Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 19 No. 5 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijet.v19i05.47465

PAPER

Natural Language Processing Approach to Evaluate Real-Time Flexibility of Ideas to Support Collaborative Creative Process

Ijaz Ul Haq[1], Manoli Pifarré[1](✉), Estibaliz Fraca[2]

[1]Department of Psychology, Sociology and Social Work, Avinguda Estudi General, University of Lleida, Lleida, Spain

[2]Department of Computer Science, University College of London, London, UK

manoli.pifarre@udl.cat

ABSTRACT

Natural Language Processing (NLP) has emerged as a valuable approach to assist in solving complex challenges in educational settings. This study explores NLP techniques, particularly sentence embedding models, to evaluate the flexibility dimension of divergent thinking during an open-ended collaborative creative (cocreative) process. The methodology involved a case study in which 25 secondary education students participated. The students worked in five collaborative groups to solve a real-life challenge through a cocreative process. During this study, we focus on evaluating flexibility, defined as the shift from one semantic category to another, grounded in the semantic similarity of ideas. Initially, we measured semantic similarity with eight-sentence embedding models and experts. We also conducted a correlation analysis of the experts and sentence embedding models to choose one highly correlated model. Subsequently, we evaluated the flexibility of ideas in creative techniques using experts and the high-correlated sentence embedding model. The results disclose that among the eight applied sentence embedding models to evaluate the semantic similarity of open-ended ideas, the Universal Sentence Encoder Transformer (USE-T) is highly correlated with experts. Moreover, USE-T strongly aligns with experts' evaluation to evaluate flexibility. These results will be valuable for designing and providing immediate feedback during cocreation, enabling AI-driven support to foster innovative solutions to real-world challenges.

KEYWORDS

creativity, collaboration, flexibility, divergent thinking, evaluation, automatic, AI, NLP

1 INTRODUCTION

Creativity has emerged as a 21st-century skill and a critical competence for professional and personal skills, now reflected in educational policies and curricula [1]. Creativity is a multifaceted concept [2], requiring iterative and improvisational creative processes that emerge in a social, contextual, and collaborative context in

Ul Haq, I., Pifarré, M., Fraca, E. (2024). Natural Language Processing Approach to Evaluate Real-Time Flexibility of Ideas to Support Collaborative Creative Process. *International Journal of Emerging Technologies in Learning (iJET)*, 19(5), pp. 93–107. https://doi.org/10.3991/ijet.v19i05.47465

Article submitted 2023-12-19. Revision uploaded 2024-01-31. Final acceptance 2024-01-31.

a classroom setting. The Collaborative Creative (cocreative) processes help students to generate creative solutions to real-world challenges by promoting creative thinking, such as Divergent Thinking (DT). DT refers to the ability to find creative potential solutions to open-ended problems. Thus, educational researchers have sparked interest in exploring ways to enhance DT to promote cocreativity within human groups during the cocreative process. This exploration includes evaluating students' divergent thinking and providing real-time feedback to boost DT and improve the cocreative process.

Creative potential is often evaluated using DT tests [3]. These tests focus on three dimensions of creativity: fluency (number of ideas), originality (unique or unusual ideas), and flexibility (number of distinct conceptual categories). Among these dimensions, flexibility is strongly associated with divergent capacity-making, associating thinking with different semantic categories, and possibly producing more originality [4]. Creativity research has provided empirical links between individual creative outcomes and the flexibility of their thinking [5]. However, flexibility evaluation and its role in promoting creativity still need to be explored in comparison with the other dimensions. Therefore, the present study focuses on the flexibility dimension of creativity.

Flexibility is conceptualised as the cognitive ability to explore the conceptual categories during ideation by moving from one semantic category to another to reach a more creative solution [6]. It is one of the key mechanisms of DT's underlying creative process [7], which refers to the key executive function of creative thinking [8]. It is associated with the agility of thoughts, adaptability, and avoidance of cognitive rigidity and fixation [9]. It drives individuals to follow diverse directions, dimensions, and pathways [10], which may be more likely to produce highly creative ideas [11]. Therefore, exploring and supporting the flexibility of ideas in the cocreative process could help students reach their creative potential. In this line of argument, we pursue to automatically evaluate the flexibility of the cocreative ideas, which could provide feedback to the students about their creative performance to achieve high creative goals. To achieve the goal of automatically evaluating the flexibility of ideas cocreated in open-ended contexts, the question emerges: Which Natural Language Processing (henceforth NLP) techniques are more reliable for automatic evaluation of ideas' flexibility dimension of divergent thinking in open-ended learning scenarios?

Prior studies reported different semantic distance NLP approaches to evaluate flexibility automatically [12]. The reason for using semantic distance is twofold. Firstly, semantic distances are helpful to model the organisation of knowledge in the mind [13]. In particular, if humans associate two concepts more easily, it shows that two concepts have a lower distance from each other and vice versa [14] [15]. Secondly, semantic distances are a common representation in NLP to measure the relatedness of two texts and offer a rich toolbox of distance measures for this purpose [16]. To measure semantic distance, two extensively used frameworks are semantic networks and semantic embeddings. Semantic networks propose to organise the concepts in a graph and that the shortest path distance represents semantic relatedness [17]. Thus, semantic networks are based on graph theory, consisting of nodes representing concepts and links between them signifying relationships. In contrast, semantic embeddings assume that concepts are implicitly represented in a vector space and that their distance corresponds to their semantic relatedness [15].

