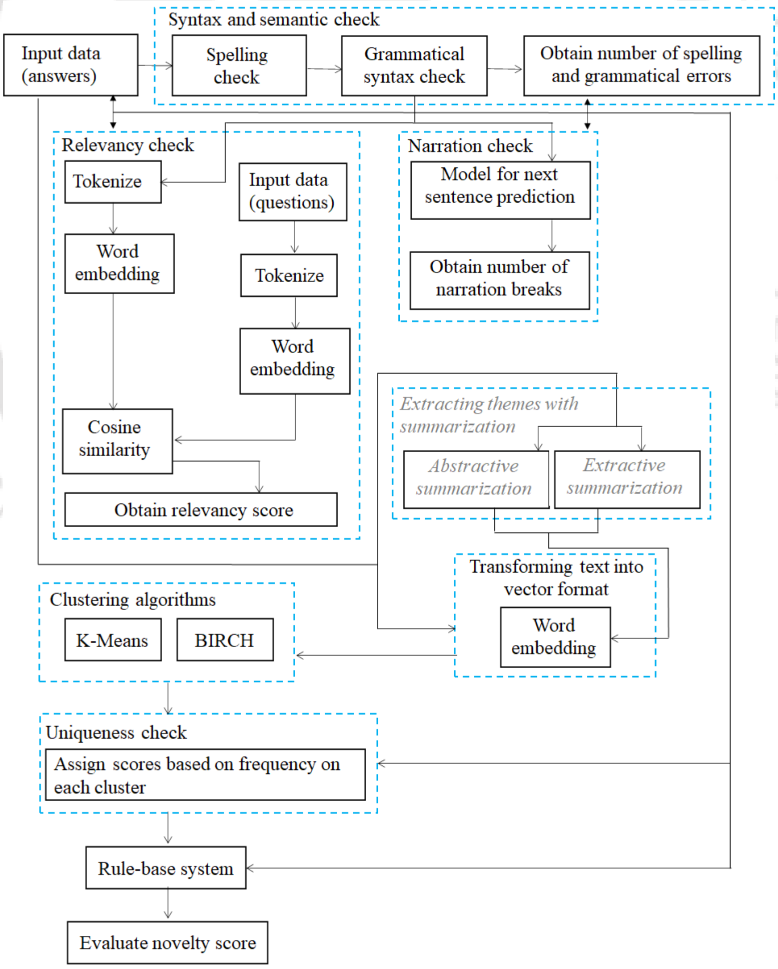


Figure 3.2: Detailed architecture of evaluating descriptive pattern of creative responses

3.3 Results and discussion

3.3.1 Data collection and pre-processing

A survey was conducted with *71* subjects to extract factors to evaluate novelty from descriptive pattern of creative responses that illustrate creative aptitude. The concise form of the questionnaire used in the survey is shown in Appendix B. The questionnaire in the survey requires providing the rating of the given factors for evaluating novelty in descriptive pattern of creative responses exhibiting creative aptitude. However, there was also a provision of appending additional factors and their rating required for evaluating novelty. It was accounted that there was an inclusion of only *5.63%* additional factors by the respondents, and rest all subjects considered the given factors as sufficient for evaluating novelty. The additional features are as follows: 1) choice of words: this factor have higher similarity with a coherence of descriptions and need not be included as an individual factor, 2) according to question: it is similar to the relevance between question and answer, 3) out-of-box: it is highly correlated with unique concept, and 4) disruptive: it focuses on thinking in a new way, hence can be considered as an overlap with the existing factors. The descriptive statistic of these factors extracted from the survey is shown in Tables 3.2-3.7.

Table 3.2: Descriptive statistic of the parameters for evaluating novelty

Statistics						
		GrammaticalMistake	*Misspellings*	*Relevancy*	*NarrationLink*	*UniqueConcept*
N	Valid	71	71	71	71	71
	Missing	0	0	0	0	0
Mean		3.42	3.15	4.73	4.20	4.76
Standard Deviation		1.19	1.05	0.53	0.8	0.52
Median		3.00	3.00	5.00	4.00	5.00
Mode		3	3	5	5	5

Table 3.3: Descriptive statistic of grammatical mistake for assessing descriptive creative responses

Grammatical Mistake					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	Not At All Important	3	4.2	4.2	4.2
	Slightly Important	14	19.7	19.7	23.9
	Important	22	31.0	31.0	54.9
	Slightly More Important	14	19.7	19.7	74.6
	Very Important	18	25.4	25.4	100.0
	Total	71	100.0	100.0	

Table 3.4: Descriptive statistic of misspellings for assessing descriptive creative responses

Misspellings						
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	Not At All Important	3	4.2	4.2	4.2
	Slightly Important	18	25.4	25.4	29.6
	Important	22	31.0	31.0	60.6
	Slightly More Important	21	29.6	29.6	90.1
	Very Important	7	9.9	9.9	100.0
	Total	71	100.0	100.0	

Table 3.5: Descriptive statistic of relevance between question and descriptive creative responses

Relevancy					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	Important	3	4.2	4.2	4.2
	Slightly More Important	13	18.3	18.3	22.5
	Very Important	55	77.5	77.5	100.0
	Total	71	100.0	100.0	

Table 3.6: Descriptive statistic of narration link for assessing descriptive creative responses

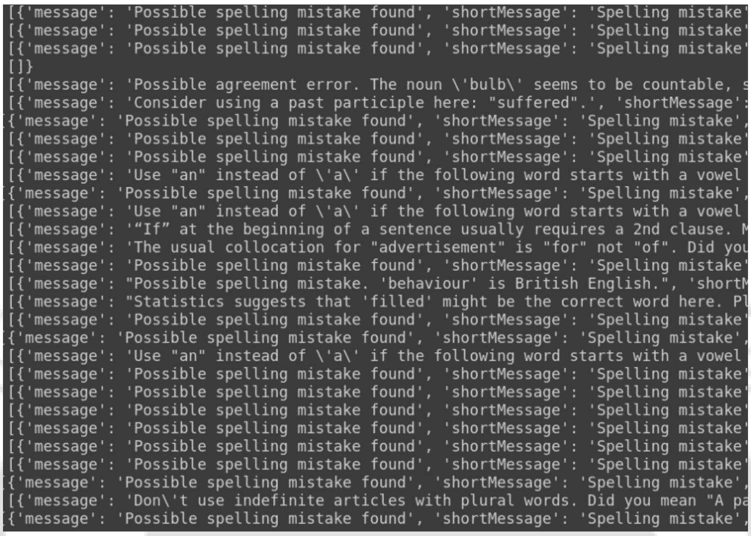
Narration Link					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	Slightly Important	1	1.4	1.4	1.4
	Important	14	19.7	19.7	21.1
	Slightly More Important	26	36.6	36.6	57.7
	Very Important	30	42.3	42.3	100.0
	Total	71	100.0	100.0	

Table 3.7: Descriptive statistic of uniqueness for assessing descriptive creative responses

Unique Concept					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	Important	3	4.2	4.2	4.2
	Slightly More Important	11	15.5	15.5	19.7
	Very Important	57	80.3	80.3	100.0
	Total	71	100.0	100.0	

3.3.2 Language processing

An API based service of an online platform Language Tool (https://languagetool.org/) (*LanguageTool – Online Grammar, Style & Spell Checker*, 2019) was used to conduct language processing of descriptive pattern of creative responses exhibiting creative aptitude. The API returned a .JSON format data for every text, which contains linguistic errors along with category of each errors (e.g. ‘grammar’, ‘duplication’, ‘non-conformance’, ‘misspelling’, ‘typographical error’, etc.). These categories of errors were segregated into two divisionsgrammatical mistakes and misspellings. Grammatical mistakes belong to ‘grammar’ and ‘non-conformance’ category whereas, misspellings belong to the ‘misspelling’ category. The errors of all other categories, such as ‘typographical error’, ‘duplication’, ‘extra whitespace’, etc. were neglected since digitized text format was considered in this context. The representation of the .JSON file is shown in Figure 3.3.



```
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[]}
[{'message': 'Possible agreement error. The noun \'bulb\' seems to be countable, s
[{'message': 'Consider using a past participle here: "suffered".', 'shortMessage':
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Use "an" instead of \'a\' if the following word starts with a vowel
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Use "an" instead of \'a\' if the following word starts with a vowel
[{'message': '“If” at the beginning of a sentence usually requires a 2nd clause. M
[{'message': 'The usual collocation for "advertisement" is "for" not "of". Did you
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': "Possible spelling mistake. 'behaviour' is British English.", 'shortM
[{'message': "Statistics suggests that 'filled' might be the correct word here. Pl
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Use "an" instead of \'a\' if the following word starts with a vowel
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Don\'t use indefinite articles with plural words. Did you mean "A pa
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
```


Figure 3.3: .JSON file returned by language tool

Subsequently, scores for grammatical mistakes and misspellings were calculated for every responses. A rule-base was formulated based on the parameters for scoring the creative responses. It is to be noted that the rules are subject to change for every examination, and it depends on numerous factors such as the type of an examination (e.g., national, institutional, etc.), level of an examination (easy, moderate, difficult, very difficult, etc.), behaviour of pedagogue (stringent, lenient), etc. Here, an illustrative example of a rule-base is shown in Table 3.8. The number of sentences present for each response was calculated. Each set of sentences had a assumed threshold value associated with it. If number of errors exceeded the threshold value, the response belonging to that particular category of the number of sentences is assigned the minimum score above the threshold (Penumatsa et al., 2006). However, if number of errors is equal to or less than the threshold value, then the score is assigned according to the formula shown in Equation 3.1.

Table 3.8: Rule-base for assigning grammatical scores

No. of Sentences (L to U)	Threshold Value (T_G)	Minimum Score Above Threshold
1 – 10	3	0.1
11 – 20	5	0.09
21-30	7	0.08
31-40	9	0.07
41-50	11	0.06

*N.B.:1) The same pattern follows for forthcoming number of sentences.
2) This rule-base is subject to change for every examination.

$$\text{Score below Treshold} = 1 - \frac{N}{1 + T_G} \tag{3.1}$$

Where N is the number of grammatical errors, and T_G is the threshold value of that category given in Equation 3.2.

$$\text{Threshold value}, T_G = 1 + \frac{2 * U}{10} \tag{3.2}$$

Where U is the upper limit of the number of sentences in each category.

The minimum score above the threshold for a category is given by Equation 3.3.

$$\text{Minimum Score above Threshold} = 0.1 - \frac{((T_G - 3) * 0.01)}{2} \tag{3.3}$$

A similar rule-based system was framed for assigning spelling scores for each response, as illustrated in Table 3.9.

Table 3.9: Rule-base for assigning spelling scores

No. of Sentences (L to U)	Threshold Value (T_S)	Minimum Score Above Threshold
1 – 10	6	0.1
11 – 20	9	0.09
21-30	12	0.08
31-40	15	0.07
41-50	18	0.06

*N.B.:1) The same pattern follows for forthcoming number of sentences.
2) This rule-base is subject to change for every examination.

Similar to the grammar checking rule-based system, this parameter also had a threshold value and a minimum score above the threshold associated with each category. The minimum score above the threshold is assigned to any response having spelling errors exceeding the specified threshold value. If the number of errors is less than or equal to the threshold value, the score assigned is given by Equation 3.4.

$$\text{Score below Threshold} = 1 - \frac{N}{1+T_S} \tag{3.4}$$

Where N is the number of spelling errors, and T_S is the threshold value of that category given by Equation 3.5.

$$\text{Threshold value}, T_S = 3\left(1 + \frac{U}{10}\right) \tag{3.5}$$

The minimum score above the threshold for a category is given by Equation 3.6.

$$\text{Minimum Score above Threshold} = 0.1 - \frac{\left(\left(\frac{T_S}{3}\right) - 2\right) * 0.01}{2} \tag{3.6}$$

Thus, a grammar and spelling score, between *0* and *1* can be assigned to each response based on the number of grammatical errors and misspellings. When number of grammatical errors increases, then it is essential to reduce the grammatical score. For example, when there was no mistake in a response, then the highest score was considered as *1*. When the upper limit of the number of lines was *10*, then a score of *0.75* score was considered with one error. Similarly, when the upper limit of the number of lines was *10*, a score of *0.5* was considered for two errors, and so on, as shown in Figure 3.4. However, when number of errors surpasses the threshold, score decrements corresponding to a range of lines.

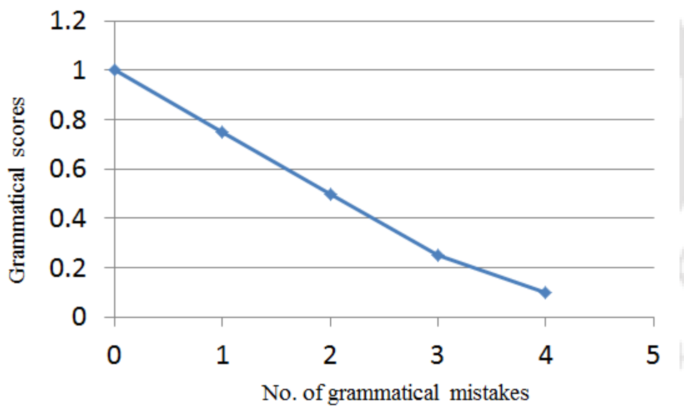


Figure 3.4: Number of grammatical mistakes vs. grammatical scores when upper limit of number of lines is 10

Similarly, evaluating misspellings of responses was essential as it supported in understanding the context by readers and thereby contributing to novelty. When number of misspellings increases, scores for spelling decreases according to rule-base. The thresholds in rule-base were marked by domain-specific experts. A response with no misspellings received the highest score of *1*. When the upper limit of lines *10* with no error, the score was 1; with one error, the score was *0.85*; with two errors, the score was *0.71*; with three errors, the score was *0.57*, and with seven errors which was above the threshold for that range of lines the score was *0.1* as illustrated in Figure 3.5. Detection of spelling errors is the most elementary and significant step in evaluating novelty. During a real-time evaluation of answer scripts, misspellings always have an adverse effect on perception of evaluators. Sometimes it might also convey a wrong meaning, which reduces the impact of novelty.

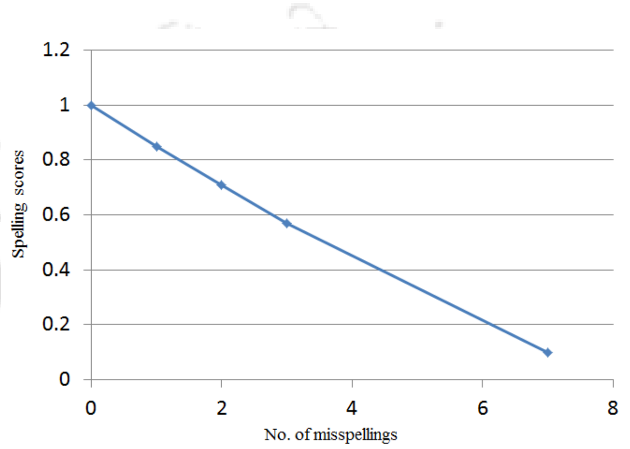


Figure 3.5: Number of misspellings vs. spelling scores when upper limit of number of lines is 10

3.3.3 Relevance between question and creative responses

For checking relevance between question and response, the textual form of a problem and its response needs to be converted into embeddings. To come up with a document embedding that captures the semantic context of a document, doc2vec (Le & Mikolov, 2014; Singhal & Bhattacharya, 2015) model was used. The doc2vec model converts each document into a 300-dimensional vector. Each dimension of the vector captures a semantic and contextual feature related to a document. This algorithm accepted labeled form of documents. It created a co-occurrence, one-hot encoded, matrix of every word in a document. Each document was given a unique id as the label. A window was set for a training process. This window determines the context of each word and is a crucial step in capturing the semantic feature of a word. Based on the window provided, the target word can be found out. This method is called the Continuous Bag of Words (CBOW) (Dynich & Wang, 2017) approach, as illustrated in Figure 3.6.

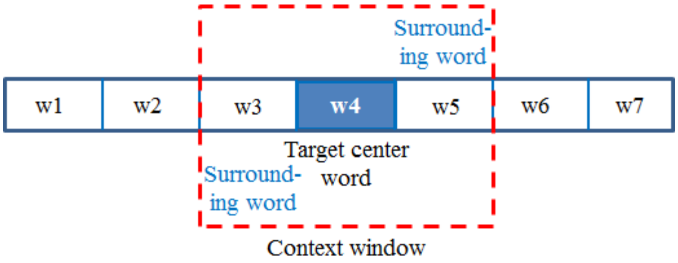


Figure 3.6: Representation of CBOW

The vectors were then fed into a network. The task of network was to optimize itself such that if given a context of words, it outputs a probability distribution of words that were likely to occur next to a provided context. The function tends to maximize the average log probability function, as illustrated below.

$$\left(\frac{1}{N}\right)\sum_{t=w}^{N-w} \log p\,(w_t \mid w_{t-w}\,, \ldots\, w_{t+w})$$

Where N is the total number of words in a text, and w is the window size.

The optimizer used here was stochastic gradient descent, since size of training corpus was considerably high. The values coming from the network were then fed into SoftMax function to obtain a probability distribution of the words. Now, weights learned during the optimization process were taken out, and considered as the embedding of the documents, as shown in Figure 3.7. The questions and the responses were separately passed through the same process, and their word embedding were obtained. Then, the cosine similarity score (Vodolazova et al., 2013) was calculated for a question and its corresponding responses, as shown in Equation 3.7.

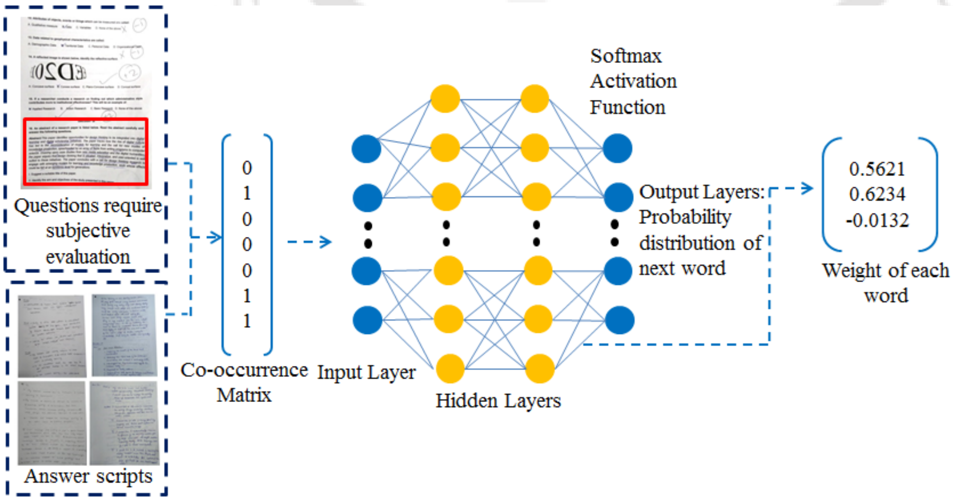


Figure 3.7: Document embedding procedure of question and creative responses

$$CosineSimilarity(a,b) = \frac{a.b}{|a|*|b|} = \frac{(\sum_{i=1}^{n} a_i\ b_i)}{\sqrt{(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)}} \tag{3.7}$$

Here, a and b represents the question and their response.

Novelty associated with any field can be judged based on its relevance to a context (Shi et al., 2021). In this context, relevance score was evaluated by cosine similarity measure which ranges between *-1* and *1*, where *-1* represented lowest relevance score and *1* represented highest relevance score. In some cases, if the relevance score was above *0.9* then it would require validation of whether the response is an exact replication of the question. Graphical representations of the relevance scores are illustrated in Figure 3.8, where twenty four descriptive pattern of creative responses demonstrated the relevance between the question and their responses.

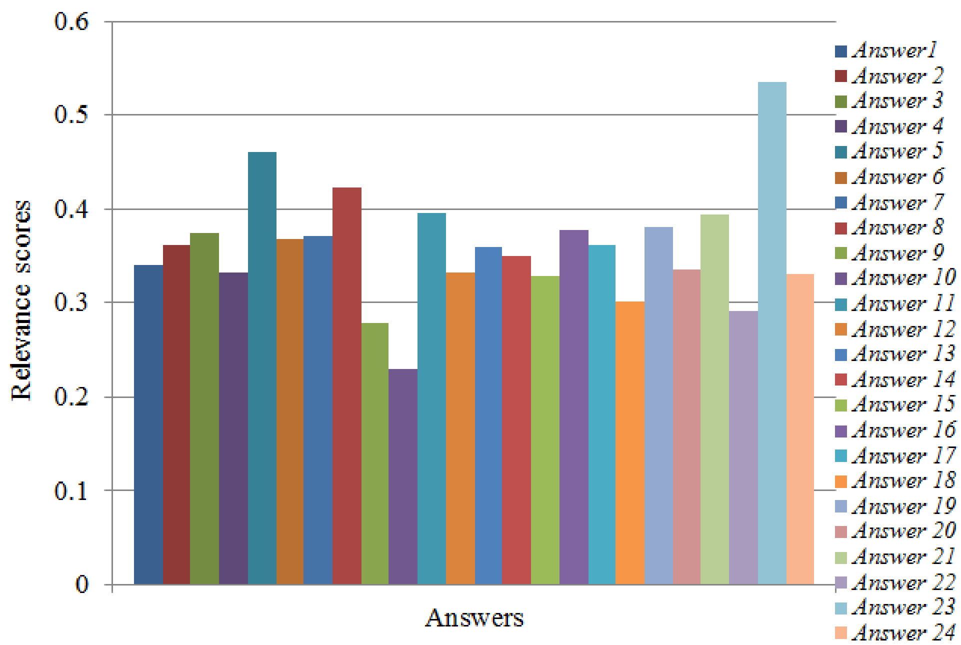


Figure 3.8: Graphical representation of relevance scores

3.3.4 Coherence in a response

Coherence or narration in a response indicates logical connectivity in a given context. Narration also demonstrates a storyline between sentences without any divergence to irrelevant context. An excellent narrative substantially contributes to novelty in a way that assists in perceiving novelty in a response. BERT is a pre-trained model which was used for identifying coherence in a response. It can be fine-tuned to perform downstream NLP tasks. BERT algorithm is based on a transformer that uses attention mechanism which learns semantic relationship between words. Here, it was deployed to find coherence between sentences in a given response. BERT processes both the left and right context, proving it completely bidirectional. The input text was tokenized and further fed into transformer encoder, which converted it into corresponding vectors. The output layer served vectors corresponding to each token, as illustrated in Figure 3.9. However, it was more towards the next word prediction and lacks contextual learning, which was initially implemented. After that, the prediction mechanism was modified in the model.

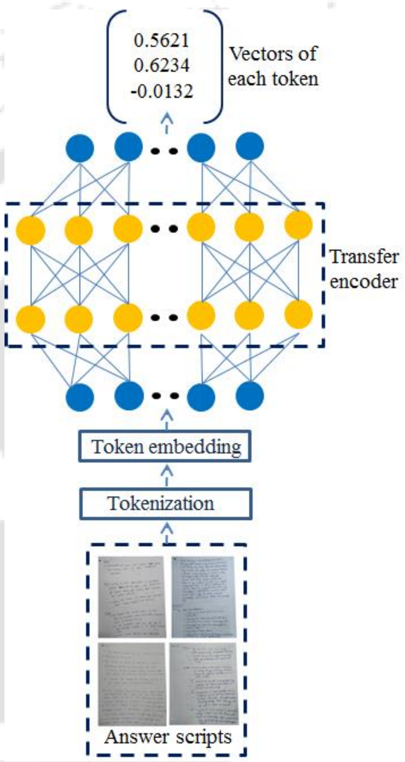


Figure 3.9: Prediction of next word

Again, a pre-trained model was considered where half of the training dataset comprised the next sentence being the next sentence, and the other half comprises the next sentence, not being the next sentence. Next, the model was fed with a pair of sentences in the form of a first and second sentence, second and third sentence, third and fourth sentence, and so on from a response, which predicted narration of the next sentence being the next sentence. The architecture of the model to identify the next sentence is shown in Figure 3.10.

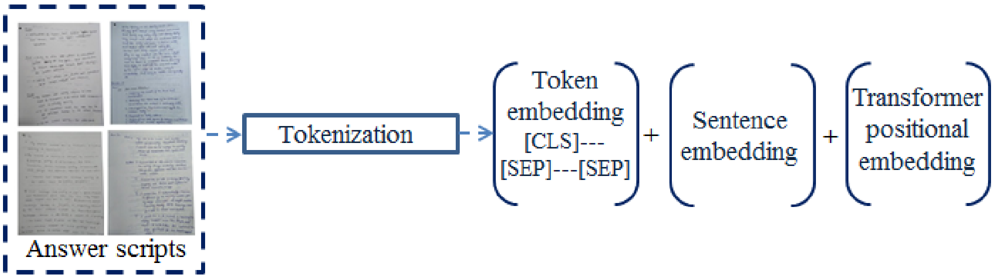


Figure 3.10: Identifying the next sentence

In this model, word and sentence embedding was performed, where [CLS] was considered as the beginning of the first sentence, and the rest of all sentences were separated by [SEP]. Subsequently, positional embedding was concatenated to each token to identify its positional sequence. The input sequence learned in the Transformer model, and the outcome of [CLS] token was a vector of weights. The SoftMax activation function was applied in the output layer to acquire the probability of the next sentence (Vodolazova et al., 2013). This model considered tensor of two values returned by the model where the first values when passes through a normalization function, provided the probability of the next sentence being the next sentence, and the second value when passes through a normalization function provided the probability of the next sentence not being the next sentence. The threshold for considering the break of narration was obtained from experienced evaluators and confirmed by manual comparison process. The threshold is subject to change depending on the type, level, and behaviour of pedagogues. A rule-based system was developed for assigning narration scores to the responses, as illustrated in Table 3.10.

Table 3.10: Rule-base for assigning narration score

No. of Sentences (L to U)	Threshold Value (T_N)	Minimum Score Above Threshold
1 – 10	2	0.1
11 – 20	4	0.09
21-30	6	0.08
31-40	8	0.07
41-50	10	0.06

*N.B.:1) The same pattern follows for forthcoming number of sentences.
2) This rule-base is subject to change for every examination.

Each set of sentences had a threshold value associated with it, which was decided by domain-experts. If the number of narration breaks exceeds the threshold value for a response belonging to that particular category of the number of sentences, was assigned the minimum score above a threshold. However, if the number of narration break was equal to or less than the threshold value, then the score was assigned according to the formula as shown in Equation 3.8.

$$\text{Score below Threshold} = 1 - \frac{N}{1 + T_N} \tag{3.8}$$

Where N is the number of narration breaks, and T_N is the threshold value of that category given by Equation 3.9.

$$\text{Threshold value}, T_N = \frac{2 * U}{10} \tag{3.9}$$

Where, U is the upper limit of the number of sentences in a response. The minimum score above the threshold for a category is given by Equation 3.10.

$$\text{Minimum Score above Threshold} = 0.1 - \frac{(T_N - 2) * 0.01}{2} \tag{3.10}$$

Thus, scores between *0* and *1* can be assigned to each response based on the number of narration breaks.

The increase in a number of narrations breaks leads to a decrease in narration score. Domain-specific experts decided the highest narration score to be considered as *1*, indicating no narration break. However, it is subject to change for every examination. The threshold of rule-base was characterized by an expert's opinion possessing experience in evaluating descriptive pattern of creative responses exhibiting creative aptitude. When the upper limit of number of lines was *10* with one narration break, the score was *0.66*; with two narration breaks, the score was *0.33*, and so on. When the narration break exceeds the threshold, the score was *0.1*, illustrated in Figure 3.11.

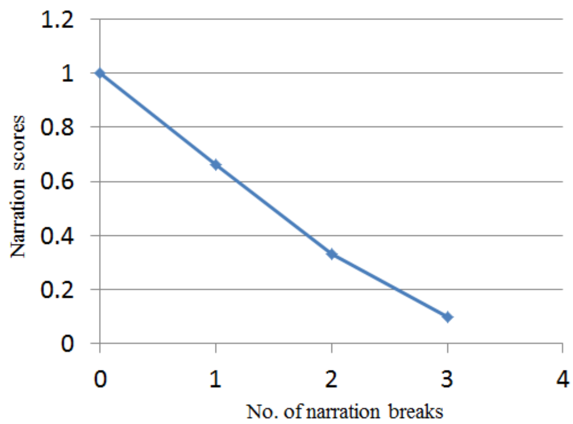


Figure 3.11: Number of narration breaks vs. narration scores

3.3.5 Consequences of summarization and without summarization

The descriptive pattern of creative responses exhibiting creative aptitude are often lengthy. Sometimes, summarization of answers is useful to capture theme of a given context. Summarization can be broadly classified as- extractive and abstractive summarization (Hsu et al., 2018). Initially, an experiment was conducted for capturing theme using extractive text summarization whose result was found to be insignificant as it selects sentences from a document based on the frequency of words that occurs in a document. Therefore, outcomes of this technique failed to generate concepts from responses, as it only extracted the weighted sentences, and the outcomes of this process were not utilized for finding unique responses.

Subsequently, abstractive text summarization was also considered in the experiment. It generated a concept from a document with a completely new set of words. The algorithm was proved to be efficient with a significant Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score. However, it requires a large dataset to train the models (Paulus et al., 2017). Also, training the models isn’t time and cost-effective. The pre-trained models were trained on domain-specific data like e-commerce, medicines, news, etc., which made it infeasible to be tested on a creativity dataset.

In this context, responses associated with evaluation of novelty were collected from Design schools. The dataset was limited since it was assembled from various Design schools where there was limited access to creative responses of entrance examinations. Very few institutes agreed on anonymously providing students responses. Training with limited dataset was infeasible for summarization algorithms. The test results highlight inclination towards the domain of trained dataset. Therefore, the summarization technique was infeasible for the proposed model, and their outcome were not further considered for processing.

3.3.6 Relative uniqueness of a creative response

To find uniqueness of a creative response, semantically similar responses were grouped together. The groups with a relatively lesser number of response were outlier and considered as unique. Initially, the K-means clustering algorithm was used to group semantically similar responses. However, it is infeasible in this scenario as the cluster varies in each run of the algorithm, as shown in Figure 3.12. Further, the BIRCH clustering algorithm was used, which provided a consistent result. BIRCH is more effective than K-means because K-means algorithm needs to be initiated with a number of clusters and iteratively moves the datapoints from one cluster to another cluster to optimize the objective measurement function. In K-means each semantically similar group are illustrated using centroid. It assures convergence to a local minimum, and characteristic of the local minimum is dependent on its elementary partition. The worst time complexity is exponential (Zhang et al., 1997). Due to this, K-means is not considered appropriate for large datasets. On the contrary, BIRCH addressed these pitfalls of K-means algorithm and scaled well with large datasets. It is important to highlight here that initial parameterization, i.e., must for k-means for desired number of clusters was not required for BIRCH (Lorbeer et al., 2018).

The clusters generated by the BIRCH clustering algorithm are illustrated in Figure 3.13. The uniqueness of responses was evaluated by calculating frequency of responses in each cluster. Cluster with relatively lowest frequency of response(s) was considered as unique. Clusters and their frequencies are illustrated in Figure 3.14, which states that the *0th* cluster has *19* elements, the *1st* cluster has *3* elements, the *2nd* cluster has *1* element, and the *3rd* cluster has *2* elements. Therefore, it indicated that the element of the *2nd* cluster is relatively unique. Further, the clusters were arranged in ascending order with respect to their frequency. A score based on uniqueness was assigned based on a scale of *0* to *1*, depending on the number of clusters. It was calculated by Equation 3.11.

```
Run 1:
Clusters by K-means:
[2 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 3 0 0 0 0 1 0 1]

Run 2:
Clusters by K-means:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 2 0 0 0 0 1 0 1]

Run 3:
Clusters by K-means:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1]
```

Figure 3.12: Clusters at multiple runs of K-means algorithm for descriptive creative responses

```
Clusters by BIRCH:
Run 1: [2 0 0 0 0 0 1 0 0 0 0 0 0 0 3
        0 0 0 0 0 0 1 0 1 3]
Run 2: [2 0 0 0 0 0 1 0 0 0 0 0 0 0 3
        0 0 0 0 0 0 1 0 1 3]
Run 3: [2 0 0 0 0 0 1 0 0 0 0 0 0 0 3
        0 0 0 0 0 0 1 0 1 3]
```

Figure 3.13: Clusters using BIRCH algorithm for descriptive creative responses

```
Counter({0: 19, 1: 3, 3: 2, 2: 1})
[(2, 1), (3, 2), (1, 3), (0, 19)]
```

Figure 3.14: Frequency of responses in cluster

$$Uniquness = 1 - \left(\frac{N}{T}\right) \tag{3.11}$$

Here, Uniqueness represents uniqueness score for descriptive pattern of creative responses, N is the number of creative responses in a cluster for a given question, T is the total number creative responses across all candidate submissions for a given question. For example, three clusters are formed based on their semantic similarity, and the elements in the cluster are as follows- *0:5*, *1:3*, *2:2* respectively and there are ten creative responses that need to be examined, then score generated are graphically illustrated in Figure 3.15.

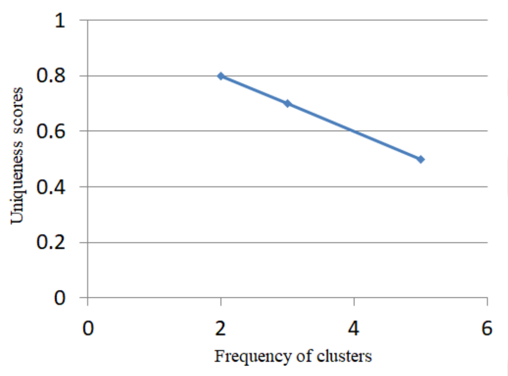


Figure 3.15: Frequency of clusters vs. uniqueness scores

3.3.7 Novelty score and model validation

The features that contributed to evaluating novelty scores were grammatical syntax, spelling of text, narration or coherence in a response, relevance between question and responses, and the relative uniqueness of a response. Further, summative assessment were conducted for evaluating novelty scores. Normalization was used to scale the score between *0.0* to *1.0*, as shown in Figure 3.16.

```
Number: 1
Relevance Score: 0.3354763
Narration Score: 0.3333333333333333
Grammar Score: 1.0
Spelling Score: 0.2857142857142857
Uniqueness Score: 0.25
Score Assigned By Model: 0.5109778008838705
Actual Score: 0.45
-----------------------------------------------------------
Number: 2
Relevance Score: 0.39898038
Narration Score: 0.6666666666666666
Grammar Score: 1.0
Spelling Score: 0.2857142857142857
Uniqueness Score: 0.25
Score Assigned By Model: 0.6786472110471294
Actual Score: 0.76
-----------------------------------------------------------
Number: 3
Relevance Score: 0.2827559
Narration Score: 1.0
Grammar Score: 0.75
Spelling Score: 0.5714285714285714
Uniqueness Score: 1.0
Score Assigned By Model: 0.7540498426358856
Actual Score: 0.8
-----------------------------------------------------------
Number: 4
Relevance Score: 0.53488994
Narration Score: 1.0
Grammar Score: 0.75
Spelling Score: 0.14285714285714285
Uniqueness Score: 0.25
Score Assigned By Model: 0.915896752950366
Actual Score: 0.78
-----------------------------------------------------------
Number: 5
Relevance Score: 0.3314595
Narration Score: 1.0
Grammar Score: 1.0
Spelling Score: 0.1
Uniqueness Score: 0.25
Score Assigned By Model: 0.6028287485313545
Actual Score: 0.5
-----------------------------------------------------------
Mean Absolute Error: 0.08540124973651518
Mean Squared Error: 0.008297932883961493
```

Figure 3.16: Novelty score evaluated by model

The model is validated by comparing the test data with gold standard collected from experts (Mayer & Butler, 1993). The performance was calculated by mean absolute error and mean squared error (Hamzacebi, 2008), as represented by Equations 3.12 and 3.13. The error is represented graphically in Figure 3.17.

$$\text{Mean Absolute Error, MAE} = 1/n \sum_{i=1}^{n} |actual\ value_i - predicted\ value_i\,| \qquad (3.12)$$

Here, n is the total number of scores awarded across all creative responses for a given question, actual score is the score awarded by human experts across all creative responses for a given question, predicted score is the score awarded by the proposed model across all creative responses for a given question.

$$\text{Mean Squared Error, MSE} = 1/n \sum_{i=1}^{n} (actual\ value_i - predicted\ value_i\ )^2 \qquad (3.13)$$

Here, n is the considered number of test datasets across all candidate submissions for a given question, the actual value is the score assigned by experts, and the predicted value indicates the score assigned by this model.

Figure 3.17: Human score vs. model score for evaluating descriptive creative responses

The mean absolute error of this model is *0.085*, and the mean squared error is *0.008*, which is acceptable, and it can be implemented for evaluating novelty in Design entrance examinations on a large scale.

3.4 Conclusion

This chapter highlighted subjective evaluation of novelty in descriptive pattern of creative responses exhibiting creative aptitude in Design education. Evaluation of novelty in creative responses like this, is still a manual process which depends on pen-and-paper-based technique. This study attempts in identifying parameters for evaluating novelty in descriptive pattern of creative responses exhibiting creative aptitude of Design education by human-centred design approach. The findings of a mixed-method study suggest that grammatical syntax, spelling of text, narration or coherence between sentences, relevance between question and responses, and uniqueness of a response are the parameters that support in evaluating novelty in mass examination. A design model is proposed that attempted in automating the manual process of evaluation based on the factors pedagogues refer to in the process of evaluating novelty. This model is implemented by algorithmic techniques. It is validated by comparative analysis of the outcomes of evaluation conducted by the proposed model and human experts, which confirms the competence of the devised model and establishes trust of pedagogues.

Chapter 4: Identifying parameters to assess novelty of labelled image-based creative responses and digitizing its evaluation process

Abstract

An inherent criterion of evaluation in Design education is novelty. Pedagogues compare and contrast responses for cohort of students in mass examination aspiring admission to Design schools to assess novelty. Large number of students participate in mass examinations, and in situations like this, examiners are confronted with multiple challenges in subjective evaluation such as- 1) Errors encountered in evaluation due to stipulated timeline, 2) Errors encountered due to prolonged working hours, 3) Errors encountered due to stress in performing repeated task on a large-scale. Pedagogues remain ever inquisitive and vigilant about the evaluation process being consistent and accurate due to monotony of repeated task. To mitigate these challenges, a computational design model is proposed for automating evaluation of novelty in labelled image-based pattern of creative responses. This type of creative responses comprises images whose parts are marked with labels. This model is developed by mixed-method research, where features for evaluating novelty of labelled image-based pattern of creative responses are investigated by conducting a survey. Further, these features were utilized to evaluate novelty and generate score for labelled image-based pattern of creative responses using Computer Vision (CV) and Deep Learning (DL) techniques. The performance metric of the model, when measured reveals a negligible difference between scores of experts and scores of the proposed model. These comparative analysis of the proposed model with human experts' confirm the competence of the devised model and would go a long way to establish trust of pedagogues by ensuring reduced error and stress during the evaluation process.

Highlights

- *Human-centred design approach to identify parameters of evaluating novelty in labelled image-based pattern of creative responses.*
- *Proposing a computational design model for evaluating novelty in labelled image-based pattern of creative responses.*
- *Implementing the model using various tools and algorithmic techniques.*
- *Validating the model by comparing its outcome with human evaluation.*

4.1 Introduction

Novelty is a significant criterion of evaluation in creative responses of Design Education (Demirkan & Afacan, 2012), which can be assessed from multiple "solution patterns" commonly followed in creative schools of Design education. This chapter highlights evaluation of novelty in labelled image-based pattern of creative responses. These responses requiring subjective evaluation comprises images or sketches whose parts are labelled. An example of labelled image-based pattern of creative response is illustrated in Figure 4.1. Pedagogue's assessing subjective answers evaluate the novelty of student's response based on their opinion, "subjective evaluation index" of their experience, analytical ability, knowledge, and persuasion (Alexiou et al., 2018; Brabb & Morrison, 1964; Furusho & Kotani, 2017; Wan et al., 2018).

