Creativity Assessment Automation

Team

Akshay Ashok Kumar

Vindya Henagama Liyanage

Wai Cho Kwan

Abhinav Singh

Table of Contents

Table of Figures

Table of Tables

Introduction

Psychologists usually define creativity as the capacity to produce ideas that are both original and adaptive. In other words, the ideas must be both new and workable. Thus, creativity enables a person to adjust to novel circumstances and to solve problems that unexpectedly arise. Obviously, such a capacity is often very valuable in everyday life. Yet creativity can also result in major contributions to human civilization. Examples include Michelangelo's Sistine Chapel, Beethoven's Fifth Symphony, Tolstoy's War and Peace, and Darwin's Origin of Species. One might conclude that creativity plays an important role in the evolution of humankind and thus more efforts need to be made to study creativity and understand it. Most research and theory-based definitions of "creativity" boil down to two components. First, creativity must represent something different, new, or innovative (Amabile, 1983; Barron 1955). Second, the component encompasses appropriateness to the task at hand.

One of the principal challenges in creativity research is assessment. From the past 40 years researchers have used objectively scored tests of creativity such as the Torrance Tests of Creativity (TTCT). The series of tests which are part of TTCT are slow and expensive to score but do offer reliability in results. Due to the drawbacks of TTCT, researchers tend to use simpler and faster self-reporting creativity measures like growth mindset and self-efficacy. Currently, multiple experts (or raters) are hired to assess the creative product by a candidate. The experts compare products to each other instead of an absolute ideal. A great deal of past research has shown that experts agree at a strikingly high rate (Amabile, 1983, 1996; Baer, 1993, 1998; Hennessey and Amabile, 1999). These experts do have some differences as shown in cross-cultural studies where it was shown that Chinese experts generally tend to give high creativity ratings than their American peers (Niu and Sternberg, 2001). It was also noted that experts from China valued novelty as did their American and Japanese peers but placed less importance to appropriateness than their peers (Paletz and Peng, 2008). This signifies that even though there is some consistency among the experts, there still needs to be an assessment mechanism that is free from these biases. This project aims to automate the scoring process of verbal creativity tests. By making use of the various advancements in the field of Machine Learning, the project aims to leverage the large datasets of creativity test results available with C3L to explore and validate a method for automating the scoring of these tests. This might enable organizations to assess the creativity of candidates in real-time or be able to take decisions regarding the creativity of individuals with more confidence. To date, however, no research appears to have emerged that uses machine learning methods drawn from the field of artificial intelligence to assess verbal tests of creativity.

Literature Review

Automating the measurement of creativity is a relatively new field. Nevertheless, measuring creativity is a field where vast research has been done. Most research is different from one another. Measuring creativity of movies and books, measuring creativity depression, measuring creativity of people are a few of them. Plucker, Makel, and Qian (2019) set out many of the important issues, ranging across all four of Rhodes's Ps: person, process, press and product (Rhodes, 1961). Two major challenges faced by creativity researchers is the subjectivity of the creativity test scores as multiple raters need to agree on the creativity of the artefact and the effort required to administer and score these tests. Together, these limitations tend to drive researchers towards methods of assessment that weaken both reliability and validity (Reiter-Palmon, Forthmann, & Barbot, 2019). This means there is no right or wrong way to say someone is absolute highly creative or has no creativity at all. And so, this calls for a methodological way to measure creativity of a person and upon further research we came across the way of measuring the verbal creativity by giving tests such as images or an object or through sensors and predicting using machine learning models.

In measuring verbal creativity there are mainly four criteria that are considered. They are effectiveness, novelty, influence, and unexpectedness [3]. Value is an important criterion in measuring creativity. It is the measurement of how an artifact (the material in which creativity is measured) is valued by the domain experts and how the artifact is expected by society. It resembles the correctness of a given artifact. Novelty is how different the artifact is compared to the known artifact in its' class. Influence is the measure of inspiration on an artifact on another artifact. Unexpectedness is the surprise effect of an artifact. The research performed in verbal creativity is done by Chieng, Liu and Wu, in the field of creativity education. Research performed by Stevens and Zebelina classified the brain states of participants into low and high creativity groups. They tested the participants using the Alternative Uses Test (AUT) and obtained their EEG (electroencephalography) data. They were able to extract relevant features from the EEG data and developed on a model based on these features that was able to successfully classify the creativity of participants. This model using SVM and QDA (Quadratic Discriminant Analysis) for classification was able to achieve an accuracy of 81.3%. With much research claiming the way of measuring creativity through tests, one of the research papers measured creativity based on responses from using brain images [16]. This led to exploring the use of machine learning models to measure creativity based on the responses from an image. For a model to predict the creativity level of a response, a lot of meaningful features must be identified that the model could potentially use to understand the image and the written response. Several steps are critical to achieve this, and these steps are part of the methodology which have been listed in the following sub-sections.

IMAGE PROCESSING

Since the verbal test that we are analysing is based on an image, a relationship between the image and the responses needs to be established. In research from Chen, S.-H., & Chen, Y.-H, it distinguishes the effectiveness of the Google Cloud Vision API compared to different efficient algorithms such as Content Based Image Retrieval method which uses classification algorithms to classify the images with joint probability distribution function and score the confidence of the classified image labels. This paper introduces the use of Google Cloud vision API to classify the images and label them with confidence score, with the use of Deep Learning and convolutional neural networks. Since there are alternate words to an object, there is a semantic gap between the labels generated by the Google Cloud Vision API and the image dataset. [10]

To fill this semantic gap, the Google Cloud Vision API is made to work with WordNet which provides synonyms for each of the labels, and it is proved to be an upgrade on the existing method with increase in accuracy of the model and the ability to train the model with the alternate words of each label from the image. In research from Shrivastava, D, CG, SA, Laha, A & Sankaranarayanan and Wei, Y "Max", Hong, J & Tellis, authors show that there should be a correlation between the creativity artifacts (in text format) and creativity levels. [3,4]

In other words, to predict creativity level, numerous features must be extracted to understand how the written text is related to the image. If the written text is only describing the image, this written text is highly correlated to the image and the novelty of the written text is low. On the other hand, if the written text is not highly correlated to the image but is still effective in describing the image, it would suggest a higher creativity level (keeping in mind that other creativity factors are satisfied like novelty, influence, and unexpectedness)

TEXT PROCESSING

The researchers developed a model for creativity assessment using Knowledge sharing Information tokenization system, a term-frequency-inverse document frequency method and support vector machines. They studied the discussion records of students during creative activities, creativity scores given by teachers and experts and developed the model that classified into high and low creativity groups. This experimental model was able to achieve an accuracy of 93%. The most important factor in the research was the use of term-frequency-inverse document frequency method to convert the text responses into machine-readable format.

Research from Kumar, V., & Subba, B demonstrated uses of 'TfidfVectorizer' to feed it into the machine learning model to predict the sentimental analysis of the textual contents in the document.[11] Before vectorizing the documents, the text documents are tokenized by a novel technique where the stop words, special characters are all eliminated. In addition, the words with similar meanings are joined together by the concept of stemming which comprises of the pre-processing steps. The TfidfVectorizer uses the inverse domain frequency and term frequency to calculate the term frequency and the weightage of the words. The parameters

set for the TfidfVectorizer includes ‘max_df’ which defines maximum value for document frequency and eliminates any words that doesn’t fall under the circumstance, another parameter ’sublinear-if’ performs scaling by replacing tf value with (1+log(tf)) to avoid dominance of certain words. Once converting documents into vector format is completed, the processed vectorized document is fed as input to the Support Vector Machine to train the model and the model will be deployed for sentimental analysis in real time. Similarly, research about author profiling from Dichiu, D. and Rancea, I., the authors use TfidfVectorizer at character level rather than the word level alone with parameters set for the function as similar as to the parameters set in Kumar, V., & Subba, B’ research.[12] The results obtained from the [3] research shows that the TfidfVectorizer along with machine learning models works best on the word level than the character level.

REALTIONSHIP BETWEEN IMAGE KEYWORDS AND TEXT

Researching more about establishing a relationship between image keywords and written text, researcher Park, K., Hong has used a methodology combining the cosine similarity and a machine learning classifier for classifying the text.[13] Another research from Kayalvizhi, S., Thenmozhi, D., & Aravindan, C showed a different approach of using Jaccard similarity and cosine similarity on the vectorized data.[14] Both the Jaccard similarity and cosine similarity return values between 0 and 1. Out of the two similarity measurements, the cosine similarity resulted in slight advantage over the Jaccard similarity when vectorized by TfidfVectorizer [14]. In research from Shrivastava, D, CG, SA, Laha, A & Sankaranarayanan and Wei, Y “Max”, Hong, J & Tellis, GJ, researchers demonstrated how to calculate similarity between different text data and analyse the correlation. This established the baseline of measurable factors in this project, and cosine similarity between text response and image descriptions/labels must be considered. Moreover, some other factors such as the word count, and the number of verb/nouns of text response are crucial features as well.

MACHINE LEARNING ALGORITHMS

Several methods were mentioned/implemented: SVM, Random Forest, Neural Network, Bayesian Regression in different papers. [1,2,3,4] From Mohammad, A. H., Alwada‘n, T., & Al-Momani, O ‘s literature, SVM algorithm is used to build models that assigns new documents into a set of predefined categories. [15] The SVM classifier has a unique identity of using hyperplanes along with the linear classifier plane. Even with the data being non-linearly separable, the SVM classifier can classify the data by adding an extra dimension to it so that become linearly separable and able to project the decision boundary back to the original dimension using mathematical transformation. This experimental model was able to achieve an accuracy of 93%, thus demonstrating the model’s feasibility for predicting creativity. The limitation of the research is that the number of participants and data points is exceptionally low.

