applied sciences

MDPI

Article

Group Assignments for Project-Based Learning Using Natural Language Processing—A Feasibility Study

Woori Kim [1] and Yongseok Yoo [2,*]

[1] Department of Special Education, Chonnam National University, Gwangju 61186, Korea; rnell777@jnu.ac.kr
[2] Department of Electronics Engineering, Incheon National University, Incheon 22012, Korea
* Correspondence: yyoo@inu.ac.kr; Tel.: +82-32-835-8453

Abstract: Group learning is commonly used in a wide range of classes. However, effective methods used to form groups are not thoroughly understood. In this study, we explore a quantitative method for creating project teams based on student knowledge and interests expressed in project proposals. The proposals are encoded to vector representations, ensuring that closely related proposals yield similar vectors. During this step, two widely used natural language processing algorithms are used. The first algorithm is based solely on the frequency of words used in the text, while the other considers context information using a deep neural network. The similarity scores for the proposals generated by the two algorithms are compared with those generated by human evaluators. The proposed method was applied to a group of senior students in a capstone design course in South Korea based on their project proposals on autonomous cars written in Korean. The results indicate that the contextualized encoding scheme produces more human-like text similarity vectors compared to the word frequency-based encoding scheme. This discrepancy is discussed from a context information standpoint in this study.

Keywords: group learning; project-based learning; group assignment; natural language processing; writing

Citation: Kim, W.; Yoo, Y. Group Assignments for Project-Based Learning Using Natural Language Processing—A Feasibility Study. *Appl. Sci.* **2022**, *12*, 6321. https://doi.org/10.3390/app12136321

Academic Editors: Adegboyega Ojo and Rizun Nina

Received: 19 April 2022
Accepted: 20 June 2022
Published: 21 June 2022

1. Introduction

1.1. Group Learning in Project-Based Courses

Group learning has become increasingly important in higher education. Courses in colleges and universities involve diverse group activities ranging from small assignments due in a few weeks to more comprehensive projects due at the end of the semester [1]. Working as a group, students have opportunities to learn from each other and receive practical peer support [2]. Group learning also provides an opportunity to practice decision-making and responsibility, which are crucial for advancing in careers after graduation in competitive environments [3].

Therefore, developing group-learning strategies has been an active research topic. Fiechtner and Davis [4] pioneered a scheme to identify group learning strategies by conducting surveys with college students to assess their perception of effective group activities. They concluded that the manner in which groups are formed and group size significantly affect learning outcomes. Specifically, groups should be formed by instructors rather than students. The group size should not be too small (less than four) or too large (larger than seven). Baer [5] further analyzed the composition of groups and its effects on cooperative learning in undergraduate courses. The results demonstrated that homogeneously grouped students significantly outperformed students grouped otherwise, which is more applicable to high- or average-achieving students. Monson [6] confirmed these findings by determining that higher group achievement led to higher individual learning gains, which was not correlated with the gender or race composition of groups.

In this study, we explore group forming strategies considering project-based learning (PBL) in Science, Technology, Engineering, and Mathematics (STEM) courses. STEM courses typically involve group activities that allow students to consolidate obtained knowledge into practical skills [7,8]. In many universities, students majoring in STEM fields are required to complete capstone design courses, where senior students work as a team to solve high-level complex problems using all the knowledge and skills acquired throughout their courses before graduation [9,10].

Despite the importance of such courses, effective methods for forming groups in capstone courses have not been investigated extensively. Currently, in most capstone courses, student choice is the most common method of designating students to teams [10], which is ineffective according to studies on group formation [4–6]. It would therefore be desirable for the instructor to form groups based on students' interests and preferences [11–13]. However, this is another challenge for the instructor because students' high-level cognitive skills are difficult to be measured and interpreted [14–16].

1.2. Contributions of the Study

Thus, we investigate a novel approach to form groups in PBL courses utilizing quantitative measures of students' backgrounds and interests. Instead of letting the students form their own groups, the students are asked to write project proposals. The written proposals demonstrate students' knowledge of the subject and highlight higher-order cognitive skills [17–19]. Thus, we use natural language processing (NLP) techniques to cluster similar proposals into groups.

Specifically, the proposed method consists of two steps. First, each text proposal is encoded into a numerical representation that captures the core semantics and context of the text. For this encoding step, we consider two NLP algorithms: a basic measure based on the frequencies of words [20] and a deep-learning-based contextual embedding scheme pretrained on large corpora [21,22]. Second, these numerical representations of the proposals are used to cluster closely related proposals [23].

The research questions addressed in this study are as follows.