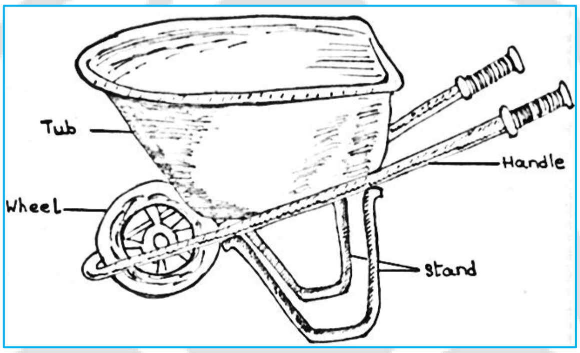


Figure 4.1: Labelled-imaged based pattern of creative response

Evaluation of novelty is not absolute, but it is relative to other responses. A response is compared with other responses to perceive its uniqueness in order to define its novelty (Sarkar et al., 2007). Such repeated tasks conducted by pedagogues increase chances of stress that might lead to errors in evaluation of novelty. Errors can also take place in a subjective evaluation process due to differences in referential metrics of individual pedagogues or maybe for unvarying repeated activity. Stress generated due to a repeated task may be another factor for introduction of error in evaluation process (Chan et al., 2010; Montgomery & Rupp, 2005). For a fair and consistent evaluation process, it is essential to address issues such as errors in novelty scoring that emerges due to stress and individual parameters of pedagogues in an assessment specifically for Design entrance examination involving subjective evaluation. This triggers two necessities viz., identifying factors that pedagogues refer to in evaluating novelty in labelled image-based pattern of creative responses exhibiting creative aptitude and proposing a model that attempts in digitizing the process of evaluation based on these features. The investigation in this chapter attempts to address the research gaps highlighted and reported in the state-of-the-art literature review presented in subsection 1.5.3, subsequently corresponding research questions and objectives reported in section 1.9 and 1.12. The research questions and the objectives are stated below again for reference.

RQ3: *While assessing creative responses of students, what are the factors that Design educators consider for assessing novelty of the responses?*

RQ4: *Can the existing process of subjective manual evaluation of creative aptitude be automated?*

Objective 6: *To identify the factors of novelty in creative aptitude evaluation.*

Objective 7: *To design a digitized system for novelty assessment in creative aptitude.*

Mixed-method research approach had been adopted in this study that attempted to investigate features of evaluating novelty. First, a survey was conducted to identify and define features for evaluating novelty. These features were then used to conceive a model that evaluates and scores novelty of creative responses. Initially, the proposed model intended at identifying whether a labelled image-based pattern of creative responses was relevant to a question. Labels from image were processed using optical character recognition (OCR) (Avadesh & Goyal, 2018). Image predictions and labels from images were mapped to vectors (Gutiérrez & Keith, 2018). Further, relevant responses were clustered to determine uniqueness based on their semantic similarity. Clusters that possessed fewer images were considered to be relatively more unique by the proposed model. Finally, a scoring mechanism was implemented based on these features. The model was validated by comparing its outcomes with human experts. Negligible differences between the outcomes indicate competence of the devised model. The overview of the model is illustrated in Figure 4.2.

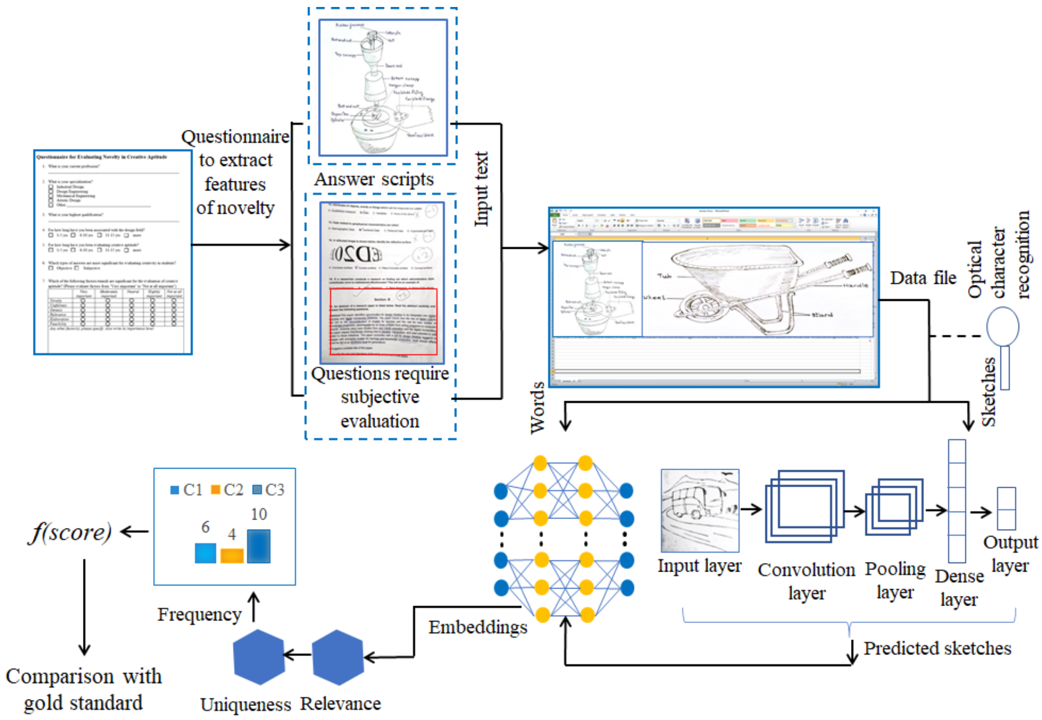


Figure 4.2: Overview of the model

4.2 Method

4.2.9 Survey plan and participants

An empirical survey was planned to identify the parameters of evaluating novelty in labelled image-based pattern of creative responses. Pedagogues associated with Design education possessing a minimum of ten years of experience were selected for this study (Demirkan & Afacan, 2012). They have expertise in evaluation for entrance examinations where students aspire admission to Design schools. The population size for this study was indefinite. A confidence interval of *0.75* with *0.05* marginal errors was considered for this study, which is usually practiced in research associated with Design and education (Krejcie & Morgan, 1970). The sample size of the survey was *71 (N=71*) subjects with an average age between *39- 62* years (*M=47.92, S.D. =5.52*), and *46%* were female (*N=33*).

The survey was conducted in renowned Design schools involving private and government organizations. Pedagogues associated with evaluating answer scripts of Design entrance examination were provided the questionnaire. The survey was conducted mostly on-site, i.e., in department of various Design schools. However, *3%* of the data were acquired by online survey form due to unavailability of the experts. Pedagogues were inquired about parameters required to evaluate novelty in mass examination where students aspire admission to Design schools. The parameters were acquired based on their opinion and a set of options provided in the questionnaire.

4.2.2 Questionnaire and parameters of evaluation

A questionnaire was designed to identify parameters to evaluate novelty in the mass examination of Design education for students aspiring admission to Design schools. In situations like this, evaluations are conducted on a large scale. Classroom-based examination is significantly distinct from mass or entrance examination (Ali, 2005; Karatas et al., 2013). Therefore, multiple items were presented in questionnaire captured from literature and post brainstorming with *5* experts (*N=5*) having a minimum of ten years of experience possessing expertise in evaluating for mass examination. The parameters of evaluation were rated using a Likert-Type scale (Croasmun & Ostrom, 2011) with labels “very important”=*1*, “slightly more important”=*2*, “important”=*3*, “slightly important”=*4*, and “not at all important”=*5*. Apart from that, provision was provided where pedagogues could include parameters of their choice depending on their knowledge and persuasion.

This questionnaire intends to identify the factors to assess labelled image-based creative responses. The questionnaires were written in the way (provided in Appendix D) to confirm whether the factors identified in literature review ascertain to the factors considered by Design pedagogues in practice. State of the art review was conducted to examine the findings from literature that contributed to the identification of factors that are referred for assessing products, solutions, ideas, etc. as illustrated in Table 3.1. User rating ratings were collected using 5-point Likert-type scale to identify the factors preferred by experts in the evaluation process, as illustrated in Appendix C.

The parameters included in questionnaire were as follows- 1) relevance, 2) uniqueness, 3) clarity, 4) choice of colours, 5) sketching ability, 6) language processing, and 7) narration (Berbague et al., 2021; Camburn et al., 2020; Chaudhuri et al., 2020, 2021b, 2021c; Demirkan & Afacan, 2012; Schumann et al., 1996; Takai et al., 2015). The explanation of these parameters are as follows. Firstly, relevance between a question and an answer verifies whether a response is appropriate for a question. Many a time, a creative response might be novel but doesn’t meet requirements of a question. In any examination, relevance between question and its response is significant (Charlet & Damnati, 2017; Gagnon et al., 2019). Secondly, novelty means newness in a response; therefore, measuring relative uniqueness in a cohort of responses is essential to determine novelty (Sarkar & Chakrabarti, 2011). Thirdly, clarity in visual representation is essential, which refers to comprehensibility and clear representation of images. It is important in this context as it empowers understanding ability (Xueqing et al., 2018). Fourthly, choice of colours contributes to aesthetic features of an image (Sangkloy et al., 2017; Z. Wang et al., 2017). Fifthly, sketching ability refers to potential of students in visually representing their ideas (Al-Homoud, 2020). Sixthly, language processing refers to checking spelling and grammar; however, it is applicable for responses consisting of descriptions. Seventhly, narrations indicate coherence between sentences (Chaudhuri et al., 2020, 2021b).

4.2.3 Detailed architecture for evaluating image with labels

The detailed architecture to evaluate novelty of image with labels specific to Design entrance examination involves multiple procedures, as shown in Figure 4.3. Initially, labelled image-based pattern of creative responses exhibiting creative aptitude were considered as input to the model. The responses were pre-processed and divided into train-and-test set for predicting the images by Visual Geometry Group (VGG-19) model. Further, it was essential to identify whether the response is relevant to the question. Therefore, question and responses were converted into high dimensional numerical vectors. Cosine similarity function was used to compare question vector with response vector. Relevance score was generated using this function. A thresholding technique was applied, which supported in identifying a value of cosine similarity function that could be accepted as a relevant response. Further, relevant responses were clustered using K-means algorithm to group semantically similar responses. However, this algorithm generated inconsistent number of clusters in each run of the algorithm (Cao et al., 2009); therefore Affinity Propagation clustering algorithm was implemented in order to get a stable set of clusters (Wei et al., 2017). Uniqueness score was generated based on relative frequency of clusters. Finally, summative assessment of relevant score and uniqueness score was performed to evaluate novelty score.

4.2.4 Dataset for image with labels

Multiple image-based datasets were investigated but couldn't be utilized in this context due to their focus on single-label classification where parts of an image were not labelled (Eitz et al., 2012). In contrast, many other datasets were not considered due to their computational complexity requiring high computational time and space (Benenson et al., 2019; Krasin et al., 2017; Kuznetsova et al., 2020; Pont-Tuset et al., 2020). Finally, to evaluate novelty in image with labels, a Real-World Web Image Dataset from National University Singapore (NUS-WIDE) dataset was selected. This dataset comprised of *25,000* Red Green Blue (RGB) images with *81* class labels. It included "low level features", "groundtruth", "tags", "concept list", "image list", and "original URLs" (Chua et al., 2009).

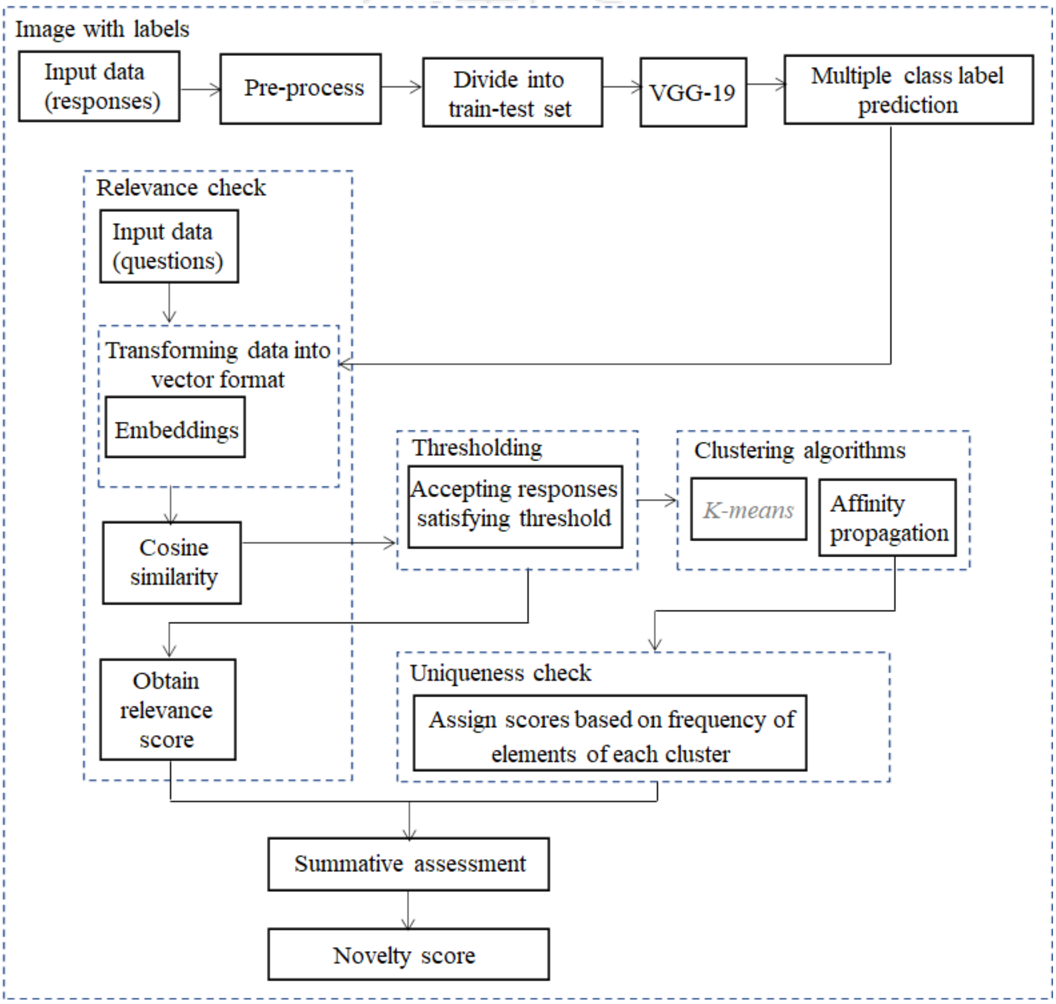


Figure 4.3: Detailed architecture for evaluation of image with labels

NUS-WIDE dataset included a bunch of local and global factors of images to experiment and correlate with results. The four demonstrative global features were as follows- 1) "64-D (dimensions) colour histogram- LAB" demonstrates colour illustration of images. LAB represents lightness and colour components of images respectively, 2) "144-D colour autocorrelogram- HSV” indicates colour distribution and spatial correlation between colour pair, 3) “73-D edge direction histogram” presents direction of the edges of an image, 4) “128-D wavelet texture” suggests multi-resolution perspective of texture investigation. The dataset represented a grid-based feature precisely a “225-D block-wise colour moment”. Colour moment illustrated mean, variance, and skewness to represent colour distribution. A “500-D bag of visual words” were present in dataset to find key-points and scale. Further, Scale Invariant Feature Transform (SIFT) was calculated where vector quantization was applied on SIFT region features to generate visual lexicon by a clustering algorithm (Chua et al., 2009).

Contributions of ground-truth in NUS-WIDE were considered from eighty-one concepts chosen widely from literature, popular tags in Flickr dataset, and common entities (Hodosh et al., 2013; Kondermann, 2013; Snoek et al., 2006). Usually, tags in a dataset were disturbed by noise and missing values. In context of this dataset, average precision and recall were above *0.5*, indicating that more than *50%* of the tags were noisy. It was infeasible to manually annotate a large volume of data. Therefore, nearest neighbour searching technique for annotation was implemented. This dataset contained six concepts and multiple sub-concepts associated with it. There were approximately *4,650* image lists in this dataset. Original-URLs in dataset comprise of the link to images from where they can be downloaded (Chua et al., 2009).

4.2.5 Pre-processing data of image with labels

The NUS-WIDE dataset was considered as an input to the model. They were pre-processed to convert images into sketches. RGB images were converted into greyscale by computing negative of an image. Further, scipy Gaussian filter was applied for transforming an image into a sketch (*Scipy.Ndimage.Gaussian_filter — SciPy v1.7.1 Manual*, 2021). All sketches were resized into *(224x224)* dimensions. After pre-processing and removing corrupt sketch files, there were *20,539* sketches of shape *(224x224x3)*. This data was randomly partitioned as *80%* and *20%* for training and validation, respectively. Specifically, *16,431* sketches having *81* class labels were used for training, and *4,108* sketches having *81* class labels were utilized for validation. The sample images for labelled image-based responses collected from online resources are illustrated in Appendix E.

4.2.6 Prediction of image with labels

A multi-label image classification model was fine-tuned on keras VGG-19 model (Simonyan & Zisserman, 2014; *VGG16 and VGG19*, 2014) pre-trained on ImageNet dataset (Russakovsky et al., 2013, 2015). Feature extraction of images was done by removing the last classification layer of VGG-19 pre-trained model, and last few layers of this feature extractor model were made trainable for fine-tuning. Further, an additional dense and dropout layer was attached at the end for the purpose of multi-label image classification. The final layer consists of a dense layer of *81* neurons having "sigmoid" activation function and "binary_crossentropy" loss. Generally, binary classification provides decent accuracy with a balanced number of examples for each class label. This architecture neither dealt with a binary nor multi-class classification. It was associated with multi-label classification, and numbers of labels were imbalanced for every class. Therefore, the metric used to measure accuracy was "f-beta" (Brownlee, 2019; Gupta & Ruebush, 2019).

4.2.7 Relevance between question and creative responses of type image with labels

For testing the model, images were scrapped from Google images, which had labels same as NUS-WIDE dataset. These types of images were considered because humans draw labelled images to make them informative for their intended audience. This test data had labels marked on images. The VGG-19 model was trained on NUS-WIDE dataset that had labels of images in separate files, hence pre-processing of this present test data was required. Hence, it was essential to extract the labels from images. An OCR Application Programming Interface (*Best Free OCR API, Online OCR, Searchable PDF – Fresh 2021 On-Premise OCR Software*, 2021) was used to recognize text on images. The labels in images were removed using inpainting, masking, and thresholding techniques (Guillemot & le Meur, 2013; Kawai et al., 2015).

A bounding box was created for each label present in images that were returned in form of .json file. Bounding boxes are white rectangles over images created using inverse masking. After masking every image, inpainting techniques were applied to remove bounding boxes from original image by using pre-defined functions in OpenCV such as cv2.INPAINT_TELEA and cv2.INPAINT_NS (*Inpainting — OpenCV 2.4.13.7 Documentation*, 2021; *OpenCV: Image Inpainting*, 2021) that fills every pixel within bounding box with the average of neighbouring pixels up to certain radius. Finally, test images were converted into sketches.

Input questions were scraped from website of CEED (Bombay, 2021a). The pre-processed test creative responses were converted into high dimensional embeddings. Numerical vectors were considered to compare question and their corresponding creative responses. In any entrance examination relevance between question and their responses are very significant. A response might be novel, but it requires to conform to the boundaries of a question; precisely responses must satisfy the requirements of questions (Chaudhuri et al., 2020; Sakata et al., 2019). Therefore, embeddings of questions and their responses were compared using cosine similarity function (Fauzi et al., 2017). It generated a relevance score between *0* and *1*. Relevant responses tend to have higher scores relative to irrelevant ones. Responses that turn out to be relevant are further considered for calculating relative uniqueness.

4.2.8 Thresholding creative responses of type image with labels

Decision-making was essential to categorize responses that can be considered relevant based on scores. A threshold value was essential that could distinguish responses as relevant or irrelevant. Multiple question-responses sets were considered where each question had multiple responses. Two experts *(N=2)* possessing expertise in evaluating answer scripts of entrance examination in Design education were selected to score image with label pattern of creative responses. Human scores were considered a golden standard. Further, scores from cosine similarity function and expert scores were considered to calculate F-measure. A threshold from *0.2* to *0.7* was considered, and their corresponding F-measure was calculated. Generally, a threshold value gets accepted, which has a high range of F-measure values (Penumatsa et al., 2006).

4.2.9 Clustering and evaluation of novelty of image with label pattern of creative responses

Clustering concept was implemented in order to group semantically similar creative responses. Initially, K-means clustering algorithm was implemented. However, it is subject to problems of generating inconsistent number of clusters and has a negative impact on noisy data points (Cao et al., 2009). Further, Affinity Propagation clustering algorithm was implemented to get a stable set of clusters. It showed significantly improved results as compared to traditional clustering algorithms. Results can be achieved in comparatively lesser time with large datasets (Wei et al., 2017). Dense clusters indicate that conceptually similar creative responses were repeated; thereby frequency of a cluster was relatively higher. Sparse clusters possess relatively lower frequency indicating less similar concepts were repeated. Clusters that had a relatively lower frequency of creative responses were considered unique. Uniqueness score was found using Equation 4.1 (Chaudhuri et al., 2021b).

$$\text{uniqueness}_{\text{scoreImagewithLabel}} = 1 - \left(\frac{NC}{TS}\right) \quad (4.1)$$

Here, $uniqueness_{scoreImagewithLabel}$ represents uniqueness score for image with label pattern of creative responses, *NC* is a total number of creative responses in a cluster for a given question, and *TS* is the total number of creative responses across all candidate submissions for a question. Relevance score was measured between values *0* and *1*. Uniqueness score was also a normalized value between *0* and *1*. Novelty score was evaluated for responses by summative assessment (Chaudhuri et al., 2020, 2021b) of relevance score and uniqueness score and further normalized within the range of *0* to *1*.

4.3 Results and discussion

4.3.1 Survey site

A survey was conducted at different Design schools associated with state and central universities. The intent was to identify factors that Design schools use to evaluate novelty of labelled image-based pattern of creative responses. Pedagogues possessing expertise in evaluating novelty were chosen as subjects for this survey. A questionnaire was framed consisting of factors for evaluating novelty in mass examination. Initially, a survey brief was introduced to the subjects. Further, the questionnaire was administered, and the respondents were asked to identify features to evaluate novelty in image-based pattern of creative responses of entrance examinations of Design education. Data collection at survey site is illustrated in Figure 4.4.

Figure 4.4: Survey site

4.3.2 Descriptive statistic of features of image with labels identified by survey

Questionnaire was framed to capture features for evaluating novelty from an image with label pattern of creative responses, as shown in Appendix D. A feature list was provided in questionnaire consisting of following items- 1) relevance, 2) uniqueness, 3) clarity, 4) sketching ability, and 5) choice of colors (Al-Homoud, 2020; Charlet & Damnati, 2017; Chaudhuri et al., 2020, 2021b; Gagnon et al., 2019; Sangkloy et al., 2017; Sarkar & Chakrabarti, 2011; Wang et al., 2017; Xueqing et al., 2018). Moreover, expert's opinion on additional features was also accepted in the survey. Initially, these features were received in a pilot study and later included in feature list of questionnaire. Four features were included namely- 1) process, 2) simplicity, 3) imaginative, and 4) versatility. Process refers to a technique of solving a problem. Simplicity of an image with label pattern of creative responses refers to a representation that is easy to perceive. Imaginative indicates a response illustrating creativity and divergent ideas. Versatility represents a response that is adaptable to newer context. Descriptive statistic of all features is illustrated in Table 4.1.

Table 4.1: Descriptive statistic of features for image with labels

Statistics										
		Relevance	*Uniqueness*	*Clarity*	*Sketching Ability*	*Choiceof Colors*	*Process*	*Simplicity*	*Imaginative*	*Versatility*
N	Valid	71	71	71	71	71	71	71	71	71
	Missing	0	0	0	0	0	0	0	0	0
Mean		1.00	1.04	4.72	4.75	4.72	4.77	4.65	4.76	4.73
Standard Deviation		0	0.2	0.56	0.52	0.53	0.45	0.63	0.52	0.53
Median		1.00	1.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Mode		1	1	5	5	5	5	5	5	5

Frequency of nine features in questionnaire captured from literature and identified from experts in a pilot study are shown in Table 4.2-4.10. It is represented in terms of very important, slightly more important, important, slightly important, and not at all important. The summarized form of captured features derived from descriptive statistics is illustrated in Figure 4.5. In context of mass examination, most of the subjects selected relevance and uniqueness of a response illustrated in dark green colour. Few subjects marked uniqueness of a response as slightly more important. Other features such as clarity, sketching ability, choice of colours, process, simplicity, imaginative, and versatility were chosen as important and slightly important by a few subjects marked in light green and yellow, respectively; while most of them chosen those as not at all important are marked in sky blue. Therefore, relevance and uniqueness of a response were considered as input to the proposed model to evaluate novelty from labelled images due to their relative higher frequency. The internal consistency of the questionnaire measured by Cronbach's alpha was found to be *0.702* (Gil-Gómez et al., 2017).

Table 4.2: Frequency of relevance in image with label pattern of creative responses

Relevance					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	very important	71	100.0	100.0	100.0

Table 4.3: Frequency of uniqueness in image with label pattern of creative responses

Uniqueness					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	very important	68	95.8	95.8	95.8
	slightly more important	3	4.2	4.2	100.0
	Total	71	100.0	100.0	

Table 4.4: Frequency of clarity in image with label pattern of creative responses

Clarity					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	4	5.6	5.6	5.6
	slightly important	12	16.9	16.9	22.5
	not at all important	55	77.5	77.5	100.0
	Total	71	100.0	100.0	

Table 4.5: Frequency of sketching ability in image with label pattern of creative responses

Sketching ability					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	12	16.9	16.9	21.1
	not at all important	56	78.9	78.9	100.0
	Total	71	100.0	100.0	

Table 4.6: Frequency of choice of colours in image with label pattern of creative responses

Choice of colours					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	14	19.7	19.7	23.9
	not at all important	54	76.1	76.1	100.0
	Total	71	100.0	100.0	

Table 4.7: Frequency of process in image with label pattern of creative responses

Process					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	1	1.4	1.4	1.4
	slightly important	14	19.7	19.7	21.1
	not at all important	56	78.9	78.9	100.0
	Total	71	100.0	100.0	

Table 4.8: Frequency of simplicity in image with label pattern of creative responses

Simplicity					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	6	8.5	8.5	8.5
	slightly important	13	18.3	18.3	26.8
	not at all important	52	73.2	73.2	100.0
	Total	71	100.0	100.0	

Table 4.9: Frequency of imaginative in image with label pattern of creative responses

Imaginative					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	11	15.5	15.5	19.7
	not at all important	57	80.3	80.3	100.0

	Total	71	100.0	100.0	

Table 4.10: Frequency of versatility in image with label pattern of creative responses

Versatility					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	13	18.3	18.3	22.5
	not at all important	55	77.5	77.5	100.0
	Total	71	100.0	100.0	

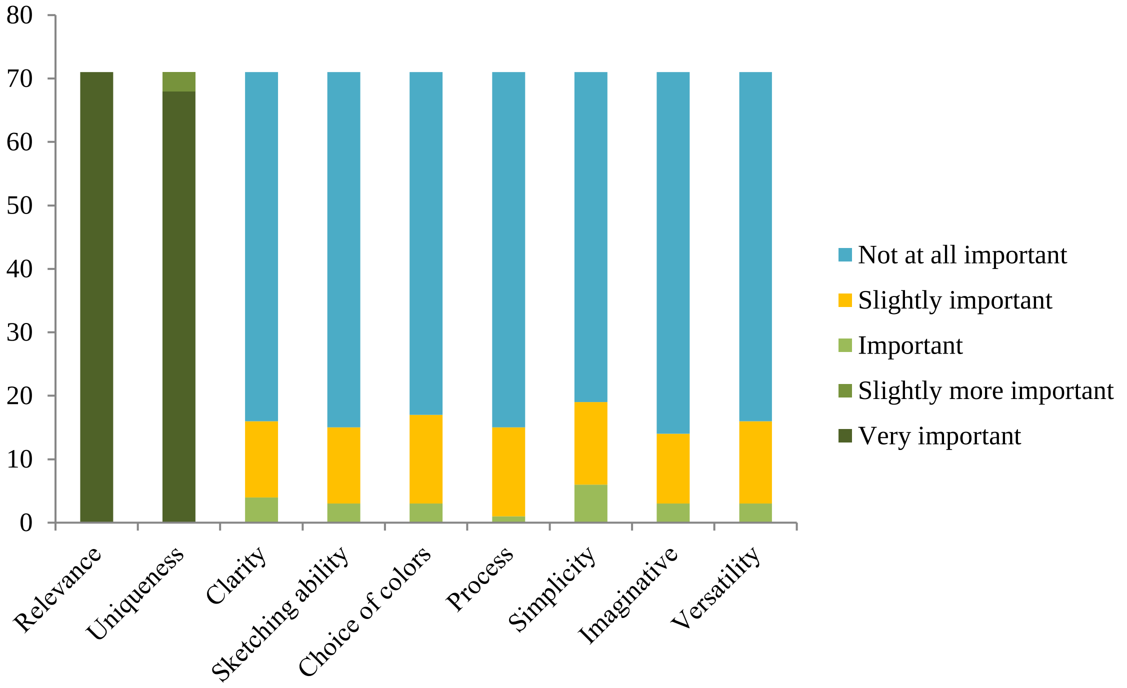


Figure 4.5: Summary of preference of features of image with labels

4.3.3 Result of prediction of image with labels

For training the data on VGG-19 model, *20,539* images were considered from NUS-WIDE dataset. They were pre-processed and converted into sketches. The sketches were randomly distributed in a proportion of *80%* and *20%* for training and validation, respectively. The number of training samples was *16,431* sketches with *81* multi-class labels, and the number of validation samples was *4,108* sketches with *81* multi-class labels. The training accuracy obtained was about *64.28%*, with a loss of *0.0773* after *30* epochs by using image augmentation technique to reduce overfitting of model and considering a batch size of *128* images. The testing accuracy of *57.31%* was obtained with a loss of *0.1114* after *30* epochs by using image augmentation technique to reduce overfitting of model and considering a batch size of *128* images. The results of few predicted test images are illustrated in Figure 4.6.

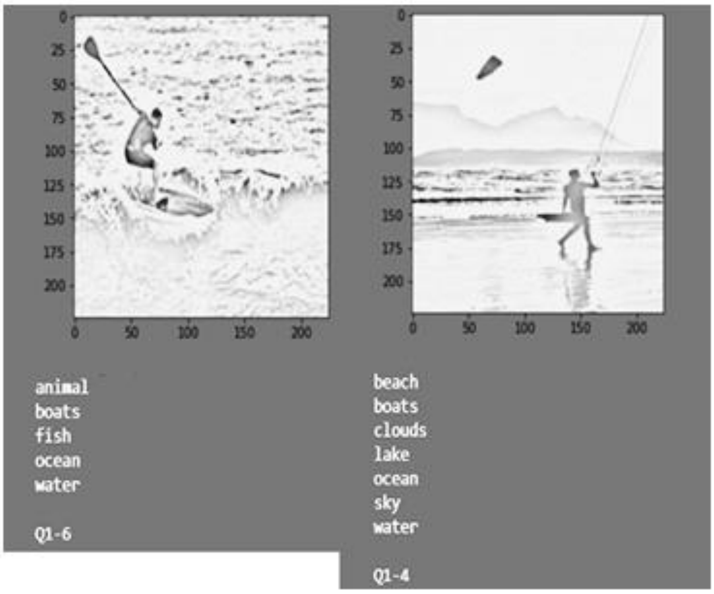


Figure 4.6: Prediction of labelled images

The prediction illustrates the images and class labels. However, input to the model was images in labelled form, precisely each part of image of image was marked. Using an online OCR application (*Best Free OCR API, Online OCR, Searchable PDF - Fresh 2021 On-Premise OCR Software*, 2021), texts were recognized in images. Bounding boxes were created around text in images using inverse masking. Further, these bounding boxes were removed using inpainting technique using pre-defined functions in OpenCV and filled the area within bounding box with the average of neighbouring pixels up to a certain radius (*Inpainting — OpenCV 2.4.13.7 Documentation*, 2021; *OpenCV: Image Inpainting*, 2021). However, the threshold of radius was pre-defined and not manipulated in present context. Outcome suggests descent results. The overall steps of this pre-processing is illustrated in Figure 4.7.

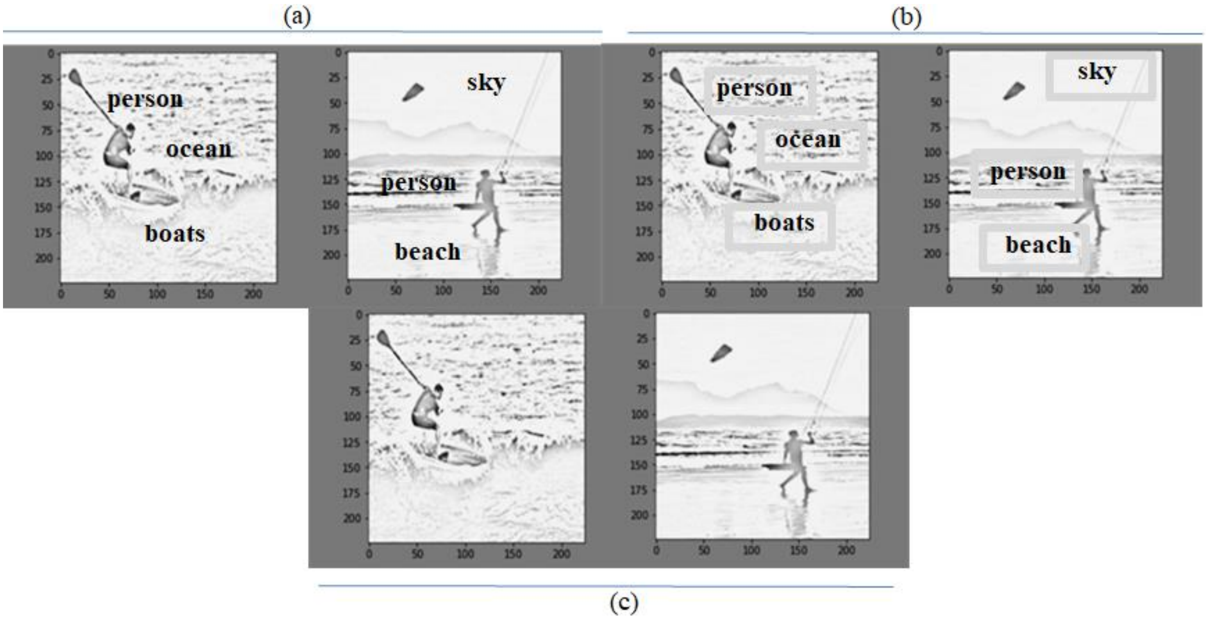


Figure 4.7: (a) Image with label responses, (b) image with bounding boxes, and (c) removal of bounding box and filling up with average of neighbouring pixels

4.3.4 Relevance score between question and image with label pattern of creative responses

Questions were converted into embeddings using doc2vec model. Image predictions and their labels were also converted into embeddings with doc2vec model using gensim (Kim et al., 2019; Trieu et al., 2017). Further, embeddings of questions and their corresponding creative responses were matched using cosine similarity function (Abdel-Basset et al., 2019; Tao et al., 2017). This function generated a score between *0* and *1*. Values close to *0* indicate less relevance between a question and a response, whereas values close to *1* indicate more relevance between a question and a response. The test results of image with labels was combined from two different sources viz., NUS-WIDE dataset and Google scraped data. There were sufficient train and test data available from NUS-WIDE dataset, but the scraped data with the same schema was supplied to the model to verify whether it can perform effectively for any random test sample possessing similar types of labels (Cesa-Bianchi et al., 2010). All samples were pre-processed, and evaluation was performed. The test results displayed here comprise of the first *50%* NUS-WIDE test dataset, and the next *50%* Google scraped test dataset. The relevance scores between a question and their corresponding responses are illustrated in Table 4.11. Ten responses are considered, and their relevance scores within the range of *0* to *1* are mapped for a given question.

Table 4.11: Relevance score between question and labelled image pattern of creative responses

Question-answer ID	**Relevance score**
Q1-1	0.612
Q1-2	0.499
Q1-3	0.548
Q1-4	0.616
Q1-5	0.548
Q1-6	0.547
Q1-7	0.304
Q1-8	0.629
Q1-9	0.481
Q1-10	0.296

Further, decision-making was required to identify a threshold value below which responses could be discarded. Relevance score of human experts and cosine similarity function was considered, where human score was considered as a golden standard. A threshold between *0.2* and *0.7* was considered, and their corresponding F-measure was found where the value corresponding to a high range of F-measure value was considered as the threshold (Penumatsa et al., 2006). The range of thresholds and their corresponding F-measure value is shown in Table 4.12.

Table 4.12: F-measure value of corresponding range of threshold

				Threshold		
	0.2	0.3	0.4	0.5	0.6	0.7
F-measure	1.0	0.99	0.97	0.96	0.77	0.19

Threshold value also depends on multiple factors such as type of examination, level of examination, etc. Type of examination indicates national, state or institutional examination (Kwon et al., 2017). Level of examination refers to difficulty of an examination (Rogers et al., 2019). Depending on selection criteria and level of proficiency of a test, an examination may have numerous difficulty levels. Initially, *0.2* was considered as the threshold in order to select relevant responses whose corresponding F-measure was *1*. It was essential to check manually whether more relevant responses were filtered out or not. With this threshold, there were only fewer matches between the question and their responses. Situations like this might increase frustration in students as most of the responses got rejected with a lesser threshold.

Result suggests that an increase of threshold leads to a decrease of F-measure values. Though *1.0* is a perfect F-measure value, due to more number of rejections of responses, the next threshold, i.e., 0.3, was considered. Manual observation suggested that with this threshold as well too many relevant responses were filtered out. There was a chance that evaluation may become too stringent and might filter out many responses that were close to the question. Often there are situations in students where they understand a concept but lack in expressing them. Pedagogues perceive these situations and may provide leniency in evaluation (Penumatsa et al., 2006). Finally, *0.4* was considered as a threshold with a corresponding *0.97* F-measure, which was similar to human-based evaluation. Image with label pattern of creative responses that were less that than *0.4* were not further considered for evaluating novelty. Responses equal to or above *0.4* were processed for finding uniqueness in a response. However, the threshold completely depends on the type and level of a test and is subject to change.

4.3.5 Results of clustering image with label pattern of creative responses and novelty scores

Initially, relevant image with label pattern of creative responses was clustered, which grouped semantically similar responses together. Dense cluster indicated similar concepts were presented by multiple students, whereas sparse cluster represented less repetition of concepts, thereby relatively novel responses. Initially, K-means algorithm was used for clustering high-dimensional vectors of image with label pattern of creative responses. Results of cluster by K-means algorithm of *3* runs by representative responses for a single question are shown in Table 4.13. Result suggests frequency of a number of responses in a cluster varies on each run of the algorithm. In first run, two clusters were generated viz., *0* and *1* with *8* elements in first cluster and *2* elements in second cluster, respectively. In second run, three clusters were generated viz. *0, 1*, and *2* with *6* elements in first cluster, *2* elements in second cluster, and *2* elements in third cluster, respectively. In third run, two clusters were generated viz. *0* and *1* with *9* elements in first cluster and *1* element in second cluster, respectively. Moreover, K-means is subject to declaration of initial cluster points. It is also affected by outliers and noisy data (Yu et al., 2018).

Table 4.13: Output of K-means at multiple runs

Run-1 {0:8, 1:2}	**Run-2 {0: 6, 1:2, 2:2}**	**Run-2 {0: 9, 1:1}**
0: Response ID = Q1-1	0: Response ID = Q1-2	0: Response ID = Q1-1
0: Response ID = Q1-2	0: Response ID = Q1-3	0: Response ID = Q1-2
0: Response ID = Q1-3	0: Response ID = Q1-4	0: Response ID = Q1-3
0: Response ID = Q1-4	0: Response ID = Q1-5	0: Response ID = Q1-4
0: Response ID = Q1-5	0: Response ID = Q1-6	0: Response ID = Q1-5
0: Response ID = Q1-6	0: Response ID = Q1-7	0: Response ID = Q1-6
0: Response ID = Q1-10	1: Response ID = Q1-1	0: Response ID = Q1-7
0: Response ID = Q1-7	1: Response ID = Q1-10	0: Response ID = Q1-8
1: Response ID = Q1-8	2: Response ID = Q1-8	0: Response ID = Q1-10
1: Response ID = Q1-9	2: Response ID = Q1-9	1: Response ID = Q1-9

Due to inconsistent number of clusters in each run, further affinity propagation algorithm was used to form semantically similar clusters. It was used to generate a stable set of clusters. This algorithm showed improved results than traditional clustering algorithms. Clusters were generated in relatively lesser time with large datasets. Literature suggests that affinity propagation has been specifically used for clustering images (Wei et al., 2017). Further, scores for responses in each cluster were calculated. Clusters possessing a relatively lesser frequency of responses scored relatively higher. The results of affinity propagation in clustering images with label pattern of creative responses and their corresponding scores are illustrated in Table 4.14.