Shrivastava, Ahmed, Laha, and Sankaranarayanan (2017) evaluated how random forest models were used to measure the creativity of movies. The authors use five different machine learning models to test and compare results. The five models are random forest, SVM, Bayesian regression, Ridge regression, and KNN. Their random forest model uses 600 trees and a minimum sample split of 2. They use 80% data for the training set and 20% for the testing set. And for their SVM model they use default settings. According to their results, random forest performs better than SVM and Bayesian regression [3].

Researchers Wu, J, Wu, B, Pan, S, Wang, H & Cai explained that Bayesian network is a simple graphical notation for conditional independence assertions. [6] This model can learn the network structure from data attributes and generate the probabilistic relationship between any cause and effect. Unlike other machine learning models, the Bayesian network will try to calculate the probability of cause-effect within attributes and predict the outcome by probabilities instead of clustering data and calculating distances between points. Since there is no current standard to measure creativity in the literature, it is essential that instead of clustering data, understanding the factors that lead to high/low creativity is imperative. Further research from Feng, G, Guo, J, Jing, B-Y & Hao demonstrates Bayesian network has over 85% of accuracy when doing text classification.[8] On the other hand, researchers Ann Sung Lee & So Young Sohn tried measure creativity on a corporate level by collecting employee's performance data and quantifying them to all 22 variables which fall under six distinct categories. They fed data into the Bayesian network model which learnt the relationships and produced a Directed Acyclic Graph (DAG) to help them identify necessary features. [9] These findings provide convincing evidence that Bayesian network could be a good option.

Image Processing

Based on the literature review, four models were selected, namely: Support Vector Machine (SVM), Random Forest, Neural networks, and Bayesian networks. As per the methodology, image processing will produce keywords by using an image analysis tool beforehand.

Four popular image processing tools namely Google Cloud Vision API, Microsoft Azure Computer Vision, AWS Rekognition and IBM Watson Visual Recognition were shortlisted and compared. All the four tools are compared on a feature basis and the comparisons are listed in table 1.

Feature	Google cloud Vision	AWS Rekognition	Microsoft Azure	IBM Watson
Object Detection	✓	✓	✓	✓
Face Detection	✓	✓	✓	✓
Scene detection (statement about image)	⊗	⊗	✓	⊗
Sentiment Detection (no separate tool)	✓	⊗	✓	⊗
Text recognition	✓	✓	✓	⊗
Written text Recognition	✓	✓	✓	⊗
Logo detection	✓	✓	✓	⊗
Landmark Detection	✓	✓	⊗	⊗
Violence Detection	✓	✓	✓	✓
Face comparison	⊗	✓	⊗	⊗
Face Search	✓	✓	⊗	⊗
Classification	✓	✓	✓	✓
Regression	✓	✓	✓	✓
Clustering	✓	✓	✓	⊗
Anomaly Detection	✓	⊗	✓	⊗
Prediction with Accuracy	✓	✓	✓	✓
Data Labelling	✓	✓	✓	✓
Provide alternate labels	✓	⊗	✓	⊗
Dominant colours detection	✓	⊗	✓	✓

Table 1 - Evaluate image processing tools

Google cloud vision AI was chosen as the best image analysis tool because of its extensive feature lists that includes its ability to acquire all the results in a single data, detecting texts of very low quality with good accuracy and the ability to provide a long list of alternate labels which will prove beneficial in filtering out many common answers.

Data Preprocessing

The dataset consisted of nine features. The caption was given by the test subjects and the JudgeBin (creativity level 1, 2, or 3) were given by experts in the field. Since the goal is to automate the prediction of JudgeBin for the verbal test using image, the rest of the features were removed from the data set because those values were scores from various other tests which were not related to verbal creativity. To predict the creativity level, caption and JudgeBin was extracted from the dataset as the initial features. Since caption which is in form of a text cannot be directly fed into the model as a feature, several other features were generated based on the caption. Features like word count, stop word count, special character count were generated based on the caption.

As pointed out earlier, Google Cloud Vision API was used to generate keywords based on the images. The result of the tool produced features like labels, text, logo, objects, and web entities found in the image. Since the list of keywords is not comprehensive enough, synonyms and antonyms of the keywords were found using the WordNet library and added to the list of keywords for the images. Adding synonyms and antonyms ensures that similarity (Jaccard or cosine) values are generated (values other than 0) for a greater number of captions. The result of the image processing tool also produced confidence percentages of the labels, text, and objects and hence each of the above object features in conjunction with similarities were divided into High Label group (confidence > 50%) and Low Label group (confidence < 50%). Table 2 outlines the list of features that were generated using the image analysis tool and caption.

Features	Description	Data Type
word_count	Word count of a caption	Integer
stop_word_count	Stop word count in a caption	Integer
special_char_count	Numbers of cardinal digit	Integer
face_j	Face recognition label Jaccard Similarity	Float
face_c	Face recognition label Cosine Similarity	Float
high_relate_text_j	High Related Text Jaccard Similarity	Float
high_relate_text_c	High Related Text Cosine Similarity	Float
low_relate_text_j	Low Related Text Jaccard Similarity	Float
low_relate_text_c	Low Related Text Cosine Similarity	Float
label_j	Image label Jaccard Similarity	Float
label_c	Image label Cosine Similarity	Float
high_relate_obj_j	High Related Object Jaccard Similarity	Float
high_relate_obj_c	High Related Object Cosine Similarity	Float
low_relate_obj_j	Low Related Object Jaccard Similarity	Float
low_relate_obj_c	Low Related Text Cosine Similarity	Float
guess_label_j	Best Guess Image label Jaccard Similarity	Float
guess_label_c	Best Guess Image Label Cosine Similarity	Float
high_relate_entities_j	High Related Web Entity Jaccard Similarity	Float
high_relate_entities_c	High Related Web Entity Cosine Similarity	Float
low_relate_entities_j	Low Related Web Entity Jaccard Similarity	Float
low_relate_entities_c	Low Related Web Entity Cosine Similarity	Float
Class	JudgeBin, the prediction target class	Integer

Table 2 - Generated features

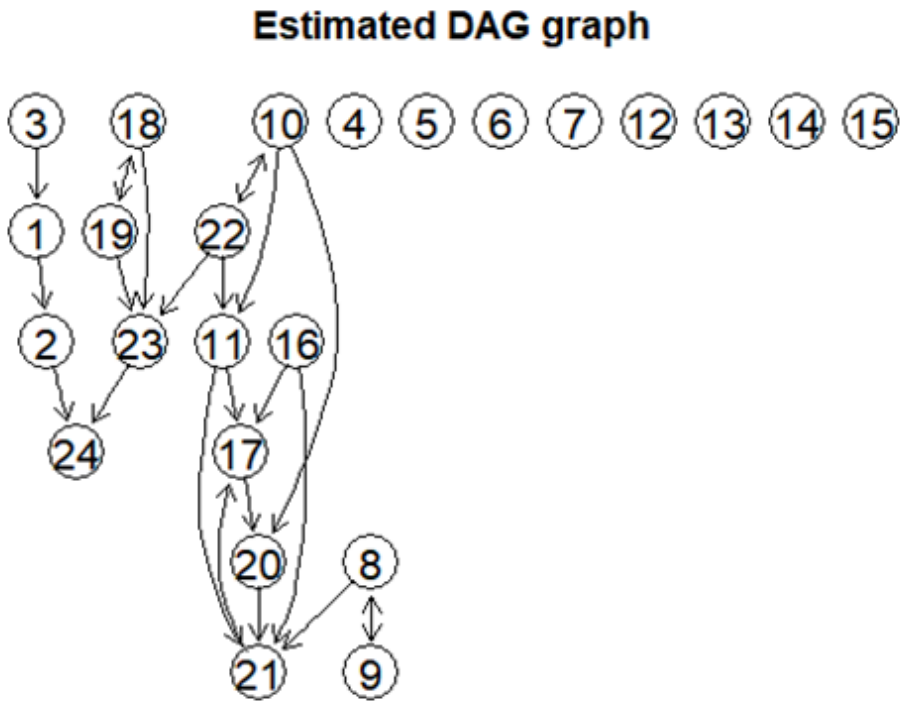

Figure 1 - Initial DAG of the features

Furthermore, Bayesian network was applied to identify if the features identified in table 2 are crucial. Figure 1 represents the DAG of the estimated cause/effect within features. The graph shows most features have some relationship with each other, which indicates that the extracted feature would provide good predictive power to the Bayesian network. However, some features have no connection or effect on other features, and the remaining features have complex relationships. More data manipulation will be performed, and another DAG will be generated to verify features importance.

Figure 2 represents the DAG when the data in the features were converted into categorical data.

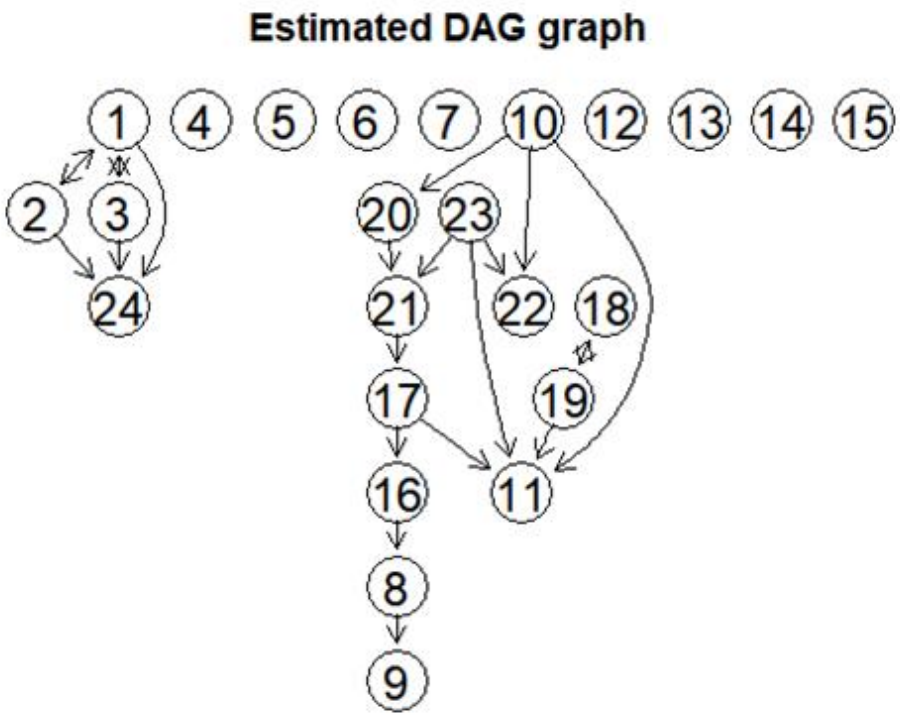

Figure 2 - DAG of the features after converting to categorical data

To convert feature data to categorical data, each value was compared with the mean of the feature column. For each observation, if the value is smaller than the mean of the column, it will be assigned as 1 else it will be assigned as 2. In case the original value of the observation is found to be 0, it will remain untouched.