1. What are the differences between the encodings using the two NLP algorithms?
2. What are the differences between the NLP algorithms and humans in comparing text proposals?
3. Which NLP algorithm is more effective in clustering text proposals?

The remainder of this paper is organized as follows. Section 2 describes data collection, NLP-based team assignments, and human validation. In Section 3, we compare the NLP-based representation schemes with human evaluation scores and present the grouping results. Then, we discuss the results and their implications in Section 4 and the limitations of this study in Section 5. In Section 6, we draw general conclusions with future research directions.

2. Materials and Methods

2.1. Data Collection

Project proposals were collected in a capstone design course for senior students in South Korea as follows. The project topic (autonomous cars) and toolkit (Nvidia JetBot) were announced in the course syllabus before the beginning of the semester. Seventeen senior students registered for the course, comprising 14 males and three females, whose ages ranged from 23 to 26, with an average of 24.5. During the first week of the course, two introductory lectures on the project topic and toolkit were provided to the students. During the second week, students were asked to write project proposals individually in their native language (Korean) and were informed that groups will be formed based on the background and interests described in their proposals. They were given one week to write the project proposals. During the third week, 17 term project proposals were collected from the participants (one proposal from each).

In terms of text lengths, the proposals were roughly divided into two groups with relatively shorter and longer samples. The average word count in a proposal was 121, with a considerably large standard deviation of 83. Figure 1 provides a histogram of the word counts. This histogram highlights a bi-modal distribution, corresponding to nine shorter and eight longer texts. The average number of words in the eight longer proposals was 200, with a standard deviation of 35. This group of eight proposals was selected as the evaluation set and was used in the human evaluation, details of which are described in Section 2.3.

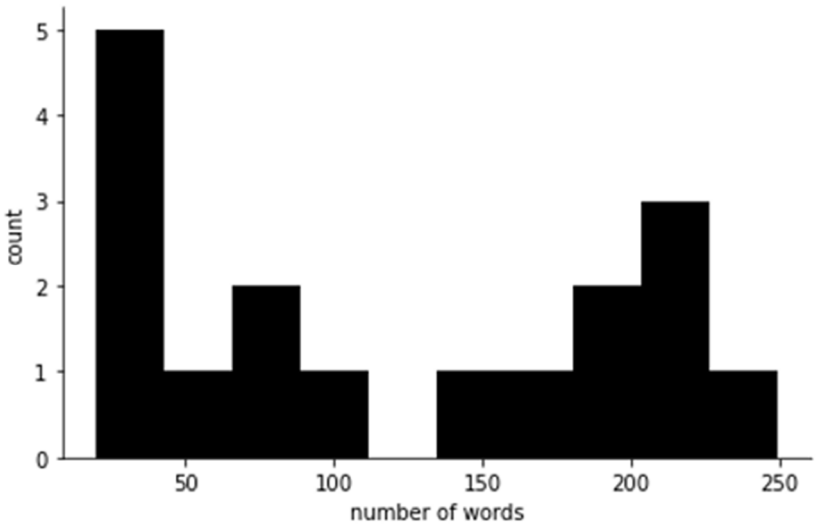

Figure 1. Histogram of proposal word counts.

2.2. NLP-Based Team Assignment

The collected text proposals were clustered into similar ones in two steps, shown as a schematic in Figure 2. First, each proposal was encoded to a numerical representation in a vector space. For this encoding step, two widely used NLP algorithms were used, and corresponding numerical representations were compared for the same set of text proposals. Second, similarity was measured for each pair of numerical representations, and this similarity value was used to group similar proposals into a group. Distributions of the similarity measures using the two NLP algorithms and clustering results were compared.

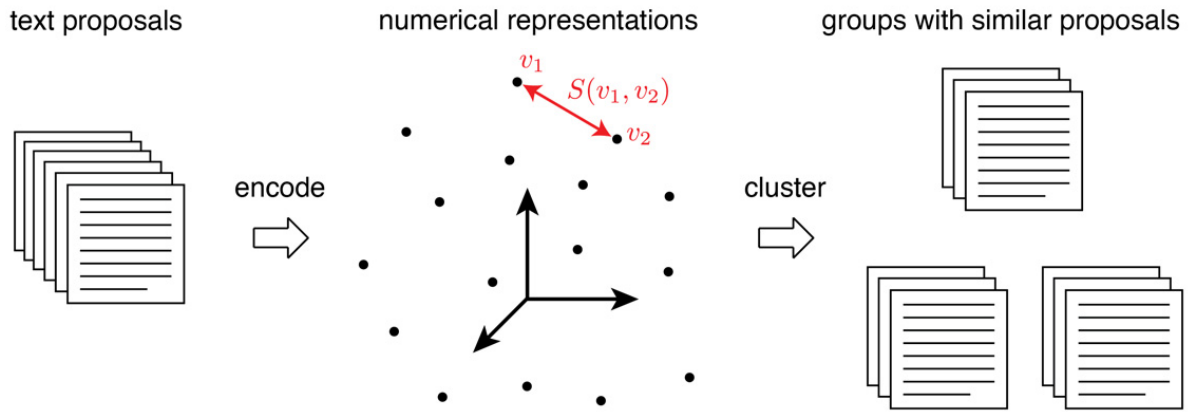

Figure 2. Schematic of the proposed method.