Two questions where each of them possessing ten creative responses were considered to display test results using affinity propagation algorithm and their corresponding scores. Four clusters were generated after the execution of the algorithm, viz., *0*, *1*, *2*, and *3* with *7* elements in first cluster, *6* elements in second cluster, *3* elements in third cluster, and *4* elements in fourth cluster. Score was calculated by number of responses in a cluster divided by total number of responses and further subtracted from *1*.

Table 4.14: Representative clusters of image with label pattern of creative response with corresponding scores

Image with label responses	Cluster number	Scores
Response ID = Q2-5	Cluster assigned = 0	Score obtained = 0.5
Response ID = Q2-1	Cluster assigned = 1	Score obtained = 0.7
Response ID = Q2-10	Cluster assigned = 1	Score obtained = 0.7
Response ID = Q2-3	Cluster assigned = 0	Score obtained = 0.5
Response ID = Q2-4	Cluster assigned = 1	Score obtained = 0.7
Response ID = Q2-6	Cluster assigned = 0	Score obtained = 0.5
Response ID = Q2-7	Cluster assigned = 2	Score obtained = 0.8
Response ID = Q2-8	Cluster assigned = 0	Score obtained = 0.5
Response ID = Q2-2	Cluster assigned = 0	Score obtained = 0.5
Response ID = Q2-9	Cluster assigned = 2	Score obtained = 0.8
Response ID = Q3-8	Cluster assigned = 3	Score obtained = 0.6
Response ID = Q3-10	Cluster assigned = 0	Score obtained = 0.8
Response ID = Q3-2	Cluster assigned = 1	Score obtained = 0.7
Response ID = Q3-7	Cluster assigned = 1	Score obtained = 0.7
Response ID = Q3-9	Cluster assigned = 1	Score obtained = 0.7
Response ID = Q3-6	Cluster assigned = 3	Score obtained = 0.6
Response ID = Q3-4	Cluster assigned = 2	Score obtained = 0.9
Response ID = Q3-5	Cluster assigned = 3	Score obtained = 0.6
Response ID = Q3-1	Cluster assigned = 3	Score obtained = 0.6
Response ID = Q3-3	Cluster assigned = 0	Score obtained = 0.8

Further, normalized novelty score for labelled image-based pattern of creative responses is shown in Table 4.15. Here, two questions where each of them possessing ten creative responses were considered to display test results of normalized novelty score for labelled image-based pattern of creative responses. Novelty score was evaluated by the summative assessment (Chaudhuri et al., 2020, 2021b) of relevance score and uniqueness score and subsequently normalized within the range of *0* to *1*. Human scores collected from experts for the same question-response pair are displayed. Negligible difference between the score generated by algorithmic computation and human-score shows the effective outcome of the model.

Table 4.15: Normalized novelty score of image with label pattern of creative responses

Image with label responses	Normalized novelty score (model)	Expert score	Model score – Expert score
Response ID = Q2-7	Score obtained = 0.40	Score obtained = 0.39	0.01
Response ID = Q2-8	Score obtained = 0.38	Score obtained = 0.38	0
Response ID = Q2-9	Score obtained = 0.45	Score obtained = 0.44	0.01
Response ID = Q2-4	Score obtained = 0.31	Score obtained = 0.30	0.01
Response ID = Q2-10	Score obtained = 0.41	Score obtained = 0.41	0
Response ID = Q2-2	Score obtained = 0.50	Score obtained = 0.53	-0.03
Response ID = Q2-3	Score obtained = 0.38	Score obtained = 0.40	-0.02
Response ID = Q2-6	Score obtained = 0.41	Score obtained = 0.42	-0.01
Response ID = Q2-5	Score obtained = 0.34	Score obtained = 0.33	0.01
Response ID = Q2-1	Score obtained = 0.36	Score obtained = 0.36	0
Response ID = Q3-8	Score obtained = 0.51	Score obtained = 0.50	0.01
Response ID = Q3-10	Score obtained = 0.57	Score obtained = 0.57	0
Response ID = Q3-2	Score obtained = 0.57	Score obtained = 0.57	0
Response ID = Q3-7	Score obtained = 0.59	Score obtained = 0.57	0.02
Response ID = Q3-9	Score obtained = 0.57	Score obtained = 0.56	0.01
Response ID = Q3-6	Score obtained = 0.53	Score obtained = 0.52	0.01
Response ID = Q3-4	Score obtained = 0.60	Score obtained = 0.62	-0.02
Response ID = Q3-5	Score obtained = 0.51	Score obtained = 0.50	0.01
Response ID = Q3-1	Score obtained = 0.46	Score obtained = 0.67	-0.21
Response ID = Q3-3	Score obtained = 0.58	Score obtained = 0.57	0.01

4.3.6 Validation of the proposed model

The proposed models to evaluate novelty from image with label require validation. It is performed by computing Mean Absolute Error (MAE) (Mayer & Butler, 1993). It is an evaluation metric that computes error between a model's predicted outcomes versus output of human scores (Cruz-Ramirez et al., 2014; Willmott & Matsuura, 2005). MAE is calculated using Equation 4.2.

$$Mean\ Absolute\ Error, MAE = 1/\mathrm{n} \sum_{i=1}^{n} (human\_score_i - model\_score_i) \qquad (4.2)$$

Where n is the total number of scores awarded across all creative responses for a given question, human_score is the score awarded by human experts across all creative responses for a given question, model_score is the score awarded by the proposed model across all creative responses for a given question. This metric was applied to test the performance of evaluating image with labels model, and MAE was found to be *0.015*, as illustrated in Figure 4.8. The difference between manual observation and model predictions was found to be minimal and acceptable in evaluating grades of students (Osmanbegovic & Suljic, 2012). Minimal error indicates that there is a negligible difference between human experts and model predictions, which further optimizes trust in both models.

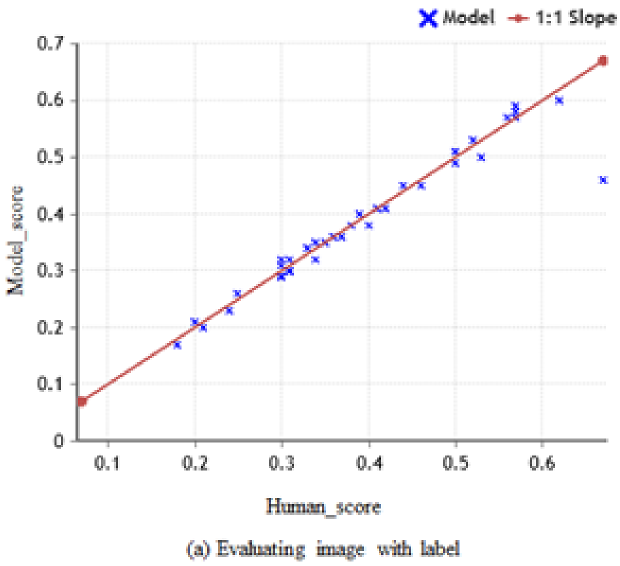


Figure 4.8: Human score vs. model score for evaluating labelled image-based creative responses

The performance-metric of this model being descent may provide support to pedagogues in reducing stress due to repeated tasks of evaluation on a large scale. Often errors are introduced in evaluation process due to monotonous and repeated tasks. Moreover, time constraint is another major factor that is a source of error as pedagogues receive stipulated time to complete their evaluation. During this process, they remain ever-inquisitive to know whether evaluation done on a large scale possesses any error or not. Moreover, any error that turns out during evaluation affects students by having a chance of losing any deserving candidate and vice versa. This may further increase frustration in students and reduce trust in examination system.

4.4 Conclusion

This chapter highlighted subjective evaluation of novelty in labelled image-based pattern of creative responses exhibiting creative aptitude in Design education. Evaluation of novelty in responses like this, is still a manual process and depends on pen-and-paper-based technique.

This study attempted in identifying parameters for evaluating novelty in labelled image-based pattern of creative responses exhibiting creative aptitude of Design education by human-centred approach. The findings of a mixed-method study suggest that relevance between question and response and uniqueness of a response are the parameters that support evaluation in mass examination of Design education. A model is proposed that attempted in digitizing the manual process of evaluation based on the factors pedagogues refer to in the process of evaluating novelty. This model is validated by comparative analysis of the outcomes of proposed model and human experts, which confirms the competence of the devised model and establishes trust of pedagogues.

Chapter 5: Identifying parameters to assess novelty of annotated image-based creative responses and digitizing its evaluation process

Abstract

Novelty is a common factor of assessment of creative responses across most of the Design schools. Pedagogues compare creative responses of understudies in mass assessment hopeful admission to Design schools. In these types of examinations, pedagogues face multiple exceptions in subjective evaluation such as errors experienced in assessment because of specified timeline, errors experienced because of long working hours, errors experienced because of stress in performing repeated tasks for a huge scope, etc. To address these type of exceptions, a computational design model is proposed that attempts for digitizing evaluation of novelty in annotated image-based pattern of creative responses. This type of creative response is an agglomeration of images or sketches and descriptions of design. This model is developed using mixed-method research, where features for evaluating novelty of annotated image-based pattern of creative responses are investigated by conducting a survey. Scores were calculated based on the features using Artificial Intelligence techniques. The estimated performance metric of this model uncovers an immaterial contrast between scores of experts and proposed model. This comparison of the proposed model with human experts affirms the strength of this model. This analysis would support in increasing trust of pedagogues by guaranteeing decreased mistakes and stress during the assessment cycle.

Highlights

- *Human-centred design approach to identify parameters of evaluating novelty in annotated image-based pattern of creative responses.*
- *Proposing a computational design model for evaluating novelty in annotated image-based pattern of creative responses.*
- *Implementing the model using various tools and algorithmic techniques.*
- *Validating the model by comparing its outcome with human evaluation.*

5.1 Introduction

Novelty is an important parameter of evaluation in Design education that pedagogues often look out for (Demirkan & Afacan, 2012). This chapter highlights evaluation of novelty in annotated image-based pattern of creative responses. These type of response is an agglomeration of images or sketches and design descriptions that require subjective evaluation. An example of annotated image-based pattern is illustrated in Figure 5.1. Pedagogues assessing subjective answers evaluate novelty of student's responses based on their opinion, "subjective evaluation index" of their experience and analytical ability, knowledge, and persuasion (Alexiou et al., 2018; Brabb & Morrison, 1964; Furusho & Kotani, 2017; Wan et al., 2018).

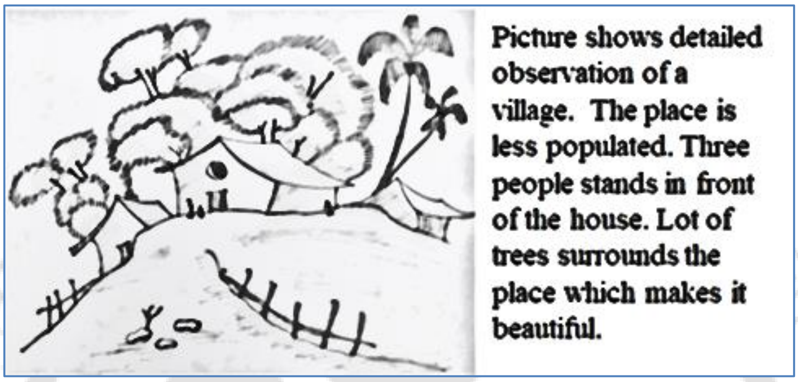


Figure 5.1: Annotated imaged based pattern of creative responses

Evaluating novelty in annotated sketches is a complex process. Generally, novelty in a response like this is defined as newness and originality and which has not been expressed earlier but maybe motivated from any point of reference (Sarkar & Chakrabarti, 2011). Evaluation of novelty is not absolute, but it is assessed relative to other responses. A creative response is compared with other responses to perceive its uniqueness in order to define its novelty (Sarkar et al., 2007). Comparing a creative response with other responses for a given question is tedious. Such repeated tasks of educators increment chances of frustration that may prompt mistakes in assessment of novelty. Errors can likewise happen in a subjective assessment process because of contrasts in referential measurements of individual educators or perhaps for unvarying repeated assessment tasks. Stress generated due to repeated undertaking task might be one more variable for entry of errors in assessment process (Chan et al., 2010; Montgomery & Rupp, 2005). It is necessary to address these issues in novelty scoring mechanism that arises due to stress and referring to experts' own referential metrics for a consistent evaluation process. This triggers two activities viz., identifying features to evaluate novelty in annotated image-based pattern of creative responses exhibiting creative aptitude and proposing a model that attempts in digitizing the process of evaluation based on these features. The investigation in this chapter attempts to address the research gaps highlighted and reported in the state-of-the-art literature review presented in subsection 1.5.3, subsequently corresponding research questions and objectives reported in section 1.9 and 1.12. The research questions and the objectives are stated below again for reference.

RQ3: *While assessing creative responses of students, what are the factors that Design educators consider for assessing novelty of the responses?*

RQ4: *Can the existing process of subjective manual evaluation of creative aptitude be automated?*

Objective 6: *To identify the factors of novelty in creative aptitude evaluation.*

Objective 7: *To design a digitized system for novelty assessment in creative aptitude.*

Mixed-method research approach had been adopted in this research that attempted in investigating features to evaluate novelty. First, a survey was conducted to identify and define features for evaluating novelty. These features were then used to conceive a model that evaluates and scores novelty of responses. The proposed model initially intended at identifying whether an annotated image-based pattern of creative responses is relevant to a question or not. Image predictions and labels from images were mapped to vectors (Gutiérrez & Keith, 2018). Further language processing was conducted on the textual descriptions that were accompanied by images. Further, relevant responses were clustered based on their semantic similarity to determine uniqueness. Clusters that possessed fewer images were considered to be relatively more unique by the proposed model. A scoring mechanism based on relative uniqueness of a response was then used for evaluating novelty. The overview of the model is illustrated in Figure 5.2.

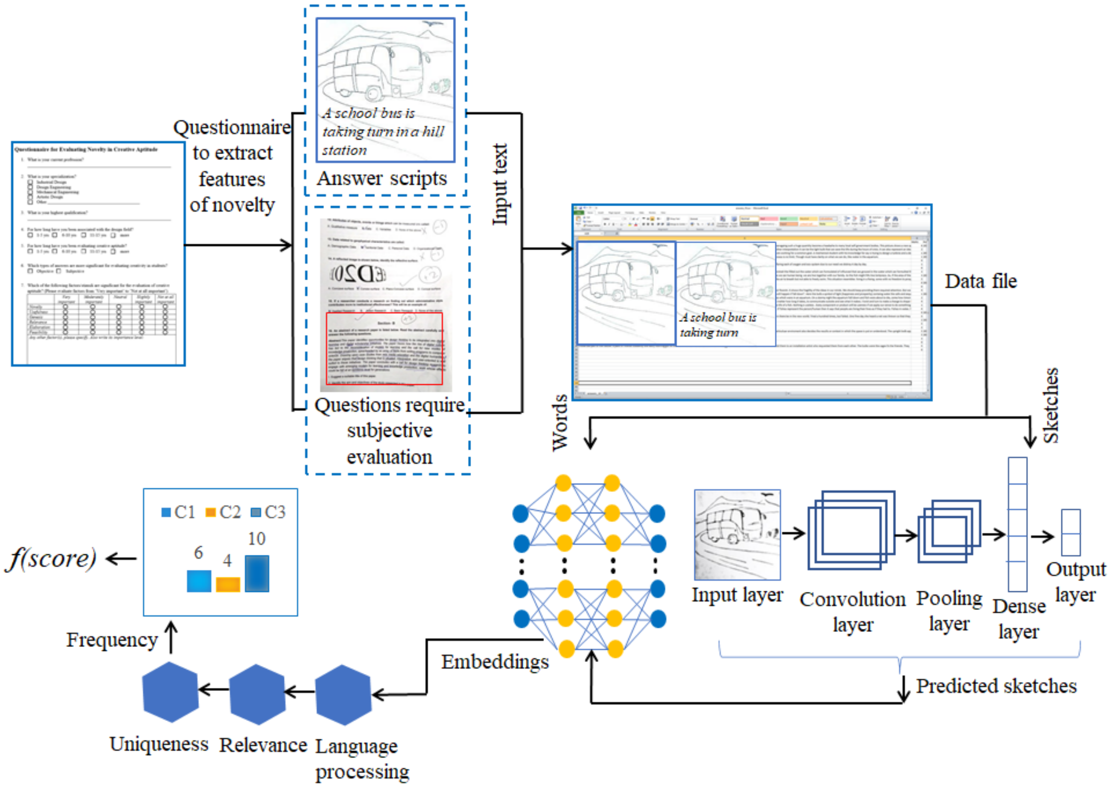


Figure 5.2: Overview of the model

5.2 Method

5.2.1 Survey plan and participants

The survey plan and demographics of subjects recruited for the survey are already described in subsection 4.2.1.

5.2.2 Questionnaire and parameters of evaluation

The details of questionnaire and the parameters of evaluating novelty are described in subsection 4.2.2. The questionnaire is shown in Appendix D. This questionnaire intends to identify the factors to assess annotated image-based creative responses. The questionnaires were written in the way (provided in Appendix D) to confirm whether the factors identified in literature review ascertain to the factors considered by Design pedagogues in practice. State of the art review was conducted to examine the findings from literature that contributed to the identification of factors that are referred for assessing products, solutions, ideas, etc. as illustrated in Table 3.1. User rating ratings were collected using 5-point Likert-type scale to identify the factors preferred by experts in the evaluation process, as illustrated in Appendix C.

5.2.3 Detailed architecture for evaluating image with annotations

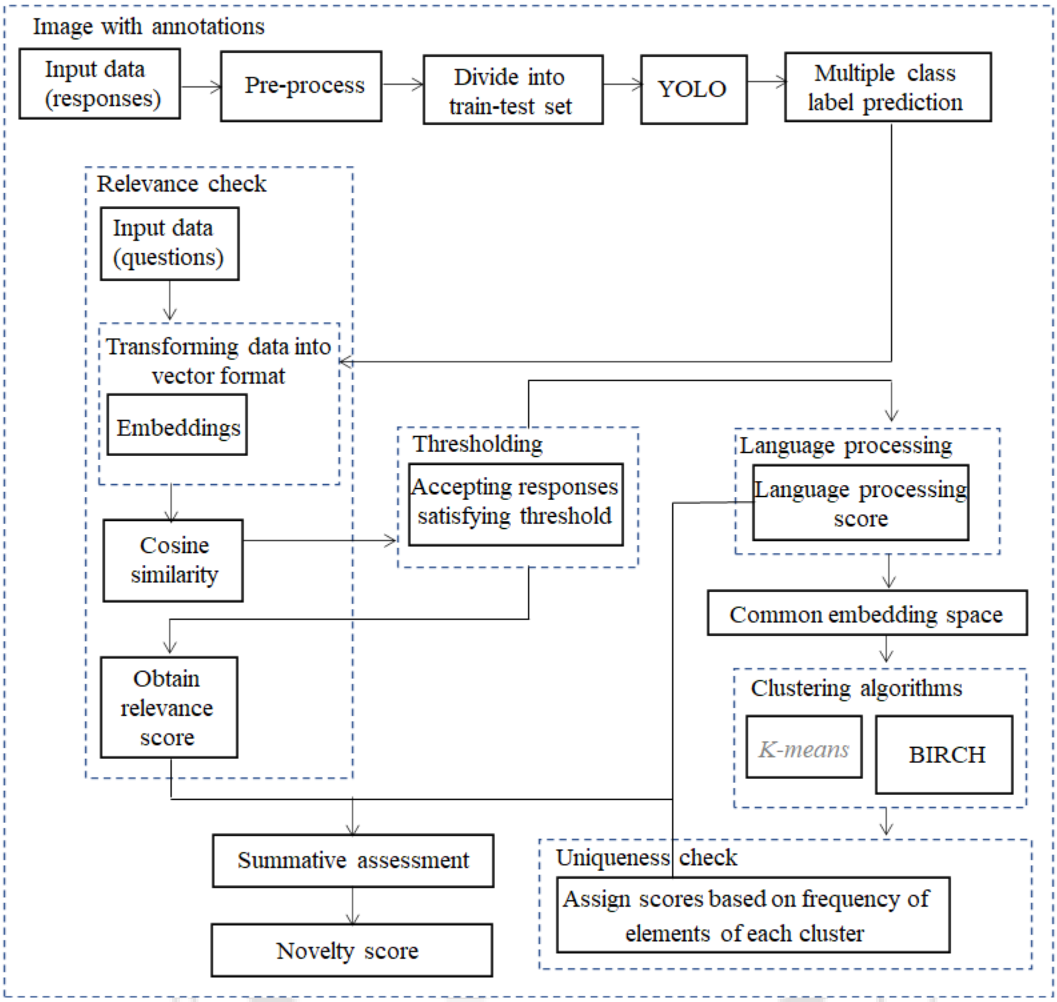


Figure 5.3: Detailed architecture for evaluation of image with annotations

The detailed architecture to evaluate novelty of image with annotations associated with Design education involve multiple procedures as illustrated in Figure 5.3. Initially, annotated image-based pattern of creative responses were considered as input to the model. The responses were pre-processed and divided into train-and-test set for predicting the responses by You Only Look Once (YOLO) architecture (Huang et al., 2018; Redmon et al., 2015). Further, it was essential to check relevance between questions and responses. Therefore, questions and responses were converted into high dimensional vectors. Cosine similarity function was used to compare question vector with response vector. A thresholding technique was applied, which supported identifying a value of the cosine similarity function that could be accepted as a relevant response. Relevant responses comprising of image and annotation were transformed into a common embedding space. Further, they were clustered using K-means algorithm to group semantically similar responses. However, this algorithm generated inconsistent number of clusters in each run of the algorithm (Cao et al., 2009). Therefore, BIRCH algorithm was implemented in order to get a stable set of clusters (Lorbeer et al., 2018). Uniqueness score was calculated based on relative frequency of clusters. Finally, summative assessment of relevant score, language processing, and uniqueness score was conducted to evaluate novelty score.

5.2.4 Dataset for image with annotations

Multiple datasets associated with images and their features were investigated but weren't utilized in this context due to lack of descriptions along with images. Repositories found were mostly single or multi-labelled classified images (Benenson et al., 2019; Eitz et al., 2012; Krasin et al., 2017; Kuznetsova et al., 2020; Pont-Tuset et al., 2020; Russakovsky et al., 2015). Images were collected from online repositories that closely resembles to the type of sketches and drawings that are generally made by students appearing for CEED entrance exams. Types of questions from CEED exam papers of previous years were considered to identify relevant images from the online repositories. MSCOCO (Microsoft Common Objects in Context) dataset is an image repository comprising approximately *1,00,000* images, where each image has five annotations. This dataset attempts to include all common objects and stuff and is similar to the responses of the Design students made in entrance examinations. It contains *80* categories of objects comprising umbrellas, handbags, tennis rackets, and many more. It also contains *91* stuff categories comprising things with indistinct boundaries like the sea, sea-shore, sky, and many more (Karpathy & Fei-Fei, 2015). An example question from CEED and their responses from manual crowdsourcing and MSCOCO responses are illustrated in Appendix E.

Further, images at a higher label were categorized as follows- 1) "iconic object images", 2) "iconic scene images", and 3) "non-iconic images". A focussed wide object in center of an image represented "iconic object images". Scenes of indoor or outdoor, usually without humans, were illustrated in "iconic scene images". "Non-iconic images" were a combination of scenes, people, animals, etc. Human resources in Amazon Turk were utilized for annotating the dataset. Due to large volume of dataset, manual annotation was computationally expensive. Therefore, annotation was categorized into three stages as follows- 1) category labelling in images, 2) instance spotting and marking instances belonging to labelled categories, and 3) object instances were segmented (Lin et al., 2014).

5.2.5 Pre-processing of data for image with annotations

The MSCOCO dataset was considered as input response consisting of an image with annotations. Images were pre-processed to convert into sketches. RGB images were transformed into greyscale by computing negative of an image. Further, scipy Gaussian filter was applied for converting an image into a sketch (*Scipy.Ndimage.Gaussian_filter — SciPy v1.7.1 Manual*, 2021). All sketches had different dimensions, due to which they were resized into a common dimension of *(224x224)*. After pre-processing and removing noisy sketch files, there were approximately *1,00,000* sketches of shape *(224x224x3)*. This data was randomly partitioned as *80%* and *20%* for training and validation, respectively. Each sketch was annotated with five descriptions. This dataset was randomly partitioned as *80%* and *20%* for training and validation, respectively. Specifically, around *80,000* sketches were used for training, and *20,000* sketches were utilized for validation, having *80* object categories and *91* stuff categories.

5.2.6 Prediction of image with annotations

Multi-label object recognition was essential to identify multiple elements in annotated image-based pattern of creative responses. YOLO architecture was used for this purpose as it is computationally fast and approximately processes *155* frames per second. It uses darknet architecture in background for multiple object detection. The network comprehends generalized representation of objects and thus becomes feasible to identify any real-time solutions. It outperformed many detection approaches like Deformable Parts model and Recurrent Convolution Neural Network. Object recognition in YOLO model is characterized as a regression problem where spatially distinct bounding boxes are created for each object and probabilities are calculated directing towards a particular class label (Huang et al., 2018; Redmon et al., 2015).

The network design comprised *24* convolution layers with a subsequent connection of *2* fully connected layers. A *1x1* reduction layer was used with subsequent connections to *3x3* convolution layers. It was a pre-trained model by ImageNet dataset where the convolutions were trained on half the resolution of *(224x224)* image, and further, the resolution was doubled for detection (Redmon et al., 2015). For the purpose of evaluating novelty of annotated images, this model was trained on MSCOCO dataset with approximately *1,00,000* images, which had features of bounding box coordinates, annotations of images, image information, and segmentation coordinates. The output layer of the network computed class label probabilities and coordinates of bounding boxes. Linear activation function was applied in final layer, and rest of the layers used "leaky rectified linear activation" function as shown in Equation 5.1 (Li et al., 2018). This function returns 'a' on receiving positive input, otherwise returns *0.1* multiplied by 'a'.

$$\varphi(a) = \begin{cases} a, & a > 0 \\ 0.1a, & otherwise \end{cases} \tag{5.1}$$

5.2.7 Relevance between question and annotated image-based pattern of creative responses

For testing the model with dataset, image with their descriptions was scrapped from Google images which had similar class labels as MSCOCO data. There was no necessity to pre-process images and their annotations as they were in separate files like they were in training set. Input questions were scraped from website of the Common Entrance Exam from Design (CEED) (Bombay, 2021a), and they were converted into high dimensional vectors. The responses were also converted into higher dimensional vectors. Question and their corresponding answer vectors were compared to measure relevance score between question and responses. In entrance examination of Design education, relevance between a question and a response is essential. A response might be novel, but it must meet the requirements of a question (Chaudhuri et al., 2020, 2021b; Sakata et al., 2019). Hence, question and their response vectors were compared by cosine similarity function (Fauzi et al., 2017). It generated a relevance score within the range of *0* and *1*. Scores that tend to *1* are relatively relevant and vice versa. Based on a threshold value, relevant responses were accepted for further processing, and irrelevant responses were discarded.

5.2.8 Thresholding of annotated image-based pattern of creative responses

Decision-making was essential to identify a score that would distinguish a response, whether it is relevant or irrelevant. A threshold value was needed to be defined for an entrance examination to identify responses as relevant or irrelevant. Multiple questions and their corresponding responses were considered for calculating threshold. Two experts *(N=2)* having expertise in evaluating answer scripts of Design entrance examination were selected to score annotated image-based pattern of creative responses. Human scores were considered a golden standard. Further, scores from cosine similarity function and expert scores were considered to calculate F-measure. A threshold value from *0.3* to *0.9* was considered, and their corresponding F-measure was calculated. A threshold corresponding to a range of high F-measure values was accepted as the threshold for segmenting relevant and irrelevant responses (Penumatsa et al., 2006).

5.2.9 Language processing of annotations

Image-based pattern of creative responses do not require any language processing; however, presence of annotations in responses requires processing it. Here, language processing refers to checking any error in a language, such as spelling and grammar. It is essential in evaluating novelty because language is a medium to convey a response and plays a significant role in persuading novelty to target audience. An online tool was used that returned .json file of mistakes in language and scores (*LanguageTool - Online Grammar, Style & Spell Checker*, 2019). It specifically identified multiple categories of errors such as grammatical error, duplication error, non-conformance of sentences, misspellings, and typological errors. Score was generated based on these categories. The returned score was normalized within the range of *0* to *1*.

5.2.10 Clustering and evaluation of novelty of annotated image-based pattern of creative responses

Clustering was essential in order to group semantically similar creative responses together. Evaluation of novelty can be calculated based on the density of clusters. But responses in this case were in two different formats, i.e., image predictions and descriptions, where one class labels were in the form of words and the other was annotations that were in the form of descriptions. Clustering responses like this format might lead to a problem of image prediction going in a cluster and corresponding annotation getting separately clustered. Further, it would lead to a problem of relative scoring of responses. Therefore, multi-model joint embedding was essential for unifying image predictions and annotations. A deep Convolution Neural Network (CNN) and a Long Short Term Memory recurrent network (LSTM) were utilized for learning image predictions and annotations, as illustrated in Figure 5.4 (Kiros et al., 2014).

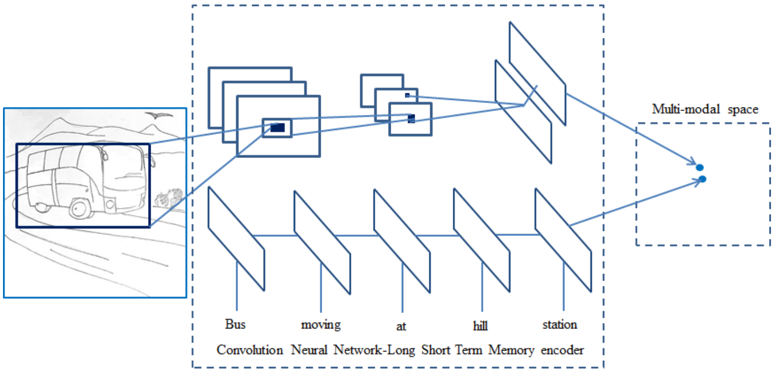


Figure 5.4: Representation of joint image-annotation embedding

The input to the networks were images and their descriptions that explains the image. Here, CNN has been widely used for representing images in higher-level feature space since it has emerged as a powerful tool for image classification and object detection. Recurrent neural networks were extensively used for language modelling. Image features of dimension *4,096* were obtained from the network trained on MSCOCO dataset, and annotations were converted into a matrix that contained all words and their frequencies. This matrix was converted into 1024-dimensional embeddings using an embedding layer provided by PyTorch. Image features were converted into 1024-dimensional embeddings using a fully connected linear layer. Descriptions were mapped to a vector, and text features were brought to a common embedding space of dimension *1024* using LSTM recurrent unit. Pairwise ranking loss was used to bring embeddings of image and description pairs closer. The loss helps penalize such that the image and its corresponding description embedding come in closer proximity. Further, clustering can be performed to assemble semantically similar creative responses.

At first, K-means clustering algorithm was implemented, which had the problem of generating an inconsistent number of clusters at each run of the algorithm. It is also affected by outlier data points (Cao et al., 2009). Therefore, BIRCH clustering algorithm was applied to get a stable set of a cluster (Lorbeer et al., 2018). This algorithm can compute a large volume of a dataset as compared to traditional clustering algorithms (Li et al., 2018). Dense cluster indicates that more number of semantically similar responses were present for a question, whereas, sparse cluster indicates less frequency of semantically similar responses. Clusters possessing a relatively lower frequency of responses were considered unique. Uniqueness score was found using Equation 5.2 (Chaudhuri et al., 2020, 2021b).

$$\text{uniqueness}_{\text{scoreImagewithAnnotation}} = 1 - \left(\frac{NC}{TS}\right) \quad (5.2)$$

Here, $Uniqueness_{scoreImagewithAnnotation}$ represents uniqueness score for annotated image-based pattern of creative responses, NC is a total number of creative responses in a cluster for a given question, and TS is the total number of creative responses across all candidate submissions for a question. Novelty score was evaluated for an annotated image-based pattern of creative solution by summation of three parameters i.e., relevance score, language score, and uniqueness score. Relevance score ranges in any value between *0* and *1*. Language and uniqueness score was also a normalized value between *0* and *1*. A summative assessment (Chaudhuri et al., 2020, 2021b) of all the three scores were conducted for evaluating novelty score, which was further normalized within the range of *0* to *1*.

5.3 Results and discussion

5.3.1 Survey site

The details of the survey site and the demographics of subjects are explained in subsection 4.3.1.

5.3.2 Descriptive statistic of features of image with annotations identified by survey

A questionnaire was framed to capture features for evaluating novelty from an annotated image-based response. A list of features was provided in questionnaire consisting of the following items- 1) relevance, 2) uniqueness, 3) clarity, 4) sketching ability, 5) choice of colours, 6) language processing, and 7) narration (Al-Homoud, 2020; Charlet & Damnati, 2017; Chaudhuri et al., 2020, 2021b; Gagnon et al., 2019; Sangkloy et al., 2017; Sarkar & Chakrabarti, 2011; Z. Wang et al., 2017; Xueqing et al., 2018). Initially, a pilot study was conducted to capture any other additional features to evaluate novelty from annotated image-based pattern of creative responses. The additional five features were as follows- 1) imaginative, 2) subject knowledge, 3) versatility, 4) presentation, and 5) refining. Imaginative indicate any annotated image-based pattern of creative responses representing creative and divergent ideas. Subject knowledge refers to adequate perception of domain of interest. A response that is adaptable based on requirement is considered versatile. Presentation is a feature in which a response is illustrated in a clear and concise manner. Refining indicates the iterative process by which a response attempts to improve gradually. Descriptive statistic of all features is illustrated in Table 5.1.

Table 5.1: Descriptive statistic of features for image with annotations

Statistics													
		Relevance	*Uniqueness*	*Clarity*	*SketchingAbility*	*ChoiceofColors*	*LanguageProces*	*Narration*	*Imaginative*	*SubjectKnowledge*	*Versatility*	*Presentation*	*Refining*
N	Valid	71	71	71	71	71	71	71	71	71	71	71	71
	Missing	0	0	0	0	0	0	0	0	0	0	0	0
Mean		1.00	1.04	4.73	4.75	4.72	1.04	4.68	4.21	4.20	4.73	4.76	4.66
Standard Deviation		0	0.2	0.56	0.52	0.53	0.2	0.6	0.86	0.8	0.53	0.52	0.58
Median		1.00	1.00	5.00	5.00	5.00	1.00	5.00	4.00	4.00	5.00	5.00	5.00
Mode		1	1	5	5	5	1	5	5	5	5	5	5

Frequency of all twelve features in questionnaire captured from literature and identified from experts in a pilot study is illustrated in Table 5.2-5.13. It was captured using Likert-type scale with labels very important=*1*, slightly more important=*2*, important=*3*, slightly important=*4*, and not at all important=*5*. The summarized form of significance of features derived from descriptive statistics is illustrated in Figure 5.5. In the context of mass examination, most of the subjects selected relevance, uniqueness of a response, and language processing feature to evaluate novelty which is represented in dark green colour. Few subjects marked uniqueness of an annotated image-based pattern of creative responses as slightly more important. Other features such as clarity, sketching ability, choice of colours, narration, imaginative, subject knowledge, versatility, presentation, and refining were chosen as important and slightly important by a few of the subjects marked in light green and yellow, respectively; while most of them chose those as not at all important are marked in sky blue. Therefore, relevance, uniqueness, and language processing of an annotated image-based pattern of creative responses were considered inputs to the proposed model due to their relative higher frequency for evaluating novelty. The internal consistency of the questionnaire measured with Cronbach's alpha was found to be *0.703* (Gil-Gómez et al., 2017).

Table 5.2: Frequency of relevance between question and annotated image-based pattern of creative responses

Relevance					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	very important	71	100.0	100.0	100.0

Table 5.3: Frequency of uniqueness in annotated image-based pattern of creative responses

Uniqueness					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	very important	68	95.8	95.8	95.8
	slightly more important	3	4.2	4.2	100.0
	Total	71	100.0	100.0	

Table 5.4: Frequency of clarity in annotated image-based pattern of creative responses

Clarity					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	4	5.6	5.6	5.6
	slightly important	11	15.5	15.5	21.1
	not at all important	56	78.9	78.9	100.0
	Total	71	100.0	100.0	

Table 5.5: Frequency of sketching ability in annotated image-based pattern of creative responses

Sketching ability					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	12	16.9	16.9	21.1
	not at all important	56	78.9	78.9	100.0
	Total	71	100.0	100.0	

Table 5.6: Frequency of choice of colours in annotated image-based pattern of creative responses

Choice of colours					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	14	19.7	19.7	23.9
	not at all important	54	76.1	76.1	100.0
	Total	71	100.0	100.0	

Table 5.7: Frequency of language processing in annotated image-based pattern of creative responses

Language processing					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	very important	68	95.8	95.8	95.8
	slightly more important	3	4.2	4.2	100.0
	Total	71	100.0	100.0	

Table 5.8: Frequency of narration in annotated image-based pattern of creative responses

Narration					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	5	7.0	7.0	7.0
	slightly important	13	18.3	18.3	25.4
	not at all important	53	74.6	74.6	100.0
	Total	71	100.0	100.0	

Table 5.9: Frequency of imagination in annotated image-based pattern of creative responses

Imagination					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	slightly more important	1	1.4	1.4	1.4
	important	17	23.9	23.9	25.4
	slightly important	19	26.8	26.8	52.1
	not at all important	34	47.9	47.9	100.0
	Total	71	100.0	100.0	

Table 5.10: Frequency of subject knowledge in annotated image-based pattern of creative responses

Subject knowledge					
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	slightly more important	1	1.4	1.4	1.4
	important	14	19.7	19.7	21.1
	slightly important	26	36.6	36.6	57.7
	not at all important	30	42.3	42.3	100.0
	Total	71	100.0	100.0	

Table 5.11: Frequency of versatility in annotated image-based pattern of creative responses

		Versatility			
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	13	18.3	18.3	22.5
	not at all important	55	77.5	77.5	100.0
	Total	71	100.0	100.0	

Table 5.12: Frequency of presentation in annotated image-based pattern of creative responses

		Presentation			
		Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
Valid	important	3	4.2	4.2	4.2
	slightly important	11	15.5	15.5	19.7
	not at all important	57	80.3	80.3	100.0
	Total	71	100.0	100.0	

Table 5.13: Frequency of refining in annotated image-based pattern of creative responses

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	important	4	5.6	5.6	5.6
	slightly important	16	22.5	22.5	28.2
	not at all important	51	71.8	71.8	100.0
	Total	71	100.0	100.0	

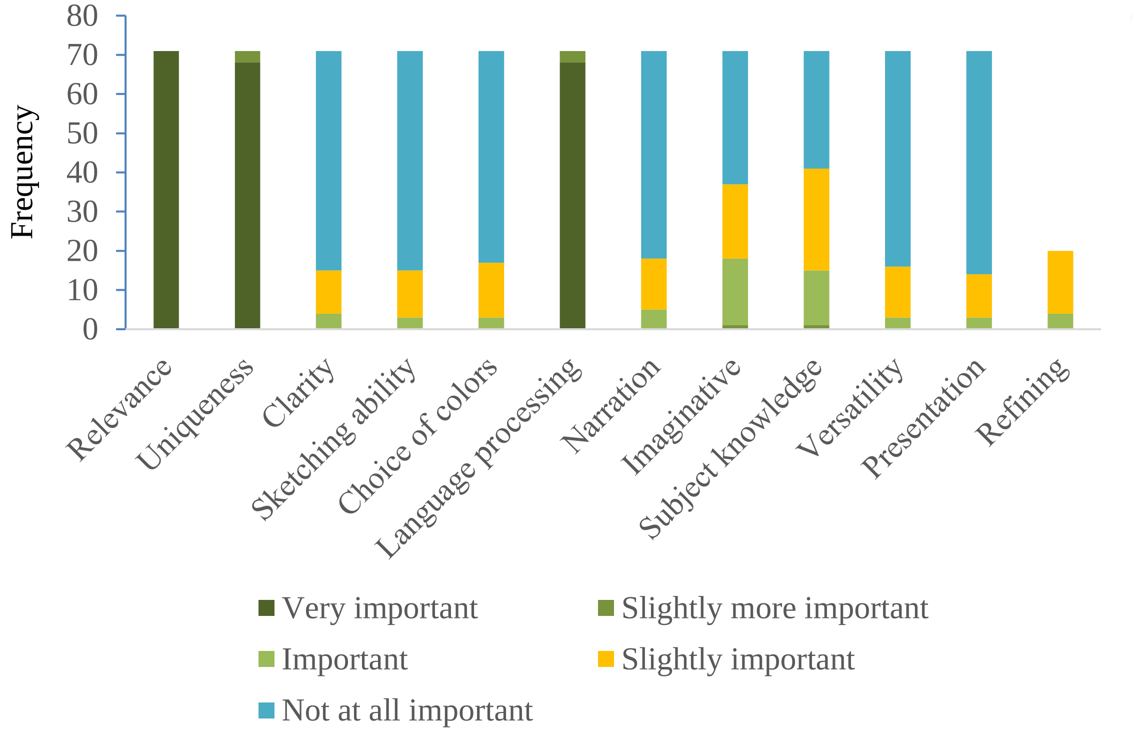


Figure 5.5: Summary of significance of features of image with annotations

5.3.3 Result of prediction of image with annotations

The YOLO model (Gordon et al., 2018) was pre-trained on ImageNet dataset. Further, approximately *1,00,000* image with annotations were considered from MSCOCO dataset. After pre-processing, sketches were randomly distributed in a proportion of *80%* and *20%* for training and validation, respectively. The dataset comprised *80* object categories and *91* stuff categories. Each image was associated with five annotations. The model contained *24* convolution layers followed by two fully connected layers. The last layer i.e., the output layer calculated probabilities of belonging to a particular class and further computed coordinates of bounding boxes of objects present in an image. The prediction of this model is illustrated in Figure 5.6.