The DAG in Figure 2 indicates a hierarchical relationship between each feature which implies that categorical data provides better predictive power. 20 new features were extracted using NLP on text response.

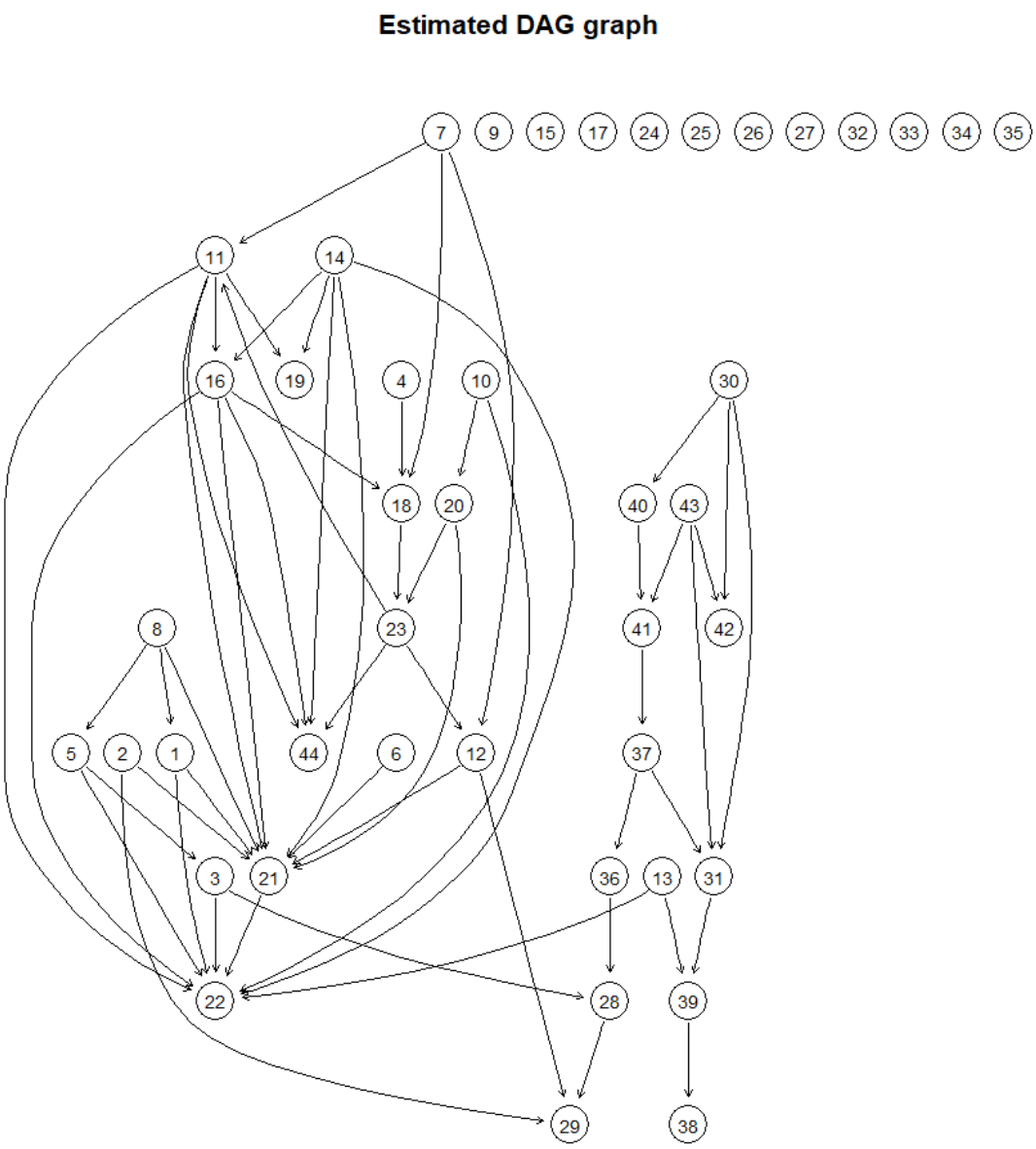

Figure 3 – DAG of the features after adding 20 new extra features

It can be observed from DAG in figure 3 that the relationships in the graph are more sophisticated. Even though there are some features that are not connected to any other feature, these features could potentially provide value when the size of the dataset increases, and the Bayesian network have more data to analyze the relationships.

DATA DICTIONARY

Table 3 shows the data dictionary of the selected data set.

Variable	Description
JudgesBin	The judge Bin value
Coordinating conjunction	Numbers of Coordinating conjunction
cardinal digit	Numbers of cardinal digit
determiner	Numbers of determiner
existential there	Numbers of existential “there”
preposition	Numbers of preposition
adjective	Numbers of adjective
modal	Numbers of modal
noun	Numbers of noun
predeterminer	Numbers of predeterminer
possessive ending	Numbers of possessive ending
personal_pronoun	Numbers of personal pronoun (hers, herself, him, himself)
adverb	Numbers of adverb
particle	Numbers of particle (about)
infinite marker	Numbers of infinite marker (to)
interjection	Numbers of interjection (goodbye)
verb	Numbers of verb
wh-determiner	Numbers of wh-determiner (that, what)
wh- pronoun	Numbers of wh- pronoun (who)
wh- adverb	Numbers of wh- adverb (how)

other	Numbers of word not in above categories
word_count	Word count of a caption
stop_word_count	Stop word count in a caption
special_char_count	Numbers of cardinal digit
face_j	Face recognition label Jaccard Similarity
face_c	Face recognition label Cosine Similarity
high_relate_text_j	High Related Text Jaccard Similarity
high_relate_text_c	High Related Text Cosine Similarity
low_relate_text_j	Low Related Text Jaccard Similarity
low_relate_text_c	Low Related Text Cosine Similarity
label_j	Image label Jaccard Similarity
label_c	Image label Cosine Similarity
high_relate_obj_j	High Related Object Jaccard Similarity
high_relate_obj_c	High Related Object Cosine Similarity
low_relate_obj_j	Low Related Object Jaccard Similarity
low_relate_obj_c	Low Related Text Cosine Similarity
guess_label_j	Best Guess Image label Jaccard Similarity
guess_label_c	Best Guess Image Label Cosine Similarity
high_relate_entities_j	High Related Web Entity Jaccard Similarity
high_relate_entities_c	High Related Web Entity Cosine Similarity
low_relate_entities_j	Low Related Web Entity Jaccard Similarity
low_relate_entities_c	Low Related Web Entity Cosine Similarity

Table 3 – Features of the selected data set

Exploratory Data Analysis

The dataset provided includes captions and creativity scores of each caption for four images. For each image, the number of responses varies relatively. The elephant image has 224 responses, the car image has 452 responses, the excavator image has 441 responses, and the stairwell image has 309 responses. In total, the dataset consists of 1426 records. Figure 4 represents the bar chart of the record count according to images.

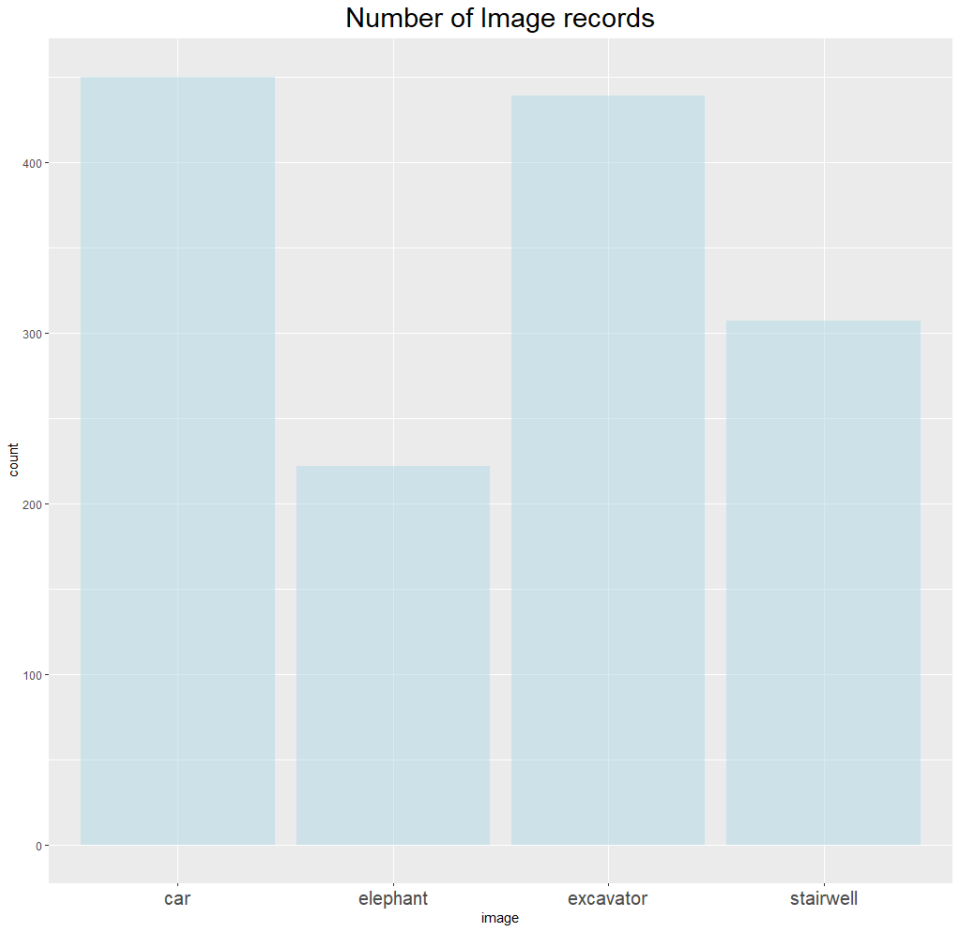

Figure 4 - Number of records according to images

The initial dataset consisted of eight features in total. Excluding the “Caption” feature, all the other features were binned to 3 categories representing various creativity levels (1- low, 2- medium, 3- high). Figure 5 represents the bar chart of the total count of records according to their creativity level.