Based on the previous two frameworks of semantic distance to evaluate flexibility, different computational techniques are used in the literature. Firstly, semantic network-based methods [18] using ontologies are used [15] [19] [20]. Ontologies represent a language in the structure and hierarchy of millions of words, such as Wordnet. However, ontologies have practical implications for open-ended ideas

that have limitations, including having a fixed hierarchical structure, limited contextual understanding, difficulty handling polysemy, and struggles with generalisation and capturing scientific language in open-ended ideation. Secondly, semantic embedding techniques, such as Latent Semantic Analysis (LSA) [21], are widely used. LSA limitations include struggles with polysemy and a need for explicit semantic context modelling. Later, word embedding models Word2Vec [22] and especially Global Vectors for Word Representation (GloVe) are currently superior to other techniques that have been previously developed [23] [24]. However, word embedding models cannot differentiate between a list of keywords and a sentence, which can lose meaningful information about the context and semantic meaning among constituents of sentences. Therefore, next, we explore the recent advancement of Artificial Intelligence (AI) techniques, which has provided opportunities for creativity evaluation research.

AI techniques, particularly pre-trained sentence embedding models, have the potential to evaluate the flexibility of open-ended ideas generated for the following three reasons: Firstly, the ideas generated in the open-ended cocreation are few and open-ended (no pre-existing domain-specific data is available); thus, sentence embedding is pre-trained over a large corpus of data, and then the learned knowledge is transferred to other downstream NLP tasks. Secondly, ideas are presented in a sentence structure (more complex than single-word ideas). Hence, sentence embedding models encode the whole sentences in the vector space, keeping the semantic and contextual meaning of words as constituents of the sentences. Thirdly, the nature of the task evaluates flexibility as category switching from one semantic concept to another based on semantic distance. The sentence embedding models tackle the limitations of previously used NLP techniques for measuring the semantic distance between textual ideas [25]. Despite their potential, there is a significant research gap, as sentence embedding models have not been extensively tested to evaluate the flexibility of ideas in PBL environments, particularly in cocreative contexts. Therefore, this paper expands this exploration by using sentence embedding models to evaluate the flexibility of the co-generated ideas during the complex cocreative process to find answers to the following research question: How could deep learning sentence embedding models evaluate the flexibility of ideas generated in a complex, open-ended, and cocreative ideation process?

Automatic flexibility evaluation in an open-ended cocreative process has wide-ranging pedagogical impacts by facilitating teachers to design and deliver real-time flexibility feedback in a classroom environment. Real-time flexibility feedback can promote cognitive adaptability, flexible mindsets, and the avoidance of rigidity in thinking [8] because moving to a different category might counteract getting stuck, which makes room for new ideas [26]. Moreover, it encourages diversity in topics, fosters the generation of novel ideas, and enhances problem-solving and cocreative skills among students. Integrating real-time flexibility evaluation and feedback can enhance teaching and learning through the cocreative processes within the school environment [27].

2 METHOD

2.1 Participants

This study involved 25 students from an urban secondary school in Lleida, Spain. The participants were organised into five small groups, each comprising five students. The teacher randomly assigned the groups to solve an open-ended

scientific challenge related to identifying the causes of pollution in the Segre River in Lleida, Spain. Among the participants, 60% were female and 40% were male, with an average age of 15 to 16 years. The research received ethical approval from the University's ethical committee.

2.2 Research context

This study adopts a case-study research design using a quantitative approach [28] to solve an open-ended scientific challenge about river's pollution in Lleida, Spain. Students work collaboratively for 18 hours to solve the challenge through technology-supported cocreative processes using the Viacocrea application [29]. The Viacocrea application is a prototype that offers a multi-user collaborative platform. Viacocrea designs the cocreative process by providing a graphical representation of the cocreative techniques, structuring the creative phases, and orchestrating small group cocreative problem-solving endeavours. In this study, students solved 13 creative techniques of the Viacocrea repertoire organised in six phases: 1) Starting up (creating a group and gaining inspiration from various resources); 2) Defining (defining and understanding the problem); 3) Designing (designing an action plan to solve the problem); 4) Building up (building up new knowledge, its organisation, and analysis); 5) Summing Up (highlighting the relevant solutions); 6) Communicating (sharing their creative solutions with others). Following the six phases of the cocreative process, students worked within a shared multi-user collaborative digital space in which all the small-group students were engaged in face-to-face and technology-based interactions. All small-group members can annotate in Viacocrea space and collaborate with group members within the group; however, there were no inter-group interactions.

In this paper, for evaluation of flexibility of ideas, researchers agreed and selected two creative techniques, such as the Learning Chain (Build Up phase) and Telescope (Sum Up phase), illustrated in Figure 1. The Learning Chain technique aims to go deeper into a topic by thinking about different key questions and answering them from different perspectives, as shown in Figure 1 (Left). Also, the Telescope technique aims to propose different ideas and then narrow them down to the most relevant ones in Figure 1 (Right).