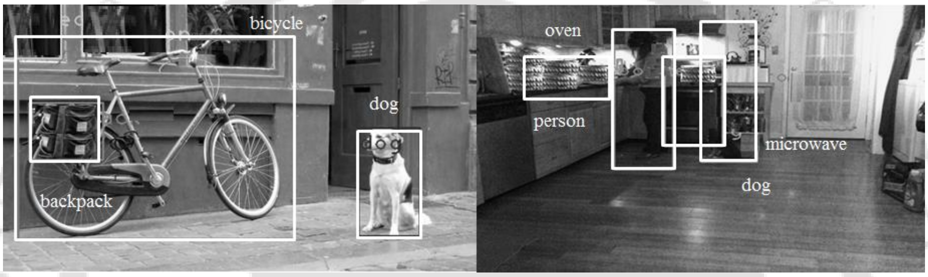


Figure 5.6: Prediction of annotated images with coordinate of bounding boxes

5.3.4 Relevance score between question and annotated image-based pattern of creative responses

A response may be novel, but it is essential to verify whether it meets the requirement of a question in an entrance examination. Therefore, it is necessary to match questions and their response and verify whether responses are intended for the particular question or not. For the purpose of matching between question-response pairs, they were converted into high dimensional embedding by Global Vectors for Word Representation (GloVe) embedding technique (Prost et al., 2019). Further, embeddings of questions and their corresponding response were matched using cosine similarity function (Fauzi et al., 2017). This function generated relevance score within the range of *0* to *1*. The values close to *0* were considered as irrelevant responses, whereas values close to *1* were considered as relevant responses. A sample relevance score between question and its corresponding descriptions is illustrated in Table 5.14. Ten responses are considered, and their relevance scores within the range of *0* to *1* are mapped with a given question.

Table 5.14: Relevance score between question and description in annotated image-based pattern of creative responses

Question-response ID	Relevance score
Q1-1	0.845
Q1-2	0.898
Q1-3	0.852
Q1-4	0.868
Q1-5	0.928
Q1-6	0.870
Q1-7	0.943
Q1-8	0.880
Q1-9	0.904
Q1-10	0.858

The test results of image with annotations was combined from two different sources viz., MSCOCO dataset and Google scraped data. There were sufficient train and test data available from MSCOCO dataset, but the scraped data with same schema was supplied to the model to verify whether it can perform effectively for any random test sample possessing similar types of labels (Cesa-Bianchi et al., 2010). All samples were pre-processed, and evaluation was performed. The test results considered for display in this study comprise the first *50%* MSCOCO test dataset, and the next *50%* Google scraped test dataset.

Major decision-making was required for a relevance score that would suggest a specific score based on which responses would be considered as irrelevant that would be discarded from further evaluation of novelty. Precisely a threshold was required that would be able to determine scores to demarcate relevant and irrelevant responses. A threshold usually depends on experts' choice of how lenient or stringent they are (Aubin et al., 2018). Many a time, if a question paper is too difficult for a cohort of students, then experts attempt to keep the evaluation lenient and vice versa. More often, threshold of evaluation depends on type and level of examination. There may be multiple types of examinations, such as national, state, institutional, and evaluation in this type of examination depend on their individual criteria. Level of examination depends on the difficulty of examination, which may be broadly classified as very easy, easy, moderate, hard, and very hard (Park et al., 2017).

Any threshold in an examination determining either success or failure of students or evaluation of any particular factor can be identified by longitudinal studies (de Vergara & Olmos, 2019). Over the years, one may study pattern of student responses, type of examination, level of examination, level of questions, etc., and further analyse and predict a threshold for evaluation. However, there are other techniques as well by which one may decide upon a threshold value. In this context, a threshold was essential to identify scores which attempt to demarcate relevant and irrelevant responses. To achieve this, experts' scores based on relevance between question and annotated image-based pattern of creative responses were collected, which was considered as a golden standard. Further, it was compared with scores generated by cosine similarity function. F-measure was calculated to evaluate model performance.

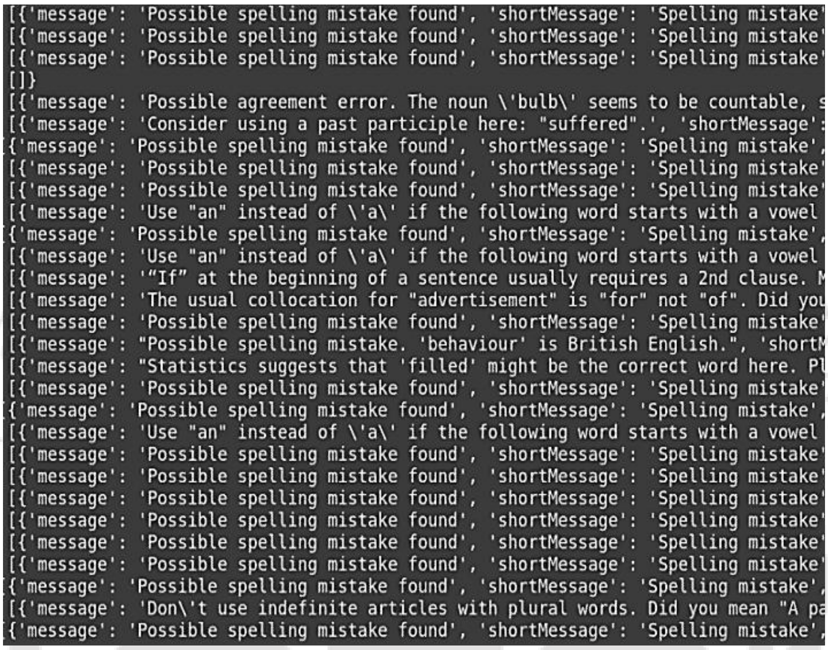
A threshold from *0.3* to *0.9* was considered, and their corresponding F-measures were computed between question and description, question and image, and image and description. The sample outcomes are shown in Table 5.15. In most question-response pairs, a threshold of *0.3* received the highest value of F-measure. In few cases, threshold of *0.4* also received higher F-measure values, but on manual verification, it was observed that approximately *10%* of relevant responses were categorized as irrelevant, which might lead to frustration in students (Penumatsa et al., 2006). Therefore, *0.3* was considered as the threshold for filtering irrelevant responses for further processing of evaluating novelty.

Table 5.15: Threshold and corresponding F-measure of annotated image-based pattern of creative responses

Question-description		**Question-image**		**Image-description**	
Threshold	F-measure	Threshold	F-measure	Threshold	F-measure
0.3	0.783	0.3	0.963	0.3	0.833
0.4	0.667	0.4	1.000	0.4	0.727
0.5	0.683	0.5	0.917	0.5	0.353
0.6	0.607	0.6	0.670	0.6	0.364
0.7	0.367	0.7	0.503	0.7	0.375
0.8	0.450	0.8	0.234	0.8	0.287
0.9	0.500	0.9	0.310	0.9	0.200

5.3.5 Results of language processing of annotations in responses

Descriptions in annotated image-based pattern of creative responses required spelling and grammatical error checking, termed as language processing in this context. Language processing though not directly associated with novelty but supports in identifying novelty in write-up of students. Therefore, it is essential in examination to process it and identify errors associated with spelling and grammar. Language exhibits the sense of a concept and understanding of novelty to its target audience. An online language processing tool (*LanguageTool - Online Grammar, Style & Spell Checker*, 2019) was used to identify multiple categories of grammatical errors such as non-conformance of sentences, duplication, typological errors, and misspellings. The tool returned .json file for mistakes and scores. A normalized score between *0* and *1* was generated for language processing. The .json file returned by the tool is illustrated in Figure 5.7 (Chaudhuri et al., 2021b).



```
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[]}
[{'message': 'Possible agreement error. The noun \'bulb\' seems to be countable, s
[{'message': 'Consider using a past participle here: "suffered".', 'shortMessage':
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Use "an" instead of \'a\' if the following word starts with a vowel
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Use "an" instead of \'a\' if the following word starts with a vowel
[{'message': '"If" at the beginning of a sentence usually requires a 2nd clause. M
[{'message': 'The usual collocation for "advertisement" is "for" not "of". Did you
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': "Possible spelling mistake. 'behaviour' is British English.", 'shortM
[{'message': "Statistics suggests that 'filled' might be the correct word here. Pl
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Use "an" instead of \'a\' if the following word starts with a vowel
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
[{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake'
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
[{'message': 'Don\'t use indefinite articles with plural words. Did you mean "A pa
{'message': 'Possible spelling mistake found', 'shortMessage': 'Spelling mistake',
```


Figure 5.7: Possible errors in annotation

5.3.6 Results of clustering annotated image-based pattern of creative responses and novelty scores

Responses that were considered relevant and post language processing require clustering them based on their semantic similarity. However, since the response was divided into two parts- image and annotation, they required common representation space. Therefore, multi-modal joint embedding was used to represent image predictions and annotations for unifying them. A deep CNN and LSTM recurrent network was used for joint image annotation embedding (Kiros et al., 2014). Further, jointly embedded responses were clustered using K-means clustering algorithm. However, in this case, K-means clustering technique provided inconsistent results in each run of the algorithm. The number of clusters as well as the density of cluster varied on repeated execution. Further, BIRCH clustering algorithm was used to get a stable set of clusters (Lorbeer et al., 2018). It is effective on large datasets. Unlike K-means it can handle outliers

(Zhang et al., 1997). The number of clusters formed after execution of both the algorithms is illustrated in Table 5.16, 5.17. Based on the density of clusters, relative uniqueness scores were computed. The score was calculated by number of creative responses in a cluster divided by the total number of creative responses corresponding to a question subtracted from *1*. Further, algorithmically computed novelty score is shown in Table 5.18.

Table 5.16: Clusters at multiple runs of K-means algorithm for image-based creative responses

Run sequence	**Clusters**
Run 1	[0 0 0 0 0 0 0 0 1 1]
Run 2	[0 0 0 0 0 0 1 1 2 2]
Run 3	[0 0 0 0 0 0 0 0 0 1]

Table 5.17: Clusters using BIRCH algorithm for image-based creative responses

Run sequence	**Clusters**
Run 1	[0 1 1 0 1 0 2 0 0 2 3 0 1 1 1 3 2 3 3 0]
Run 2	[0 1 1 0 1 0 2 0 0 2 3 0 1 1 1 3 2 3 3 0]
Run 3	[0 1 1 0 1 0 2 0 0 2 3 0 1 1 1 3 2 3 3 0]

Table 5.18: Normalized novelty score for annotated image-based pattern of creative responses

QR-ID	**Cluster**	**Normalized novelty score**
ResponseID = Q1-1	Cluster assigned = 0	Relevance Score: 0.467, Uniqueness Score: 0.467, Language Score: 1.0, Normalized novelty Score:0.644
ResponseID = Q1-2	Cluster assigned = 0	Relevance Score: 0.427, Uniqueness Score: 0.467, Language Score: 1.0, Normalized novelty Score:0.631
ResponseID = Q1-3	Cluster assigned = 1	Relevance Score: 0.835, Uniqueness Score: 0.667, Language Score: 1.0, Normalized novelty Score: 0.834
ResponseID = Q1-4	Cluster assigned = 0	Relevance Score: 0.327, Uniqueness Score: 0.467, Language Score: 0.929, Normalized novelty Score: 0.574
ResponseID = Q1-5	Cluster assigned = 2	Relevance Score: 0.330, Uniqueness Score: 0.867, Language Score: 1.0, Normalized novelty Score: 0.732
ResponseID = Q1-6	Cluster assigned = 0	Relevance Score: 0.339, Uniqueness Score: 0.467, Language Score: 0.917, Normalized novelty Score: 0.574
ResponseID = Q1-7	Cluster assigned = 1	Relevance Score: 0.603, Uniqueness Score: 0.667, Language Score: 1.0, Normalized novelty Score: 0.757
ResponseID = Q1-8	Cluster assigned = 1	Relevance Score: 0.616, Uniqueness Score: 0.667, Language Score: 1.0, Normalized novelty Score: 0.760

ResponseID = Q1-9	Cluster assigned = 1	Relevance Score: 0.340, Uniqueness Score: 0.667, Language Score: 1.0, Normalized novelty Score: 0.669
ResponseID = Q1-10	Cluster assigned = 1	Relevance Score: 0.690, Uniqueness Score: 0.667, Language Score: 1.0, Normalized novelty Score: 0.786
ResponseID = Q1-11	Cluster assigned = 0	Relevance Score: 0.497, Uniqueness Score: 0.467, Language Score: 1.0, Normalized novelty Score: 0.655
ResponseID = Q1-12	Cluster assigned = 2	Relevance Score: 0.345; Uniqueness Score: 0.867, Language Score: 1.0, Normalized novelty Score: 0.737
ResponseID = Q1-13	Cluster assigned = 0	Relevance Score: 0.367, Uniqueness Score: 0.467, Language Score: 1.0, Normalized novelty Score: 0.611
ResponseID = Q1-14	Cluster assigned = 0	Relevance Score: 0.374, Uniqueness Score: 0.467, Language Score: 1.0, Normalized novelty Score: 0.0.614
ResponseID = Q1-15	Cluster assigned = 0	Relevance Score: 0.467, Uniqueness Score: 0.467, Language Score: 1.0, Normalized novelty Score: 0.644
ResponseID = Q1-16	Cluster assigned = 1	Relevance Score: 0.605, Uniqueness Score: 0.667, Language Score: 1.0, Normalized novelty Score: 0.757

The displayed test result considered a single question with corresponding sixteen annotated image-based pattern of creative responses. Three clusters were formed viz., *0*, *1*, and *2* with *8* elements in the first cluster, *6* elements in the second cluster, and *2* elements in the third cluster, respectively. Novelty score was measured by the summative assessment (Chaudhuri et al., 2020, 2021b) of relevance score, language, score, and uniqueness score. It was further normalized within the range of *0* to *1*.

5.3.7 Validation of the proposed model

The proposed models to evaluate novelty from an annotated image-based pattern of creative responses required validation. It was performed by computing Mean Absolute Error (MAE) (Mayer & Butler, 1993). It is an evaluation metric that computes error between a model's predicted outcomes versus output of human scores (Cruz-Ramirez et al., 2014; Willmott & Matsuura, 2005). MAE is calculated using Equation 5.3.

$$Mean\ Absolute\ Error, MAE = 1/\mathrm{n} \sum_{i=1}^{n} (human\_score_i - model\_score_i) \quad (5.3)$$

Where, n is the total number of scores awarded across all creative responses for a given question, human_score is the score awarded by human experts across all creative responses for a given question, model_score is the score awarded by the proposed model across all creative responses for a given question. This metric was applied to test the performance of evaluating image with annotation model, and MAE of it was found to be *0.009*, as illustrated in Figure 5.8. The difference between manual observation and model predictions was minimal and acceptable in evaluating scores of students (Osmanbegovic & Suljic, 2012). The training of the models have been satisfactory from the fact that during validation negligible differences were observed between this course of the human experts and the proposed models. Minimal error indicated that there was a negligible difference between human experts and model predictions, which further optimized trust in this model.

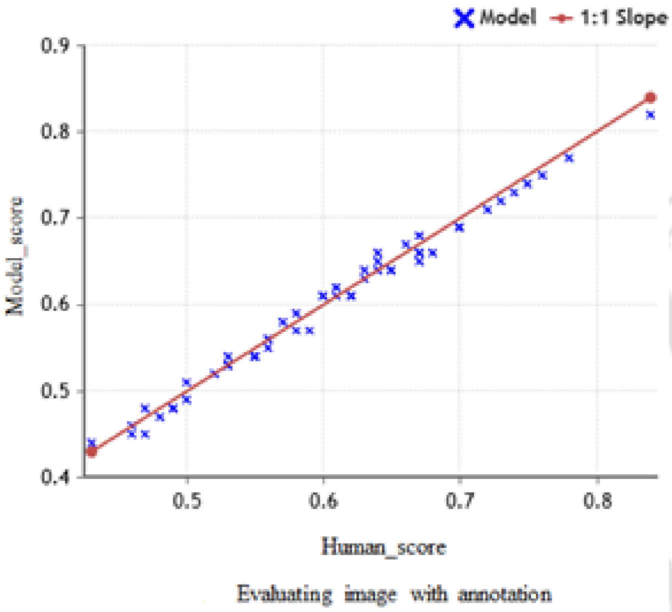


Figure 5.8: Human score vs. model score for evaluating annotated image-based creative responses

The performance-metric of this proposed model being descent may provide support to pedagogues in reducing stress due to repeated tasks of evaluation on a large scale. Often errors are introduced in evaluation process due to monotonous and repeated tasks. Moreover, time constraint is another major factor that is a source of error as pedagogues receive stipulated time to complete their evaluation. During this process, they remain ever-inquisitive to know whether evaluation conducted on a large scale possesses any error or not. Moreover, any error that occurs during evaluation affects students by having a chance of losing any deserving candidate and vice versa. This may further increase frustration in students and reduce trust in examination system. Therefore, this model would support pedagogues in large scale evaluation process.

5.4 Conclusion

This chapter highlighted subjective evaluation of novelty in annotated image-based pattern of creative responses exhibiting creative aptitude in Design education. Evaluation of novelty in creative responses like this, is still a manual process and depends on pen-and-paper-based technique. This study attempts in identifying parameters for evaluating novelty in annotated image-based pattern of creative responses exhibiting creative aptitude in Design entrance examinations. The findings of a mixed-method study suggest that relevance between question and answer, language processing, and uniqueness of a response are the parameters that support assessment of novelty in mass examination. A model is proposed that attempted to automate the manual process of evaluation based on the factors pedagogues refers to in evaluating novelty. This model is validated by comparative analysis of the outcomes of proposed model and human experts, which confirms the competence of the devised model and establishes trust of pedagogues.

Chapter 6: Discussion and conclusion

Abstract

Experimental results and findings of the investigation from "Chapter-2" to "Chapter-5" have been briefly explained in the present chapter with key findings of the thesis. This chapter presents the overall discussion of identifying creative questions and assessing their responses. This chapter also highlights the kind of exceptions handled by the proposed systems. Further, it presents the implications drawn from this research. Accordingly, recommendations are provided to pedagogues of Design education in identifying creative questions that have the potential to instigate creative response from students and evaluating their responses. These suggestions intends to prepare Design education community specifically Design pedagogues to embrace changes in existing ways of framing creative questions and assessing their responses. This chapter also lists the novelties in contribution of this research from the perspective of knowledge-base, methods, design, and design education. This chapter ends with limitations, future scope, and an overall conclusion of the thesis.

Highlights

- *Illustrating the key findings of the thesis.*
- *Elucidating the fulfillment of the objectives.*
- *Demonstrating the validation of models.*
- *Providing implications from multiple perspectives.*
- *Highlighting the novelties in the contribution of the research work.*
- *Highlighting limitations and future scope.*

6.1 Discussion

This thesis highlights the creative aptitude assessment process for Design based educational institutes. Specifically, the assessment is from two major perspectives-1) identifying creative questions that instigate creative response from students, and 2) evaluating responses to these questions that are illustrated by multiple patterns of creative responses. This investigation is directed towards systematically identifying variables of creative questions and evaluation of creative responses. Previous studies pointed out the stress factors of pedagogues due to huge workload in institutions (Boyle et al., 1995; Naghieh et al., 2015; Prilleltensky et al., 2016;

Sanetti et al., 2020; Skaalvik & Skaalvik, 2016, 2017), but hardly any reported literatures were found that addressed frustrations of pedagogues generated due to large scale evaluation of creative responses and issues related to consistency and errors during creative question formulation and solution evaluation. This research addressed this gap by development of a digitized system that attempts at automating the assessment process specifically for large scale Design entrance examinations.

The proposed systems were validated by comparing their outcomes with human evaluation scores. The reliability estimate of the examiners and the proposed model in identifying creative questions was considerably high *(α=0.96)*. The MAE of descriptive, labelled image-based, and annotated image-based pattern of creative responses are *0.085*, *0.015*, and *0.009*, respectively. These experimental results indicate that there is a negligible difference between the evaluation of the proposed systems and their human-based evaluation scores.

Assessment conducted in real-time environment encounters multiple exceptions which were identified and handled by the proposed models. For example, relevance between the question and a response is a significant factor in evaluating creative aptitude. In the proposed model, relevance score between question and response was obtained by using cosine similarity function. However, in some situations it might happen that a student intentionally reproduce the same question and presented it as a response. In that case, the relevance score would be extraordinarily high, leading to miscalculation in the computational model. Therefore, rules are essential to verify a very high relevance score. An overwhelming score requires a matching of string length between question and its corresponding response. Pattern matching of string would also further support the verification process. This case is illustrated in Figure 6.1, where a question-response pair receives extraordinary scores. A response, when passed through a rule-base comprising of measuring string length and pattern matching may subject to change in relevance score.

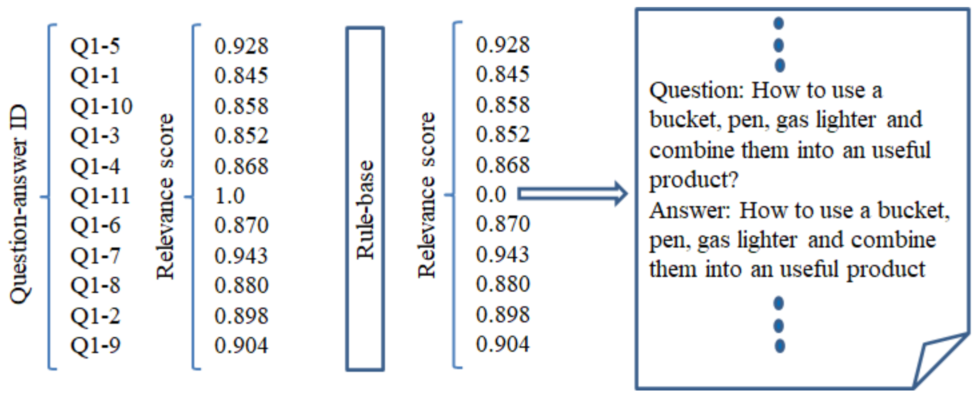


Figure 6.1: Handling faulty relevance score

Another case has been studied, where a student might just put a line or a joke as response to a given question. In that case, embeddings of the question and their corresponding responses were matched using cosine similarity function (Fauzi et al., 2017). This function generated a relevance score within the range of *0* to *1*. Any unrelated line or a joke for a given question generated a value almost closer to *0* and lower than the threshold. Therefore, the responses getting scores closer to *0* could be considered irrelevant. These kinds of cases were undertaken to handle exceptions in the proposed models.

In case educational practitioners would like to replicate this study, then the first and foremost criteria are to identify domain-specific features of subjective evaluation of creative aptitude. A mixed-method approach is essential to capture the evaluation process of the particular field of interest. Any new domain-specific features for assessing questions or answers requires to be included in the model. An appropriate scoring function needs to be defined depending on the scoring mechanism to be followed in an examination. The major points to be considered to replicate the entire study are:

i. These studies are specifically meant for identifying questions that has the potential to instigate creative responses from students, and subjective evaluation of novelty in creative responses.

ii. A human-centred design approach is essential to investigate the identification of creative questions and evaluating their creative responses.

iii. A scientific and systematic study is essential to identify the features for identifying creative questions and evaluating creative responses.

iv. Investigating evaluation mechanism is required for any new domain-specific feature.

v. Decision-making is required to formulate a scoring function to measure the intended score.

vi. Updating the proposed model is required to imbibe new characteristics in it.

vii. It is essential to validate the proposed model to establish trust of pedagogues.

6.1.1 Key findings of the thesis

The salient findings of the thesis work described in different chapters are listed as follows.

i. Presently, subjective evaluation of creative aptitude in India is based on pen-and-paper-based techniques. This manual evaluation process conducted on a large scale leads to inconsistencies and errors in assessment. This thesis therefore, focussed on identifying features of evaluation to automate the assessment process.

ii. Creative question has the potential to instigate creative response. While framing creative questions examiners often self-evaluate, compare, and contrast their ideas before finally phrasing the question. During this process, they remain ever-inquisitive to know whether questions framed by them are really creative; to be more precise Design pedagogues often critically examine their framed questions by asking, do the questions framed can really capture creative responses? In order to avoid human bias, it is essential to identify features of questions that has the potential of instigating creative response from students. Further, automating the process of identifying creative questions may support pedagogues in decision-making of whether a question reformulation is required or not.

iii. Twenty two variables were systematically identified in creative questions that has the potential to instigate creative response among students. These include ‘question_verify_intent’,‘question_communicational’,‘question_expect_short_answer’,‘question_seek_fact’,‘question_novel_answer’,‘question_interest_others’,‘question_interest_self’,‘question_multi_interpretation’,‘question_verify’,‘question_seek_opinion’,‘question_choice_type’,‘question_compare_type’,‘question_consequence_action’, ‘question_definition’,‘question_entity’, ‘question_instructions’, ‘question_procedure’, ‘question_seek_reason’,‘question_spelling’,‘question_well_written’,‘question_subject ivity',and 'question_ polarity. These features supported in selecting creative questions from a bunch of non-creative questions.

iv. A computational model was designed and developed to identify creative questions from non-creative questions. This would ensure support to pedagogues in deciding whether a question confirms the features of a creative question and whether question reformulation is required or not.

v. Identifying a question that has the potential to instigate creative response among students is subjective and usually depends on experts' opinion. Therefore, inter-rater reliability of subjective judges and outcome of the model was measured. The reliability estimate between subjective judges and the proposed model was satisfactory *(α=0.96)*. This higher rate of alpha highlight high agreement among Design experts and the models, we believe this would ensure increased trust in the proposed model.

vi. While creative question plays a major role in triggering creative responses, assessment of the responses play the most important role to ensure appropriate evaluation of the creative aptitude. Novelty is an important parameter that Design pedagogues look out for in these type of responses. Systematically five dimensions were identified that explains novelty in descriptive pattern of creative responses. The dimensions were grammatical mistakes and misspellings in responses, finding relevance between questions and responses, coherence in responses, and measuring relative uniqueness among responses. Though language processing revealed minor deviation from the mean value, but it has been considered as a parameter for evaluating novelty (Schumann et al., 1996). This parameter was considered as it supported pedagogues in perceiving novelty. These dimensions would support pedagogues in a consistent evaluation process of responses during mass examinations.

vii. A computational model was designed and developed to evaluate novelty in descriptive pattern of creative responses. This model would support pedagogues to consistently measure novelty in mass examinations.

viii. Literature findings suggests seven dimensions of evaluating image-based pattern of creative responses (Berbague et al., 2021; Camburn et al., 2020; Chaudhuri et al., 2020, 2021b; Demirkan & Afacan, 2012; Schumann et al., 1996; Takai et al., 2015), where systematically two parameters were identified by Design pedagogues for specifically assessing labelled image-based pattern of creative responses. These features support finding relevance between question and labelled image-based pattern of creative responses and uniqueness of a response.

ix. A computational model was designed and developed to evaluate novelty in labelled image-based pattern of creative responses. This model would support pedagogues to consistently measure novelty in mass examinations.

x. Literature findings suggests seven dimensions of evaluating image-based pattern of creative responses (Berbague et al., 2021; Camburn et al., 2020; Chaudhuri et al., 2020, 2021b; Demirkan & Afacan, 2012; Schumann et al., 1996; Takai et al., 2015), while systematically three parameters were identified by Design pedagogues for specifically assessing annotated image-based pattern of creative responses in mass examination of Design education. These features would support language processing, finding the relevance between question and annotated image-based pattern of creative responses, and uniqueness of a response.

xi. A computational model was designed and developed to evaluate novelty in annotated image-based pattern of creative responses. This model would support pedagogues to consistently measure novelty in mass examinations of Design education.

xii. The performance metric of the proposed models for evaluating descriptive, labelled image-based, and annotated image-based pattern of creative responses were measured. The MAE of these models were satisfactory as negligible difference was found between the outcome of the models and experts.

xiii. The proposed model comprise multiple self-contained pre-defined models. A comparative study was conducted among the baseline models, and the best-performing model was considered and made part of the proposed model.

6.2 Fulfillment of objectives

Objective 1: *To identify questions that has the potential to instigate creative responses among students.*

Creative question is a significant component in Design entrance examination that triggers creativity in students. There are multiple types of questions and majority of them try to extract learning and knowledge from students. Creative question is an art and science that instigates creative response among students. Literature review and mixed-method research technique was applied to identify creative questions that invokes creative response, which was elaborated in Chapter 1 and 2. Thus, research objective 1 was fulfilled.

Objective 2: *To identify variables of questions that has the potential to instigate creative responses among students.*

Systematically, twenty-two variables of questions were identified that have the potential to instigate creative response among students. Initially, interview technique was applied to identify the features. Subsequently, their data was transcribed into open codes, clustered, and labelled. Finally, frequency analysis was conducted to identify the repetition of codes to establish their significance. This is described in Chapter 2. Thus, research objective 2 was fulfilled.

Objective 3: *To design a digitized system to identify creative questions that has the potential to instigate creative responses among students.*

An automated system was designed and developed to identify creative questions that has the potential of capturing creative aptitude. The system is based on the variables identified for creative questions. The system was validated by estimating the inter-rater reliability of the outcome of the model and experts. This study was elaborated in Chapter 2. Thus, research objective 3 was fulfilled.

Objective 4 & 5: *To examine the role of novelty in assessment of creative aptitude & to examine types of responses in evaluating novelty in creative aptitude.*

Novelty is associated with newness of a response. Understanding novelty and the manual process of evaluating aptitude for choosing students appearing in mass examination of Design entrance tests were elaborated in Chapter 1. This chapter also highlighted the objective and subjective types of questions. Subjective types of questions expect responses based on one's choice and persuasion. This thesis focussed on subjective questions and assessing their open-ended responses based on certain constraints. Further, it revealed the pattern of creative responses viz. descriptive, labelled image-based, and annotated image-based pattern of creative responses. Thus, research objective 4 and 5 was fulfilled.

Objective 6: *To identify the factors of novelty in creative aptitude evaluation.*

Initially, the commonly referred factors of novelty were identified from literature review. Further, a questionnaire-based survey was conducted to find the subfactors of novelty in creative aptitude evaluation. Five subfactors of novelty in mass examination were identified for descriptive pattern of creative response viz., grammatical mistakes, misspellings, relevance between question and their responses, coherence in responses, and relative uniqueness of a response. Two subfactors of novelty in mass examination were identified for labelled image-based pattern of creative responses viz., relevance between question and a response, and relative uniqueness of a response. Three subfactors of novelty in mass examination were identified for annotated image-based pattern of creative responses viz., language processing, relevance between question and their responses, and relative uniqueness of a response. The systematic identification of the factors of evaluating novelty associated with various pattern of creative responses has been reported in Chapter 3, 4, and 5. Thus, research objective 6 was fulfilled.

Objective 7: *To design a digitized system for novelty assessment in creative aptitude.*

An automated system was designed and developed to evaluate creative aptitude based on the factors of novelty. A system is designed for evaluating multiple pattern of creative responses viz., descriptive pattern of creative response, labelled image-based pattern of creative response, and annotated image-based pattern of creative response. The models are validated by comparing their outcomes with human experts, which exhibited negligible differences. The computational models for various creative responses have been briefly described in Chapter 3, 4, and 5. Thus, research objective 7 was fulfilled.

6.3 Validation of the models

Validation of model for identifying creative questions

A computational design model was proposed that attempts to identify creative questions from a bunch of other questions. Validation of the model was conducted by measuring inter-rater reliability among the subjective judges or examiners (gold standard data) and models. The evaluation procedure involved *21* subjective examiners *(N=21)* from different design institutes within the age group of *32-55* years *(M=37.76; S.D.=5.03)*, years of experience within *3-21* years *(M=9.04; S.D.=4.24)*, and *66.7%* were male, *28.6%* female and *4.8%* preferred not to say. Results highlight higher agreement *(α=0.96)* among examiners and the outcome of the proposed computational model. The higher rate of alpha is a significant predictor of trust in the proposed model. The entire architecture includes multiple computational models. Baseline comparison was conducted and the best-performing model was considered as benchmark for this study.

Validation of evaluating creative responses

The models presented in chapters 3, 4, and 5 were validated by comparing the outcome of the proposed models with human experts evaluation of creative responses (Design pedagogues). The validation of the models is highlighted by measuring the differences in assessment between models and experts, which was found negligible (Ling & Mahadevan, 2013). Two questions and their ten responses for each of the questions were considered for comparing the assessment scores given by models and experts. The human expert evaluation process was conducted by ten Design pedagogues *(N=10)*, within the age group of *32-55* years *(M=39.7, S.D.=7.24)*, with *3-21* years of experience *(M=11.8, S.D.=6.92)*, and possessing expertise in assessing creative aptitude in Design educational settings. Twenty Design students were randomly selected and one of the three creative questions that were formulated specifically for conducting this test was administered one at a time. They were briefed about the procedure and were allotted 120 minutes to propose solutions. The students were mostly within the age group of *17-22* years *(M=18.5, S.D.=1.35)*, and mostly belonging to first year of Indian Design schools. A sample set of questions and their responses are illustrated in Appendix E. The responses received from the twenty students were then evaluated by Design pedagogues. They were briefed to evaluate the responses based on novelty of the proposed solutions. The data collected from experts were considered a golden standard. Scores of the responses evaluated by experts for the twenty creative responses from two questions *(10 each)* for each pattern of questions were compared with the scores of the proposed models, as illustrated in Appendix F.

Validation of model for evaluating descriptive pattern of creative response

A computational design model was proposed that evaluated novelty in descriptive pattern of creative response illustrating creative aptitude. The algorithmically computed novelty score generated from the models were compared with human awarded scores. There was minimum difference between the actual scores awarded by experts and the predicted scores by this model. The MAE of the model was found to be *0.085*, which is acceptable in these types of studies for evaluation in entrance examination (Cox et al., 2009). This is a significant predictor of trust in the proposed model.

Validation of model for evaluating labelled image-based pattern of creative response

A computational design model was proposed that evaluated novelty in labelled image-based pattern of creative responses illustrating creative aptitude. The algorithmically computed novelty score generated from the models were compared with human awarded scores. There was minimum difference between the actual scores awarded by experts and the predicted scores by this model. The MAE of the model was found to be *0.015*, which is acceptable in these types of studies for evaluation in mass examination (Osmanbegovic & Suljic, 2012). This comparative analysis of the proposed model with human experts' confirm the competence of the devised model and would go a long way to establish trust of pedagogues by ensuring reduced error and stress during the evaluation process.

Validation of model for evaluating annotated image-based pattern of creative response

A computational design model was proposed that evaluated novelty in annotated image-based pattern of creative responses illustrating creative aptitude. The algorithmically computed novelty score generated from the models were compared with human awarded scores. There was minimum difference between the actual scores awarded by experts and the predicted scores by this model. The MAE of the model was found to be *0.009*, which is acceptable in these types of studies for evaluation in mass examination (Osmanbegovic & Suljic, 2012). This negligible error would ensure trust of pedagogues in the proposed model and would lead to consistency in the evaluation process of annotated image-based pattern of creative responses.

6.4 Implications

6.4.1 Implication from the perspective of design process

Cross's model informs four staged design process that includes 1) exploration that comprises of initial investigation, 2) generation involving creation of conceptual design, 3) evaluation, where the ideas are assessed and final versions of design are decided and developed, and 4) communication, where manuals or documents are prepared for manufacturing (Cross, 2000). There are other classical models for software design process such as Waterfall model, which involves iterative steps of initial requirement analysis, prototypical design, development, evaluation or testing, deployment at client's site, and maintenance (Bassil, 2012).

Majority of the design process in literature highlights a common phase of design process that is the analysis, design, and evaluation phase. The evaluation phase is mainly associated with assessing person, product, ideas, etc. This thesis focussed on the analysis, design and evaluation stage of the design process, which is from the perspective of creative aptitude assessment in entrance examination of Design education. Initially, a research gap has been identified in the analysis phase that is situated in the context of entrance examination conducted on a large scale by Design educational institutes. Further, research questions were formulated, and research objectives were identified to address the research questions. A research design was planned focussing on the evaluation phase of creative questions and assessment of various pattern of creative responses common in mass examination of Design education. The existing manual process of evaluation was transformed and designed into an automated one.

Specifically, the design phase encountered empathizing, which involved understanding the existing manual method of assessing creative questions and their responses. The manual evaluation was conducted to understand the experience. Further, a new system was designed, where the traditional pen-and-paper-based evaluation process migrated into an automated one. The migration process involved identifying the features of creative questions and evaluation of novelty in creative responses. Further, computational models were proposed to automate the evaluation process. Finally, the proposed models were validated by estimating the performance of the models to establish trust in the system. These phases are illustrated in Figure 6.2.

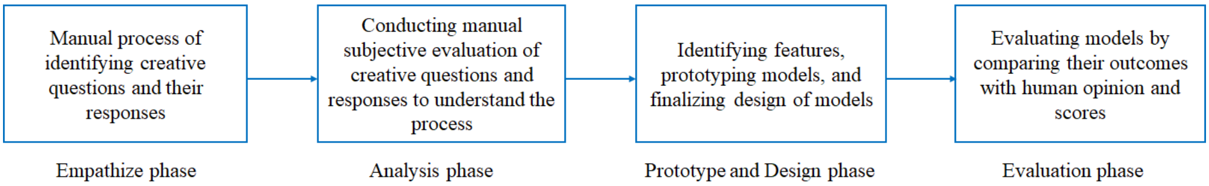


Figure 6.2: Implications from the perspective of design process

6.4.2 Implication from the perspective of human-centred approach

Human-centred approach is primarily associated with fields such as ergonomics, human-computer interaction, and artificial intelligence. International standards such as ISO 9241-210 "Ergonomics of human-centred system interaction" described human-centred design as an "approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors/ergonomics and usability knowledge and techniques" (Giacomin, 2014).

This thesis focused on the recommendations of international standards of human-centred approach. Firstly, the studies that are part of this thesis emphasized the adoption of multidisciplinary skills and perspectives of design and artificial intelligence. The design and development of the system involved mixed-method research approaches. Interview technique was applied for identifying the features of questions that has the potential to instigate creative response from students. Further, survey-based studies was conducted for identifying the features of evaluating novelty in mass examination of Design education. Computational measures were conceived and implemented for designing and developing the models for identifying creative questions and evaluating novelty in mass examination of Design education.