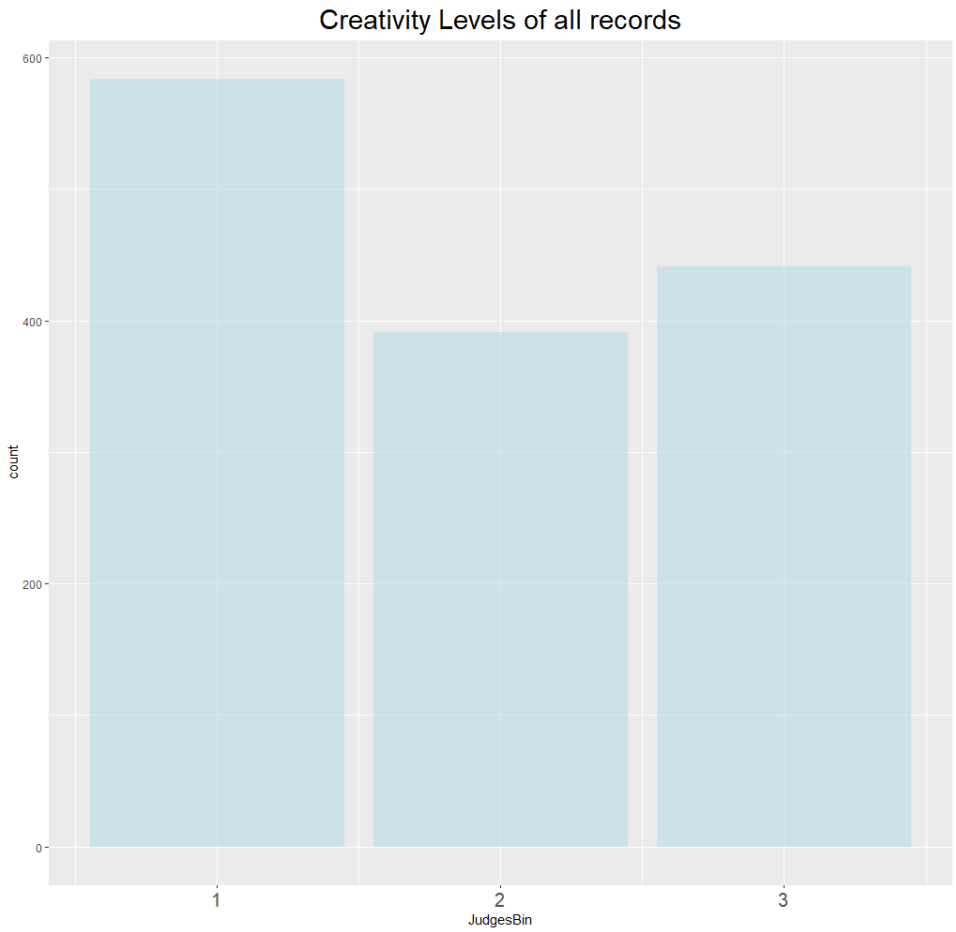

Figure 5 - Total records by creativity level

Table 4 shows the features provided in the initial dataset.

Feature	Description	Type
Caption	Caption for the image written by test subjects.	String
OpenBin	Candidate's ability to be open to new experiences	Ordinal
IRTBin	Candidate's ability for Intellectual Risk Taking	Ordinal
GrowthBin	Candidate's Growth Mindset	Ordinal
CSEBin	Candidate's Baseline Self-Efficacy	Ordinal
CPIBin	Candidate's Creative Personal Identity	Ordinal
JudgesBin	Candidate's creativity level for the caption given by judges.	Ordinal
DTBin	Candidate's composite score of Divergent Thinking	Ordinal

Table 4 - Features in the initial data file

	OpenBin	*IRTBin*	*GrowthBin*	*CSEBin*	*CPIBin*	*JudgesBin*	*DTBin*
OpenBin	*1*						
IRTBin	*0.272*	*1*					
IRTBin	*0.174*	*0.339*	*1*				
CSEBin	*0.305*	*0.361*	*0.329*	*1*			
CPIBin	*0.195*	*0.235*	*0.256*	*0.562*	*1*		
JudgesBin	*0.100*	*0.079*	*0.076*	*0.141*	*0.081*	*1*	
DTBin	*0.100*	*0.109*	*0.143*	*0.200*	*0.176*	*0.095*	*1*

Table 5 - Correlation matrix of initial features

Table 5 represents the correlation matrix based on the initial features available in the dataset. The correlation matrix represents the linear association between two variables with a value between 1 or -1. A 0 represents that there is no relationship while an absolute value of 1 indicates a strong relationship between the variables. As can be observed from figure 3, the maximum value of 0.562 is present between CPIBin and CSEBin which indicates that the two

variables are moderately associated with each other. All other relationships seem to be between moderate and weak and hence these features can be potentially used for creating a model. Since there are no strong correlations between features, none of the features were removed based on the correlation matrix. As mentioned in the data preprocessing section, except Caption and JudgesBin, all the other features were not considered for the model.

With the addition of new features from the data, we were able to visualize the relationship between these features. We started exploring the effects of the features “word_count” and “special_char_count” on the JudgesBin creativity levels.

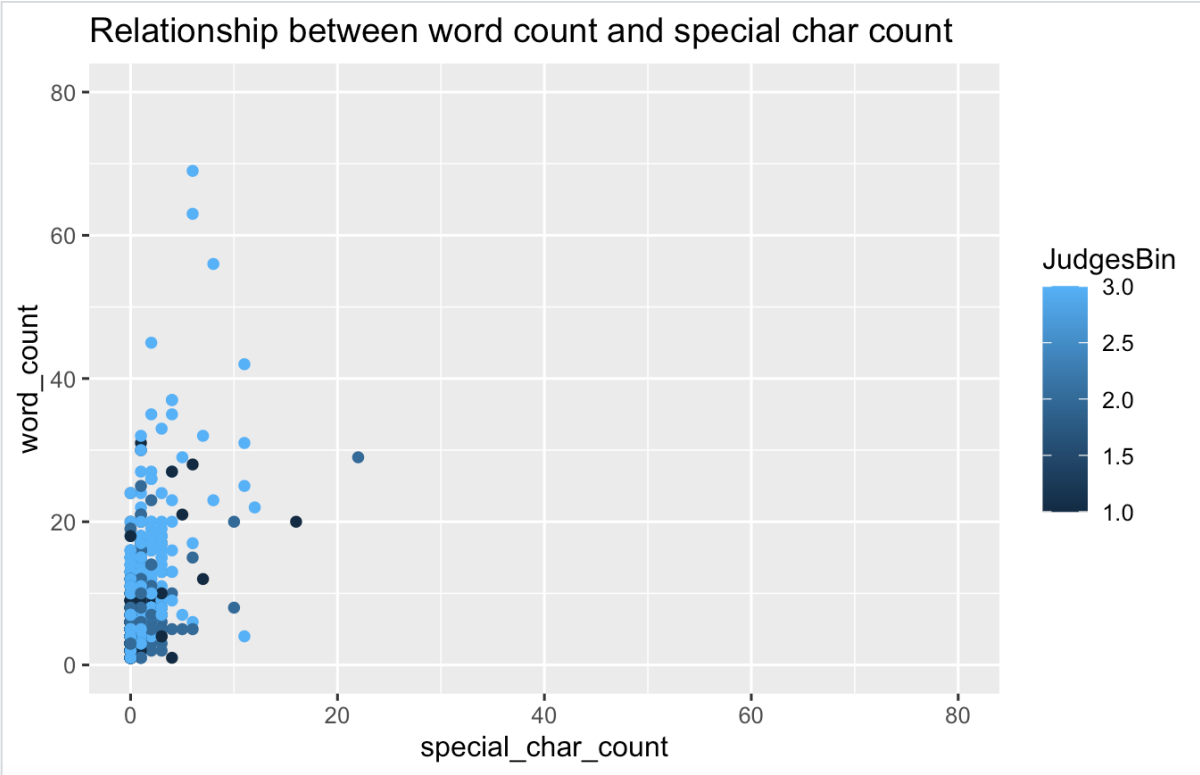

Figure 6 - Special character count and word count according to creativity level

Figure 6 represents the scatter plot of special character count and word count with corresponding JudgeBin (Creativity Level). Both high and medium creative captions have less word count with more special characters compared to captions with low creativity. According to the scatter plot, higher number of word count in the caption doesn’t necessarily imply that the caption will be highly creative.