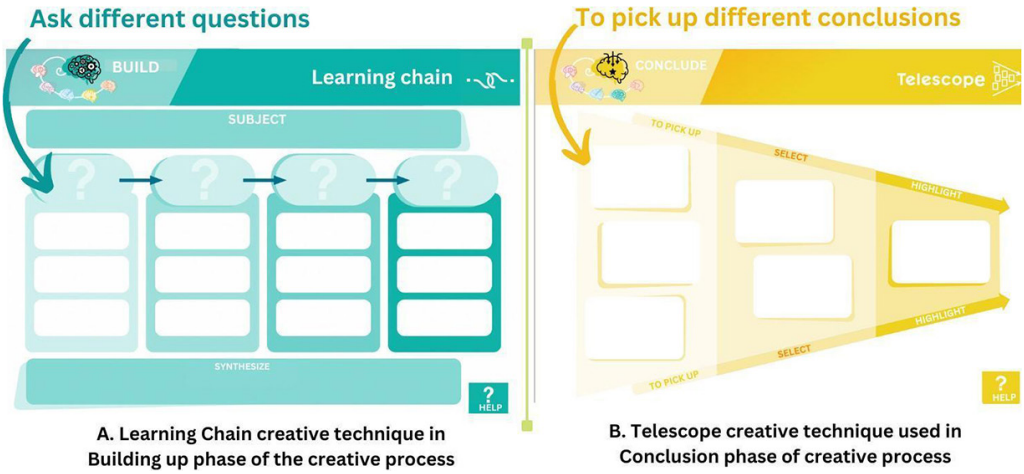

Fig. 1. Learning Chain and Telescope creative techniques used in the Cocrea application for flexibility evaluation

2.3 Study procedure

In this section, firstly, we provide insights into the collection and preparation of our dataset, which originates from the creative techniques employed during the cocreative process. Secondly, we compute ideas' semantic similarity using eight-sentence embedding models and expert evaluations. This semantic similarity computation is the basis for the subsequent flexibility evaluation. We perform this step because flexibility evaluations are based on semantic similarity. Subsequently, we identify a sentence embedding model that correlates highly with experts' scores. Thirdly, we conduct a flexibility evaluation of the cocreative ideas using the highly correlated model and experts' scores. Lastly, the data analysis methodology employed in this study is presented.

Dataset collection and preparation. As described in Research Context 2.2, students shared their cocreative ideas using creative techniques. We collected data in small groups using two creative techniques: Learning Chain and Telescope in the Building Up and Sum Up phases, respectively. For the Learning Chain technique, we focused on the ideas generated in the three sections: subject, question asked, and conclusion, highlighted in Figure 1 Left). In the subject section, eight ideas were introduced, resulting in 36 similarity comparisons after comparing each idea with the remaining seven ideas, excluding duplicates. In the questions asked section, participants shared 20 ideas, yielding 210 similarity comparisons from comparing each idea with the other 19, excluding duplicates. In the conclusion section, 11 ideas were shared, resulting in 66 similarity comparisons from comparing each idea with the other 10. Overall, the Learning Chain technique produced similarity comparisons of a total of 282 (36 + 210 + 66 = 282). For the Telescope technique, participants shared 23 ideas, so by comparing each idea with the rest of the 22 ones, excluding the duplicates, leads to 276 comparisons of ideas. Combining both techniques, we obtained 588 (282 + 276 = 588) ideas. This dataset is used for computing the semantic similarity of ideas by experts and sentence embedding models, described in Section 2.3.2.

For flexibility evaluation, we created response pairs from the list of responses of each collaborative group to assess the flexibility of thinking. Similar to previous studies [6, 30], we paired the ideas based on the design of the two creative techniques, Learning Chain and Telescope. For example, each group shared four ideas in the question-asked section of the Learning Chain. So, within the group, we compare the first idea with the second, the second with the third, and so on. Also, each group shared at least two ideas in the selected section of the Telescope, so we compared the similarity between the consecutive ideas. Flexibility evaluation is further described in Section 2.3.3.

****Text similarity with experts and sentence embedding models.** Flexibility evaluation depends on the semantic distance between the ideas. Therefore, first, we need to compute the semantic similarity of ideas with expert scores and sentence embedding models. This evaluation determines which sentence embedding model highly correlates with experts' scores. Therefore, we accomplished this through the following two actions.

Firstly, in the experts' flexibility evaluation, three experts evaluated the ideas based on semantic similarity by following two key criteria: 1) the general meaning of the ideas and 2) the use of key concepts, key topics, and details of the ideas (details can change the meaning). Experts rate each idea according to a similarity scale from 0 (completely dissimilar) to 1 (completely similar). The three experts individually score each idea based on semantic similarity with other ideas. After individual scoring, the three experts revised the ideas and discussed their scores to achieve

agreed-upon expert scores. The agreed expert scores will be used to compare correlation with sentence embedding models.