The studies were designed from multiple perspectives such as from the aspect of pedagogues, students, and optimization of logistics. The designed models attempted in supporting pedagogues to conduct subjective evaluation of creative aptitude consistently. Further, it aided in automatically identifying creative questions. The models were proposed from the perspective of students aspiring admission to Design schools. A consistent evaluation process would support the deserving student to get selected. Furthermore, traditional pen-an-paper-based evaluation technique involves a large volume of logistics and manpower. An automated system attempts in reducing logistics and manpower, thereby saving time and money.

Secondly, an explicit understanding of users, tasks, and environment was gained. The active subjects of these studies were the pedagogues. The evaluation process of pedagogues were perceived, their pain points and stress generators were identified. There are multiple challenges of evaluation in this environment. These are errors encountered due to stipulated timeline for assessment, errors encountered due to prolonged working hours, and errors encountered due to stress in performing repeated tasks on a large scale. Further, pedagogues often formulates creative questions to capture creativity from students. During this process, they remain ever-inquisitive to know whether the questions framed by them really instigates creative responses. Often, it causes self-doubt in pedagogues.

Thirdly, human-centred evaluation driven design was considered for the studies presented here. This system tried to imbibe the characteristics of how pedagogues evaluate in mass examinations of Design education. The proposed computational models attempted to mimic human-like procedure of identifying creative questions and assessing their responses. Threshold value determined for features was based on the way human experts conduct assessment. The threshold was made variable and susceptible to be changed depending on the type of examination, level of the examination, stringency, or leniency of a pedagogue. Moreover, a particular design of a model was accepted depending on the performance of the proposed model. A human-centred evaluation was conducted, where the proposed models' outcome was compared with experts' outcomes. Design of the model was confirmed depending on its closeness to human-based assessment outcomes.

6.5 Recommendations

A list of recommendations is illustrated Figure 6.3. The elements of the diagram is clustered based on its association with the creative questions and their type of responses. Initially creative questions and their responses needs to be considered. Further, studies suggests features that are recommended for identifying creative question that instigate creative responses (illustrated using green colour). Features for assessing descriptive, labelled image-based, and annotated image-based creative responses are marked with light blue colour, grey, and yellow colour, respectively. The features that is practiced by a combination of patterns for evaluation is represented by combination of these colours. However, selecting these features for evaluating creative aptitude depends on the type (institutional, nationalized, etc.) and level of a test (easy, moderate, difficult, very difficult, etc.) based on which the scores are assessed.

The list of recommendations proposed are as follows:

i. This research work recommends using twenty-two variables to identify questions that triggers creative responses, five variables to evaluate novelty in descriptive creative responses, two variables to evaluate novelty in labelled image-based creative responses, and three variables to evaluate novelty in annotated image-based creative responses.

ii. The threshold for each parameters of the model is adaptable and thus threshold values are recommended to be changed based on the type of examinations, difficulty level of examinations, stringency or leniency of pedagogues.

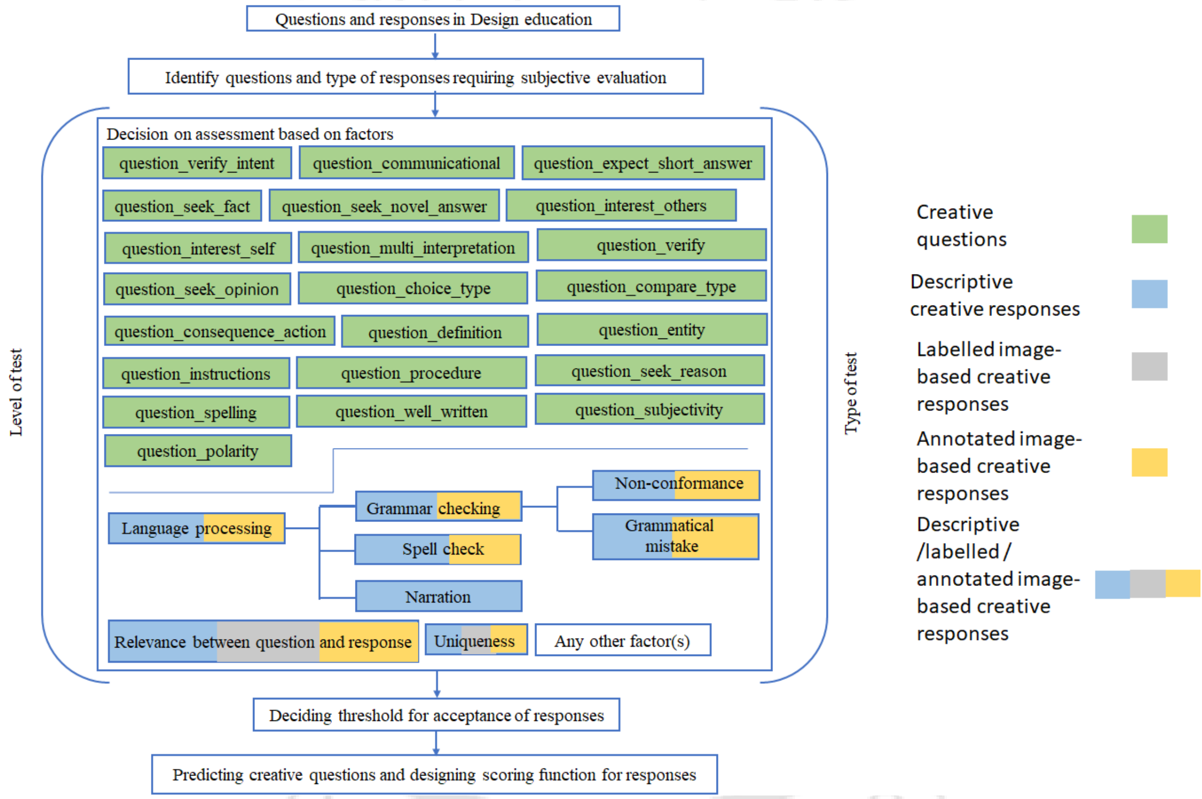


Figure 6.3: Overview of recommendations

6.6 Novelties in contribution of the present research

Outcomes of this research have significant contribution to the domain of both knowledge-base and design and development of digitized human-centred assessment of creative aptitude in questions and responses in Indian scenarios. The novelty in contribution of this research are as follows.

6.6.1 Contribution to knowledge-base

Reference questionnaire to identify creative questions and evaluate novelty in creative responses in Design entrance examination

The questionnaire is designed for identifying factors of creative questions that triggers creative responses and factors of evaluating novelty in these responses. This questionnaire is validated and is a contribution to the knowledge-base in this field that could be used by future researchers.

Database of dimensions of creative questions and evaluating novelty in creative responses

There is a dearth of information related to features or dimensions of creative questions that triggers creative responses and evaluating novelty for multiple patterns of creative responses. This research work contributes the specific dimensions of assessing creative aptitude that would support education technology researchers in future.

Database of creative questions and their responses

Specifically, there is a dearth of dedicated database of creative questions that triggers creative responses and a repository of creative responses. Questions and responses has been identified at various repositories, but their primary association was not with creativity aptitude assessment. Therefore, multiple datasets were collaborated from various repositories (Bombay, 2021a, 2021b; Chua et al., 2009; Karpathy & Fei-Fei, 2015; McAuley, 2018), public and private institutions. These data could be used by instructional designers and data scientists to further explore possibilities for better training and evaluation.

6.6.2 Contribution toward methods

A significant contribution from the perspective of methods is the proposed architectures to identify creative questions and evaluate novelty in creative responses. The architectures highlight a structured paradigm and represent a step-by-step problem-solving approach. Therefore, this architecture could act as a scaffold to identify creative questions and evaluate novelty of creative responses for a large scale Design entrance examination. The proposed architecture was validated by comparing its outcome with human experts' score. A negligible difference indicates that the proposed architecture is consistent and can be trusted by pedagogues as a substitute for the existing manual evaluation process.

The outcome of the research investigation is significant as it contributes by defining a technique (a model) that not only extracts quantitative values from creative responses, but also devises a way to propose a score that objectively evaluates novelty. Apart from this, the research investigation also contributes a set of formal evaluation indicators that can be used during assessment of questions and their responses by Design pedagogues. This research would provide knowledge to data scientists community in the direction of utilizing multiple data resources and effectively use in scientific techniques, methods, and algorithms for a wide spectrum of applications. Another major contribution would be to the student community who aspires to get admission in Design schools through entrance examinations by a consistent evaluation process. Examinations conducted on a large scale might encounter inconsistency in evaluation process, but the application of a standardized automated system like this attempts to reduce errors in evaluation. This may result in increasing the chance of getting the desired student being selected through the examination.

The methods presented in this thesis is robust that is proved by comparing the results with manual methods. Reliability is also measured among the outcomes of the proposed method and human experts. This method adopted systematic data acquisition and reporting of descriptive statistic, feature extraction, proposal of models, and validation of models. This method is proposed to maintain a structured paradigm and scientific replicability. This method would support the designer and research community in understanding the process of capturing features of evaluating creative responses and creative questions associated with mass examination of Design education. The proposed architectures would support education technology researchers and influence their assessment design process.

Creativity is involved and practiced in all domains. This methods in the future could act as a scaffold for identifying the features of creative questions and evaluating their responses for other domains such as engineering, business, etc. However, an appropriate questionnaire needs to be formulated to capture creativity feature for other domains and evaluate their responses. Analyzing the acquired data would provide features of creative questions and evaluating novelty of responses of domain of interest. In order to automate the process of identifying creative questions and responses of domain of interest, DL models needs to be trained with similar types of questions and responses. Further, categorized outcomes of DL models could be compared with the categories made by human experts. Higher agreement between the models and humans would indicate the success of experiments for any domain of interest.

6.6.3 Contribution to Design and Design education

The research investigation reported here focuses on addressing challenges faced by the Design educators community and can be directly related as a contribution from the perspective of Design Praxiology as proposed by (Cross, 1999; Gasparski, 1979). The proposals made in this research work possesses an approach that intends to prepare Design education community specifically Design pedagogues to embrace changes in existing ways of framing creative questions and assessing their responses. This study has addressed a Design problem, which is specifically based on addressing human-errors in the process of evaluation and assessment of creative responses.

Stress is often responsible for errors and inconsistencies in evaluation. To address these problems, this thesis suggested the design of an automation process of identifying creative questions and evaluating novelty from descriptive, labelled-image, and annotated image-based pattern of creative responses. This is a significant contribution to Deign education as it would support pedagogues on a large scale evaluation of creative responses and identifying creative questions, thereby reducing their workload and frustrations of a repeated task.

Consistency in assessment is an inherent criterion in any examination process. It supports in awarding impartial scores to students and unbiased identification of creative questions. Design education involves human experts who manually evaluate creative aptitude on a large scale in mass examinations. During this process, pedagogues often conduct subjective evaluation based on their own referential metrics. This leads to inconsistency in an evaluation process. Moreover, examinations like this require creative questions that attempt in capturing creativity of students. During formulating these type of questions, examiners compare and contrast their ideas in order to frame questions that can capture creativity of students. However, during this process they are often biased by their past experiences. Therefore, computational models presented in this thesis is a contribution to Design education as they would support conducting assessment of creative aptitude and identification of creative questions based on pre-defined feature set. The methods involved in this models would ensure scientific replicability, thereby confirming consistency in assessment process. The outcomes of this thesis are significant as it would support increasing trust in subjective evaluation in Design education.

This thesis contributes in design process of assessment system as it support in systematically identifying features of evaluating creative aptitude and identifying creative questions that instigate creative responses. The design of the computational models would support in consistent evaluation and identifying creative questions. The overall findings of this thesis would have high potential to support designers and researchers in designing an assessment process for Design education.

6.7 Limitation and future scope

There was a limitation of availability of creative questions that were collected from various institutions and web repositories associated with Design entrance examinations. Furthermore, limited creative responses (descriptive, labelled image, and annotated image) collected from Design schools be it private or public compelled the usage of using similar synthetic datasets which were not directly from Design schools. For training the proposed models, an average of 50,000 datasets were utilized. Collecting such a huge number of annotated and labelled images specifically from Design schools is extremely difficult given the fact that negligible number of students in India appear for Design entrance exams and are currently studying Design as compared to other disciplines. To address this issue, a specific technique was adopted. Images were collected from various online repositories that closely resembles to the type of sketches and drawings that are generally made by Design students. Moreover, pre-processing were conducted to the image files collected and were converted into sketches. The detailed procedure of converting online resources into patterns similar to responses of Design entrance examination is illustrated in Section 4.2.4, 4.2.5, 5.2.4, and 5.2.5.

A limitation of the computational design models presented in this thesis is that it is domain-specific for Design education. The computation in these models intends to specifically support Design pedagogues. Presently, this model does not capture the capability of adapting itself to a newer environment or domain. It is also not tested for any other domains. It is also essential to extend the architecture in case any new feature needs to be appended to the existing models.

Hand-tuning method was used to optimize the values of the hyperparameters. We have used basic recommendations highlighted in literature (Bengio, 2012). We specifically did not use any algorithms or libraries for hyperparameter optimization. Data augmentation was performed where data from multiple sources were added to the training set to avoid overfitting. We validated our models in Chapter 2, 3, 4, and 5 by comparing the assessment scores of the proposed models with human-based assessment.

Extensive work in the field of hand writing recognition has been carried out in literature and evident from the works of (Loudon et al., 2005; Memon et al., 2020; Sharma & Kaushik, 2020). In this research therefore no attempt was made to investigate and develop any such model for hand writing recognition of students responses. We believe that since the focus of the study is to identify factors and conceive a model for novelty evaluation of students creative responses, existing techniques of hand recognition systems can be incorporated along with the model that we have proposed for novelty assessment of the creative aptitude. This may be considered as a limitation of this existing study. We therefore propose that further studies can be conducted to verify the efficacy of hand-writing recognition systems with our model of novelty assessment for evaluating creative aptitude.

This research suggested a systematic procedure of identifying creative questions. In future, there is a scope that researchers may investigate the process of automatically formulating creative questions. This might reduce bias of including examiners individual characteristics and past experiences. This might also support pedagogues in formulating creative questions on a large scale required in mass examinations of Design education. In future, one may quantify the reduction of stress of pedagogues using the support systems for identifying creative questions and assessing their responses. Further, limited resource and time constraints compelled this research to only focussed the evaluation of novelty in mass examinations of Design education. However, in future researchers may look into other aspects of evaluation, which might be significant for Design education and the Design community.

6.8 Conclusion

This was probably the first attempt to the best of our knowledge in digitizing the identification of creative questions and evaluation of creative aptitude for entrance examination of Design education. The results of this research could be further used by designers and researchers for designing and developing a comparatively optimized system that would support pedagogues in the assessment process. The results found from the individual studies were descent when compared with human experts and would support in ensuring trust in the overall research. This induces scientific replicability and consistency of the research.

The present research is first of its kind to the best of our knowledge to suggest dimensions to evaluate novelty and identify creative questions in entrance examination of Design education. The number of features identified were twenty two, five, two, and three for identifying creative questions, and evaluating novelty in descriptive pattern of creative responses, labelled image-based pattern of creative responses, and annotated image-based pattern of creative responses, respectively. Datasets were scrapped from multiple websites, and synthetic datasets were identified for all these types of questions and creative responses. The models are well tested and this provided a provision for researchers to replicate the study and further explore this domain. Educational technology development organizations or start-ups may find this thesis endeavour highly beneficial for their design process of assessment systems. This thesis significantly contributes to the knowledge-base, methods, design, and design education.

References

Abacha, A. ben, & Demner-Fushman, D. (2019). On the Role of Question Summarization and Information Source Restriction in Consumer Health Question Answering. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2019*, 117–126. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568117/

Abdel-Basset, M., Mohamed, M., Elhoseny, M., Chiclana, F., Zaied, A. E.-N. H., & others. (2019). Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar disorder diseases. *Artificial Intelligence in Medicine*, *101*, 101735. https://doi.org/10.1016/j.artmed.2019.101735

Aburas, R., & Nurunnabi, M. (2019). The relationship between admission exam and students' performance on an undergraduate interior design program. *International Journal of Technology and Design Education*, *29*(4), 961–986. https://doi.org/10.1007/s10798-018-9474-z

Aksoy, C., Can, F., & Kocberber, S. (2012). Novelty Detection for Topic Tracking. *Journal of the American Society for Information Science and Technology*, *63*(4), 777–795. https://doi.org/10.1002/asi.21697

Albitar, S., Fournier, S., & Espinasse, B. (2014). An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification. *International Conference on Web Information Systems Engineering*, *8786 LNCS*, 105–114. https://doi.org/10.1007/978-3-319-11749-2_8

Alexiou, E., Ebrahimi, T., Bernardo, M. v, Pereira, M., Pinheiro, A., Cruz, L. A. D. S., Duarte, C., Dmitrovic, L. G., Dumic, E., Matkovics, D., & others. (2018). Point cloud subjective evaluation methodology based on 2d rendering. *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6. https://doi.org/10.1109/QoMEX.2018.8463406

Al-Homoud, M. (2020). ARCHITECTURE AND DESIGN PLACEMENT EXAM-CONTEXT AND CONTENT. *Proceedings of INTED2020 Conference*. https://doi.org/10.21125/inted.2020.2000

Ali, A. (2005). Analysis Of Turkish high-school physics-examination questions and university entrance exams questions according to Blooms' taxonomy. *Journal of Turkish Science Education*, *2*(2), 68.

All India Council for Technical Education- Approval Process Handbook 2021-22. (2021). https://www.aicte-india.org/sites/default/files/approval/Approval%20Process%20Handbook_2021-22.pdf

Almeida, L. S., Prieto, L. P., Ferrando, M., Oliveira, E., & Ferrándiz, C. (2008). Torrance Test of Creative Thinking: The question of its construct validity. *Thinking Skills and Creativity*, *3*(1), 53–58. https://doi.org/10.1016/j.tsc.2008.03.003

Alshanqiti, A., Bajnaid, A., Gilal, A. R., Aljasir, S., Alsughayyir, A., & Albouq, S. (2020). Intelligent Parallel Mixed Method Approach for Characterising Viral YouTube Videos in Saudi Arabia. *Int. J. Adv. Comput. Sci. Appl*. https://doi.org/10.14569/IJACSA.2020.0110382

Al-Zahrani, A. M. (2015). From passive to active: The impact of the flipped classroom through social learning platforms on higher education students' creative thinking. British journal of educational technology, 46(6), 1133-1148.

Amabile, T. M. (1988). A model of creativity and innovation in organizations. *Research in Organizational Behavior*, *10*(1), 123–167.

Amarbayasgalan, T., Jargalsaikhan, B., & Ryu, K. H. (2018). Unsupervised novelty detection using deep autoencoders with density based clustering. *Applied Sciences*, *8*(9), 1468. https://doi.org/10.3390/app8091468

Amini, A., Gabriel, S., Lin, P., Koncel-Kedziorski, R., Choi, Y., & Hajishirzi, H. (2019). Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *ArXiv Preprint ArXiv:1905.13319*.

Amorim, M., Bortoloti, F. D., Ciarelli, P. M., Salles, E. O. T., & Cavalieri, D. C. (2019). Novelty detection in social media by fusing text and image into a single structure. *IEEE Access*, *7*, 132786–132802. https://doi.org/10.1109/ACCESS.2019.2939736

Anantrasirichai, N., & Bull, D. (2021). Artificial intelligence in the creative industries: a review. *Artificial Intelligence Review*, 1–68. https://doi.org/10.1007/s10462-021-10039-7

Annamoradnejad, I., Fazli, M., & Habibi, J. (2020). Predicting Subjective Features from Questions on QA Websites using BERT. *2020 6th International Conference on Web Research (ICWR)*, 240–244. https://doi.org/10.48550/arXiv.2002.10107

Arsyad, N., Rahman, A., & Ahmar, A. S. (2017). Developing a self-learning model based on open-ended questions to increase the students' creativity in calculus. *Global Journal of Engineering Education*, *9*(2), 143–147. https://doi.org/10.26858/gjeev19i2y2017p143147

Asunda, P., & Hill, R. (2007). Features of engineering design in technology education. *Journal of Industrial Teacher Education*, *44*(1), 1–13. https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1004&context=ncete_cstudies

Aubin, A.-S., St-Onge, C., & Renaud, J.-S. (2018). Detecting rater bias using a person-fit statistic: A Monte Carlo simulation study. *Perspectives on Medical Education*, *7*(2), 83–92. https://doi.org/10.1007/s40037-017-0391-8

Avadesh, M., & Goyal, N. (2018). Optical character recognition for sanskrit using convolution neural networks. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 447–452.

Aziza, M. (2018). An analysis of a teacher's questioning related to students' responses and mathematical creativity in an elementary school in the UK. *International Electronic Journal of Elementary Education*, *10*(4), 475–487. https://doi.org/10.26822/iejee.2018438138

Baloche, L., & Platt, T. J. (1993). Sprouting Magic Beans: Exploring Literature throuth Creative Questioning and Cooperative Learning. *Language Arts*, *70*(4), 264–271.

Baron, P. (2018). Heterarchical Reflexive Conversational Teaching and Learning as a Vehicle for Ethical Engineering Curriculum Design. *Constructivist Foundations*, *13*(3).

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or Surprise. *Frontiers in Psychology*, 1–15. https://doi.org/10.3389/fpsyg.2013.00907

Basias, N., & Pollalis, Y. (2018). Quantitative and qualitative research in business \& technology: Justifying a suitable research methodology. *Review of Integrative Business and Economics Research*, *7*, 91–105.

Bassil, Y. (2012). A simulation model for the waterfall software development life cycle. *ArXiv Preprint ArXiv:1205.6904*. https://doi.org/10.48550/arXiv.1205.6904

Basu, S., Mooney, R. J., Pasupuleti, K. v., & Ghosh, J. (2001). Evaluating the novelty of text-mined rules using lexical knowledge. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–238. https://doi.org/10.1145/502512.502544

Bayer-Hohenwarter, G. (2010). Comparing Translational Creativity Scores of Students and Professionals: Flexible Problem-Solving and. New approaches in translation process research, 39, 83.

Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., & Neubauer, A. C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. Thinking Skills and Creativity, 21, 75-84.

Benenson, R., Popov, S., & Ferrari, V. (2019). Large-scale interactive object segmentation with human annotators. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11700–11709.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural networks: Tricks of the trade (pp. 437-478). Springer, Berlin, Heidelberg.

Berbague, C. E., Karabadji, N. E., Seridi, H., Symeonidis, P., Manolopoulos, Y., & Dhifli, W. (2021). An overlapping clustering approach for precision, diversity and novelty-aware recommendations. *Expert Systems with Applications*, *177*, 114917. https://doi.org/10.1016/j.eswa.2021.114917

Best Free OCR API, Online OCR, Searchable PDF - Fresh 2021 On-Premise OCR Software. (2021). https://ocr.space/

Bodesheim, P., Freytag, A., Rodner, E., & Denzler, J. (2015). Local novelty detection in multi-class recognition problems. *2015 IEEE Winter Conference on Applications of Computer Vision*, 813–820. https://doi.org/10.1109/WACV.2015.113

Bombay, I. (2021a). *CEED 2022*. http://www.ceed.iitb.ac.in/2022/

Bombay, I. (2021b). *UCEED 2022*. http://www.uceed.iitb.ac.in/2022/

Bonnici, A., Gross, C., McGuire, P. C., Ormö, J., Walter, S., & Wendt, L. (2010). The Cyborg Astrobiologist: compressing images for the matching of prior textures and for the detection of novel textures. *European Planetary Science Congress (EPSC)*, *5*.

Boracchi, G., Carrera, D., & Wohlberg, B. (2014). Novelty detection in images by sparse representations. *2014 IEEE Symposium on Intelligent Embedded Systems (IES)*, 47–54. https://doi.org/10.1109/INTELES.2014.7008985

Borg, A., Boldt, M., Rosander, O., & Ahlstrand, J. (2021). E-mail classification with machine learning and word embeddings for improved customer support. *Neural Computing and Applications*, *33*(6), 1881–1902. https://doi.org/10.1007/s00521-020-05058-4

Boyle, G. J., Borg, M. G., Falzon, J. M., & Baglioni Jr, A. J. (1995). A structural model of the dimensions of teacher stress. *British Journal of Educational Psychology*, *65*(1), 49–67. https://doi.org/10.1111/j.2044-8279.1995.tb01130.x

Brabb, G. J., & Morrison, E. D. (1964). The evaluation of subjective information. *Journal of Marketing Research*, *1*(4), 40–44. https://doi.org/10.1177/002224376400100406

Brownlee, J. (2019). *Multi-Label Classification of Satellite Photos of the Amazon Rainforest*. https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-to-classify-satellite-photos-of-the-amazon-rainforest/?photos-of-the-amazon-rainforest

Camburn, B., He, Y., Raviselvam, S., Luo, J., & Wood, K. (2020). Machine learning-based design concept evaluation. *Journal of Mechanical Design*, *142*(3), 31113. https://doi.org/10.1115/1.4045126

Cao, F., Liang, J., & Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers & Mathematics with Applications*, *58*(3), 474–483. https://doi.org/10.1016/j.camwa.2009.04.017

Cesa-Bianchi, N., Hardoon, D. R., & Leen, G. (2010). Guest Editorial: Learning from multiple sources. *Mach Learn-Guest Editorial: Learning from Multiple Sources*, *79*, 1–3. https://doi.org/10.1007/s10994-010-5169-8

Chakrabarti, A. (2006). Defining and supporting design creativity. In DS 36: Proceedings DESIGN 2006, the 9th International Design Conference, Dubrovnik, Croatia (pp. 479-486).

Chan, A. H. S., Chen, K., & Chong, E. Y. L. (2010). *Work stress of teachers from primary and secondary schools in Hong Kong*. http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp1903-1906.pdf

Charlet, D., & Damnati, G. (2017). Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 315–319. https://doi.org/10.18653/v1/S17-2051

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage. http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Charmaz_2006.pdf

Charnley, F., Lemon, M., & Evans, S. (2011). Exploring the process of whole system design. *Design Studies*, *32*(2), 156–179. https://doi.org/10.1016/j.destud.2010.08.002

Chaudhuri, N. B., Dhar, D., & Yammiyavar, P. G. (2020). A computational model for subjective evaluation of novelty in descriptive aptitude. In *International Journal of Technology and Design Education* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s10798-020-09638-2

Chaudhuri, N. B., Dhar, D., & Yammiyavar, P. G. (2021a). A human-centred deep learning approach facilitating design pedagogues to frame creative questions. *Neural Computing and Applications*, 1–28. https://doi.org/10.1007/s00521-021-06511-8

Chaudhuri, N. B., Dhar, D., & Yammiyavar, P. G. (2021b). Automating assessment of Design exams: A case study of novelty evaluation. *Expert Systems with Applications*, 116108. https://doi.org/https://doi.org/10.1016/j.eswa.2021.116108

Chaudhuri, N. B., Dhar, D., & Yammiyavar, P. G. (2021c). Do Design Entrance Exams in India Really Test Creative Aptitude? An Analytical Study of Design Tests Conforming Creativity Benchmarks. *Design for Tommorrow*, 371–383. https://doi.org/10.1007/978-981-16-0119-4_30

Cheung, C. K., Rudowicz, E., Yue, X., & Kwan, A. S. (2003). Creativity of university students: What is the impact of field and year of study?. The Journal of Creative Behavior, 37(1), 42-63.

Chew, S. W., Lin, I.-H., & Chen, N.-S. (2019). Using Socratic questioning strategy to enhance critical thinking skill of elementary school students. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, *2161*, 290–294. https://doi.org/10.1109/ICALT.2019.00088

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 539–546. https://doi.org/10.1109/CVPR.2005.202

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: a real-world web image database from National University of Singapore. *Proceedings of the ACM International Conference on Image and Video Retrieval*, 1–9. https://doi.org/10.1145/1646396.1646452

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. *ArXiv Preprint ArXiv:1905.10044*. https://doi.org/10.48550/arXiv.1905.10044

Clarkson, J., & Eckert, C. (2010). *Design process improvement: a review of current practice*. https://link.springer.com/book/10.1007/978-1-84628-061-0

Constantinou, F. (2021). How novel can examination questions really be? Exploring the boundaries of creativity in examination question writing. *Research Papers in Education*, *00*(00), 1–19. https://doi.org/10.1080/02671522.2021.1961297

Cox, A., Duncheon, N., & McDavid, L. (2009). Peers and teachers as sources of relatedness perceptions, motivation, and affective responses in physical education. *Research Quarterly for Exercise and Sport*, *80*(4), 765–773. https://doi.org/10.1080/02701367.2009.10599618

Croasmun, J. T., & Ostrom, L. (2011). Using likert-type scales in the social sciences. *Journal of Adult Education*, *40*(1), 19–22.

Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper Review*, *23*(2), 72–79. https://doi.org/10.1080/02783190009554069

Cross, N. (1999). Design research: A disciplined conversation. *Design Issues*, *15*(2), 5–10.

Cross, N. (2000). Strategies for Product Design. In *Engineering design methods: strategies for product design, 3rd edn. The Open University, Milton Keynes*. Nigel Cross. The open University, Milton Keynes, UK, Ed. Wiley.

Cruz-Ramirez, M., Hervás-Martiinez, C., Sánchez-Monedero, J., & Gutiérrez, P. A. (2014). Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing*, *135*, 21–31. https://doi.org/10.1016/j.neucom.2013.05.058

Csikszentmihalyi, M., & Wolfe, R. (2014). New Conceptions and Research Approaches to Creativity: Implications of a Systems Perspective for Creativity in Education. In *The Systems Model of Creativity* (pp. 161–184). Springer Netherlands. https://doi.org/10.1007/978-94-017-9085-7_10

Cutumisu, M., & Guo, Q. (2019). Using Topic Modeling to Extract Pre-Service Teachers' Understandings of Computational Thinking from Their Coding Reflections. *IEEE Transactions on Education*, *62*(4), 325–332. https://doi.org/10.1109/TE.2019.2925253

Daly, S. R., Mosyjowski, E. A., & Seifert, C. M. (2014). Teaching creativity in engineering courses. *Journal of Engineering Education*, *103*(3), 417–449. https://doi.org/10.1002/jee.20048

Davis, D. C., Gentili, K. L., Trevisan, M. S., & Calkins, D. E. (2002). Engineering design assessment processes and scoring scales for program improvement and accountability. *Journal of Engineering Education*, *91*(2), 211–221. https://doi.org/10.1002/j.2168-9830.2002.tb00694.x

de Vergara, J. E., & Olmos, R. (2019). Reliability score evaluation of continuous assessment tests: A longitudinal study. *The International Journal of Electrical Engineering & Education*. https://doi.org/10.1177/0020720919879375

Demir, S., & Sahin, F. (2014). Assessment of open-ended questions directed to prospective science teachers in terms of scientific creativity. *Procedia-Social and Behavioral Sciences*, *152*, 692–697. https://doi.org/10.1016/j.sbspro.2014.09.264

Demirkan, H., & Afacan, Y. (2012). Assessing creativity in design education: Analysis of creativity factors in the first-year design studio. *Design Studies*. https://doi.org/10.1016/j.destud.2011.11.005

Désir, C., Bernard, S., Petitjean, C., & Heutte, L. (2013). One Class Random Forests. *Pattern Recogn.*, *46*(12), 3490–3506. https://doi.org/10.1016/j.patcog.2013.05.022

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Dippo, C., & Kudrowitz, B. (2013). Evaluating the alternative uses test of creativity. 2013 NCUR.

D'Souza, R. (2021). What characterises creativity in narrative writing, and how do we assess it? Research findings from a systematic literature search. Thinking Skills and Creativity, 42, 100949.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful?. Psychology of Aesthetics, Creativity, and the Arts, 9(1), 35.

Dynich, A., & Wang, Y. (2017). Analysis of novelty of a scientific text as a basis for assessment of efficiency of scientific activities. *Journal of Organizational Change Management*, *30*(5), 668–682. https://doi.org/10.1108/JOCM-10-2016-0226

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on Graphics (TOG)*, *31*(4), 1–10. https://doi.org/10.1145/2185520.2185540

Ellis, K. (1993). Teacher Questioning Behavior and Student Learning: What Research Says to Teachers. *Convention of The Western States Communication Association*.

Esposito, M., Damiano, E., Minutolo, A., de Pietro, G., & Fujita, H. (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Information Sciences*, *514*, 88–105. https://doi.org/10.1016/j.ins.2019.12.002

Fauzi, M. A., Utomo, D. C., Setiawan, B. D., & Pramukantoro, E. S. (2017). Automatic essay scoring system using N-gram and cosine similarity for gamification based E-learning. *Proceedings of the International Conference on Advances in Image Processing*, 151–155. https://doi.org/10.1145/3133264.3133303

Fiorineschi, L., & Rotini, F. (2021). Novelty metrics in engineering design. In *Journal of Engineering Design*. https://doi.org/10.1080/09544828.2021.1928024

Florez, L., & Castro-Lacouture, D. (2013). Optimization model for sustainable materials selection using objective and subjective factors. *Materials and Design*, *46*, 310–321. https://doi.org/10.1016/j.matdes.2012.10.013

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, *3*(4), 648–666. https://doi.org/10.1111/j.1756-8765.2011.01149.x

Furusho, Y., & Kotani, K. (2017). Objective & Subjective Evaluation Models of Pencil Still Drawings for Art Education. *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–5. https://doi.org/10.1109/DICTA.2017.8227473

Gagnon, V., Labrie, A., Bhatnagar, S., & Desmarais, M. C. (2019). Filtering non-relevant short answers in peer learning applications. *EDM*.

Gamon, M. (2020). Graph-based text representation for novelty detection. *Proceedings of TextGraphs: The 1st Workshop on Graph-Based Methods for Natural Language Processing*, *June*, 17–24. https://doi.org/10.3115/1654758.1654762

Garaigordobil, M. (2006). Intervention in creativity with children aged 10 and 11 years: Impact of a play program on verbal and graphic–figural creativity. Creativity Research Journal, 18(3), 329-345.

Garcia-Esteban, S. (2017). Do Video Learning Objects Develop Digital Competence in Teacher Training? *RaeL: Revista Electronica De*. http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=18859089&AN=128258576&h=fMIchnjI%2Be3%2FRnYjTpwAv%2B5Muemr0JlGPuDXSqR6kcr1NuAUhlbJbeAlA%2FTEKtOKkwfnY3rOo8fOh%2BJ5uCZRPA%3D%3D&crl=c

Garcia-Pablos, A., Perez, N., & Cuadros, M. (2020). Sensitive data detection and classification in spanish clinical text: Experiments with bert. *ArXiv Preprint ArXiv:2003.03106*. https://doi.org/10.48550/arXiv.2003.03106

Gasparski, W. W. (1979). Praxiological—systemic approach to design studies. *Design Studies*, *1*(2), 101–106. https://doi.org/10.1016/0142-694X(79)90006-1

Gemmell, B., & Vyas, P. (2016). *The Future of Design Education in India*. www.britishcouncil.in

Getzels, J. W., & Jackson, P. W. (1962). *Creativity and intelligence: Explorations with gifted students.*

Ghosal, T., Edithal, V., Ekbal, A., Bhattacharyya, P., Tsatsaronis, G., & Chivukula, S. S. S. K. (2018). Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection. *Proceedings of the 27th International Conference on Computational Linguistics*, 2802–2813. https://www.aclweb.org/anthology/C18-1237

Ghosh, S. (1993). An exercise in inducing creativity in undergraduate engineering students through challenging examinations and open-ended design problems. *IEEE Transactions on Education*, *36*(1), 113–119. https://doi.org/10.1109/13.204828

Giacomin, J. (2014). What is human centred design? *The Design Journal*, *17*(4), 606–623. https://doi.org/10.2752/175630614X14056185480186

Gigerenzer, G. (1994). Where do new ideas come from. *Dimensions of Creativity*, 53–74. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.7721&rep=rep1&type=pdf

Gil-Gómez, J.-A., Manzano-Hernández, P., Albiol-Pérez, S., Aula-Valero, C., Gil-Gómez, H., & Lozano-Quilis, J.-A. (2017). USEQ: a short questionnaire for satisfaction evaluation of virtual rehabilitation systems. *Sensors*, *17*(7), 1589. https://doi.org/10.3390/s17071589

Goldschmidt, G. (2009). The dialectics of sketching. *Creativity Research Journal*, *4*(2), 123–143. https://doi.org/10.1080/10400419109534381

Goldschmidt, G. (2014). Modeling the role of sketching in design idea generation. In *An anthology of theories and models of design* (pp. 433–450). Springer. https://doi.org/10.1007/978-1-4471-6338-1_21

Goldschmidt, G., & Sever, A. L. (2011). Inspiring design ideas with texts. *Design Studies*, *32*(2), 139–155. https://doi.org/10.1016/j.destud.2010.09.006

Gonzalez, A., Peters, M. L., Orange, A., & Grigsby, B. (2017). The influence of high-stakes testing on teacher self-efficacy and job-related stress. *Cambridge Journal of Education*, *47*(4), 513–531. https://doi.org/10.1080/0305764X.2016.1214237

González-Cutre, D., Sicilia, Á., Sierra, A. C., Ferriz, R., & Hagger, M. S. (2016). Understanding the need for novelty from the perspective of self-determination theory. *Personality and Individual Differences*, *102*, 159–169. https://doi.org/10.1016/j.paid.2016.06.036

Google QUEST Q&A Labeling | Kaggle. (2020). https://www.kaggle.com/c/google-quest-challenge/overview

Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2018). Iqa: Visual question answering in interactive environments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4089–4098. https://doi.org/10.48550/arXiv.1712.03316

Guillaumin, M., Verbeek, J., & Schmid, C. (2009). Is that you? Metric learning approaches for face identification. *2009 IEEE 12th International Conference on Computer Vision*, 498–505. https://doi.org/10.1109/ICCV.2009.5459197

Guillemot, C., & le Meur, O. (2013). Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine*, *31*(1), 127–144. https://doi.org/10.1109/MSP.2013.2273004

Gupta, A., & Ruebush, E. (2019). AquaSight: Automatic Water Impurity Detection Utilizing Convolutional Neural Networks. *ArXiv Preprint ArXiv:1907.07573*. https://doi.org/10.48550/arXiv.1907.07573

Gutiérrez, L., & Keith, B. (2018). A Systematic Literature Review on Word Embeddings. *International Conference on Software Process Improvement*, 132–141. https://doi.org/10.1007/978-3-030-01171-0_12

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, *2*, 1735–1742. https://doi.org/10.1109/CVPR.2006.100

Haik, Y., Sivaloganathan, S., & Shahin, T. M. (2015). *Engineering design process*. Cengage Learning.

Halim, A., Mazlina, H., & others. (2018). Questioning skill of science teacher from the students perscpective in senior high school. *Journal of Physics: Conference Series*, *1088*(1), 12109. https://doi.org/10.1088/1742-6596/1088/1/012109

Hamzacebi, C. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences*, *178*(23), 4550–4559. https://doi.org/10.1016/j.ins.2008.07.024

Hargie, O. (2006). *The handbook of communication skills*. https://dl.uswr.ac.ir/bitstream/Hannan/141245/1/9781138219120.pdf

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Hayashi, T., & Ohsawa, Y. (2015). Estimation of novelty assessment of strategic scenarios using relativeness. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2015-Janua*(January), 441–446. https://doi.org/10.1109/ICDMW.2014.68

Hilgart, M. M., Ritterband, L. M., Thorndike, F. P., & Kinzie, M. B. (2012). Using instructional design process to improve design and development of Internet interventions. *Journal of Medical Internet Research*, *14*(3), e89. https://doi.org/10.2196/jmir.1890

Hoddy, E. T. (2019). Critical realism in empirical research: employing techniques from grounded theory methodology. *International Journal of Social Research Methodology*, *22*(1), 111–124. https://doi.org/10.1080/13645579.2018.1503400

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899. https://doi.org/10.1613/jair.3994

Hoffmann, H. (2007). Kernel PCA for novelty detection. *Pattern Recognition*, *40*(3), 863–874. https://doi.org/10.1016/j.patcog.2006.07.009

Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2017). Modelling e-learner comprehension within a conversational intelligent tutoring system. *IFIP World Conference on Computers in Education*, 251–260. https://doi.org/10.1007/978-3-319-74310-3_27

Hossain, N., Krumm, J., Sajed, T., & Kautz, H. (2020). Stimulating creativity with funlines: A case study of humor generation in headlines. *ArXiv Preprint ArXiv:2002.02031*. https://doi.org/10.48550/arXiv.2002.02031

Hsu, W.-T., Lin, C.-K., Lee, M.-Y., Min, K., Tang, J., & Sun, M. (2018). A unified model for extractive and abstractive summarization using inconsistency loss. *ArXiv Preprint ArXiv:1805.06266*. https://doi.org/10.48550/arXiv.1805.06266

Huang, R., Pedoeem, J., & Chen, C. (2018). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. *2018 IEEE International Conference on Big Data (Big Data)*, 2503–2510. https://doi.org/10.1109/BigData.2018.8621865

Inpainting — OpenCV 2.4.13.7 documentation. (2021). https://docs.opencv.org/2.4/modules/photo/doc/inpainting.html

Isman, A. (2011). Instructional Design in Education: New Model. *Turkish Online Journal of Educational Technology-TOJET*, *10*(1), 136–142.