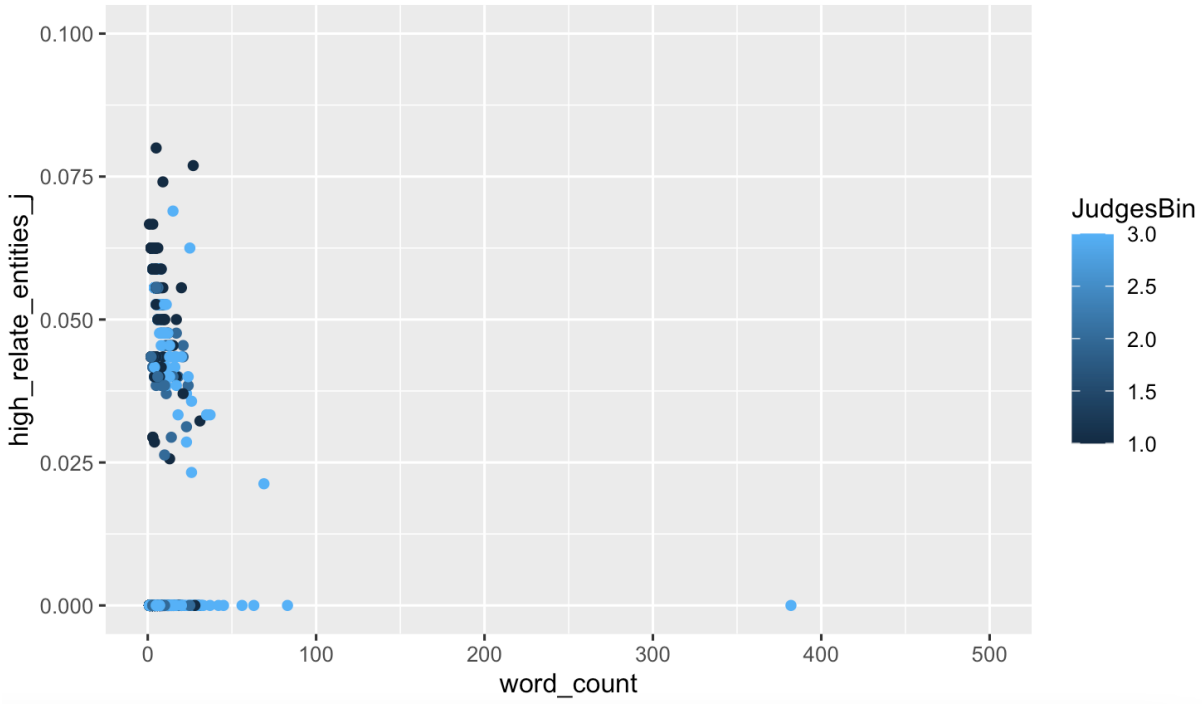

Figure 7 - Relationship between word count and high related entities(keywords) with creativity level

Figure 7 represents the relationship between word count and high related entities with corresponding creativity level. Regardless of the word count the captions with high related entities falls to high creativity and medium creativity bins.

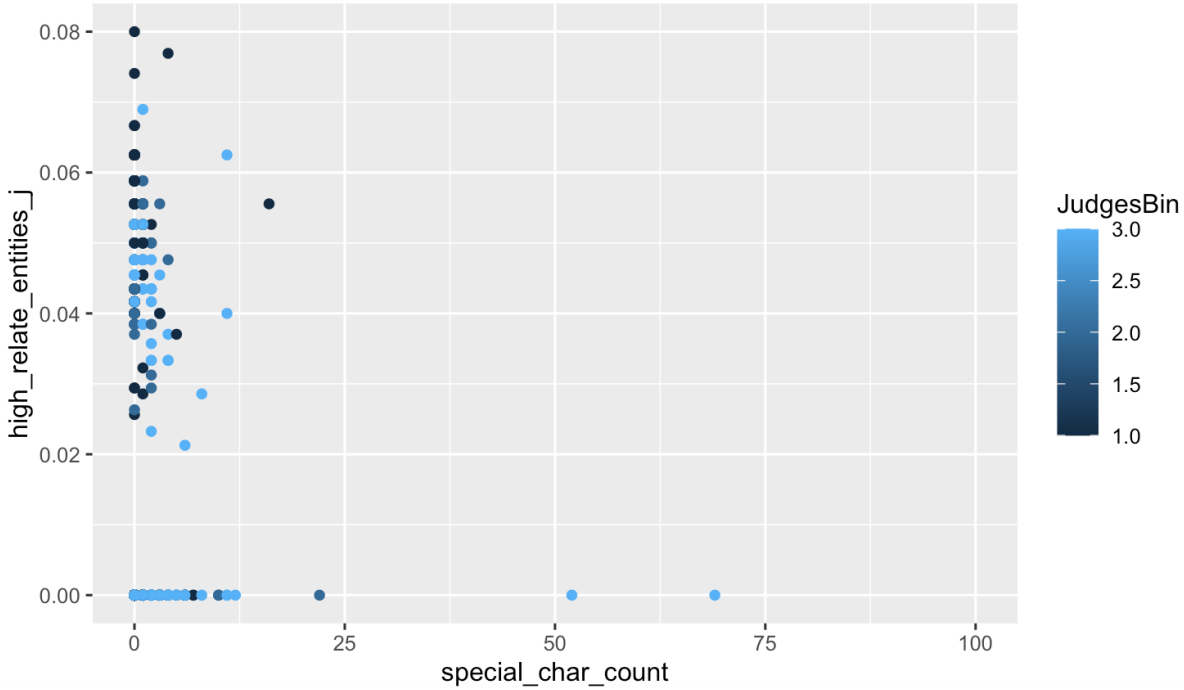

Figure 8 - Relationship between special character count and high related entities(keywords) with creativity level.

Figure 8 represents the relationship between special character count and high related entities (keywords) with corresponding creativity levels. It can be observed that the “high related entities” have a huge impact on the creativity level. Next, investigating the wordcloud of the captions.

Figure 9 - Word cloud of Caption from dataset

Figure 9 above represents the word cloud of all the captions in the dataset. Word clouds are graphical representation of word frequencies that give greater prominence to words that occur more frequently. The word cloud was created after remove most common English stop words to analyse the caption text in the dataset. It can be observed that some of the most prominent words highlighted are stairs, fence, life, two, big, bottom, dunk and Asian.

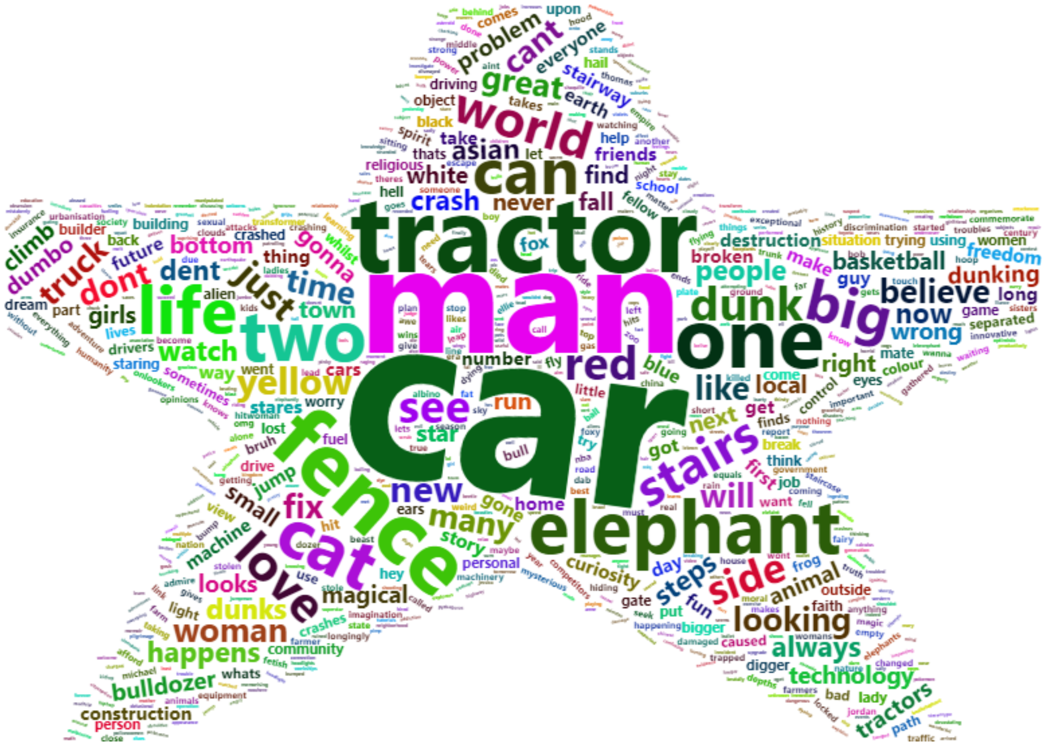

Figure 10 - Word cloud of Caption that is high creativity

Figure 11 - Word cloud of Caption that is low creativity

Figures 10 and 11 represent word cloud of high and low creativity captions. It can be observed that most words of high frequency appear in both the word clouds like fence, man, elephant, tractor and one. This indicates that even though the creativity levels are different for the word clouds, the words with highest frequency are similar and hence these highlighted words cannot be used to distinguish between creative levels.

Methods

SUPPORT VECTOR MACHINE (SVM)

Classification and regression problems are both common applications for the Support Vector Machine (SVM). An important purpose of the SVM method is the creation of the best decision boundary or line that can separate n-dimensional space into classes, so that fresh data points may be readily placed in the correct category in the future. A hyperplane denotes the optimal choice boundary. SVM selects the hyperplane's extreme points and vectors. Support vectors are extreme situations, and so the algorithm is known as a Support Vector Machine (SVM). Figure 12 represents an SVM.