Secondly, we employ eight off-the-shelf, widely used sentence embedding models to test which model correlates highly with experts' evaluation in automatic evaluation, which include: a) The Language Model (ELMo) [31] has been selected, which has a bi-directional Long Short-Term Memory (LSTM) architecture trained on the one billion word benchmark; b) Two InferSentence embedding models [32], namely, InferSent with GloVe [33] and InferSent with FastText [34], have the architecture of Bi-directional LSTM with softmax trained on the Stanford Natural Language Inference (SNLI) dataset [35]; c) two unsupervised general-purpose universal sentence encoders [36], namely, Universal Sentence Encoder (USE) Transformer and USE Deep Averaging Network (DAN), trained on large corpora of Wikipedia, SNLI, web news, and web questions and answers, are used; and d) three Sentence-BERT were taken into account, namely, all-MiniLM-L6-v2 [37], all-mpnet-base-v2 [38], and SRoBERTa-large [39], that are fine-tuned on NLI [40].

From the eight embedding models above, the one with the highest correlation with experts' scores will be used for flexibility evaluation.

Flexibility evaluation. Flexibility evaluation involves counting the number of switches from one concept to another between two successive ideas generated during the cocreative process. Therefore, in this study, we measure the semantic similarity of ideas from 0 (completely dissimilar) to 1 (completely similar) in the Learning Chain and Telescope creative techniques. Semantic similarity identifies any change in a category switch (dissimilar or closer to dissimilar means the ideas are related to different concepts) or if there is a category stay (e.g., similar or close to similar means the ideas are related to the same concept). A similarity threshold of 0.5 is applied because it shows the 50% probability of category stay or category switch if the semantic similarity scores span from 0 to 1(0, 0.1,0.2,0.3,0.4,0.5,0.6,0.7, 0.8,0.9,1.0).

A similar methodology has also been applied in other studies related to creativity [6]. If values lie between to 0 and 0.5 or equal to 0.5 (closer to 0 means related to a different category), there is a category switch, and if the value lies between 0.5 and 1 (closer to 1 means related to a similar category), there is a category stay. Each idea was coded for category switch (1) and category stay (0) [9], and finally, the number of switches as a flexibility score.

Data Analysis. We performed the following two analyses. Firstly, we adopted Pearson and Spearman correlation analysis to measure the correlation between experts and automatic scores to examine which sentence embedding model highly correlates with experts' scores. Secondly, we use descriptive statistics to compare the flexibility score of experts and the model, which is highly correlated with experts' scores.

3 RESULTS

Flexibility evaluation relies on semantic concept similarity to assess category shifts or continuations across two consecutive responses. The flexibility evaluation involved a two-step approach with two results. Initially, we employed eight sentence embedding models alongside expert ratings to evaluate semantic similarity among ideas extracted from the two creative techniques. The correlation analysis between the embedding models and experts' scores to identify the most correlated model for flexibility computation is presented in Section 3.1. Subsequently, Section 3.2 presents

the flexibility evaluation scores of ideas generated during the Learning Chain and Telescope creative techniques.

3.1 Comparing automatic scores from sentence embedding models with experts' scores

In the automatic evaluation of flexibility, firstly, we tested eight sentence embedding models to select which embedding model correlates highly with the experts' score on the dataset used in this study. So, referring to the correlation between the sentence embedding models and experts' scores, the results are illustrated in Table 1. Our results revealed that the USE-T model displayed a substantial correlation with the expert scores (r = .860, Spearman = .784), followed by USE DAN (r = .827, and Spearman = .728), all-MiniLM (r = .839, and Spearman = .753), and all-mpnet (r = .804, and Spearman = .713). The results show that USE-T can be used for flexibility evaluation, a highly correlated model among the eight-sentence embedding models. This result aligns with those of another study that found that USE-T highly correlates with experts' score-based semantic similarity criteria [41].

Table 1. Pearson and Spearman's correlations of pre-trained models with experts' scores

Model	Experts	
	Pearson	Spearson
ELMo	.565**	.565**
InferSet-GloVe	.106*	108**
InferSet-FastText	.627**	.570**
USE-T	**.860****	**.784****
USE-DAN	.827*	.728**
SBERT		
+ SRoBERTa-NLILarge	.608**	.606**
+ *all*-MiniLM	.839**	.753**
+ *all*-MPnet	.804**	.713**

The results show that among the eight sentence embedding models to evaluate the semantic similarity of open-ended ideas, USE-T has a high correlation with the experts' scores. Therefore, we can use USE-T to evaluate the flexibility dimension of divergent thinking based on semantic similarity. Next, we present the flexibility scores based on semantic similarity using USE-T and expert evaluations.

3.2 Flexibility evaluation

We performed a flexibility evaluation using expert evaluation and USE-T, a model that strongly correlates with expert evaluations. The flexibility evaluation scores for co-generated ideas from five collaborative groups of students engaged in Learning Chain and Telescope creative techniques are presented in Table 2. The findings reveal a remarkable consistency between the flexibility scores computed by experts and those obtained using USE-T, as depicted in Table 2. However, it is worth noting a

slight discrepancy in the results for group 4, where experts identified three instances of flexibility switches, while USE-T detected two flexibility switches.