Jagtap, S. (2019). Design creativity: refined method for novelty assessment. In *International Journal of Design Creativity and Innovation* (Vol. 7, Issues 1–2, pp. 99–115). https://doi.org/10.1080/21650349.2018.1463176

Jobanputra, M. (2019). Unsupervised Question Answering for Fact-Checking. *ArXiv Preprint ArXiv:1910.07154*. https://doi.org/10.48550/arXiv.1910.07154

Jodhi, V., Kulkarni, S., Jaypruiya, P., Kulkarni, S., & Padture, J. (2018). *Quillionz - World's First AI-Powered Question Generator*. https://www.quillionz.com/

Jordanous, A. (2012). A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. *Cognitive Computation*, *4*(3), 246–279. https://doi.org/10.1007/s12559-012-9156-1

Karatas, H., Alci, B., & Aydin, H. (2013). Correlation among high school senior students test anxiety, academic performance and points of university entrance exam. *Educational Research and Reviews*, *8*(13), 919–926. https://doi.org/10.5897/ERR2013.1462

Karimi, P., Maher, M. lou, Davis, N., & Grace, K. (2019). Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System. *ArXiv Preprint ArXiv:1906.10188*. https://doi.org/10.48550/arXiv.1906.10188

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137. https://doi.org/10.48550/arXiv.1412.2306

Kasravi, K., & Risov, M. (2009). Multivariate patent similarity detection. *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, HICSS*, 1–8. https://doi.org/10.1109/HICSS.2009.318

Kaufman, J. C., Plucker, J. A., & Russell, C. M. (2012). Identifying and assessing creativity as a component of giftedness. In *Journal of Psychoeducational Assessment* (Vol. 30, Issue 1, pp. 60–73). https://doi.org/10.1177/0734282911428196

Kaufman, J. C., & Sternberg, R. J. (2010). *The Cambridge handbook of creativity*. Cambridge University Press.

Kawai, N., Sato, T., & Yokoya, N. (2015). Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics*, *22*(3), 1236–1247. https://doi.org/10.1109/TVCG.2015.2462368

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., & Farhadi, A. (2016). A diagram is worth a dozen images. *European Conference on Computer Vision*, 235–251. https://doi.org/10.48550/arXiv.1603.07396

Kerner, H. R., Wagstaff, K. L., Bue, B. D., Wellington, D. F., Jacob, S., Horton, P., Bell, J. F., Kwan, C., & Amor, H. ben. (2020). Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions. *Data Mining and Knowledge Discovery*, *34*(6), 1642–1675. https://doi.org/10.1007/s10618-020-00697-6

Kerner, H. R., Wellington, D. F., Wagstaff, K. L., Bell, J. F., Kwan, C., & Amor, H. ben. (2019). Novelty Detection for Multispectral Images with Application to Planetary Exploration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 9484–9491. https://doi.org/https://doi.org/10.1609/aaai.v33i01.33019484

Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF--IDF, LDA, and Doc2Vec. *Information Sciences*, *477*, 15–29. https://doi.org/10.1016/j.ins.2018.10.006

Kim, E., & Horii, H. (2015). A study on an assessment framework for the novelty of ideas generated by analogical thinking. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S1877042815039142. https://doi.org/10.1016/j.sbspro.2015.06.435

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv Preprint ArXiv:1411.2539*. https://doi.org/10.48550/arXiv.1411.2539

Kondermann, D. (2013). Ground Truth Design Principles: An Overview. *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*. https://doi.org/10.1145/2501105.2501114

Kornish, L. J., & Jones, S. M. (2021). Raw Ideas in the Fuzzy Front End: V erbosity Increases Perceived Creativity. Marketing Science, 40(6), 1106-1122.

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., … Murphy, K. (2017). *Open Images V6*. https://storage.googleapis.com/openimages/web/index.html

Krejcie, R. v, & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, *30*(3), 607–610. https://doi.org/10.1177/001316447003000308

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Kubat, M. (2017). *An introduction to machine learning*. Springer.

Kuniavsky, M. (2010). *Smart things: ubiquitous computing user experience design*. Elsevier.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., & others. (2020). The open images dataset v4. *International Journal of Computer Vision*, 1–26. https://doi.org/10.1007/s11263-020-01316-z

Kwon, S. K., Lee, M., & Shin, D. (2017). Educational assessment in the Republic of Korea: lights and shadows of high-stake exam-based education system. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 60–77. https://doi.org/10.1080/0969594X.2015.1074540

Landa, R. (2018). *Graphic design solutions*. Cengage Learning.

LanguageTool - Online Grammar, Style & Spell Checker. (2019). https://languagetool.org/

Le, H. T., & Le, T. M. (2013). An approach to abstractive text summarization. *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, 371–376. https://doi.org/10.1109/SOCPAR.2013.7054161

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, 1188–1196. https://doi.org/10.48550/arXiv.1405.4053

Li, J., Su, Z., Geng, J., & Yin, Y. (2018). Real-time detection of steel strip surface defects based on improved yolo detection network. *IFAC-PapersOnLine*, *51*(21), 76–81. https://doi.org/10.1016/j.ifacol.2018.09.412

Li, M., Li, W., Wang, F., Jia, X., & Rui, G. (2021). Applying BERT to analyze investor sentiment in stock market. *Neural Computing and Applications*, *33*(10), 4663–4676. https://doi.org/10.1007/s00521-020-05411-7

Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gotzsche, P., Loannidis, J., Clarke, M., Devereaux, P., Kleijnen, J., & Moher, D. (2018). The novelty 'sweet spot' of invention. *Nature*, *388*, 539–547. https://doi.org/10.1017/dsj.2017.23

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Linder, R., Webb, A. M., & Kerne, A. (2013). Searching to Measure the Novelty of Collected Ideas. *CHI 2013 Evaluation Methods for Creativity Support Environments Workshop*, *36*.

Ling, Y., & Mahadevan, S. (2013). Quantitative model validation techniques: New insights. Reliability Engineering & System Safety, 111, 217-231.

Lorbeer, B., Kosareva, A., Deva, B., Softić, D., Ruppel, P., & Küpper, A. (2018). Variations on the clustering algorithm BIRCH. *Big Data Research*, *11*, 44–53. https://doi.org/10.1016/j.bdr.2017.09.002

Loudon, G. H., Wu, Y. M., & Pittman, J. A. (2005). U.S. Patent No. 6,956,969. Washington, DC: U.S. Patent and Trademark Office.

Ma, H., Chu, X., Xue, D., & Chen, D. (2017). A systematic decision making approach for product conceptual design based on fuzzy morphological matrix. In *Expert Systems with Applications* (Vol. 81, pp. 444–456). https://doi.org/10.1016/j.eswa.2017.03.074

Madabushi, H. T., Lee, M., & Barnden, J. (2018). Integrating question classification and deep learning for improved answer selection. *Proceedings of the 27th International Conference on Computational Linguistics*, 3283–3294.

Maguire, M. (2001). Methods to support human-centred design. *International Journal of Human-Computer Studies*, *55*(4), 587–634. https://doi.org/10.1006/ijhc.2001.0503

Markou, M., & Singh, S. (2003a). Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, *83*(12), 2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018

Markou, M., & Singh, S. (2003b). Novelty detection: a review—part 2:: neural network based approaches. *Signal Processing*, *83*(12), 2499–2521. https://doi.org/10.1016/j.sigpro.2003.07.019

Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., & Lopez, A. M. (2018). Metric learning for novelty and anomaly detection. *ArXiv Preprint ArXiv:1808.05492*. https://doi.org/10.48550/arXiv.1808.05492

Mayer, D. G., & Butler, D. G. (1993). Statistical validation. Ecological modelling, 68(1-2), 21-32.

McAuley, J. (2018). *Amazon question/answer data*. http://jmcauley.ucsd.edu/data/amazon/qa/

McAuley, J., & Yang, A. (2016). Addressing complex and subjective product-related queries with customer reviews. *Proceedings of the 25th International Conference on World Wide Web*, 625–635. https://doi.org/10.1145/2872427.2883044

McCarthy, M., Chen, C. C., & McNamee, R. C. (2018). Novelty and usefulness trade-off: Cultural cognitive differences and creative idea evaluation. Journal of Cross-Cultural Psychology, 49(2), 171-198.

McDaniel, R., Fanfarelli, J. R., & Lindgren, R. (2017). Creative Content Management : Importance , Novelty , and. *IEEE Transactions on Professional Communication*, *60*(2), 204–205. https://doi.org/10.1109/TPC.2017.2656578

Mehta, P., & Ingole, N. (2015). Mechanics of paper setting: Being a paper setter. *The Art of Teaching Medical Students*, *3*, 276–287.

Mehta, R., & Dahl, D. W. (2019). Creativity: Past, present, and future. *Consumer Psychology Review*, *2*(1), 30–49. https://doi.org/10.1002/ARCP.1044

Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 8, 142642-142668.

Merriam-Webster. (2021). *Novelty | Definition of Novelty by Merriam-Webster*. https://www.merriam-webster.com/dictionary/novelty

Mihaylova, T., Karadjov, G., Atanasova, P., Baly, R., Mohtarami, M., & Nakov, P. (2019). SemEval-2019 task 8: Fact checking in community question answering forums. *ArXiv Preprint ArXiv:1906.01727*.

https://doi.org/10.48550/arXiv.1906.01727

Mihaylova, T., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Mohtarami, M., Karadzhov, G., & Glass, J. (2018). Fact checking in community forums. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Milanowski, A. T., & Heneman, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, *15*(3), 193–212. https://doi.org/10.1023/A:1012752725765

Miljković, D. (2010). Review of novelty detection methods. *The 33rd International Convention MIPRO*, 593–598.

Mirabito, Y., & Goucher-Lambert, K. (2020). The Role of Idea Fluency and Timing on Highly Innovative Design Concepts. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 83976, p. V008T08A035). American Society of Mechanical Engineers.

Montgomery, C., & Rupp, A. A. (2005). A meta-analysis for exploring the diverse causes and effects of stress in teachers. *Canadian Journal of Education/Revue Canadienne de l'éducation*, 458–486. https://doi.org/10.2307/4126479

Montiel, J., Read, J., Bifet, A., & Abdessalem, T. (2018). Scikit-multiflow: A multi-output streaming framework. *The Journal of Machine Learning Research*, *19*(1), 2914–2915. https://doi.org/10.48550/arXiv.1807.04662

Naghieh, A., Montgomery, P., Bonell, C. P., Thompson, M., & Aber, J. L. (2015). Organisational interventions for improving wellbeing and reducing work-related stress in teachers. *Cochrane Database of Systematic Reviews*, *4*. https://doi.org/10.1002/14651858.CD010306.pub2

Nam, C. S., & Smith-Jackson, T. L. (2007). Web-Based Learning Environment: A Theory-Based Design Process for Development and Evaluation. *Journal of Information Technology Education: Research*, *6*(1), 23–43. https://www.learntechlib.org/p/111406. https://doi.org/10.28945/200

Nappi, J. S. (2017). The importance of questioning in developing critical thinking skills. *Delta Kappa Gamma Bulletin*, *84*(1), 30.

Oman, S. K., Tumer, I. Y., Wood, K., & Seepersad, C. (2013). A comparison of creativity and innovation metrics and sample validation through in-class design projects. *Research in Engineering Design*, *24*(1), 65–92. https://doi.org/10.1007/s00163-012-0138-9

Omari, A., Carmel, D., Rokhlenko, O., & Szpektor, I. (2016). Novelty based ranking of human answers for community questions. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 215–224. https://doi.org/10.1145/2911451.2911506

OpenCV: Image Inpainting. (2021). https://docs.opencv.org/4.x/df/d3d/tutorial_py_inpainting.html

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3–12.

Östlund, U., Kidd, L., Wengström, Y., & Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: a methodological review. *International Journal of Nursing Studies*, *48*(3), 369–383. https://doi.org/10.1016/j.ijnurstu.2010.10.005

Oxman, R. (2004). Think-maps: teaching design thinking in design education. *Design Studies*, *25*(1), 63–91. https://doi.org/10.1016/S0142-694X(03)00033-4

Park, D., Nam, J., & Park, J. (2020). Novelty and influence of creative works, and quantifying patterns of advances based on probabilistic references networks. *EPJ Data Science*, *9*(1), 2. https://doi.org/10.1140/epjds/s13688-019-0214-8

Park, I. S., Suh, Y. O., Park, H. S., Kang, S. Y., Kim, K. S., Kim, G. H., Choi, Y.-H., & Kim, H.-J. (2017). Item development process and analysis of 50 case-based items for implementation on the Korean Nursing Licensing Examination. *Journal of Educational Evaluation for Health Professions*, *14*. https://doi.org/10.3352/jeehp.2017.14.20

Park, J. Y. (2011). Design education online: Learning delivery and evaluation. *International Journal of Art & Design Education*, *30*(2), 176–187. https://doi.org/10.1111/j.1476-8070.2011.01689.x

Park, N. K., Chun, M. Y., & Lee, J. (2016). Revisiting individual creativity assessment: Triangulation in subjective and objective assessment methods. Creativity Research Journal, 28(1), 1-10. https://doi.org/10.1080/10400419.2016.1125259

Paul, R., & Elder, L. (2019). *The Thinker's Guide to Socratic Questioning*. Rowman & Littlefield.

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. *ArXiv Preprint ArXiv:1705.04304*. https://doi.org/10.48550/arXiv.1705.04304

Penumatsa, P., Ventura, M., Graesser, A. C., Louwerse, M., Hu, X., Cai, Z., & Franceschetti, D. R. (2006). The right threshold value: What is the right threshold of cosine measure when using latent semantic analysis for evaluating student answers? *International Journal on Artificial Intelligence Tools*, *15*(05), 767–777.

PÉrez, R. P. Ý., & Sharples, M. (2001). MEXICA: A computer model of a cognitive account of creative writing. Journal of Experimental & Theoretical Artificial Intelligence, 13(2), 119-139.

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., & Ferrari, V. (2020). Connecting vision and language with localized narratives. *European Conference on Computer Vision*, 647–664. https://doi.org/10.48550/arXiv.1912.03098

Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

Prilleltensky, I., Neff, M., & Bessell, A. (2016). Teacher stress: What it is, why it's important, how it can be alleviated. *Theory Into Practice*, *55*(2), 104–111. https://doi.org/10.1080/00405841.2016.1148986

Pritzker, S. R., & Runco, M. A. (2011). *Encyclopedia of creativity*. Academic Press/Elsevier.

Prost, F., Thain, N., & Bolukbasi, T. (2019). Debiasing Embeddings for Reduced Gender Bias in Text Classification. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 69–75. https://doi.org/10.18653/v1/W19-3810

Rack, O., Zahn, C., & Mateescu, M. (2018). Coding and Counting-Frequency Analysis for Group Interaction. *The Cambridge Handbook of Group Interaction Analysis*.

Ranasinghe, T., Zampieri, M., & Hettiarachchi, H. (2019). BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. *FIRE (Working Notes)*, 199–207.

Ranjan, B. S. C., Siddharth, L., & Chakrabarti, A. (2018). A systematic approach to assessing novelty, requirement satisfaction, and creativity. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, *32*(4), 390–414. https://doi.org/10.1017/S0890060418000148

Rao, A. S., Avadhani, P. S., & Chaudhuri, N. B. (2017). Detecting targeted malicious e-mail using linear regression algorithm with data mining techniques. In *Computational intelligence in data mining* (pp. 23–35). Springer. . https://doi.org/10.1007/978-981-10-3874-7_3

Rao, G., Huang, W., Feng, Z., & Cong, Q. (2018). LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*, *308*, 49–57. https://doi.org/10.1016/j.neucom.2018.04.045

Rashid, S., & Qaisar, S. (2016). Developing Critical Thinking through Questioning Strategy among Fourth Grade Students. *Bulletin of Education and Research*, *38*(2), 153–168.

Ray, S. (2019). A quick review of machine learning algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39. https://doi.org/10.1016/10.1109/COMITCon.2019.8862451

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You look only once: unified real-time object detection. *ArXiv Preprint ArXiv:1506.02640*. https://doi.org/10.48550/arXiv.1506.02640

Reinhold, J. C., He, Y., Han, S., Chen, Y., Gao, D., Lee, J., Prince, J. L., & Carass, A. (2020). Finding novelty with uncertainty. *Medical Imaging 2020: Image Processing*, *11313*, 113130H. https://doi.org/10.48550/arXiv.2002.04626

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics*, *19*(1), 159–177.

Richards, K. A. R., Washburn, N., Carson, R. L., & Hemphill, M. A. (2017). A 30-Year Scoping Review of the Physical Education Teacher Satisfaction Literature. *Quest*, *69*(4), 494–514. https://doi.org/10.1080/00336297.2017.1296365

Richey, R. C., Fields, D. C., & Foxon, M. (2001). *Instructional design competencies: The standards*. ERIC.

Rogers, J., Smith, A. M., Ham, Y., & Russell, J. M. (2019). Course Delivery Format and Exam Performance Stratified by Question Difficulty. *The FASEB Journal*, *33*(S1), 592–598. 10.1096/fasebj.2019.33.1_supplement.598.2

Rothwell, W. J., & Kazanas, H. C. (2011). *Mastering the instructional design process: A systematic approach*. John Wiley & Sons.

Rowland, G. (1993). Designing and instructional design. *Educational Technology Research and Development*, *41*(1), 79–91. https://doi.org/10.1007/BF02297094

Ruiz-Pastor, L., Chulvi, V., Mulet, E., & Royo, M. (2021). The relationship between personal intrinsic factors towards a design problem and the degree of novelty and circularity. *Research in Engineering Design*, *0123456789*. https://doi.org/10.1007/s00163-021-00374-9

Runco, M. A., & Plucker, J. A. (2001). *Commemorating Guilford's 1950 Presidential Address: A Special Double Issue of creativity Research Journal*. Routledge. https://doi.org/10.4324/9781410608604

Russakovsky, O., Deng, J., Huang, Z., Berg, A. C., & Fei-Fei, L. (2013). Detecting Avocados to Zucchinis: What Have We Done, and Where Are We Going? *2013 IEEE International Conference on Computer Vision*, 2064–2071. https://doi.org/10.1109/ICCV.2013.258

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Saaksjarvi, M., & Goncalves, M. (2018). Creativity and meaning: including meaning as a component of creative solutions. *Artificial Intelligence ForEngineering Design, Analysisand Manufacturing*. https://doi.org/10.1017/S0890060418000112

Sahamid, H. (2016). Developing Critical Thinking through Socratic Questioning: An Action Research Study. *International Journal of Education and Literacy Studies*, *4*(3), 62–72.

Said-Metwaly, S., den Noortgate, W., & Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. *Creativity. Theories--Research-Applications*, *4*(2), 238–275. https://doi.org/10.1515/ctra-2017-0013

Sakata, W., Shibata, T., Tanaka, R., & Kurohashi, S. (2019). FAQ retrieval using query-question similarity and BERT-based query-answer relevance. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1113–1116. https://doi.org/10.1145/3331184.3331326

Salehi, M., Eftekhar, A., Sadjadi, N., Rohban, M. H., & Rabiee, H. R. (2020). Puzzle-AE: Novelty detection in images through solving puzzles. *ArXiv Preprint ArXiv:2008.12959*.

Sanetti, L. M. H., Boyle, A. M., Magrath, E., Cascio, A., & Moore, E. (2020). Intervening to Decrease Teacher Stress: a Review of Current Research and New Directions. *Contemporary School Psychology*, 1–10.

Sangkloy, P., Lu, J., Fang, C., Yu, F., & Hays, J. (2017). Scribbler: Controlling Deep Image Synthesis with Sketch and Color. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6836–6845. https://doi.org/10.1109/CVPR.2017.723

Sarkar, P., & Chakrabarti, A. (2011). Assessing design creativity. *Design Studies*, *32*(4), 348–383. https://doi.org/10.1016/j.destud.2011.01.002

Sarkar, P., Chakrabarti, A., & others. (2007). Development of a method for assessing design creativity. *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007*, 349–350.

Sbaraini, A., Carter, S. M., Evans, R. W., & Blinkhorn, A. (2011). How to do a grounded theory study: a worked example of a study of dental practices. *BMC Medical Research Methodology*, *11*(1), 1–10. https://doi.org/10.1186/1471-2288-11-128

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, *98*, 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, *13*(7), 1443–1471. https://doi.org/10.1162/089976601750264965

Schumann, J., Strothotte, T., Laser, S., & Raab, A. (1996). Assessing the effect of non-photorealistic rendered images in CAD. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 35–41. https://doi.org/10.1145/238386.238398

scipy.ndimage.gaussian_filter — SciPy v1.7.1 Manual. (2021). https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.gaussian_filter.html

Self, J. A. (2019). Communication through design sketches: Implications for stakeholder interpretation during concept design. *Design Studies*, *63*, 1–36. https://doi.org/10.1016/j.destud.2019.02.003

Sharma, A., & Kirloskar, V. (2008). *India Design Report*. https://www.cii.in/webcms/Upload/a2.pdf

Sharma, B. (2016). A focus on reliability in developmental research through Cronbach's Alpha among medical, dental and paramedical professionals. *Asian Pacific Journal of Health Sciences*, *3*(4), 271–278. https://doi.org/10.21276/apjhs.2016.3.4.43

Sharma, R., & Kaushik, B. (2020). Offline recognition of handwritten Indic scripts: A state-of-the-art survey and future perspectives. Computer Science Review, 38, 100302.

Shedko, A. Y. (2018). Semantic-map-based assistant for creative text generation. *Procedia Computer Science, 123*, 446–450. https://doi.org/10.1016/j.procs.2018.01.068

Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, *2*(1), 1–37. https://doi.org/10.1145/3419106

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Silvia, P. J. (2008). Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan. *Creativity Research Journal*, *20*(1), 34–39. https://doi.org/10.1080/10400410701841807

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Singhal, S., & Bhattacharya, A. (2015). Abstractive text summarization. *Department of Computer Science IIT Kanpur*, 1–11. https://www.studocu.com/en-us/document/harvard-university/introduction-to-computer-science/report-grade-a/5662797

Sintini, L., & Kunze, L. (2020). Unsupervised and Semi-Supervised Novelty Detection using Variational Autoencoders in Opportunistic Science Missions. *Proceedings of the British Machine Vision Conference (BMVC)*.

Skaalvik, E., & Skaalvik, S. (2016). Teacher Stress and Teacher Self-Efficacy as Predictors of Engagement, Emotional Exhaustion, and Motivation to Leave the Teaching Profession. *Creative Education*, *07*, 1785–1799. https://doi.org/10.4236/ce.2016.713182

Skaalvik, E., & Skaalvik, S. (2017). Teacher Stress and Teacher Self-Efficacy: Relations and Consequences. *Aligning Perspectives on Health, Safety and Well-Being. Springer, Cham.* https://doi.org/https://doi.org/10.1007/978-3-319-53053-6_5

Skansi, S. (2018). *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer.

Sluis-Thiescheffer, W., Bekker, T., Eggen, B., Vermeeren, A., & de Ridder, H. (2016). Measuring and comparing novelty for design solutions generated by young children through different design methods. In *Design Studies* (Vol. 43, pp. 48–73). https://doi.org/10.1016/j.destud.2016.01.001

Smith, P. L., & Ragan, T. J. (2004). *Instructional design*. John Wiley & Sons.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., & Smeulders, A. W. M. (2006). The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. *Proceedings of the 14th ACM International Conference on Multimedia*, 421–430. https://doi.org/10.1145/1180639.1180727

Soborof, I., & Harman, D. (2005). Novelty detection: The TREC Experience. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 105–112.

Sohoni, M. (2016). The AICTE review: An opportunity for engineering education reform. *Current Science*, *110*(2), 159–166. https://doi.org/10.18520/cs/v110/i2/159-166

Song, J., Song, Y.-Z., Xiang, T., & Hospedales, T. (2017). Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. *The British Machine Vision Conference (BMVC 2017), Bmvc*. https://dx.doi.org/10.5244/C.31.45

Srinivasan, V., & Chakrabarti, A. (2008). DESIGN FOR NOVELTY-A FRAMEWORK? In *designsociety.org*. https://www.designsociety.org/download-publication/26735/design_for_novelty_–_a_framework

Srinivasan, V., & Chakrabarti, A. (2010). Investigating novelty--outcome relationships in engineering design. *AI EDAM*, *24*(2), 161–178. https://doi.org/10.1017/S089006041000003X

Stefanuk, B., Pascual, A., Skonieczny, K., McIsaac, K., Raimalwala, K., Smal, E., & Faragalli, M. (2020). DETECTING NOVELTIES ON PLANETARY SURFACES WITH AUTOENCODERS. *I-SAIRAS2020*.

Still, A., & d'Inverno, M. (2016). A history of creativity for future AI research. *Proceedings of the 7th International Conference on Computational Creativity, ICCC 2016*, 147–154.

Stoycheva, K. (2010). Tolerance for ambiguity, creativity, and personality. *Bulgarian Journal of Psychology*, *1–4*, 178–188.

Stronge, J. H. (2018). *Qualities of effective teachers*. ASCD. https://eric.ed.gov/?id=ED582452#:~:text=A%20research%2Dbased%20framework%20for,%3B%20and%20(6)%20Professionalism.

Sung, S. H., Li, C., Chen, G., Huang, X., Xie, C., Massicotte, J., & Shen, J. (2021). How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology*, *30*(2), 210–226. https://doi.org/10.1007/s10956-020-09856-2

Swenson, J. (2016). Teaching By Design Using a Question Template To Transform Student Writing. *Business Education Innovation Journal*, *8*(2).

Szymkowski, M., Saeed, E., Omieljanowicz, M., Omieljanowicz, A., Saeed, K., & Mariak, Z. (2020). A Novelty Approach to Retina Diagnosing Using Biometric Techniques With SVM and Clustering Algorithms. *IEEE Access*, *8*, 125849–125862. https://doi.org/10.1109/ACCESS.2020.3007656

Taborda, E., Chandrasegaran, S. K., Kisselburgh, L., Reid, T., & Ramani, K. (2013). Enhancing Visual Thinking in a Toy Design Course Using Freehand Sketching. *Proceedings of the ASME Design Engineering Technical Conference*, *7*, 267–276. https://doi.org/10.1115/DETC2012-71454

Takai, S., Esterman, M., & Midha, A. (2015). An approach to study associations between design concepts and design outcomes. *Journal of Mechanical Design*, *137*(4), 41101. https://doi.org/10.1115/1.4028992

Tan, J. S. T. (2017). Factors Affecting Stress among Faculty Members of Public Universities in the Philippines: A Multiple Regression Analysis. *International Journal of Psychological Studies*, *9*(3), 64. https://doi.org/10.5539/ijps.v9n3p64

Tao, Z., Liu, H., Fu, H., & Fu, Y. (2017). Image Cosegmentation via Saliency-Guided Constrained Clustering with Cosine Similarity. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4285–4291.

Thomas, A., Kowar, M. K., & Sharma, S. (2008). Intelligent Fuzzy decision making for subjective answer evaluation using utility functions. *2008 First International Conference on Emerging Trends in Engineering and Technology*, 587–591. https://doi.org/10.1109/ICETET.2008.216

Trieu, L. Q., Tran, H. Q., & Tran, M.-T. (2017). News Classification from Social Media Using Twitter-Based Doc2Vec Model and Automatic Query Expansion. *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 460–467. https://doi.org/10.1145/3155133.3155206

Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information Processing and Management*, *48*(2), 303–325. https://doi.org/10.1016/j.ipm.2011.07.006

VGG16 and VGG19. (2014). https://keras.io/api/applications/vgg/#vgg19-function

Vodolazova, T., Lloret, E., Muñoz, R., & Palomar, M. (2013). Extractive text summarization: can we use the same techniques for any text? *International Conference on Application of Natural Language to Information Systems*, 164–175. https://doi.org/10.1007/978-3-642-38824-8_14

Wachs, J., Daróczy, B., Hannák, A., Páll, K., & Riedl, C. (2018). And now for something completely different: Visual novelty in an online network of designers. *Proceedings of the 10th ACM Conference on Web Science*, 163–172. https://doi.org/10.1145/3201064.3201088

Wan, M., & McAuley, J. (2016). Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 489–498. https://doi.org/10.1109/ICDM.2016.0060

Wan, W., Wu, J., Shi, G., Li, Y., & Dong, W. (2018). Super-Resolution Quality Assessment: Subjective Evaluation Database and Quality Index Based on Perceptual Structure Measurement. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME.2018.8486519

Wang, H. (2013). A commentary on China’s new curriculum and the programs to design subjects for the College Entrance Examination. *Chinese Education & Society*, *46*(1), 32–40. https://doi.org/10.2753/CED1061-1932460103

Wang, S., & Jiang, J. (2015). Learning natural language inference with LSTM. *ArXiv Preprint ArXiv:1512.08849*. https://doi.org/10.48550/arXiv.1512.08849

Wang, S., Li, H., Li, J., Zhang, Y., & Zou, B. (2018). Automatic analysis of lateral cephalograms based on multiresolution decision tree regression voting. *Journal of Healthcare Engineering*, *2018*. https://doi.org/10.1155/2018/1797502

Wang, W., & Lu, Y. (2018). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. *IOP Conference Series: Materials Science and Engineering*, *324*(1), 12049. https://doi.org/10.1088/1757-899X/324/1/012049

Wang, Y., Chang, L., Cheng, Y., Jin, L., Cheng, Z., Deng, X., & Duan, F. (2018). TEXT2SKETCH : LEARNING FACE SKETCH FROM FACIAL ATTRIBUTE TEXT. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 669–673. https://doi.org/10.1109/ICIP.2018.8451236

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615. https://doi.org/10.18653/v1/D16-1058

Wang, Z., Liu, D., Chang, S., Dolcos, F., Beck, D., & Huang, T. (2017). Image aesthetics assessment using Deep Chatterjee's machine. *2017 International Joint Conference on Neural Networks (IJCNN)*, 941–948. https://doi.org/10.1109/IJCNN.2017.7965953

Ward, T. B., & Kolomyts, Y. (2010). Cognition and creativity. *The Cambridge Handbook of Creativity*, 93–112.

Watts, L. S., & Blessinger, P. (2016). *Creative learning in higher education: International perspectives and approaches*. Routledge.

Wei, W., Ming, Z., Nie, L., Li, G., Li, J., Zhu, F., Shang, T., & Luo, C. (2016). Exploring heterogeneous features for query-focused summarization of categorized community answers. *Information Sciences*, *330*, 403–423. https://doi.org/10.1016/j.ins.2015.10.024

Wei, Z., Wang, Y., He, S., & Bao, J. (2017). A Novel Intelligent Method for Bearing Fault Diagnosis Based on Affinity Propagation Clustering and Adaptive Feature Selection. *Know.-Based Syst.*, *116*(C), 1–12. https://doi.org/10.1016/j.knosys.2016.10.022

Weisbrod, G., & Kroll, E. (2018). Idea-configuration-evaluation (ICE): development and demonstration of a new prescriptive model of the conceptual engineering design process based on parameter analysis and C--K theory. *Research in Engineering Design*, *29*(2), 203–225. https://doi.org/10.1007/s00163-017-0263-6

Welch, M., Barlex, D., & Lim, H. S. (2000). Sketching: Friend or foe to the novice designer? *International Journal of Technology and Design Education*, *10*(2), 125–148. https://doi.org/10.1023/A:1008991319644

Wentling, T. L., & Johnson, S. D. (1999). The Design and Development of an Evaluation System for Online Instruction. *Proceedings of WebNet World Conference on the WWW and Internet 1999*, 1149–1154. https://www.learntechlib.org/p/7340

Wiesche, M., Jurisch, M. C., Yetton, P. W., & Krcmar, H. (2017). Grounded theory methodology in information systems research. *MIS Quarterly*, *41*(3), 685–701. https://doi.org/10.25300/MISQ/2017/41.3.02

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Wu, Y. (2013). An empirical study of narrative imagery in implicit and explicit contexts. Computers in Human Behavior, 29(4), 1580-1589.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. *ArXiv Preprint ArXiv:1904.02232*. https://doi.org/10.48550/arXiv.1904.02232

Xueqing, Z., Pavlos, M., Tobias, S., & Matthias, R. (2018). A high-performance approach for enhancing the clarity of hand-drawn sketch images. *Basic Sciences Journal of Textile Universities/Fangzhi Gaoxiao Jichu Kexue Xuebao*, *31*(2).

Yaribeygi, H., Panahi, Y., Sahraei, H., Johnston, T. P., & Sahebkar, A. (2017). The impact of stress on body function: A review. *EXCLI Journal*, *16*, 1057. https://doi.org/10.17179/excli2017-480

Yilmaz, S., & Toklu, S. (2020). A deep learning analysis on question classification task using Word2vec representations. *Neural Computing and Applications*, 1–20. https://doi.org/10.1007/s00521-020-04725-w

Yu, S.-S., Chu, S.-W., Wang, C.-M., Chan, Y.-K., & Chang, T.-C. (2018). Two improved k-means algorithms. *Applied Soft Computing*, *68*, 747–755. https://doi.org/10.1016/j.asoc.2017.08.032

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data--which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*(1), 1–10. https://doi.org/10.1186/s12874-016-0200-9

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2019.1648300

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. ACM sigmod record, 25(2), 103-114.

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, *1*(2), 141–182. https://doi.org/10.1023/A:1009783824328

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zolfaghari, A. R., Fathi, D., & Hashemi, M. (2011). Role of creative questioning in the process of learning and teaching. *Procedia-Social and Behavioral Sciences*, *30*, 2079–2082. https://doi.org/10.1016/j.sbspro.2011.10.404

Appendix

Appendix A. Interview questions for identifying dimensions of questions that has the potential to instigate creative responses from students

1. Suppose you have to formulate questions for Design aptitude testing. There are different types of questions one can frame to test creativity such a Mathematical, Logical Artistic, Hidden Mystery, Metaphoric, etc which one do you use normally ? Which one would you be interested in learning more about ?
2. In your view, how do you define or explain a creative question?
3. Can you educate us on how a Creative question becomes different from a non-creative routine one ?
4. Do other types of Question tests creative aptitude ?
5. What are the qualities/ingredients must a question have to be classified as creative?
6. Which type of questions are suitable for framing a creative question? a) Open-ended b) Close-ended.
7. Why do think open-ended/close-ended question is significant for framing creative questions?
8. Which type of questions are suitable for framing a creative question? a) Subjective b) Objective.
9. Why do you think subjectivity/objectivity is significant in framing creative questions?
10. How important is it for you to verify the extent to which intent of a question is understood? Why? Explain.
11. How important is it for you to verify the extent to which a question is conversational? Why so? Explain
12. How important is it for you to expect creativity associated with facts? (Example- Design a chair with material specification, fulfilling ergonomic criteria, etc.)
13. As a design educator is it important to include some ‘ application’ aspect in a creative question?
14. How important is it for you to frame questions to get uncommon answers? (Going to school by road, rail, water is common, but ropeway is an uncommon solution)
15. How important is it for you to check your questions whether they look interesting to others? (If you check, then how?)

16. How much do you rate your questions in terms of interesting in a scale of 1 to 5 where 1 being the least interesting and 5 is most interesting?

17. Which type of questions do you generally prefer as creative questions: a) Questions conveying same interpretation across all respondents b) Questions conveying multiple interpretation. (Example: Have you smoked at least 100 cigarettes in your entire life ? One may make multiple interpretations for this question. a) Cigarettes that are only inhaled b) Any cigarettes, whether or not you inhaled c) Cigarettes that are completely finished d) Cigarettes that are partially smoked e) Cigarettes that only took a puff or two inhaled f) It may be manufactured cigarettes g) It may be hand-rolled cigarettes h) It may be marijuana cigarettes i) It may be cigars j) It may be clove cigarettes.

18. How important is it for you to check whether a question can really be reported as a creative question in mass examination? If important, how?

19. Does your creative questions seek opinion from students?

20. Does your question look for a comparison of alternatives of solutions?

21. Does your question look for a consequence of a particular action? (What are preferences if you sell this car?)

22. Does your question sometimes look for procedures?

23. How important is it for you to frame question that seeks a well-explained solution?

24. How important is it for you to check the narration while framing questions?

25. How do you identify creative questions among a bunch of questions? Explain the criteria you use to identify them.

26. What are the factors of creative questions, in your opinion?

27. Explain the process of framing creative questions?

28. After framing a question, how do you assess whether the question is creative?

29. How do you assess the degree of creativity in a question in terms of low, medium, high, and very high?

30. What would be your suggestion for framing creative questions?

31. What are your recommendations for identifying creative questions from a bunch of other questions?

Appendix B. Concise form of questionnaire to capture the parameters to evaluate novelty in descriptive creative responses

Concepts that will help to fill up the form are as follows:

Novelty

Dictionary meaning: The quality of being new, original, or unusual.

Definition: Novelty encompasses both new and original. It does not resemble something formerly known. It may also be defined with reference to the previous ideas of the individual concerned.

Grammatical mistake

Meaning: Mistakes associated with sentence construction.

Misspellings

Meaning: Mistakes associated with spelling of a word.

Relevance between question and solution

Meaning: The answer must ft within the task constraints.

Narration link between sentences

Meaning: Able to convey a narration without invoking a break or diverging the concept. *Unique concept of solution*

Meaning: New and original concept, possessing least similarity among other solutions. *Objective questions*

Meaning: Objective answers are restricted to factual knowledge. E.g. Multiple choice questions, fill in the blanks, match the following, etc.

Subjective questions

Meaning: Subjective answers reflect personal opinions, advices, preferences, and experiences. E.g. Descriptive short/long questions, essays, etc.

Descriptive responses

Meaning: This type of *responses* comprises of sentences and paragraphs.

E.g. There are multiple ways by which I can go to school. I can go by car, bike, or school bus. Sometimes, I think of teleportation so that I can reach from one place to another within a fraction of second. However, teleportation is not a feasible idea right now. What if I can design a cost effective flying machine to bypass the traffic? This design with start with a feasible structure of a vehicle. Further, conceptualization would take place where there might be combination of ideas from multiple domains.

1. Which types of *responses* are more significant for evaluating creativity in students? (subjective/objective/both)

2. Which of the following factors/stimuli is important for you during the evaluation of novelty in descriptive *responses*? Please evaluate factors from 'Very important' to 'Not at all important'. (grammatical mistake, misspellings, relevancy between question and solution, narration link between sentences, unique concept of solution, any additional factor(s))

3. Rank the following factors in terms of your prioritization to evaluate novelty in descriptive *responses*. The ranks can range from 1 to 5 and two or more factors can get the same rank depending on your choice. Rank 1 is the most prioritized factor whereas rank 5 is the least prioritized factor. (grammatical mistake, misspellings, relevancy between question and solution, narration link between sentences, unique concept of solution, any additional factor(s))

4. Personal details were collected like current profession, specialization, highest qualification, years of association with design field, experience in selecting candidates, parameters effecting the evaluation, experience in selecting candidates based on creativity

Appendix C. Illustration of items in the questionnaire

	Very important	*Slightly more important*	*Important*	*Slightly important*	*Not at all important*
Grammatical mistake	○	○	○	○	○
Misspellings	○	○	○	○	○
Fluency	○	○	○	○	○
Flexibility	○	○	○	○	○
Usefulness	○	○	○	○	○
Clarity	○	○	○	○	○
Relevant to the question	○	○	○	○	○
Choice of colours	○	○	○	○	○
Sketching ability	○	○	○	○	○
Narration link between sentences	○	○	○	○	○
Unique concept	○	○	○	○	○

Appendix D. Concise form of questionnaire to capture the parameters to evaluate novelty in labelled image-based and annotated image-based pattern of creative responses

The following concepts might help in answering the survey questions.