Using an example, the SVM algorithm's operation can be better understood. Let's say we have a dataset with two tags (green and blue) and two features (x1 and x2). A classifier that can classify the pair of coordinates (x1, x2) as either green or blue is what we're looking for. With 2-d space, we can readily distinguish between these two groups by drawing a simple straight line between them. However, these groups can be divided up along more than one line. Because of this, the SVM method helps to discover an optimal boundary or region, which is referred to as a hyperplane. The SVM algorithm locates the intersection of the two classes' lines nearest to a given point. Support vectors are the technical term for these sites. Margin refers to the distance between the vectors and the hyperplane. SVM's objective is to increase this margin to the maximum possible.

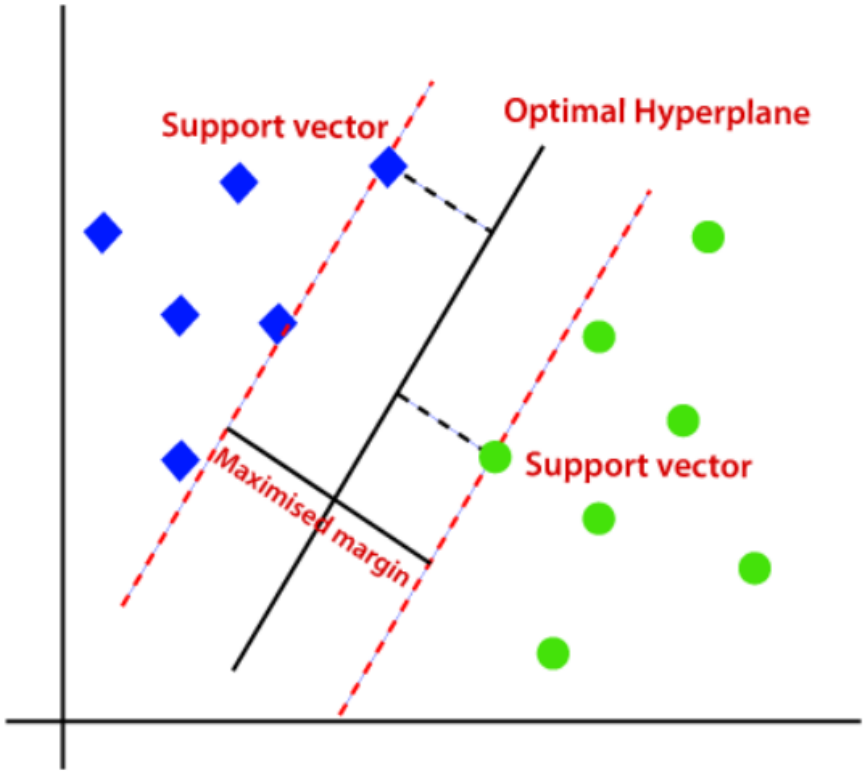

Figure 12 – Explanation of SVM

The final dataset is provided as an input to the model and the dataset is split into 70% training data and 30% test data. The results of the SVM model are explained in the table 6.

	Predicted Class = Low	Predicted Class = Medium	Predicted Class = High
Actual Class = Low	158	8	15
Actual Class = Medium	90	8	16
Actual Class = High	76	14	41

Table 66 - SVM confusion matrix

Metric	Score
Accuracy	48.59
F1 Score	38.02
Recall	41.87
Precision	44.12

Table 77 - SVM results

The SVM model achieves a total prediction accuracy of 48.59% as per table 7. The F1 Score of the model is 38% which is majorly due to the low prediction of the “Medium” level creativity.

RANDOM FOREST

Random Forest (RF) is used as one of the machine learning models for the prediction of creativity level. RF model is made up of several decision trees. Each decision tree predicts the creativity level and after following a majority voting, the result of the RF is the class with most votes by the decision trees. Since collective decision-making outperforms a single decision-making process, RF models are vastly superior to decision trees. Figure 13 shows RF model with three decision trees.

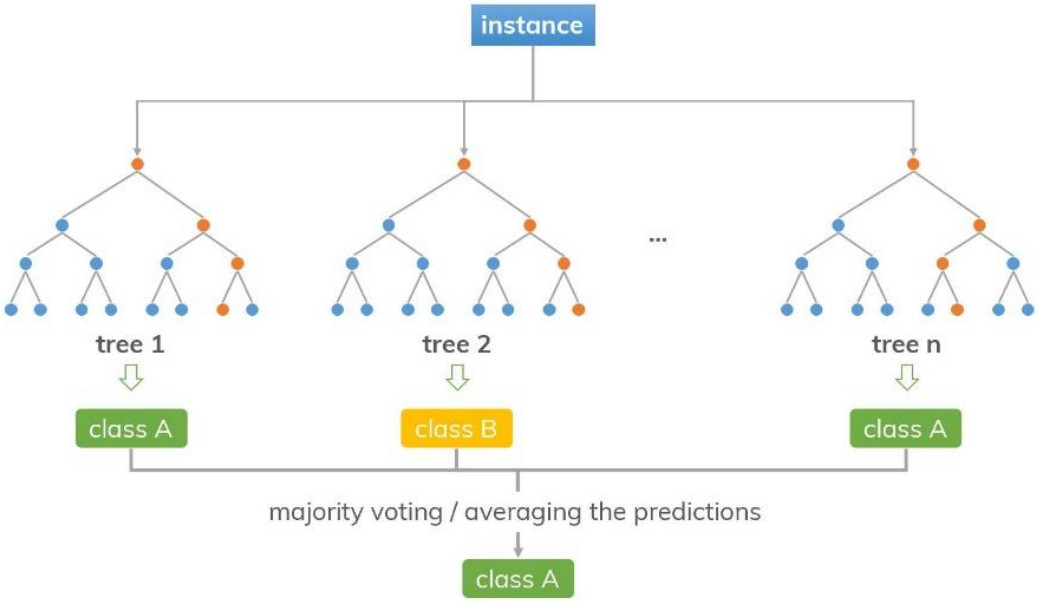

Figure 13 - Random Forest model

The number of trees used in the RF model are 300. Increasing the number of trees can make the prediction rate higher in random forest but may overfit the data.

When testing the model with the final dataset, the accuracy of the model is between 44% to 50%. This is because of the segmenting of testing and training data set randomly assigned as 40% for testing and 60% for training data.

Tables 8 and 9 show the results of classification. Table 8 shows the confusion matrix for the classification. The model has low f1 score for medium creativity as compared to low and high creativity levels.

	Predicted Class = Low	Predicted Class = Medium	Predicted Class = High
Actual Class = Low	144	41	33
Actual Class = Medium	81	34	41
Actual Class = High	55	34	104

Table 88 - Confusion matrix of random forest results

Metric	Score
Accuracy	49.73
F1 Score	46.66
Recall	47.33
Precision	46.66

Table 99 - Result for classification using random forest model for the test data set

BAYESIAN NETWORK

Bayesian network calculates the probability cause-effect of features to predict outcomes. Figure 14 represents the Bayesian network based on the NPC algorithm

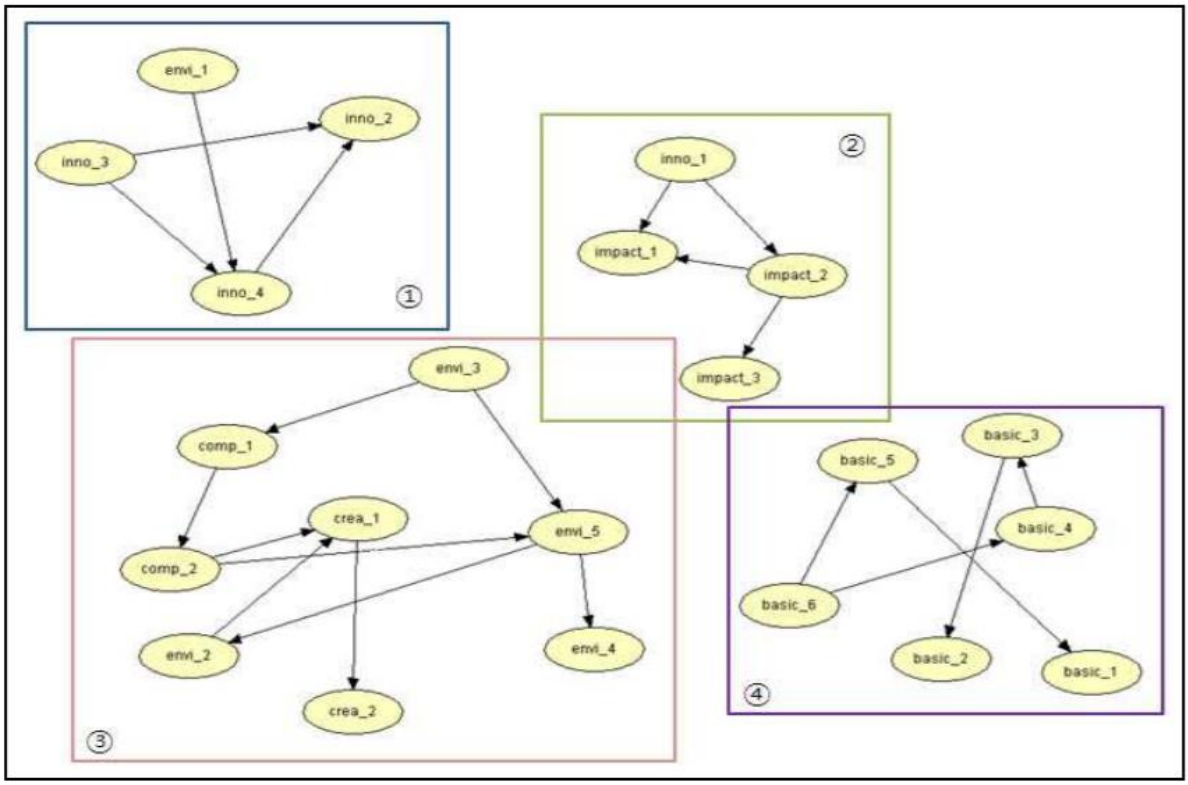

Figure 14 - Bayesian network based on NPC algorithm

The DAG represents the cause-effect between each feature from the data. Unlike other machine learning models, the Bayesian network doesn't need to interpret the relation between each cluster, instead it will calculate the probability and learn the relationships from the data.