Table 2. Flexibility scores (number of category switches) calculated by experts and USE-T in two creative techniques, i.e., Learning Chain and Telescope

	Learning Chain		Telescope	
	Experts	USE-T	Experts	USE-T
Group 1	2	2	1	1
Group 2	2	2	1	1
Group 3	2	2	1	1
Group 4	3	2	1	1
Group 5	2	2	1	1

4 DISCUSSION

Our study claims that sentence embedding models can be a valuable technique to automatically evaluate idea flexibility cogenerated in open-ended learning contexts. Firstly, and in contrast to existing manual approaches in creativitiy assessment, this approach effectively addresses the challenges associated with subjective creativity evaluation of open-ended ideation, such as concerns regarding reliability (e.g., different experts can have different opinions and scoring), cost (needs several experts to do manual evaluation), and time constraints [42] [43] [44].

By contrast, sentence embedding models provide consistent and objective scoring, reduce resource expenditure without involving human resources, and automatically produce quick evaluations. These capabilities enable sentence embedding models to be integrated into technology-supported environments to evaluate and provide timely feedback on students' creativity performance. Secondly, our approach outperforms previous research in automatic creativity evaluation, which has already explored various techniques, including statistical methods (such as LSA), knowledge-based similarity (ontology-based approaches), and deep learning methods such as GloVe and other word embedding models. These computational techniques exhibit limitations, e.g., LSA is a counting-based method, knowledge-based presents ideas in a graph structure, and word embedding models are based on word embedding. Therefore, when ideas are generated in sentence structure, these techniques consider those lists of keywords and lose the semantic and contextual meaning among the words that constitute a meaningful sentence. Therefore, our study shows that sentence embedding outperforms the previously existing techniques, which are able to vectorise the whole idea (sentence) into numerical vector space, keeping the semantic and contextual meaning of words' constituents in the sentences. In sum, this capability addresses the shortcomings of previous techniques used in creativity research and sets a new standard in evaluating creativity in educational settings.

Transitioning into the specific learning context of open-ended cocreation developed in our study, we discuss three arguments underscoring that sentence embedding models emerge as highly effective solutions for addressing flexibility evaluation challenges. Firstly, the cocreative ideas are expressed in sentence structures and sentence embedding models are specialised to effectively encode the entire sentence while preserving its syntactic and semantic meanings [45]. This capability ensures

a comprehensive representation of the ideas, making the models well-suited for assessing sentence-level creativity. Secondly, the limited availability of datasets in cocreative scenarios is efficiently handled by the unsupervised nature of sentence embedding models, which excel in zero-shot tasks, allowing them to generalise to new and unseen data without extensive training [46]. This adaptability enables accurate evaluations even with small datasets. Lastly, as flexibility evaluation relies on semantic similarity tasks, sentence embedding models outperform earlier approaches for text similarity computation [47]. In conclusion, these effective solutions through sentence embedding models further solidify their role as a robust tool in evaluating open-ended cocreation within educational settings.

Furthermore, our study reveals that four sentence embedding models exhibit high correlations with expert evaluations Specifically, the USE and SBERT models, such as USE-T, USE-DAN, all-MiniLM, and all-mpnet, stand out in performance (see Table 1). In our view, several factors could contribute to these promising high correlation results with experts' scores. Firstly, these models are trained using a combination of unsupervised training on unlabeled datasets and fine-tuning on supervised datasets like the SNLI dataset. The inclusion of SNLI training is crucial for obtaining higher-quality sentences. Secondly, the nature of the datasets used for training these four models proves advantageous for our context. These models are trained on vast and diverse data sources, including Wikipedia, SNLI, web news, web questions, and the answers book corpus, which align well with our dataset. Lastly, the variation in results can be attributed to the different abilities and architectures of the pre-trained embedding models [48], essential for learning contextual representations of words and the semantic vectorisation of whole sentences. Architectures like transformers, deep averaging networks, and finSoftmax BERT on SNLI with Softmax have demonstrated superior performance in our evaluations. In conclusion, among these four models, we selected the USE-T to compute the flexibility further because it is most highly correlated with experts' scores on computing the semantic similarity of ideas.

Finally, our research reveals that the USE-T model consistently produced flexibility scores that closely matched expert evaluations in Table 2. However, there was a difference in flexibility scores on only one score. For example, experts calculated two categories while USE-T computed three, which can be observed in Group 4 of the Learning Chain technique in Table 2. This minute variation occurs with ideas having scores closer to 0.5, whereas ideas with similarity scores significantly different from 0.5 exhibit clear distinctions in similarity or dissimilarity. We argue that this variability in the results is not dependent on the choice of the similarity threshold (0.5). Even if we were to select a different threshold, we might still encounter similar minute variations in the scores. In conclusion, the high validity of USE-T's scores, and similar to the experts' scores, demonstrates that USE-T can effectively assess the flexibility of open-ended ideas in a real classroom setting, facilitating the cocreation process.