Image with label responses

Image with labels refers to drawings or sketches marked with texts.

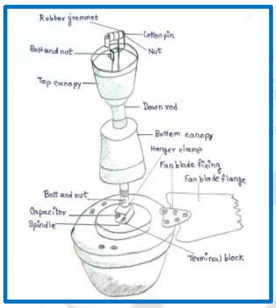
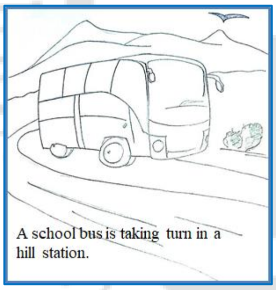
Example. 

Image with annotation responses

Annotated images are a solution pattern that exhibits drawings or sketches agglomerated with textual descriptions.

Example. 

Novelty

Novelty encompasses both new and original. It does not resemble something formerly known. It may also be defined with reference to the previous ideas of the individual concerned.

Relevance

The solution fits within task constraints. The solution does what it is supposed to do.

Uniqueness

New and original concept, possessing least similarity among other solutions.

Clarity

Clarity in visual solution is essential which indicates comprehensibility and clear representation. It is important in images as it empowers understanding ability.

Sketching ability
Sketching ability refers to potential of students in visually representing their ideas.

Choice of colors
Choice of colours is associated with slecting a set of colors over another. It also attempts to optimize aesthetic feature of an image.

Process
Process refers to steps of solving a problem.

Simplicity
Simplicity of an image with label solution refers to a representation that is easy to perceive.

Imaginative
Imaginative associated with any annotated image-based solution represents creative and divergent ideas.

Versatility
A solution which is adaptable based on requirement is considered as versatile.

Language processing
Language processing refers to checking of spelling and grammar, however, it is applicable for solutions consisting of descriptions.

Narration
Narrations indicate coherence between sentences.

Subject Knowledge
Subject knowledge refers to adequate perception of a domain of interest.

Presentation
Presentation is a feature in which a solution is illustrated in a clear and concise manner.

Refining
Refining indicates the iterative process by which a solution attempts to improve gradually.

Survey questions:

1. What is your current profession?
2. Gender?
3. Age?
4. What is your specialization?

a) Industrial design, b) Design engineering, c) Mechanical engineering, d) Artistic design, e) Other

5. For how long have you been associated with the design field?

6. How long have you been evaluating novelty in design aptitude?

7. Which types of responses are more significant for evaluating creativity in students?

   a) Subjective, b) objective, c) both

8. Which of the following factors/stimuli are significant for evaluation of novelty in image with label solutions? (Please evaluate factors from ‘Very important’ to ‘Not at all important’).

   a) Relevance, b) Uniqueness, c) Clarity, d) Sketching ability, e) Choice of colors, f) Process,

   g) Simplicity, h) Imaginative, i) Versatility, j) other

9. Which of the following factors/stimuli are significant for evaluation of novelty in image with annotation responses? (Please evaluate factors from ‘Very important’ to ‘Not at all important’).

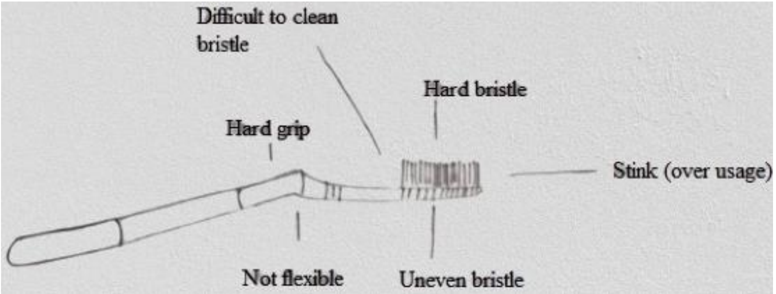
   a) Relevance, b) Uniqueness, c) Clarity, d) Sketching ability, e) Choice of colors, f) Language processing, g) Narration, h) Imaginative, i) Subject Knowledge, j) Versatility, k) Presentation, l) Refining, m) other

Appendix E. Sample creative questions and their responses

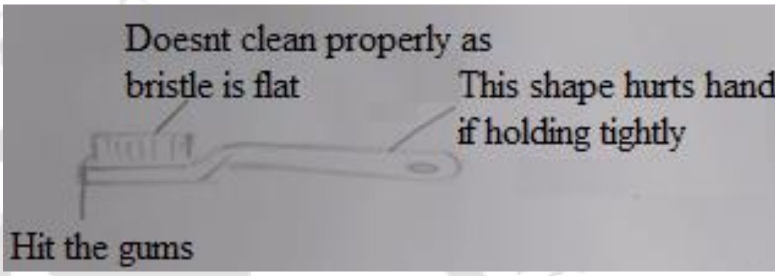
Questions	Responses
Two bulbs are in close proximity. Each of these bulbs is half-filled with water and a fish. Do you find anything interesting that will build up a small narrative? Tell your story.	Response 1: Fishes are part of the nature. A pairs of friends suddenly one early morning got on the fishermen's net and found themselves cagged in two different bulbs. The fishermen sold them to an installation artist who requested them from each other. The bulbs were the cages fo the friends. They are loved each other so much, so they are having communication by signals. Design inventions could help this installation in a way that both art and nature could co-exist. Response 2: In a family there lived timon and pumba who brothers. Timon is quite serious always, but pumba a little bit pranky and naughty. Timon and pumba used to love two gold fishes which were in an aquarium. On a stormy night the aquarium fell down and fish were about to die, some how timon saved them by putting them in a water placed, but pumba being a naughty child thought of a plan to store these fishes in two bulbs. This made them ascured and happy as they can carry their loved fishes in their hands. Response 3: The narration is a brilliant example of recycling or reusing products. We know every year, hundreds of filament bulbs are being converted into huge waste, and managing such a huge quantity has become a headache for many local self-government bodies. This question shows a new application of otherwise environmental hazard. It is a brilliant idea to remove the filament and fill it with fish. It may turn into a mini or pocket aquarium. The value can further be added by spray painting the surface with multiple column colors, which enhances the visual appeal. So the photograph clearly depicts an interesting design thought by converting waste into a fascinating new product. It has clearly good intentions too. It reduces environmental waste and introduces a new visually appealing product.
Sketch a tooth brush. According to you what are the 6 most important problems that users might face while using such a	

toothbrush? Label these problems in the part of the brush that they may occur in few words. Do not suggest solutions.

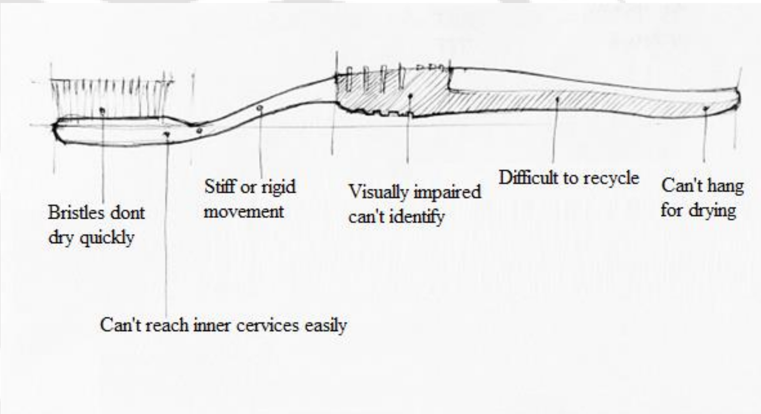
Response 1:



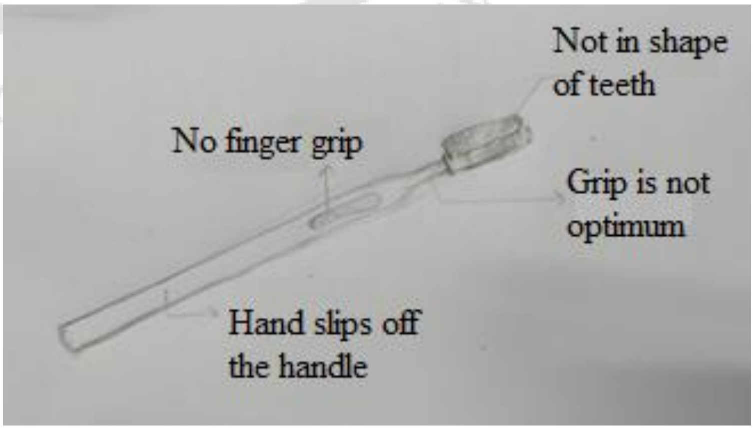
Response 2:



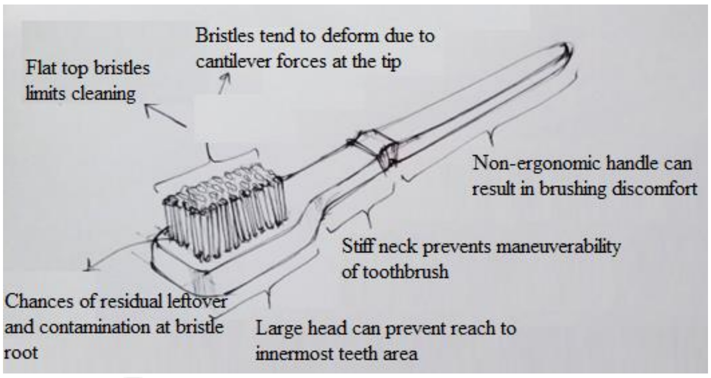
Response 3:



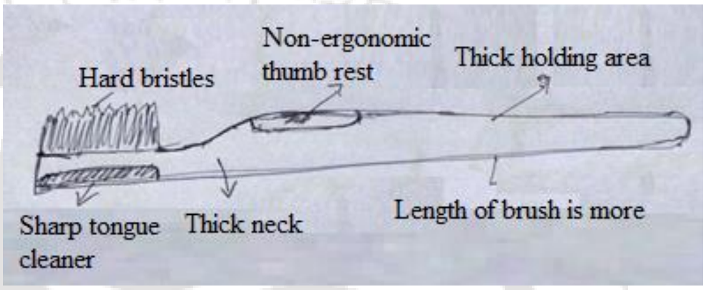
Response 4:



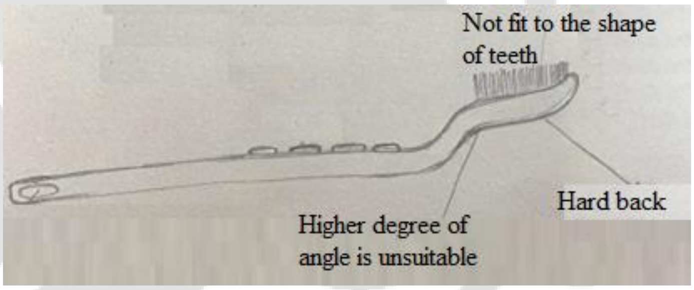
Response 5:



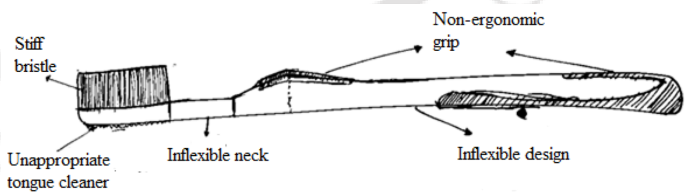
Response 6:



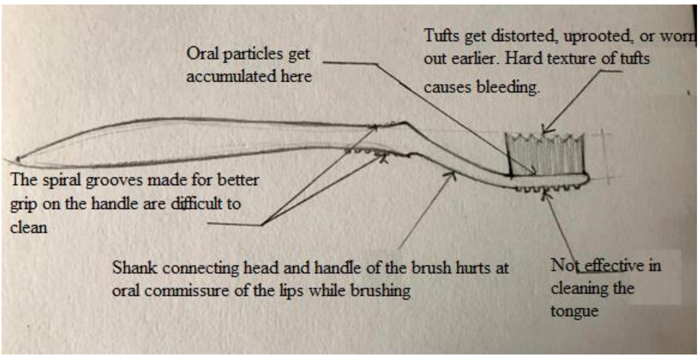
Response 7:



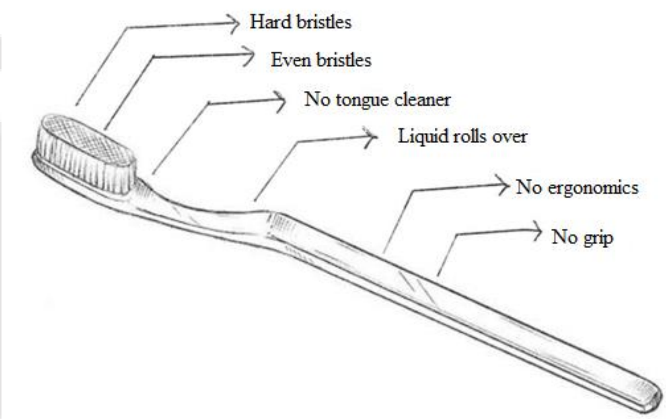
Response 8:



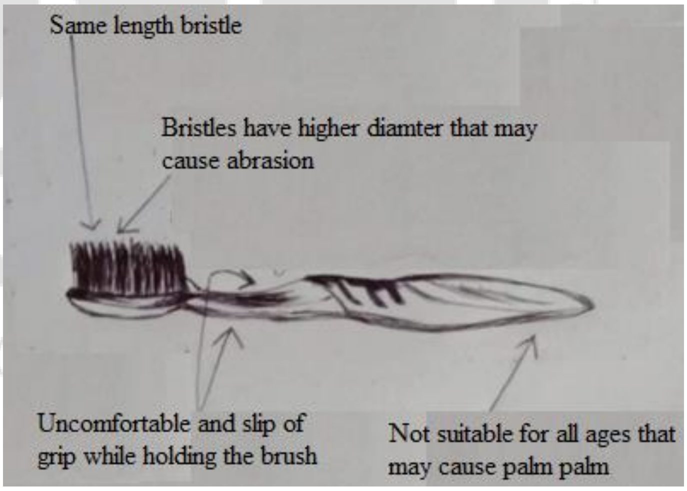
Response 9:



Response 10:



Response 11:



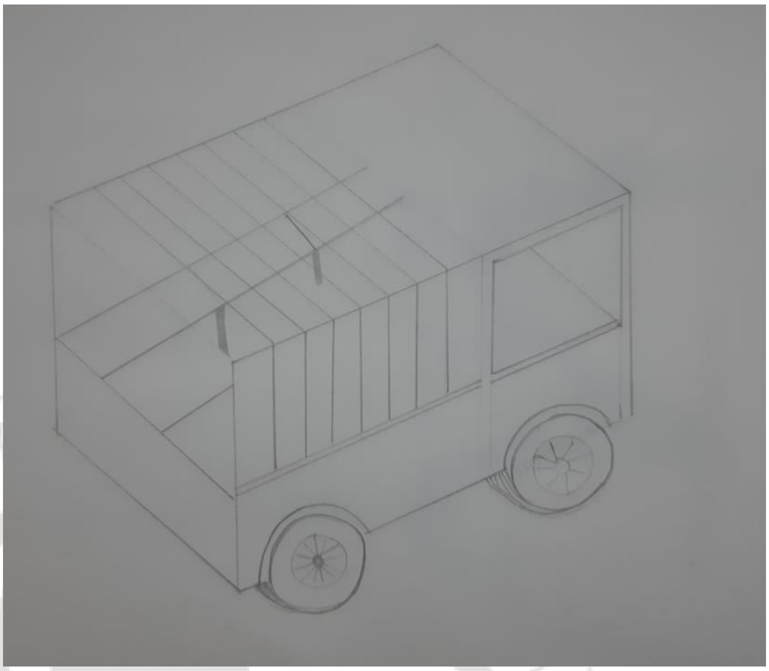
There are many local innovative transportation solutions that mostly operate in rural India. These vehicles are flexible to carry large number of people, smaller machine equipments, agro-machinery and even livestock like cattle, goat, etc. These transport systems are neither safe nor comfortable but are indispensible in rural scenarios. At times, these are the only modes of transportation available. You are expected to think about the context, the target users (needs and expectations), and the purpose. Your observations and insights should lead to the development of original solutions. You are free to choose one concept over others, develop the final rendering and detail the key features of the final concept.

Task: Design a context sensitive, safe, workable and pleasing 4-wheeler vehicle for hilly regions that fulfils the following criteria: a) Carrying capacity of at least 6 passengers, including the driver. b) Has a convertible space that can be used to carry at least 1 cattle/small machine/milk cans/fuel drums and a tool kit. If no goods are transported, the space can also be used to take passengers. c) Provides protection with respect to weather conditions (such as sunshades). d) Has provisions for easy maintenance.

Present your design proposal as follows:

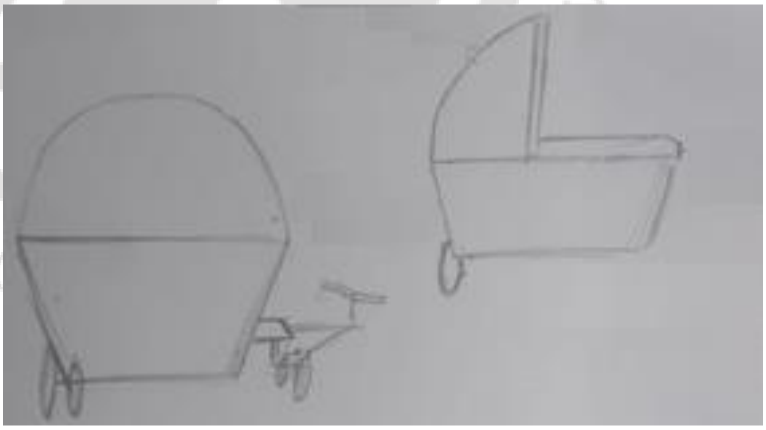
a) Sketch a unique design concept

Response 1:



The four-wheel mobile has a compatible design. It could accommodate 6 people. Mostly the passengers could be seated at the back of the mobile. The Seats can be removed to make way for the cattle and goods. The upper cover is made up of a trampoline tightened through ropes so it can be removed to accommodate cattle. The design can even help in safe travel towards the hilly areas.

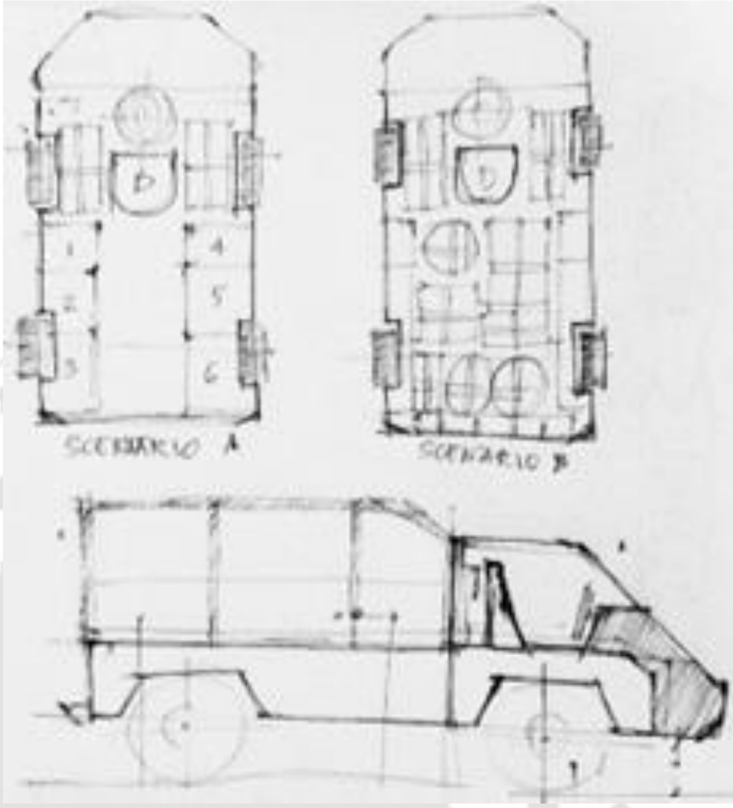
Response 2:



A convertible 4-wheel rickshaw designed for rural areas with big wheels and protected areas can be used as a storing place as well as for passengers.

b) Write a passage/description of the sketch and how it justifies the problem statement.

Response 3:

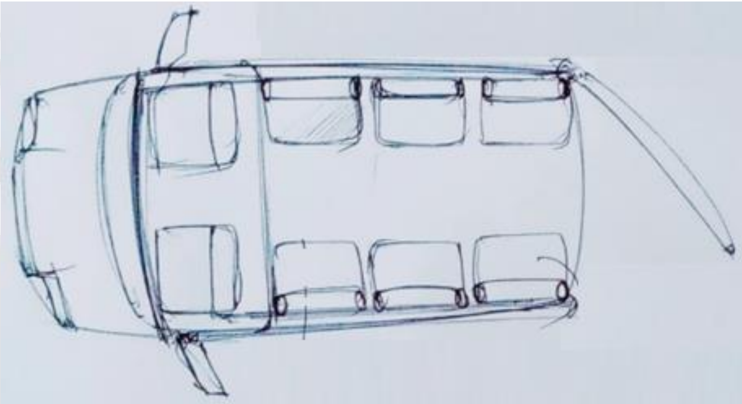


This conceptual mobility device with large wheels and high ground clearance is designed to travel on hilly terrain. The driver is placed at the center of the vehicle with storage spaces on either side. The rear seats are placed in a parallel setting which can be folded to make space for more storing goods/cattle. Additionally, the rear space has a soft-top which is supported by a frame structure.

Response 4:

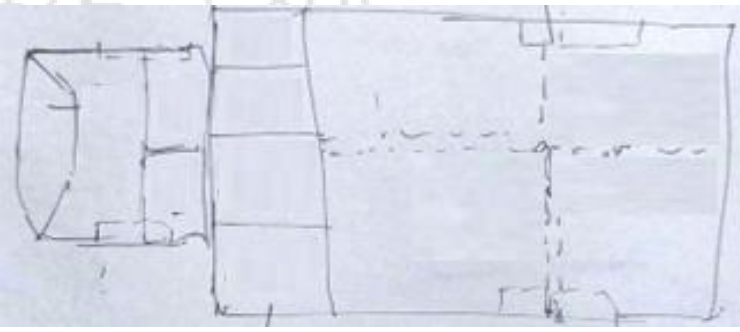
This is a four wheel drive with large threaded wheels made for uneven dirt roads. The vehicle has extra lights at the top to illuminate dark rural roads during the night. The front seat of the vehicle can hold upto 3 people while the back seaters can hold another 5. The back seats are foldable such that it can make room in the back for cattle. There is a ramp to help cattle climb on and off the vehicle. The seats are accessible both via the ramp and doors on the side. There is a long sliding window for the passengers in the back but also in case cattle are being transported. The ramp aides in carrying farm produce on and off the vehicle.

Response 5:



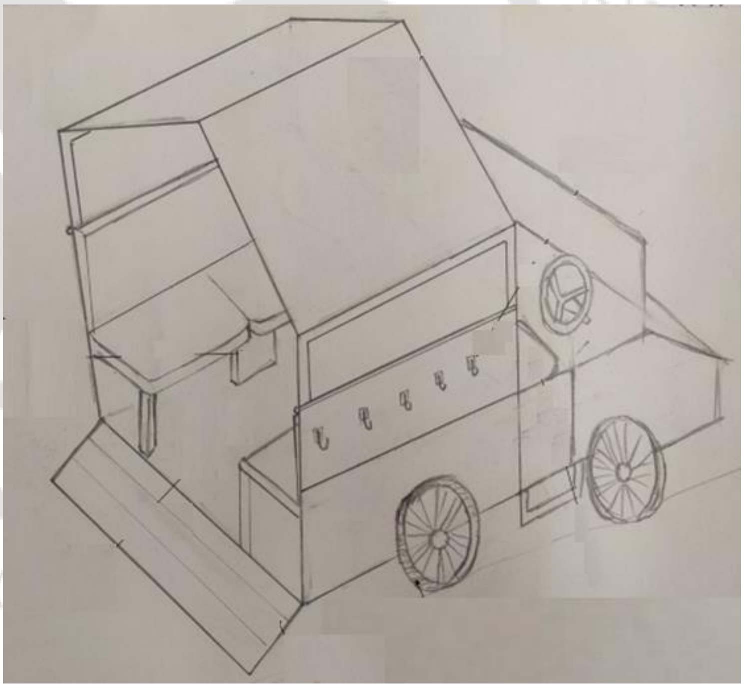
The concept above represents an interior layout of a four-wheeler. The layout shows 2 front facing seats and 6 side mounted seats for carrying 8 persons together. The rear side mounted seats are foldable as shown in the figure above. When moving with luggage or machinery, cattle or firm produces some or all of the seats can be folded to utilize the large rear portion of the 4-wheeler. During all seats folded setup the car can carry 3 large size cattles. There is a glass partition door between 2 front facing seats and the rear portion of the 4-wheeler.

Response 6:



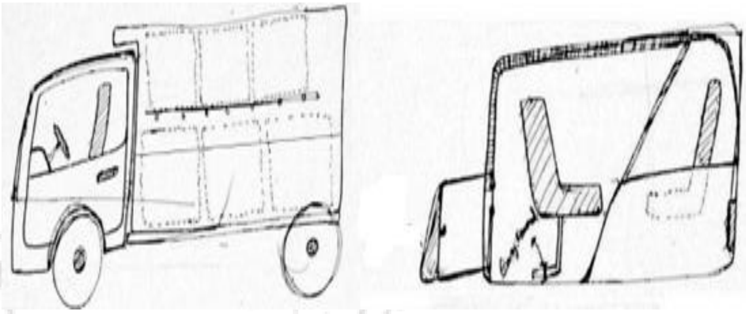
The truck can be designed in such a way that the same truck can be used for transporting people and transporting goods or cattle. a) The front portion of the truck is arranged with two seats, as usual. Whereas, the back space of the truck can be made convertible for 4 more seats, just by placing a foldable cushion seat attached to the wall of the truck. The foldable cushion should be like in the train berth, but to avoid the space occupancy, it can be arranged in a way that it can be folded downwards when not in use, just like a straight wall. When it has to be used as a seat for the passengers, it can be folded up and hooked with supporting latches. b) The same truck can be fitted with detachable steel rod frames vertically and horizontally to transport cattle and also allocate separate standing areas for the cattle to balance themselves while being transported. c) The same truck can be used for transporting goods when the detachable rods are removed and space is made to carry goods. At times goods can be carried with the rods in place also.

Response 7:



It is a compact vehicle that will carry 6 people including the driver. It will have thick tires similar to ATB (All-terrain bikes). But the body of the vehicle will be at a low height to maintain the centre of gravity during the transport. It will consist of foldable shade for uncertain weather conditions in hilly areas. Seats are foldable to make more space for goods transport and have a slide at the back side which will work as a ramp to load cattle. Hooks are also given for binding the load with the lorry.

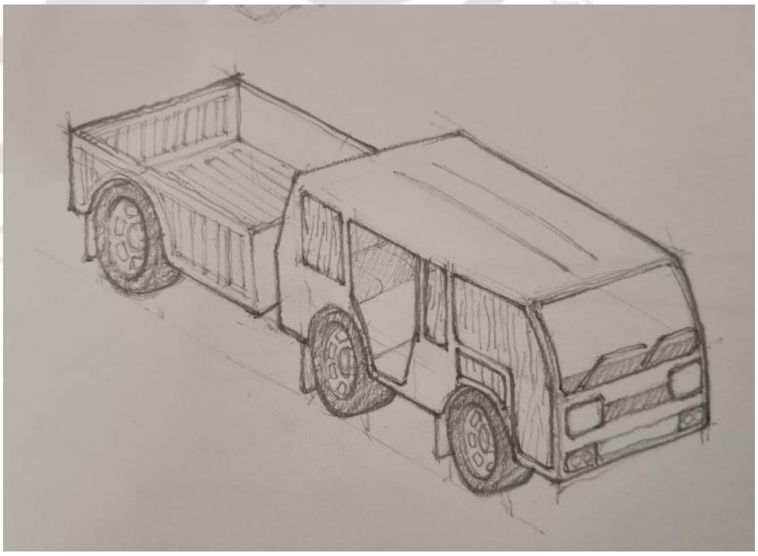
Response 8:



The seating in the back compartment should be made with a simpler fitting mechanism which makes the seating removed easily whenever it is needed. However, the lower body in both cases is the same one.

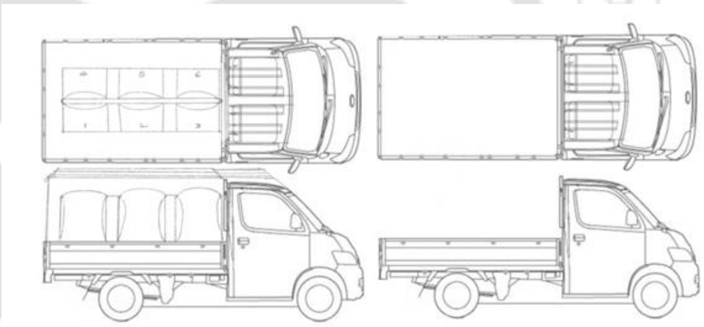
The upper shade is designed in a way that it fits the bottom easily. Also, the upper body is provided with small holes that allow the metal rods to pass through which can act as the base for the separating plane to keep much carriable luggage sorted, these separators can be used to separate one cattle from another. According to the requirement, the size of the upper shade can be selected.

Response 9:



The four wheeled multi-purpose vehicle is designed for seven passengers plus one driver. It has also been provided with easily attachable/detachable trolley for additional load carrying such as sufficiently large number of people, livestock like cattle, farm produces, machineries and equipment and other goods, as and when needed. Therefore, it offers flexible mobility solution in hilly rural area. The engine of the vehicle is mounted on the frame of the vehicle, that is on the bottom part of the vehicle, giving it necessary stability in hilly terrain. The passenger seating area is designed for comfortable seating of the people. And when not used for the people, it can be converted into material carrying space. The vehicle runs on the Liquefied Petroleum Gas (LPG) as well as normal fuel, making it economically viable option. The vehicle is aesthetically pleasing and also serves the functional requirements.

Response 10:

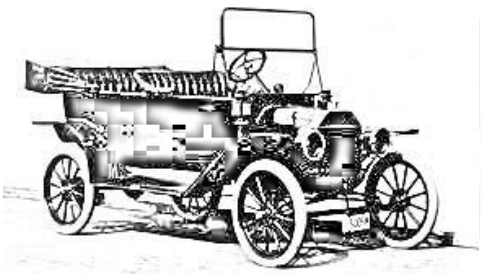


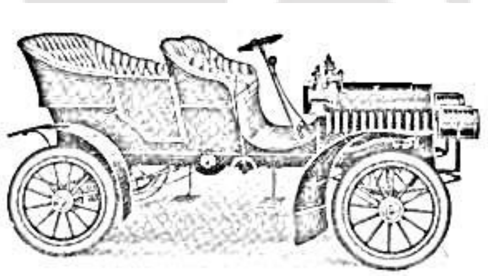
The four-wheeler is a multi role passenger and goods mini truck. The back compartment has a shed and 6 seats all inwards so that the passengers can sit together and have a good center of gravity. the extra leg room allows to keep luggage. In goods mode, the back seats can be removed to add luggage. The van has a roof to protect from rain and harsh windy climate of hilly region.

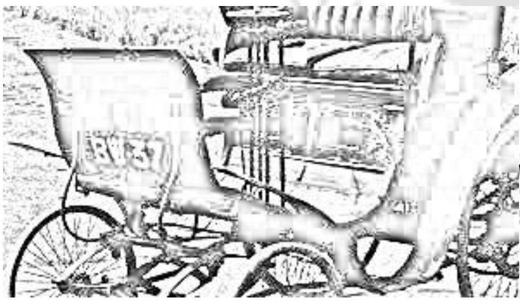
Response 11:

The four wheeler is designed in such a way that it is governed by the concept of convertible space for cattle and goods. The four wheeler has been provided with good quality small truck tyres at back as well as heavy doors and vehicle body is provided for the 4 wheeler. Therefore this makes the truck context sensitivity and safe and workable . Since it is an electric four wheeler and is automatic therefore it is very easy to maintain.

The sample images for labelled image-based responses collected from online resources.

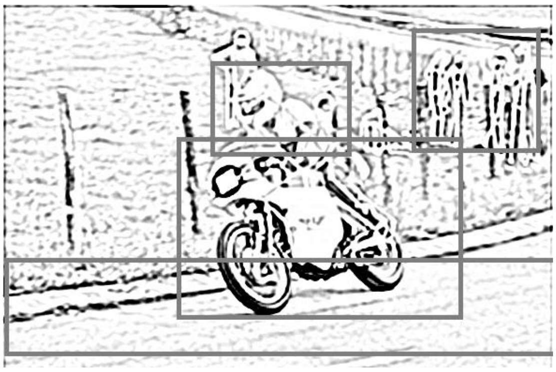




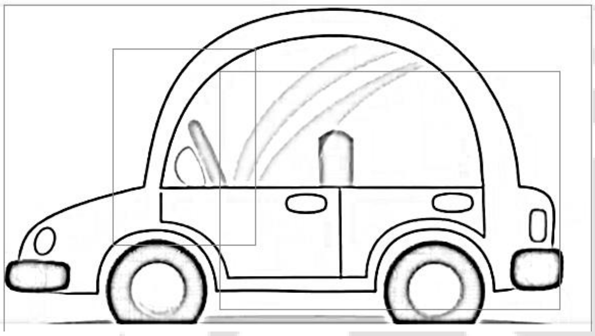


The example of images collected from online dataset for annotated image-based responses.

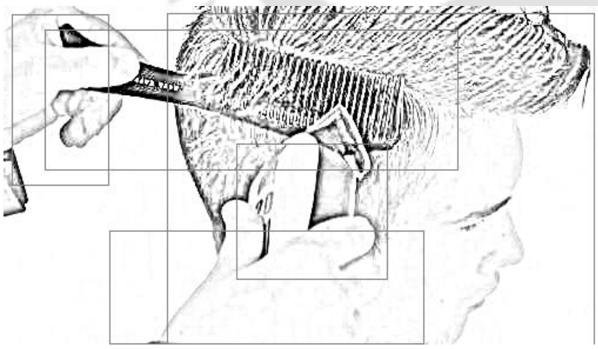
The scenario of kitchen environment is detailed. A man in a kitchen instructing a woman on what to do. A woman observing something on a kitchen stove and listening to the man. The man and woman appears to be making something in their kitchen. There are lots of kitchen stuffs around. The cupboard displays the necessary kitchen items.



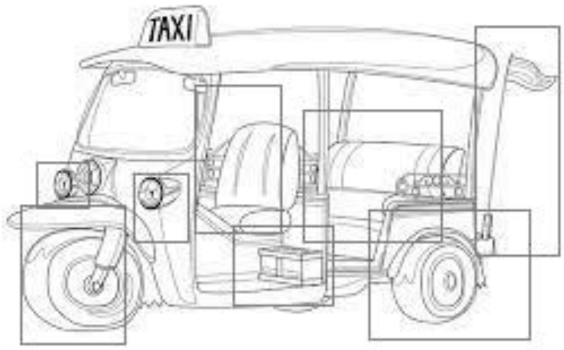
The bike illustrated can be driven by bikers having minimum expertise. The sitting height is flexible that supports in accommodating variable height of bikers. The bike is capable of running smoothly in different types of roads. It can bend up to forty five degree while riding that is significant in racing.



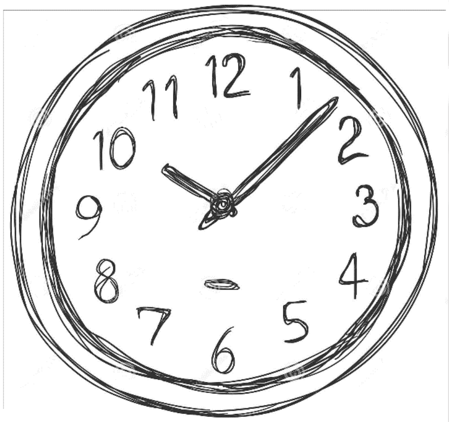
One could start experimenting with different styles and try submitting them online to see if anyone likes it or not. If it's good, one is ready to go. If it's not, keep trying. You're still young. You have a lot more time to think of more cool designs. One can mix red and blue to make purple. Add a little bit of white to get some beautiful shades of lavender. Secondary colors vary with the chemical composition of the paints. Acrylic paints, watercolor paints, and oil paints have different chemical compositions. So one get different secondary colors by mixing them. For example, one get different colors by mixing cerulean blue instead of ultramarine blue with red and yellow.



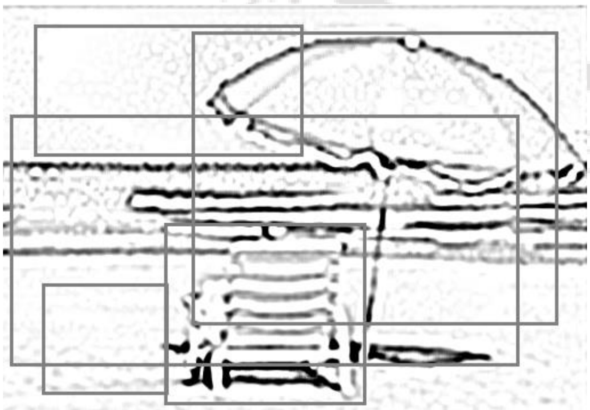
Haircut is a significant part of human identity. It reveals one’s personality, hobby, culture, etc. It also impacts one’s personality. Haircut are done by hairdressers who uses different instruments to cut the hair into a perfect shape. They use comb, trimmer, scissor, and other similar instruments for a haircut. Haircut has a significant impact on one’d emotion also. A hairdressers mistake often makes a man or woman unconfident until their hair grows to a desired shape and size. A bad haircut can make a person sob and vice versa. Therefore, a good haircut is essential for a confident human-being.



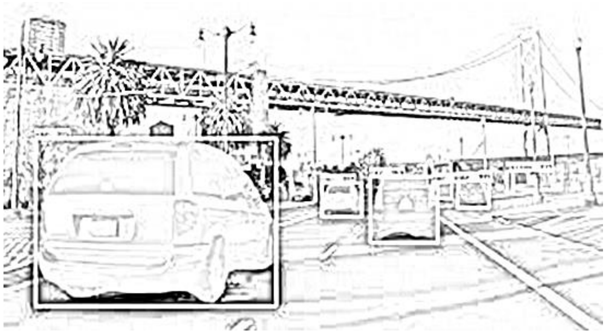
This three-wheeler taxi has innovative three-sixty degree rotatable wheels. The taxi can park in less parking places due to the facility of flexible wheel rotation. It makes it convenient to be used in small lanes and roads that are flooded with traffic every time. The taxi is decorated with flag that gives it a traditional look. The driver has a comfortable cushioned seat. The back-seat has a long foot rest area.



One day Tom bought a clock from a shop. He was very much fascinated by the aesthetics of the clock. After reaching home from the shop, he started staring at the clock and felt asleep. He started dreaming about time travel where he was able to go back in past as well as in future. He started to see his old grandparents who were no more alive, past relationships, etc. Then he started seeing future where cars were flying in the sky, and finally a humanoid asked him for a lemon juice and suddenly he woke up. After waking up, he first looked up at the clock and thought whether it was really a time travel. And, then he realized that it was a dream and smiled at himself.



A beach is always a place of fun and games. Resort beside beach often adds comfort essentials to make the beach a more favourable place for fun. Sitting arrangements are made for relaxation, umbrellas are provided for providing shadow. The sun lounger make the visitor look the sea, sky, sand for long time and cherish the beauty of the beach.

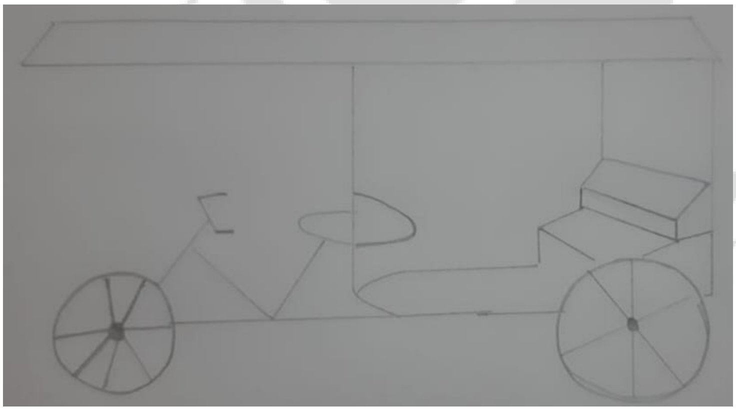


This is the daily scene of a country highway. Lots of cars passes by in a busy day. Everyone passes each other and each of them possess a unique story. The stories are never published or known to others. A writer once thought that he would collect the story of every driver in a highway and publish them with their consent. He started asking for lift from drivers and communicate with them. He used to ask about the incidents during driving. In this way, he collected the best stories among them and published and gifted the society with many incidents that drivers experience in a day-to-day basis.