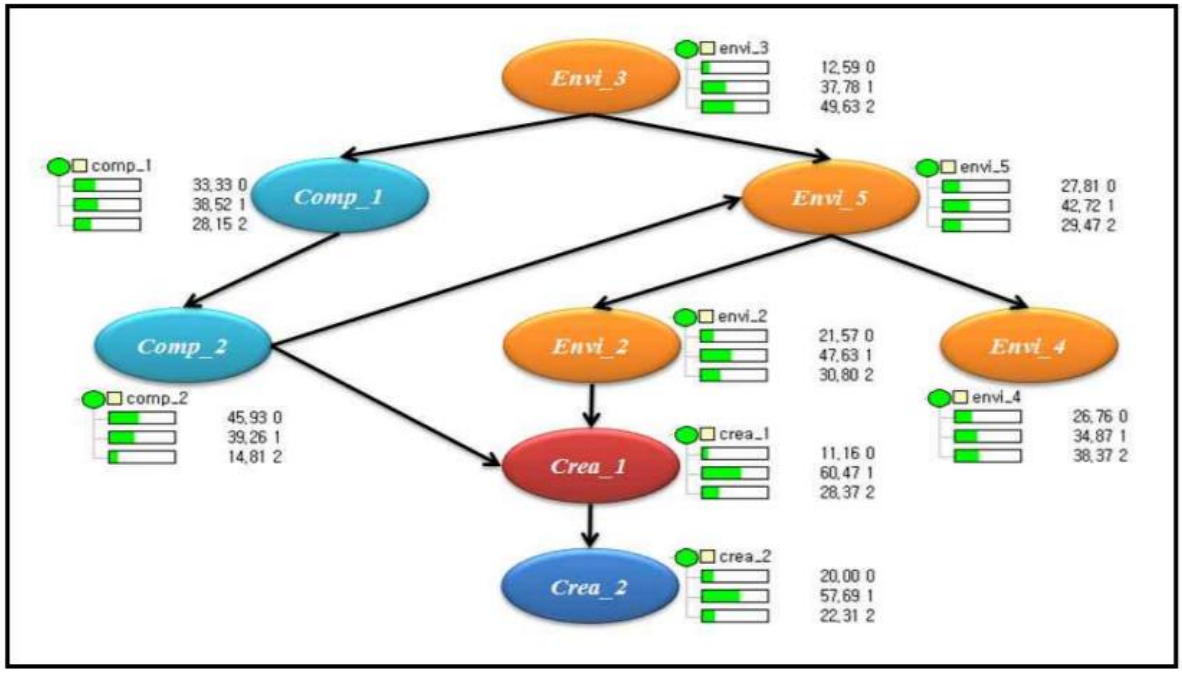

Figure 15 - Bayesian sub-network based on the NPC algorithm

Beside the Directed Acyclic Graph, Bayesian network also calculates the probability and generates a probability table. With this table, Bayesian network would be able to learn which feature is holds more predictive power. Figure 15 represents a Bayesian sub-network based on the NPC algorithm.

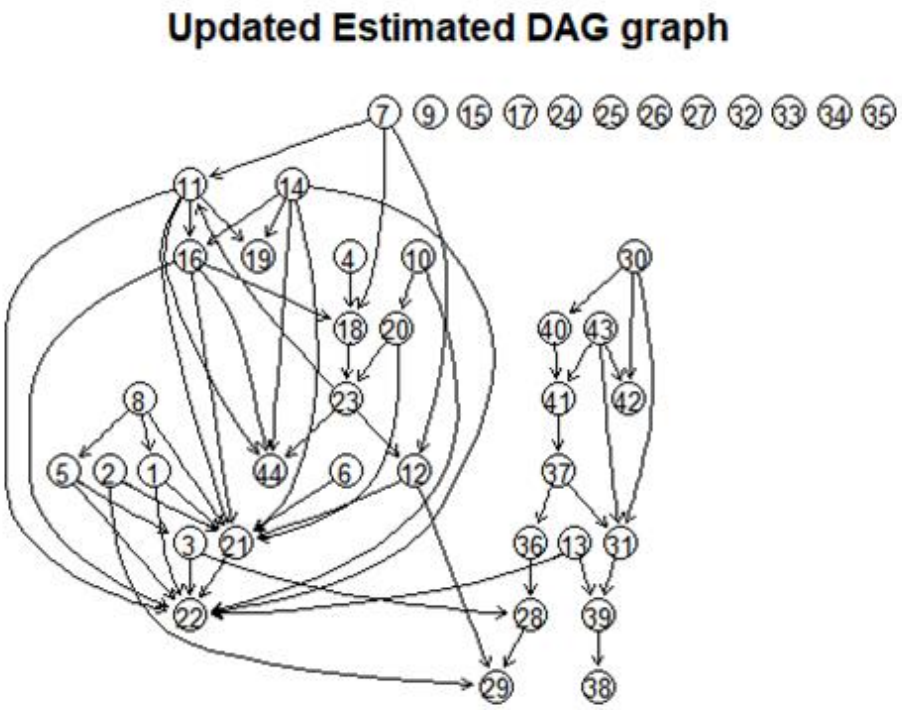

Figure 16 - DAG generated after data pre-processing

After the data pre-processing, the DAG represented in figure 16 was generated. It indicates a more detailed relationship of each feature from the dataset. This dataset was then fed into Bayesian network, and prediction outcomes were observed as in Table 10.

	Class 1	Class 2	Class 3
Class 1	448	231	162
Class 2	30	46	45
Class 3	106	114	234

Table 1010 - Prediction result for Bayesian network

Metric	Score
Accuracy	0.5141
95% CI	(0.4877, 0.5405)
P-Value	7.8444-15
Kappa	0.2307
Mcnemar's Test P-Value	< 2.2e-16

Table 1111 - Confusion Matrix and Statistics for Bayesian network

	Class 1	Class 2	Class 3
Sensitivity	0.7671	0.11765	0.5306
Specificity	0.5276	0.92683	0.7744
Pos Pred Value	0.5327	0.38017	0.5154
Neg Pred Value	0.7635	0.73359	0.7848
Prevalence	0.4124	0.27613	0.3114
Balanced Accuracy	0.6474	0.52224	0.6525

Table 1212 - Statistics by class for Bayesian network

From the result above the accuracy was 51.4%, and P-value shows this model covers most data which is efficient. Sensitivity for Class 2(middle) and Class 3(High) are 12% and 53%. For Class1(low), Sensitivity is 76.7% which is great. Even the sensitivity for Class 2 and 3 are low, but recall(specificity) is 92.6% and 77% which is great. The result indicates high potential of the model. For a such small dataset that doesn't have any obvious pattern, Bayesian networks still manage to have a good prediction accuracy. Hence, if the dataset increases, this model could perform even better.

NEURAL NETWORK

A neural network is built of many Neurons that are arranged in multiple layers which work together to give a desired output based on the inputs and the weights of each neuron. A neuron is the building block of a neural network. It takes two or more input (X_1 and X_2), applies the weight of each input (f(x)), and produces an output (y). A neuron is represented in figure 17.

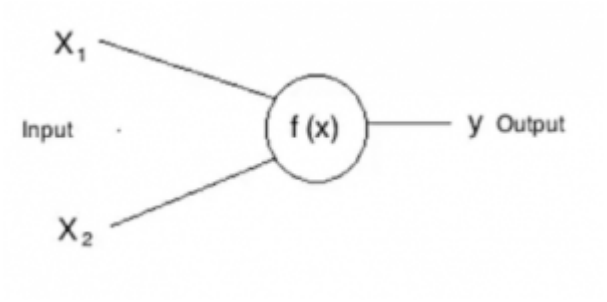

Figure 17 - Representation of Neuron

This is also known as a simple perceptron or single layer perceptron. When multiple neurons are stacked together in layers, we get a network of neurons which is popularly known as a neural network (NN) or a multi-layer perceptron. A neural network is represented in figure 18.

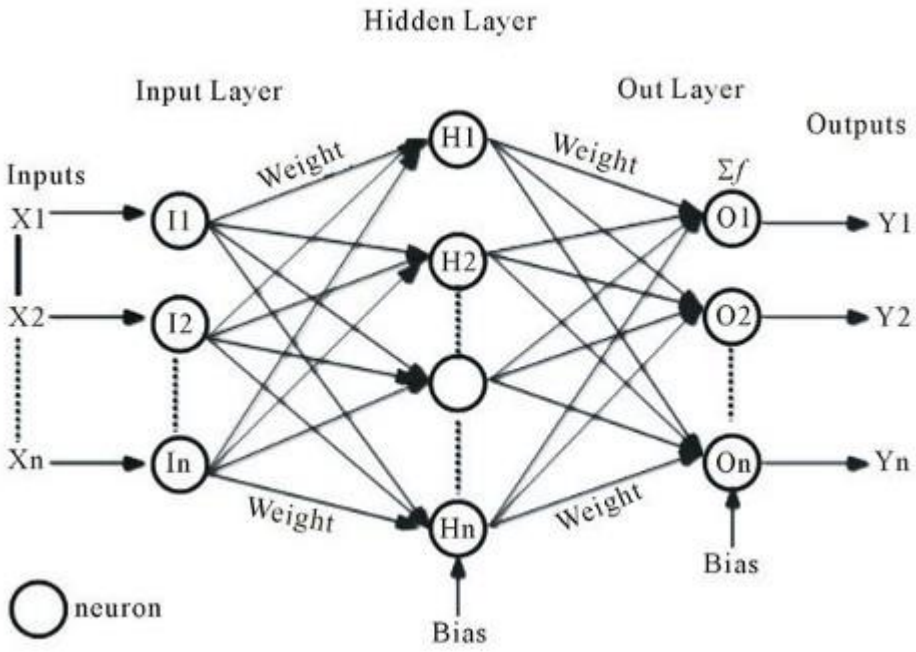

Figure 18 - Representation of Neural Network

The final dataset was then used to train the neural network model. 70% of the data was used for training the model and 30% of the data was used for testing purposes. The model accuracy on the testing data was found to be between 38% and 56%. This behavior is observed since the division of training and testing data has been performed randomly in the 10-cross validation testing method.