To sum up, our study has provided experimental evidence regarding using sentence embedding models, especially USE-T, to evaluate the flexibility of open-ended ideas generated during the cocreative process. The study contributes to ecological and pedagogical value in learning and teaching environments. This AI technique could be integrated into the Viacocrea technology-supported creative process and in other digitalised educational settings to provide feedback to teachers and particularly to a group of students. Flexibility feedback could suggest new categories to avoid fixating on one category, which could help improve divergent thinking. Moreover, real-time flexibility feedback during the cocreative process would facilitate, orchestrate,

and regulate students' cocreation actions, which would grow into fruitful group creative-thinking mechanisms [49].

5 CONCLUSION

This paper aims to automatically evaluate the flexibility of open-ended ideas generated while solving a real-life challenge. To meet this objective, we explore our research question: "How could deep learning sentence embedding models evaluate the flexibility of ideas generated in the context of a complex, open-ended, and cocreative ideation process?". To answer the research question, this study made mainly the following two contributions.

Firstly, we proposed using sentence embedding models for flexibility evaluation as a semantic category switch between consecutive ideas. The use of sentence embedding models tackles the computational challenges faced by previously developed techniques (e.g., semantic networks, LSA, word embedding models) for automatically evaluating the flexibility of open-ended ideas. Furthermore, our study revealed that sentence embedding models effectively evaluate the flexibility of open-ended ideas generated during the cocreative process in the real classroom context.

Secondly, we conducted a case study to validate the reliability of the sentence embedding models to evaluate the flexibility of open-ended ideas. Our results disclosed that among the eight applied sentence embedding models considered, four models, namely, USE-T, USE-DAN, all-MiniLM, and all-mpnet, produced highly correlated results with experts' scores on evaluating the semantic similarity of ideas. Furthermore, we used the highly correlated model USE-T and experts' scoring to evaluate the flexibility of ideas during the cocreative process. We found that USE-T yielded significantly similar results to the experts' scores when evaluating flexibility based on the number of category switches during cocreation. These contributions are based on psychometric evidence that provides scalable, valid, replicable, and cost-effective tools for evaluating the flexibility of open-ended ideas during the cocreative process.

Our study contributions hold significant ecological, pedagogical and practical implications in real classroom settings. The design of technology capable to support cocreative problem solving is becoming increasingly prevalent in educational settings. Our study makes strides towards AI-supported orchestration of cocreation processes by examining the possibilities that sentence embedding models offer to evaluate the flexibility of ideas cogenerated during the process of solving a complex and open-ended challenge. This contribution also situates its application in other educational technologies and e-learning platforms by incorporating creativity and AI into teaching and learning environments. Simultaneously, our study could be integrated into these platforms to facilitate AI assistance in real-time flexibility feedback on students' solutions to assist them to generate diverse solutions and improve their divergent thinking. These contributions aim to equip individuals with the competencies to generate creative solutions for complex economic, environmental, and social challenges, leading towards a more innovative and sustainable future.

5.1 Limitations and future work

In our view, our research exhibits certain limitations that offer opportunities for future enhancements. In creativity research, flexibility can be evaluated through two primary approaches: one involves category switching, which signifies a shift

from one semantic idea to another, while the other entails the identification of semantic categories or topics. Our study specifically focused on evaluating flexibility through the lens of category switching. Nevertheless, it is worth noting that flexibility can also be assessed by identifying semantic categories or topics within a broader and more open-ended cocreative process. However, this approach necessitates to predefine or limit the number of categories. Therefore, it can constrain the emergence of new categories, as students' thoughts are compartmentalised within predetermined categories, hindering the natural evolution of innovative categories.

Furthermore, our study employed eight widely used pre-trained sentence embedding models with a relatively limited dataset. However, it is important to note that, in this particular case study, we used pre-trained models and did not involve model training to learn from the large data. Nevertheless, the dataset, generated within a real-world context, retains its value in reaching our research objectives by successfully evaluating the flexibility of ideas generated in the cocreative techniques. This research aims to enhance the synergies between AI and creativity support to facilitate the development of effective teaching and learning methodologies within the cocreation context.

6 CLOSING

6.1 Acknowledgment

This research has been funded by the Ministry of Science and Innovation of the Government of Spain under Grants PDC2022-133203-I00 and PID2022-139060OB-I00.

6.2 Disclosure statement

The authors have no conflicts of interest to disclose.

6.3 Data availability

The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

6.4 Ethics approval

All authors declare that ethical standards have complied with the approval of the study by the university's Institutional Review Board.

6.5 Informed consent

All the participants have given their written informed consent.