An example question from CEED and their responses from manual crowdsourcing and MSCOCO responses.

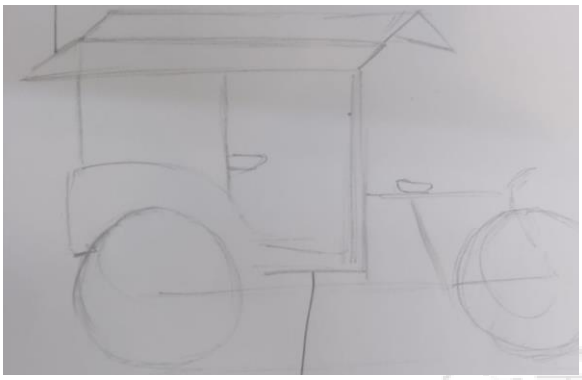
Q: An example of a question considered is as follows: A company is planning to introduce a three-wheeler taxi in the city of Banaras (Varanasi) for local commuters and pilgrims. Design a three wheeler taxi by identifying three distinct facility provided by it. The evaluation is based on the factor of novelty. A typical response must contain a unique design concept accompanied with a passage describing the concept. Response from manual crowdsourcing data is as follows:

Response 1:



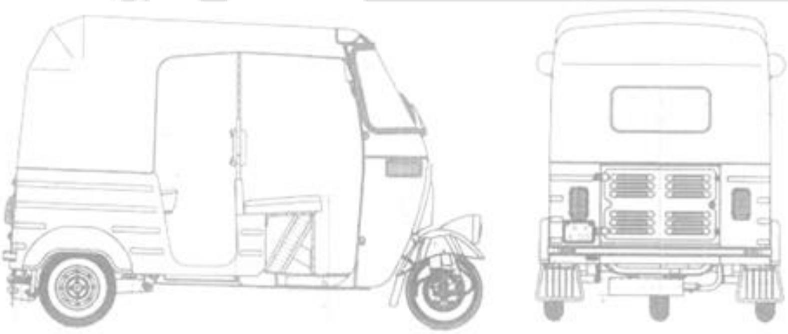
This three-wheel taxi works on two factors. Keeping in mind the hot region of Banaras this taxi will protect the customer and the driver from the heat and sunlight. It is more spacious to keep luggage and can be handy in caring for small goods.

Response 2:



Designed for the weather of Varanasi the extra flaps will provide protection from Rain + Summer. The wheels are big hence the space is above ground which will make it easier for the tourist to travel in the streets with water-logged roads. And the structure of rickshaw is supposed to be thin for making it convenient for it to pass through the thin streets of Varanasi. Inspired from the old transport it is build in a dome like structure which has a full covering. And a gate opening from above.

Response 3:

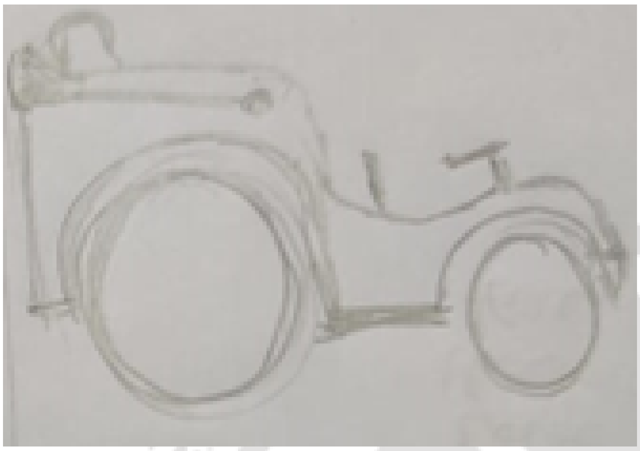


Three wheel taxi offering protection comfort safety and reliability in a city of small roads and quick turns. The vehicle has front and back suspension for the roads of Varanasi. There is a protection on the roof for rain. The engine and a passenger or rear mounted so that the centre of gravity is below and the driving is easier.

Response 4:

This three-wheeled vehicle is a fixed structure with cushioned seats in the back for passenger comfort and a flat shade on top to protect the passengers from the harsh sun of the Indian summers. The front of the vehicle is a standard cycle mechanism supported with a gear system for easy mobility. The vehicle has lower clearance to make it easier for passengers to get on and off the vehicle.

Response 5:



This concept is inspired by a buggy vehicles. It has a larger wheel base in the back and a smaller wheel base in the front. The vehicle uses a motorised engine to decrease the load of the driver. It uses a door mechanism through which passengers can get on and off the vehicle. The seats are elevated with the larger wheel base which allows passengers a clear view of the sights around them. At the back is an extendible soft top which can be put through an automated mechanism to protect passengers from rain and sunshine.

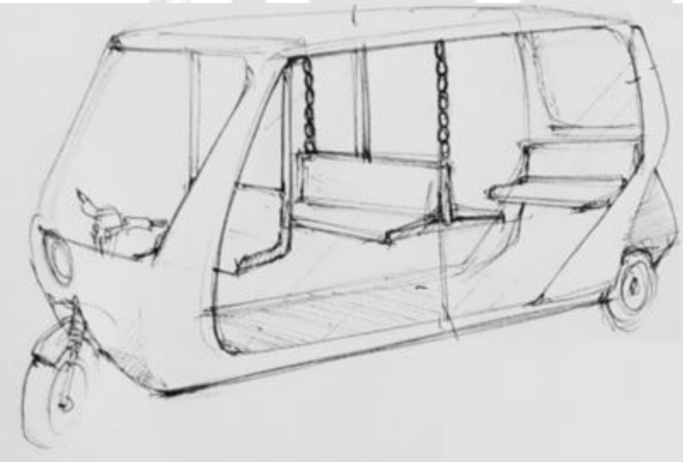
Response 6:

A golf cart inspired three wheel vehicle. It has a complete covered top which has solar panels mounted on it and runs similar to an e-bike. It also runs on CNG in case of rainy or cloudy days. The seats are cushioned and designed to be comfortable for both the passenger and the driver. It is an ideal vehicle for the Indian summers when the sun is out. The front has a windshield similar to a golf cart. This vehicle closely resembles the autos found in India but runs more on like a 3wheeled e bike.

Response 7:

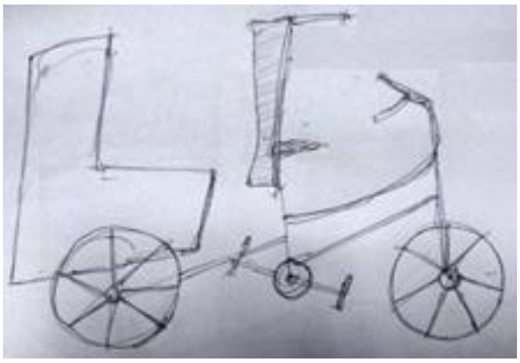
The last concept is similar to a traditional three-wheeled vehicle, but instead of a bicycle front it is fitted with a motorcycle engine. The back end is designed to match the old-school aesthetics of hand-pulled or bicycle based vehicles found in Calcutta. The wheels are thicker to accommodate the engine power rather than the usual bicycle tired. Passengers climb in and out of the vehicle like the old fashioned style from the front. There is a back cover to protect passengers from the sun and rain.

Response 8:



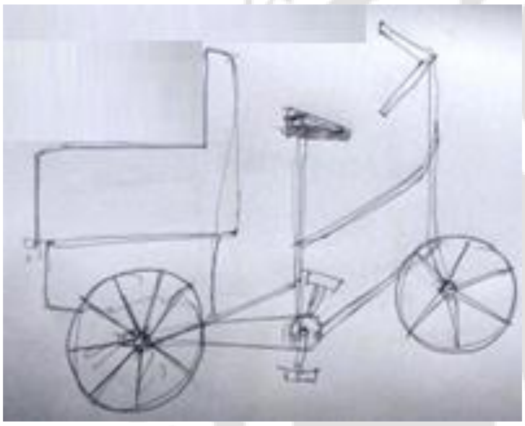
The distinct facility of the above 3-wheeler concept are follows: a) It provide a large seating space i.e. 12 seats in total. b) The seats are top mounted, so entire bottom space is available or luggage and misc. purpose. c) To provide better user experience the roof and the side doors were kept transparent and are large in size. d) The all enclosed concept protect passengers from wind and rain.

Response 9:



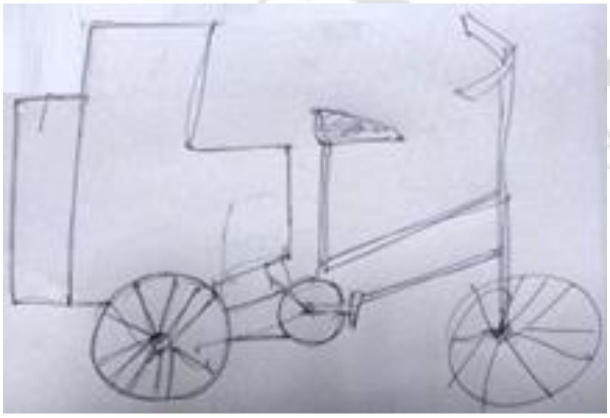
The taxi can be fitted with a lightweight frame with attached transparent sheet or polythene sheet in view of Covid-19 situation to avoid direct contact with the driver.

Response 10:



The taxi's seat can be arranged away from the driver in the opposite direction to provide broader view to the tourists. This way the tourists can enjoy site seeing in a better way.

Response 11:

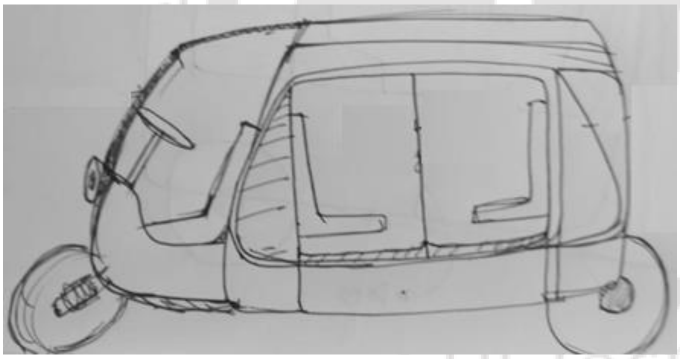


In the traditional taxis, the luggage will be kept in the foot-rest area, which causes discomfort for the passengers. Or at times, the passengers place the luggage in their lap. In order to avoid this, a small luggage carriage area can be provided behind the taxi so that the passengers can place their luggage.

Response 12:

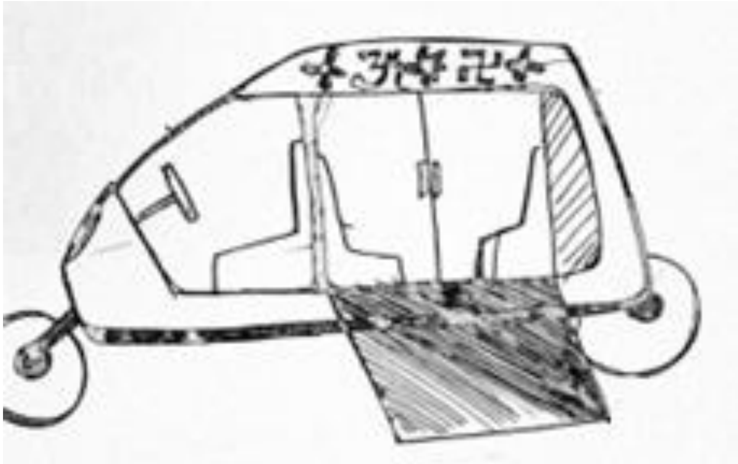
As Varanasi is a highly crowded city and the roads are always full of vehicles. So, if the vehicles will take up less space on the roads and in parking areas. Also, to carry people comfortably and follow some kind of modular shape that can stack two or more vehicles. So the shape of the taxi is designed by keeping in mind to stack the taxis in lesser space. The followings are the key features of the design: a) Flexible front seat without partition. b) Modular shape of the vehicle. c) Open body to view cityscape.

Response 13:



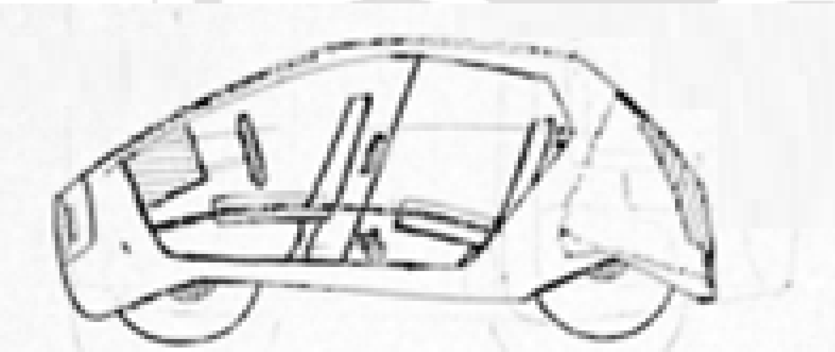
As it is a cab that needs to satisfy certain needs of tourists like being able to see everything, also to look novel the doors and windows should be a single entity with transparent material so that it looks authentic. At the back, there should be enough luggage carrying space so tourists can carry their travel bags. Front glass of the driver should be very similar to the material that is used for the doors provided for tourists so can be in synergy. Also, the colors should be blue and saffron.

Response 14:



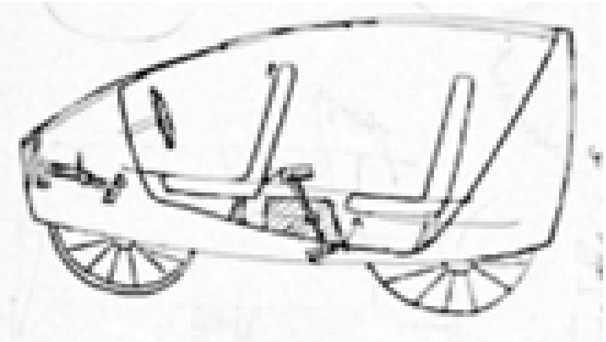
As it is a taxi that needs to satisfy certain needs of tourists like being able to see everything, and also looks the doors and windows should be of single material using the law of similarity. At the back, there should be enough luggage carrying space so tourists can carry their emergency travel bags. The front glass of the driver should be very similar to the material that is used for the doors provided for tourists so can be in synergy. Also, the colors should be blue and saffron. The diesel engine provided in this concept can support normal scenarios, also support is provided for all the commuters to get in easily.

Response 15:



The top of the car is provided with glass which can become transparent and opaque according to the need it does not provide a broad vision on the side, but also in the top.

Response 16:

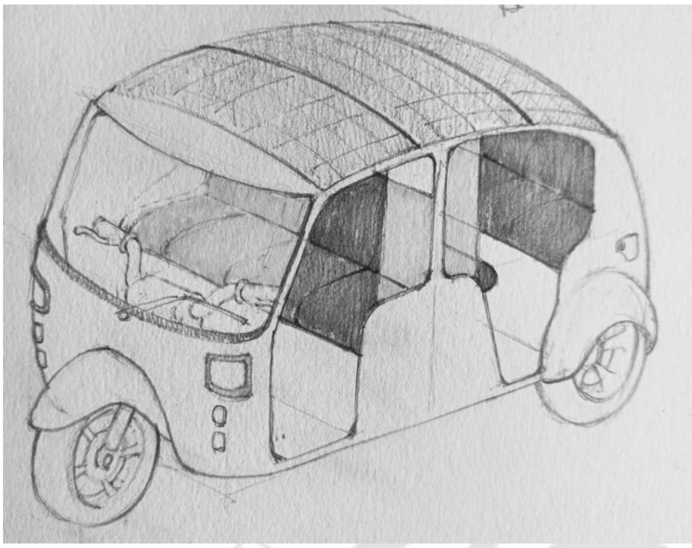


This concept is for self-driving provided with both manual (pedaling) and electric power. Pedaling provides engagement and fun to certain tourists who wish to see Varanasi according to their liking.

Response 17:

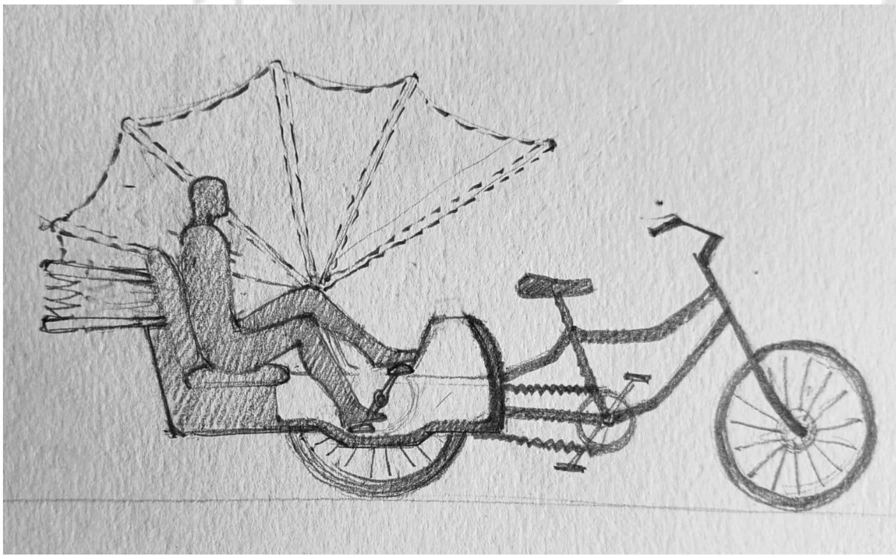
Here the doors to the rikshaw is provided with doors that are translucent, blocking the sun if needed. The door when opened looks like butterfly wings making the rikshaw more interesting.

Response 18:



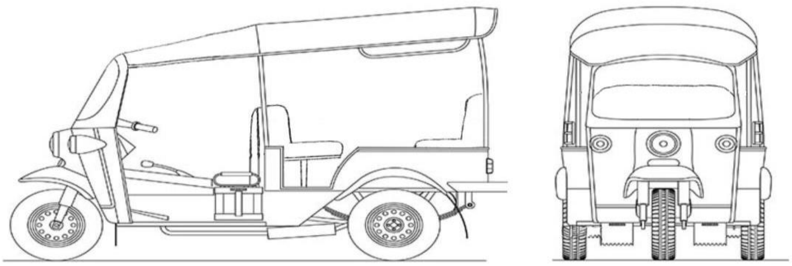
The three – wheeled taxi is designed in such a way that it can accommodate six to seven passengers. It is particularly conceived by keeping needs of Varanasi city transportation in mind, compact as well as spacious enough to navigate on the city roads. The design is especially innovative in a way that it has a solar panel roof which helps converting the solar energy into electric energy stored in batteries of the taxi. It also has an electric charging point, making it dual – solar and electric powered taxi. Thereby, also making it sustainable taxi and cost effective solution in long run. As it is a battery operated taxi, it contains less moving engine parts and less complicated power transmission mechanism, hence it has less maintenance cost compared with other conventional options. The three wheeler has sufficient leg room and comfortable cushion seating arrangement, for the passengers to enjoy the ride.

Response 19:



The three – wheeled cycle rickshaw is a creative example of sustainable mobility solution designed for the needs of Varanasi city commuters. The novelty of the design lies in the pedalling mechanism, which also allow passengers to pedal the rickshaw along with the rider. It is designed by keeping health conscious travellers in mind who would be happy to pedal. Those who don't want to pedal can also ride, in this case the rider will only pedal the rickshaw. Additionally, battery assistance is also given in the rickshaw, which can store any extra energy generated during pedalling by passengers/rider, and aid the riding when needed. It doesn't produce any emission, making it eco-friendly solution. It has elegant umbrella cover and comfortable seating arrangement for three passengers for short or medium range travel in the intricate part of Varanasi city.

Response 20:



This is an electric tuk-tuk three-wheeler for the narrow and steep turning streets of Banaras. The vehicle can support 1+4 people. The open vehicle is good for the city with humid and hot temperature. The shed is to protect from rain and sun.

Q: An example of a CEED question, Draw a perspective view of a kitchen interior with a stove, kitchen utensils (such as pressure cooker, cooking pans, sauce pans etc.), dining utensils (such as ceramic plates, cups, glasses etc.), a wash basin, storage racks with stored cooking ingredients (such as spices in small plastic bottles), fresh cut vegetables kept beside the stove and at least two kitchen gadgets. Write a short paragraph describing your sketch. The sketches from MSCOCO dataset are as follows:

Response 1:

The scenario of kitchen environment is detailed. A man in a kitchen instructing a woman on what to do. A woman observing something on a kitchen stove and listening to the man. The man and woman appears to be making something in their kitchen. There are lots of kitchen stuffs around. The cupboard displays the necessary kitchen items.

Response 2:

A woman is in a kitchen with her cat. She has a vegetable bag that is filled with vegetables. The kitchen is equipped with a refrigerator. There are several containers beside the refrigerator. The space is filled with kitchen cabinet. There are many utensils like plates, bottles, bowls filled vegetables, etc. over the slab. It also has a wash basin where many utensils are kept.

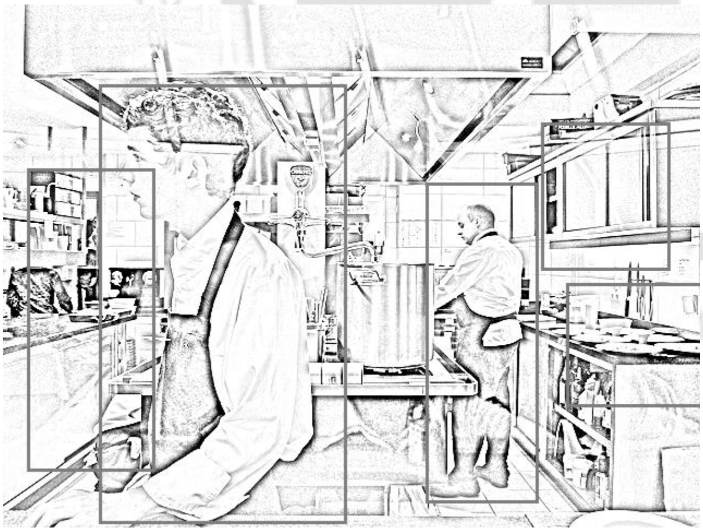
Response 3:

This kitchen is full of utensils. A chef is using the gas stove. Beside the stove, it has cooking pans, sauce pans, and various types of spoons hanging from the walls. The pans are of different shapes and sizes. The man has cooked a curry and placed it on the table. There are spoons, fresh cut vegetables placed on the table. The storage racks is holding all the essential items of the kitchen.

Response 4:

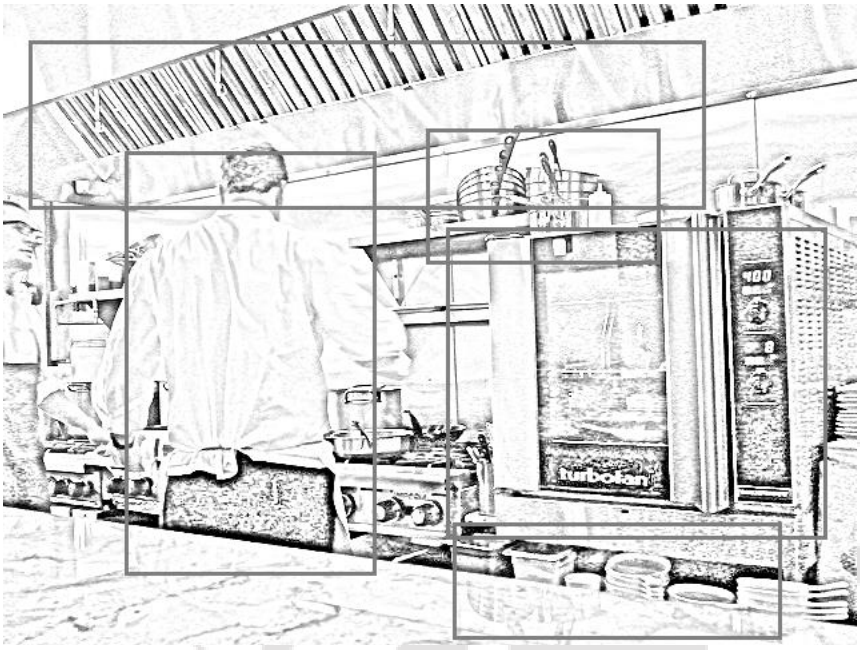
The kitchen illustrated is well equipped with all facilities of modern cooking arrangements. The kitchen interior is having storage racks, chimney, dish washer, baking oven, microwave oven. The table contains fresh cut vegetables, containers, pizza placed on it. There are several containers that supports in storing food items, grocery, etc.

Response 5:



A master chef and a cook is preparing dishes in a kitchen. They are preparing dishes individually. There are chimneys in the kitchen that makes it well-ventilated. There are lots of plates, bowls, crockeries on the table. The cooked dishes are placed on the table.

Response 6:



Refrigerator is an important part of any kitchen that keeps the food and raw vegetables fresh. The refrigerators in this kitchen stores all the essentials for cooking. There is a gas stove in the kitchen, where a man is cooking. Beside the stove, there are fresh cut vegetables that are used for cooking. The are containers, various types of spoons, spatula, sauce pans, etc. that are kept on the kitchen rack.

Response 7:

The displayed kitchen is a store-house of all types of crookeries, utensils, and modern cookware. There dishwashers, gas stove, and baking oven used for cleaning and cooking. There are scissors, knifes placed on the table for cutting vegetables. There are small plastic bottles used for filling water, oil, sauces. Apart from that, there are jugs, plates, mixer grinders, mugs, etc. in the kitchen. The wall rack is filled with lot of plates. All these items supports in healthy cooking stuffs.

Response 8:

This is a narrow view of a kitchen. There is a door for entry in the kitchen. There are racks and slabs on both left and right side of the kitchen. One side of the kitchen is equipped with wash basin. Beside that, there are plates kept on the slab. Vegetables are kept beside the wall. There are closed racks at the top where the stored items are not visible. The kitchen slab has attached drawers where utensils can be stored for immediate usage during cooking. On the other side of the kitchen, there is a gas stove and baking oven. Presently, a sauce pan is placed on the gas stove indicating that something is being cooked. There is a double-door refrigerator placed near the door. There are many containers that are placed on the refrigerator.

Response 9:

The kitchen is combined with dining area. The dining area has a dining table, chairs, and sofa. The table has food served on it. The kitchen area has microwave where something is being cooked. This area is decorated with flower vase. It has laptop placed on a table. There is a dustbin, chair with cushion, and many other essential items in the table. A kitchen rack is transparent that is filled with necessary food stuffs. Other racks are wooden and is opaque in nature. There are plates, containers, and vegetables in the kitchen slab.

Response 10:

The woman has a plate and a spoon in her hand. There is a food item in the plate and she is ready to eat with the help of the spoon. Probably, she has just cooked the food in the gas oven and the provided cooking pan. There is a bottle and other foods in the slab. The racks in the wall are wooden, where food items are stored.

Response 11:

The kitchen has a beautiful view of a garden from its window. These windows has beautiful curtains. There in induction cook top, plastic bottles, placed on the slab. There is a big wash basin. There is a dish washer and many drawers at the bottom of the slab.

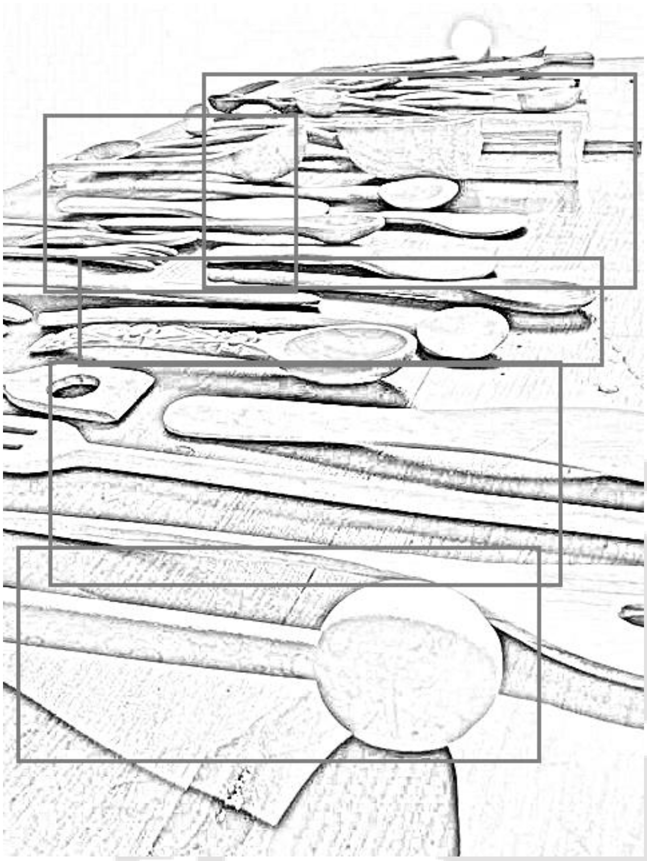
Response 12:

This is a very small kitchen. It has a compact table-based attached ovens. There are multiple ovens, which includes a four-burner gas stove, a baking oven at the bottom. Also, there is a wall mounted microwave oven. There are also several wall mounted racks and a refrigerator at the corner of the kitchen.

Response 13:

There is a door at the entrance of the kitchen. There is a round table at the center of the kitchen. The table has fruits on it. There is big refrigerator and multiple wall-mounted kitchen cabinets The gas stove is surrounded by many small plastic bottles that are used while cooking. It also has a baking oven at the bottom of the unit.

Response 14:



This is a display of various types of spoons in the kitchen that are used while cooking. Some of them are also used for eating. There is a oil-poring spoon, serving spoon, dinner fork, desert fork, dinner knife, desert knife, sugar spoon, tea spoon, fish knife, butter knife, and soda spoon, ice-cream spoon.

Response 15:

The are many people in the kitchen who are discussing and cooking food. There are several plastic bottles, jugs, serving bowls, spoons on the table. The is focussing light on the roof that supports in better vision while cooking. There is a wash basin and multiple wall-mounted kitchen cabinets.

Response 16:

A lady is happily cooking inside a kitchen. She has a big cooking pan in her hand, which she would place on the gas stove. There are many wall mounted kitchen cabinets to store utensils and food items. There are many small plastic bottles, fresh cut vegetables on the kitchen slab. There are lights hanging from the roof that provides better visibility of foods in the pan while cooking. There is a wash basin where utensils might be washed after cooking.

Response 17:

A woman has prepared a big cake in her kitchen. Now she is trying to cut it with the help of a knife. In the back side, there are several contains which probably she has used while baking. There is also a big wash basin where half of it is visible and rest is hidden behind the wall.

Appendix F. Sample of normalized scores of human experts and models for assessing creative responses.

Descriptive creative response (DCR)	Expert score (DCR)	Predicted score (DCR)	Labelled image-based creative response (DIBR)	Expert score (DIBR)	Predicted score (DIBR)	Annotated image-based creative response (AIBR)	Expert score (AIBR)	Predicted score (AIBR)
Response ID = Q1-1	0.50	0.51	Response ID = Q2-7	0.39	0.40	Response ID = Q1-5	0.67	0.66
Response ID = Q1-2	0.76	0.74	Response ID = Q2-8	0.38	0.38	Response ID = Q1-9	0.49	0.48
Response ID = Q1-3	0.80	0.80	Response ID = Q2-9	0.44	0.45	Response ID = Q1-2	0.46	0.45
Response ID = Q1-4	0.90	0.91	Response ID = Q2-4	0.30	0.31	Response ID = Q1-10	0.46	0.46
Response ID = Q1-5	0.6	0.6	Response ID = Q2-10	0.41	0.41	Response ID = Q1-7	0.47	0.45
Response ID = Q1-6	0.70	0.71	Response ID = Q2-2	0.53	0.50	Response ID = Q1-1	0.48	0.47
Response ID = Q1-7	0.61	0.60	Response ID = Q2-3	0.40	0.38	Response ID = Q1-4	0.55	0.54
Response ID = Q1-8	0.73	0.72	Response ID = Q2-6	0.42	0.41	Response ID = Q1-3	0.56	0.55
Response ID = Q1-9	0.89	0.89	Response ID = Q2-5	0.33	0.34	Response ID = Q1-8	0.58	0.59

Response ID = Q1-10	0.67	0.68	Response ID = Q2-1	0.36	0.36	Response ID = Q1-6	0.56	0.56
Response ID = Q2-1	0.87	0.86	Response ID = Q3-8	0.50	0.51	Response ID = Q2-3	0.50	0.49
Response ID = Q2-2	0.72	0.71	Response ID = Q3-10	0.57	0.57	Response ID = Q2-10	0.50	0.51
Response ID = Q2-3	0.92	0.90	Response ID = Q3-2	0.57	0.57	Response ID = Q2-5	0.49	0.48
Response ID = Q2-4	0.55	0.52	Response ID = Q3-7	0.57	0.59	Response ID = Q2-2	0.43	0.44
Response ID = Q2-5	0.93	0.91	Response ID = Q3-9	0.56	0.57	Response ID = Q2-6	0.47	0.48
Response ID = Q2-6	0.62	0.62	Response ID = Q3-6	0.52	0.53	Response ID = Q2-9	0.65	0.64
Response ID = Q2-7	0.94	0.95	Response ID = Q3-4	0.62	0.60	Response ID = Q2-1	0.62	0.61
Response ID = Q2-8	0.67	0.66	Response ID = Q3-5	0.50	0.51	Response ID = Q2-4	0.67	0.68
Response ID = Q2-9	0.92	0.92	Response ID = Q3-1	0.67	0.46	Response ID = Q2-8	0.68	0.66
Response ID = Q2-10	0.68	0.68	Response ID = Q3-3	0.57	0.58	Response ID = Q2-7	0.64	0.64

Appendix G. List of publications

Journal Papers

1. **Chaudhuri, N. B**., Dhar, D., & Yammiyavar, P. G. (2020). A computational model for subjective evaluation of novelty in descriptive aptitude. *International Journal of Technology and Design Education*, 1-38. https://doi.org/10.1007/s10798-020-09638-2 [SCIE/Scopus indexed, IF: 2.177]

2. **Chaudhuri, N. B.**, Dhar, D., & Yammiyavar, P. G. (2021). Automating assessment of Design exams: A case study of novelty evaluation. *Expert Systems With Applications*, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.116108
[SCIE/Scopus indexed, IF: 6.954]

3. **Chaudhuri, N. B.**, Dhar, D., & Yammiyavar, P. G. (2021). A human-centred deep learning approach facilitating design pedagogues to frame creative questions. Neural Computing and Applications, 1-28. https://doi.org/10.1007/s00521-021-06511-8
[SCIE/Scopus indexed, IF: 5.606]

4. **Chaudhuri, N. B.**, & Dhar, D.(2021). Designing deep-network based novelty assessment model in Design education (In Review)

5. **Chaudhuri, N. B.**, & Dhar, D.(2021). Digitizing creativity evaluation in Design education: A systematic literature review (In Review)

Conference

1. **Chaudhuri, N. B.**, Dhar, D., & Yammiyavar, P. G. (2021). Do Design Entrance Exams in India Really Test Creative Aptitude? An Analytical Study of Design Tests Conforming Creativity Benchmarks. In Design for Tomorrow—Volume 2 (pp. 371-383). Springer, Singapore.
[Scopus indexed, ***Distinguished paper award***]

Patent

1. **Chaudhuri, N. B.**, Dhar, D., & Yammiyavar, P. G. (2021). A System for Evaluating Novelty of Creative Write-Up. Application No. 202131006753 [Patent filed in India]

Appendix H. Permission of journals for reusability of content

1. Permission from the journal titled "International Journal of Technology and Design Education".

Dear Nandita,

Thank you for your email.

Your first query is not for this department – please redirect your question to the journal.

As for your second question:

Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.
Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]'. https://www.springer.com/gp/rights-permissions/obtaining-permissions/882

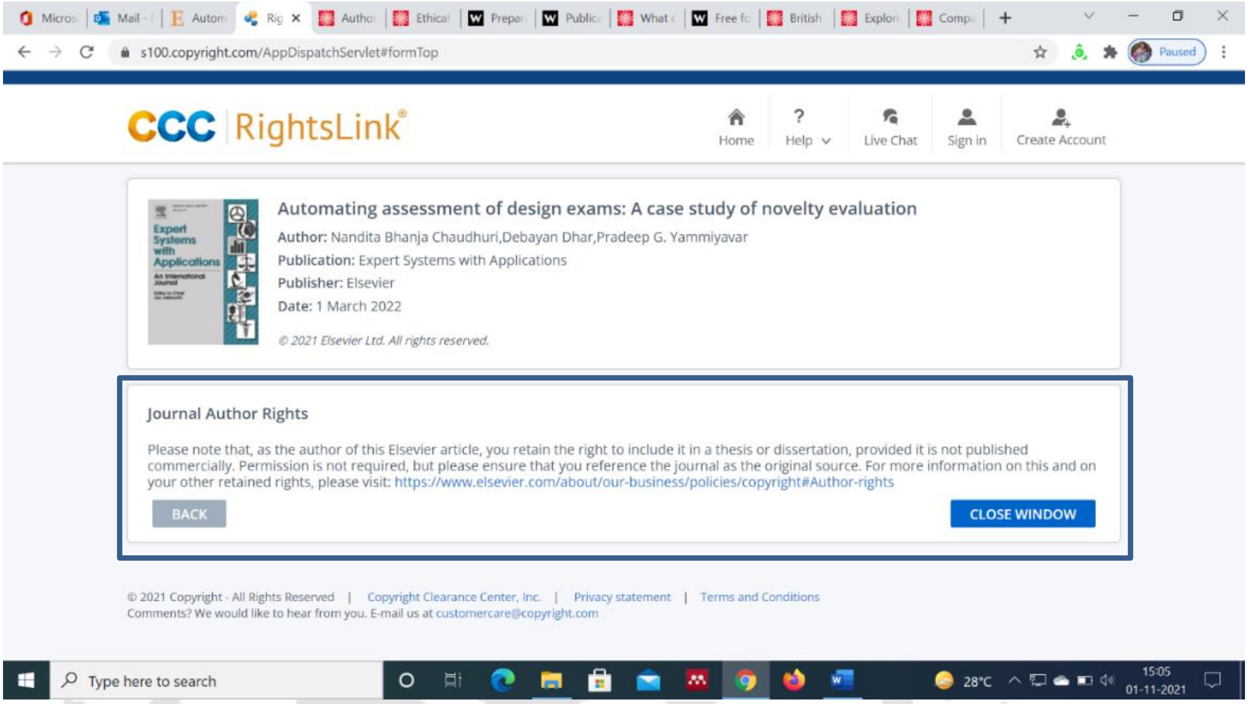
If you require formal permission, please locate your article on link.springer.com or nature.com, scroll to the end of the article page, click on the reprints and permissions tab, and input the details of your request via our CCC RightsLink service. Be sure to select 'reuse in a dissertation/thesis' and to click 'yes' to 'I am the author'.

Good luck completing your degree!

Best wishes,
Bod

Bod Adegboyega
Permissions Assistant
Springer Nature
4 Crinan Street, London N1 9XW, UK
T +44 (0) 442034263235
Bod.adegboyega.1@springernature.com

2. Permission from the journal titled "Expert Systems with Applications".



3. Permission from the journal titled “Neural Computing and Applications”.

Dear Nandita,

Thank you for your email.

Authors have the right to reuse their article’s Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.
Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]’. https://www.springer.com/gp/rights-permissions/obtaining-permissions/882

If you require formal permission, please locate your article on link.springer.com or nature.com, scroll to the end of the article page, click on the reprints and permissions tab, and input the details of your request via our CCC RightsLink service. Be sure to select ‘reuse in a dissertation/thesis’ and to click ‘yes’ to ‘I am the author’.

Good luck completing your degree!

Best wishes,
Bod

Bod Adegboyega
Permissions Assistant
Springer Nature
4 Crinan Street, London N1 9XW, UK
T +44 (0) 442034263235
Bod.adegboyega.1@springernature.com