Table 14 shows the results from the neural network model when classifying. The accuracy of the model is 47%. When comparing the f1 scores for different classes in the model, the f1

score for both low creativity and high creativity is greater than 0.5. So, the model does a good job when classifying these two classes. But since the f1 score is lower in medium creativity, it seems the model is having trouble classifying captions that belong to medium creativity. Overall, the f1 score is 0.47 which is medium and indicates the model has not performed good overall when classifying into the three categories. Table 13 shows the confusion matrix for the classification.

	Predicted Class = Low	Predicted Class = Medium	Predicted Class = High
Actual Class = Low	133	35	25
Actual Class = Medium	83	30	46
Actual Class = High	56	25	125

Table 1313 - Confusion matrix using Neural Network

Metric	Score
Accuracy	47%
F1 Score	0.47
Recall	0.41
Precision	0.45

Table 1414 - Results of Neural Network

Model Comparison

After training the four models with the selected dataset, the next step was the model comparison to find which model fits best to automate the assessment of creativity. All four models were trained and tested using the same dataset and same features to do a proper comparison among the models. Table 15 shows the performance evaluation of each model. The initial plan was to find a single model with the highest performance. But because of the lack of data, it was hard to train the models to get higher performance than their current performance. All the models achieve an accuracy of around 50%. So, it is hard to determine what the best model is. But compared to SVM the other three models have slightly better performance. Random forest and Bayesian network models perform better than the neural network. Both precision and recall are in the range of 40% – 50% in all four models. According to the obtained results, the random forest model and the Bayesian network model is recommended in the assessment of creativity.

Model	Accuracy	Precision	recall
SVM	48.59%	44%	42%
Random Forest	49.78 %	46.66%	47.33%
Bayesian Network	51.41%	47%	47%
Neural Network	49.12%	45%	44.66%

Table 15 15 - Model comparison

Conclusion

After performing the various steps in the analysis like data processing, model training, model testing, and comparing the performance of each model, several conclusions can be drawn.

As can be observed from the model comparison table 16, each of the models had an accuracy around 50% which makes all these models unsuitable for use in real world applications. One of the possible reasons this deficiency in accuracy is observed is due to the small volume of dataset that was available for training and testing the machine learning models. The initial list of features in the model was expanded and a total of 44 features were used. Even after drastically increasing the number of meaningful features for the models to be trained on, the deficiency in accuracy remains which might hint that either the dataset volume is very small, or the features are not useful when predicting the creativity level of individuals. Both issues indicate that there is a massive negative impact of low dataset volume to successfully generate and train the models.

As per the methodology followed in the analysis, it is of utmost importance that the model recognizes the objects present in the images based on which the candidates write a creative caption. To achieve this, several third-party Image analysis tools were examined, and the best performance was achieved by Google Cloud Vision API. Even though the tool was able to recognize multiple objects in any given image, the accuracy of these objects and the number of objects being identified was questionable especially when the object has some features involved which are different than the normal set of features for the object. For instance, in one of the images where an elephant (an object, which is the central theme of the image) was flying (unusual feature for the object), the tool was not able to identify the object at all. In the same image, the tool misclassified a cartoonish fox as rabbit which is not acceptable as well. This complication was mainly affected the vector image that was provided, and the other images had better keywords compared to the mentioned vector image. This process of keyword generation using Image analysis tool is pivotal to the automation of the creativity scoring process and the above issues are an impediment to the automation involved. For the

models to be trained with high accuracy, the keyword list for an image needs to be extensive and accurate. Until better image analysis tools are available, the models will be performing poorly. One of the possible options is to either train an image recognition model from scratch specifically for our tools or the keyword generation needs to be a manual component in the process when new images are being added to the system for creativity analysis.

Overall, all models have similar prediction results, but SVM has the lowest accuracy, precision, and recall. This confirms that SVM is a relatively simple prediction model, and it can't handle data with too many features. SVM is a simple binary classification algorithm and in this case, SVM performs poorly classifying the multiclass target variable which is one of the shortcomings of the algorithm. These shortcomings due to the multiclass classification can be improved using certain techniques such as One vs One (OVO) approach and One vs All (OVA) approach.

The random forest model also has an accuracy of around 50%. But it can be considered better than SVM because it gives a result of collective decision making. The random forest model could improve the results if it had a larger dataset and alter the features such as keyword count and antonym count as a density instead of a count. Because the dataset was small the tree count of the forest was selected as 300 because it gives more precise results for the selected dataset. But when using the model, the tree count should be tweaked properly using trial and error as higher number of trees in the model might give better results and need to make sure overfitting is not present.

As Bayesian network, even the accuracy of this model is slightly better than others. However, the precision and recall are also low which makes random forest is still the best performing model. The only specific feature of this model that others could not do is this model is able to identify the cause and effect of each feature. This unique ability of the model provides significant value when extracting/identifying features from the dataset. In this experiment, the dataset is limited, and this feature was used to identify which feature could provide predictive power. When the dataset increases, this feature can be used again to investigate each parameter more precisely. By doing that, unnecessary parameters could be removed, and accuracy, precision, and recall could be improved.

When investigating the various metrics for Neural Networks, it can be observed that the model. Neural networks need much more data to be effective than the traditional machine learning algorithms. As concluded before, due to the various limitations in the dataset, this model is severely lacking and could be improved by enriching the dataset further. Another way to improve the neural network model could be to use TensorFlow rather than using the fast-prototyping API used in this analysis namely, Keras API. TensorFlow allows for more customizations for building the model according to the needs of the problem at hand but

takes far longer time to train and finetune the model. It is recommended to build the model using Tensor flow library as part of the future work.

Reference

1. Muldner, K. and Burleson, W., 2015. Utilizing sensor data to model students' creativity in a digital environment. *Computers in Human Behavior*, 42, pp.127-137.
2. Stevens, C. and Zabelina, D., 2020. Classifying creativity: Applying machine learning techniques to divergent thinking EEG data. *NeuroImage*, 219, p.116990.
3. Shrivastava, D, CG, SA, Laha, A & Sankaranarayanan, K 2017, 'A Machine Learning Approach for Evaluating Creative Artifacts'.
4. Wei, Y "Max", Hong, J & Tellis, GJ 2021, 'EXPRESS: Machine Learning for Creativity: Using Similarity Networks to Design Better Crowdfunding Projects', Journal of Marketing, p. 2224292110054.
5. Park, K, Hong, JS & Kim, W 2020, 'A Methodology Combining Cosine Similarity with Classifier for Text Classification', Applied Artificial Intelligence, vol. 34, no. 5, pp. 396–411.
6. Wu, J, Wu, B, Pan, S, Wang, H & Cai, Z 2015, 'Locally Weighted Learning: How and When Does it Work in Bayesian Networks?', International Journal of Computational Intelligence Systems, vol. 8, no. Supplement 1, p. 63.
7. Xu, S 2018, 'Bayesian Naïve Bayes classifiers to text classification', Journal of Information Science, vol. 44, no. 1, pp. 48–59.
8. Feng, G, Guo, J, Jing, B-Y & Hao, L 2012, 'A Bayesian feature selection paradigm for text classification', Information Processing & Management, vol. 48, no. 2, pp. 283–302.
9. Ann Sung Lee & So Young Sohn 2015, 'Bayesian Network Analysis for Organizational Creativity', ISPIM Conference Proceedings, p. 1.
10. Chen, S.-H., & Chen, Y.-H. (2017),"A content-based image retrieval method based on the google cloud vision API and WordNet", *Intelligent Information and Database Systems*, 651–662,.
11. Kumar, V., & Subba, B, (2020), "A TfidfVectorizer and SVM based sentiment analysis framework for Text Data Corpus", *2020 National Conference on Communications (NCC)*, https://doi.org/10.1109/ncc48643.2020.9056085
12. Dichiu, D. and Rancea, I., 2021, "Using Machine Learning Algorithms for Author Profiling In Social Media", *Notebook for PAN at CLEF 2016*.
13. Park, K., Hong, J. S., & Kim, W, (2020), "A methodology combining cosine similarity with classifier for text classification", *Applied Artificial Intelligence*, *34*(5), 396–411, https://doi.org/10.1080/08839514.2020.1723868
14. Kayalvizhi, S., Thenmozhi, D., & Aravindan, C, (2019), "*Legal Assistance Using Word Embeddings*"
15. Mohammad, A. H., Alwada'n, T., & Al-Momani, O, (2016), "Arabic text categorization using support vector machine, Naïve Bayes and neural network", *GSTF Journal on Computing (JoC)*, *5*(1), https://doi.org/10.7603/s40601-016-0016-9
16. H. Takeuchi *et al.*, "Failing to deactivate: The association between brain activity during a working memory task and creativity," *NeuroImage*, vol. 55, no. 2, pp. 681–687, Mar. 2011, doi: 10.1016/j.neuroimage.2010.11.052.
17. D. Johnson, "POS (Part-Of-Speech) Tagging & Chunking with NLTK," *Guru99.com*, 11-Jun-2019. [Online]. Available: https://www.guru99.com/pos-tagging-chunking-nltk.html. [Accessed: 02-Oct-2021]
18.