7 REFERENCES

[1] J. A. Plucker, M. S. Meyer, S. Karami, and M. Ghahremani, "Room to run: Using technology to move creativity into the classroom," in *Creative Provocations: Speculations on the Future of Creativity, Technology & Learning*, Springer, 2023, pp. 65–80. https://doi.org/10.1007/978-3-031-14549-0_5

[2] R. K. Sawyer, "The iterative and improvisational nature of the creative process," *Journal of Creativity*, vol. 31, p. 100002, 2021. https://doi.org/10.1016/j.yjoc.2021.100002

[3] A. Leroy, M. Romero, and L. Cassone, "Interactivity and materiality matter in creativity: Educational robotics for the assessment of divergent thinking," *Interactive Learning Environments*, vol. 31, no. 4, pp. 2194–2205, 2023. https://doi.org/10.1080/10494820.2021.1875005

[4] Y. N. Kenett and M. Faust, "A semantic network cartography of the creative mind," *Trends in Cognitive Sciences*, vol. 23, no. 4, pp. 271–274, 2019. https://doi.org/10.1016/j.tics.2019.01.007

[5] S. Mastria, S. Agnoli, M. Zanon, S. Acar, M. A. Runco, and G. E. Corazza, "Clustering and switching in divergent thinking: Neurophysiological correlates underlying flexibility during idea generation," *Neuropsychologia*, vol. 158, p. 107890, 2021. https://doi.org/10.1016/j.neuropsychologia.2021.107890

[6] K. Grajzel, S. Acar, D. Dumas, P. Organisciak, and K. Berthiaume, "Measuring flexibility: A text-mining approach," *Frontiers in Psychology*, vol. 13, p. 1093343, 2023. https://doi.org/10.3389/fpsyg.2022.1093343

[7] N. Boot, M. Baas, E. Mühlfeld, C. K. de Dreu, and S. van Gaal, "Widespread neural oscillations in the delta band dissociate rule convergence from rule divergence during creative idea generation," *Neuropsychologia*, vol. 104, pp. 8–17, 2017. https://doi.org/10.1016/j.neuropsychologia.2017.07.033

[8] S. Acar, K. Berthiaume, K. Grajzel, D. Dumas, C. lemister, and P. Organisciak, "Applying automated originality scoring to the verbal form of torrance tests of creative thinking," *Gifted Child Quarterly*, vol. 67, no. 1, pp. 3–17, 2023. https://doi.org/10.1177/00169862211061874

[9] S. Acar, M. A. Runco, and U. Ogurlu, "The moderating influence of idea sequence: A re-analysis of the relationship between category switch and latency," *Personality and Individual Differences*, vol. 142, pp. 214–217, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191886918303428.

[10] S. Acar, A. M. Abdulla Alabbasi, M. A. Runco, and K. Beketayev, "Latency as a predictor of originality in divergent thinking," *Thinking Skills and Creativity*, vol. 33, p. 100574, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1871187119300422.

[11] W. Zhang, Z. Sjoerds, and B. Hommel, "Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking," *NeuroImage*, vol. 210, p. 116572, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920300598.

[12] J. D. Patterson *et al.*, "Multilingual semantic distance: Automatic verbal creativity assessment in many languages," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 17, no. 4, p. 495, 2023. https://doi.org/10.1037/aca0000618

[13] A. Kovalkov, B. Paaßen, A. Segal, N. Pinkwart, and K. Gal, "Automatic creativity measurement in scratch programs across modalities," *IEEE Transactions on Learning Technologies*, vol. 14, no. 6, pp. 740–753, 2021. https://doi.org/10.1109/TLT.2022.3144442

[14] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3668–3678. https://doi.org/10.1109/CVPR.2015.7298990

[15] N. Kenett, "What can quantitative measures of semantic distance tell us about creativity?" *Current Opinion in Behavioral Sciences*, vol. 27, pp. 11–16, 2019. https://doi.org/10.1016/j.cobeha.2018.08.010

[16] N.-J. Akpinar, A. Ramdas, and U. Acar, "Analyzing student strategies in blended courses using clickstream data," *arXiv preprint arXiv:2006*, p. 00421, 2020. https://doi.org/10.48550/arXiv.2006.00421

[17] G. V. Georgiev and D. D. Georgiev, "Enhancing user creativity: Semantic measures for idea generation," *Knowledge-Based Systems*, vol. 151, pp. 1–15, 2018. https://doi.org/10.1016/j.knosys.2018.03.016

[18] S. Acar and M. A. Runco, "Assessing associative distance among ideas elicited by tests of divergent thinking," *Creativity Research Journal*, vol. 26, no. 2, pp. 229–238, 2014. https://doi.org/10.1080/10400419.2014.901095

[19] K. Beketayev and M. A. Runco, "Scoring divergent thinking tests by computer with a semantics-based algorithm," *Europe's Journal of Psychology*, vol. 12, no. 2, p. 210, 2016. https://doi.org/10.5964/ejop.v12i2.1127

[20] A. L. Cosgrove, Y. N. Kenett, R. E. Beaty, and M. T. Diaz, "Quantifying flexibility in thought: The resiliency of semantic networks differs across the lifespan," *Cognition*, vol. 211, p. 104631, 2021. [Online]. Available: https://doi.org/10.1016/j.cognition.2021.104631.

[21] K. Dunbar and E. Forster, "Creativity evaluation through latent semantic analysis," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2009, vol. 31, no. 31.

[22] Y.-T. Sung, H.-H. Cheng, H.-C. Tseng, K.-E. Chang, and S.-Y. Lin, "Construction and validation of a computerized creativity assessment tool with automated scoring based on deep-learning techniques." *Psychology of Aesthetics, Creativity, and the Arts*, 2022. https://doi.org/10.1037/aca0000450

[23] D. Dumas, P. Organisciak, and M. Doherty, "Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 15, no. 4, p. 645, 2021. https://doi.org/10.1037/aca0000319

[24] R. E. Beaty and D. R. Johnson, "Automating creativity assessment with semdis: An open platform for computing semantic distance," *Behavior Research Methods*, vol. 53, no. 2, pp. 757–780, 2021. https://doi.org/10.3758/s13428-020-01453-w

[25] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," *ACM Computing Surveys*, vol. 54, no. 2, 2021. [Online]. Available: https://doi.org/10.1145/3440755.

[26] Y. Wu and W. Koutstaal, "Creative flexibility and creative persistence: Evaluating the effects of instructed vs autonomous choices to shift vs, dwell on divergent and convergent thinking," *Consciousness and Cognition*, vol. 105, p. 103417, 2022. https://doi.org/10.1016/j.concog.2022.103417

[27] L. Kaminskiene, V. Žydžiunaite, V. Jurgile, and T. Ponomarenko, "Co-creation of learning: A concept analysis," *European Journal of Contemporary Education*, vol. 9, no. 2, pp. 337–349, 2020. https://doi.org/10.13187/ejced.2020.2.337

[28] G. Thomas, *How to Do Your Case Study*. New York, NY: Sage Publications Ltd., pp. 1–320, 2021.

[29] M. Pifarré, "Designing, implementing and evaluating a co-creative support technology," in *EDULEARN23 Proceedings*, IATED, 2023, pp. 4364–4367. https://doi.org/10.21125/edulearn.2023.1146

[30] D. R. Johnson, A. S. Cuthbert, and M. E. Tynan, "The neglect of idea diversity in creative idea generation and evaluation." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 15, no. 1, p. 125, 2021. https://doi.org/10.1037/aca0000235

[31] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, 2018, vol. 1, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202

[32] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," arXiv preprint arXiv:1802.05365v2, 2018. https://doi.org/10.48550/arXiv.1802.05365

[33] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. https://doi.org/10.3115/v1/D14-1162

[34] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *TACL*, vol. 5, 2017. https://doi.org/10.1162/tacl_a_00051

[35] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015. https://doi.org/10.48550/arXiv.1508.05326

[36] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018. https://doi.org/10.48550/arXiv.1803.11175

[37] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019. https://doi.org/10.48550/arXiv.1908.10084

[38] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020. https://doi.org/10.48550/arXiv.2002.10957

[39] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020. https://doi.org/10.48550/arXiv.2004.09297

[40] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017. https://doi.org/10.48550/arXiv.1705.02364

[41] J. B. Kenworthy, S. Doboli, O. Alsayed, R. Choudhary, A. Jaed, A. A. Minai, and P. B. Paulus, "Toward the development of a computer-assisted, real-time assessment of ideational dynamics in collaborative creative groups," *Creativity Research Journal*, pp. 1–16, 2023. https://doi.org/10.1080/10400419.2022.2157589

[42] S. Said-Metwaly, W. V. den Noortgate, and E. Kyndt, "Approaches to measuring creativity: A systematic literature review," *Creativity. Theories – Research – Applications*, vol. 4, no. 2, pp. 238–275, 2017. https://doi.org/10.1515/ctra-2017-0013

[43] C. Olivares Rodrıguez, M. Guenaga, and P. Garaizar, "Automatic assessment of creativity in heuristic problem-solving based on query diversity, *Dyna*, vol. 92, pp. 449–455, 2017. https://doi.org/10.6036/8243

[44] S. Doboli, J. Kenworthy, P. Paulus, A. Minai, and A. Doboli, "A cognitive inspired method for assessing novelty of short-text ideas," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206788

[45] A. Kalinowski and Y. An, "Exploring sentence embedding structures for semantic relation extraction," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–7. https://doi.org/10.1109/IJCNN52387.2021.9534215

[46] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. https://doi.org/10.48550/arXiv.2005.14165

[47] O. Shahmirzadi, A. Lugowski, and K. Younge, "Text similarity in vector space models: A comparative study," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, 2019, pp. 659–666. https://doi.org/10.48550/arXiv.1810.00664

[48] Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, and J. Jiang, "Federated learning from pre-trained models: A contrastive learning approach," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 332–19 344, 2022. https://doi.org/10.48550/arXiv.2209.10083

[49] M. Pifarre, "Using interactive technologies to promote a dialogic space for creating collaboratively: A study in secondary education," *Thinking Skills and Creativity*, vol. 32, pp. 1–16, 2019. https://doi.org/10.1016/j.tsc.2019.01.004

8 AUTHORS

Ijaz Ul Haq, Department of Psychology, Sociology and Social Work, Avinguda Estudi General, 4, University of Lleida, 25001, Lleida, Spain.

Manoli Pifarré, Department of Psychology, Sociology and Social Work, Avinguda Estudi General, 4, University of Lleida, 25001, Lleida, Spain (E-mail: manoli.pifarre@udl.cat).

Estibaliz Fraca, Department of Computer Science, University College of London, London, UK.