The first step is to encode each proposal into a numerical representation using either term frequency–inverse document frequency (TF-IDF) [20] or universal sentence embedding (USE) [21,22]. For TF-IDF, each text was normalized, with nouns extracted using the open-source Korean text processor [24]. For processing Korean texts, it is a customary

preprocess step to use normalized nouns for efficiency and robustness [25,26]. The number of unique nouns was 478. Thus, each text was represented by a 478-dimensional vector, where each element corresponded to the occurrence of each word. Specifically, TF-IDF signifies the product of the term frequency (TF) and inverse document frequency (IDF), defined as follows:

$$\text{TF-IDF}(\text{t}, \text{d}) = \text{TF}(\text{t}, \text{d})\ \text{IDF}(\text{t})\text{IDF}(\text{t}) = \log \frac{1+\text{n}}{1+\text{df}(\text{t})} + 1 \quad (1)$$

Here, $\text{TF}(\text{t}, \text{d})$ is the term frequency of the given term (word) t in the document (text) d, n is the total number of texts, and $\text{df}(\text{t})$ is the number of documents (texts) that contain term (word) t. The resulting TF-IDF vectors were normalized using the Euclidean norm.

With the USE representation, each text was encoded to a 512-dimensional vector using a deep neural network as follows. First, input texts were tokenized using SentencePiece [27]. Then, the encoder component of the transformer architecture [28] with bi-directional self-attention produced context-aware token representations. A pretrained model for this encoder was downloaded from the TF HUB [29], trained using the Stanford Natural Language Inference (SNLI) corpus [30], with question–answer and translation pairs mined from the Internet [21,22]. Finally, these token representations were averaged to obtain a text-level representation for each text.

The key difference between TF-IDF and USE is the use of the context information. TF-IDF is solely based on the frequency of each word. In contrast, the deep neural network used for USE is trained to take neighboring words into account. This difference would lead to qualitatively different numerical representations and grouping results.

Using the vector representations obtained via either TF-IDF or USE, the similarity between a given text pair was calculated using the cosine similarity of their corresponding vector representations:

$$S(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1|\ |v_2|} \quad (2)$$

where $\cdot$ denotes the inner product of two vector representations (v_1 and v_2) and $|\ |$ represents the Euclidean norm. The similarity measures based on TF-IDF and USE are denoted as $S_{\text{TF-IDF}}$ and S_{USE}, respectively.

Based on these similarity measures, texts were clustered into groups through agglomerative clustering [23]. Initially, each text forms a cluster; then, similar clusters are recursively merged.

2.3. Human Evaluation

The NLP-based similarity measures were compared with those obtained based on human evaluations. In the evaluation set, eight texts corresponded to 28 ($={}_8\text{C}_2$) possible pairs. For simplicity, 14 pairs were randomly selected among the 28 pairs for the human evaluation. Three human evaluators (research assistants who completed prerequisite courses) were recruited. Each evaluator received two sessions of training on the topic and scoring criteria. For the evaluation, a web-based evaluation tool was used, where each screen presents two texts in a random order followed by a question on the relevance of each pair on a scale of 1 to 5. Then, the average scores were compared with the NLP-based similarity measures. The similarity scores provided by the human evaluators are denoted as S_{human}.

3. Results

3.1. Similarity Measures Obtained Using TF-IDF vs. USE

The $S_{\text{TF-IDF}}$ and S_{USE} values were modestly correlated. Figure 3A,B show the heatmaps of $S_{\text{TF-IDF}}$ (A) and S_{USE} (B) values for all text pairs, respectively. Figure 3C shows the scatter plots of $S_{\text{TF-IDF}}$ and S_{USE} with the identify line (dashed) presented as a reference. Text pairs with smaller (larger) $S_{\text{TF-IDF}}$ values tended to correspond to smaller (larger)

S_{USE} values. The Pearson correlation coefficient between S_{TF-IDF} and S_{USE} was 0.54. Texts 7 and 8 produced the highest similarity values for both S_{TF-IDF} and S_{USE}.

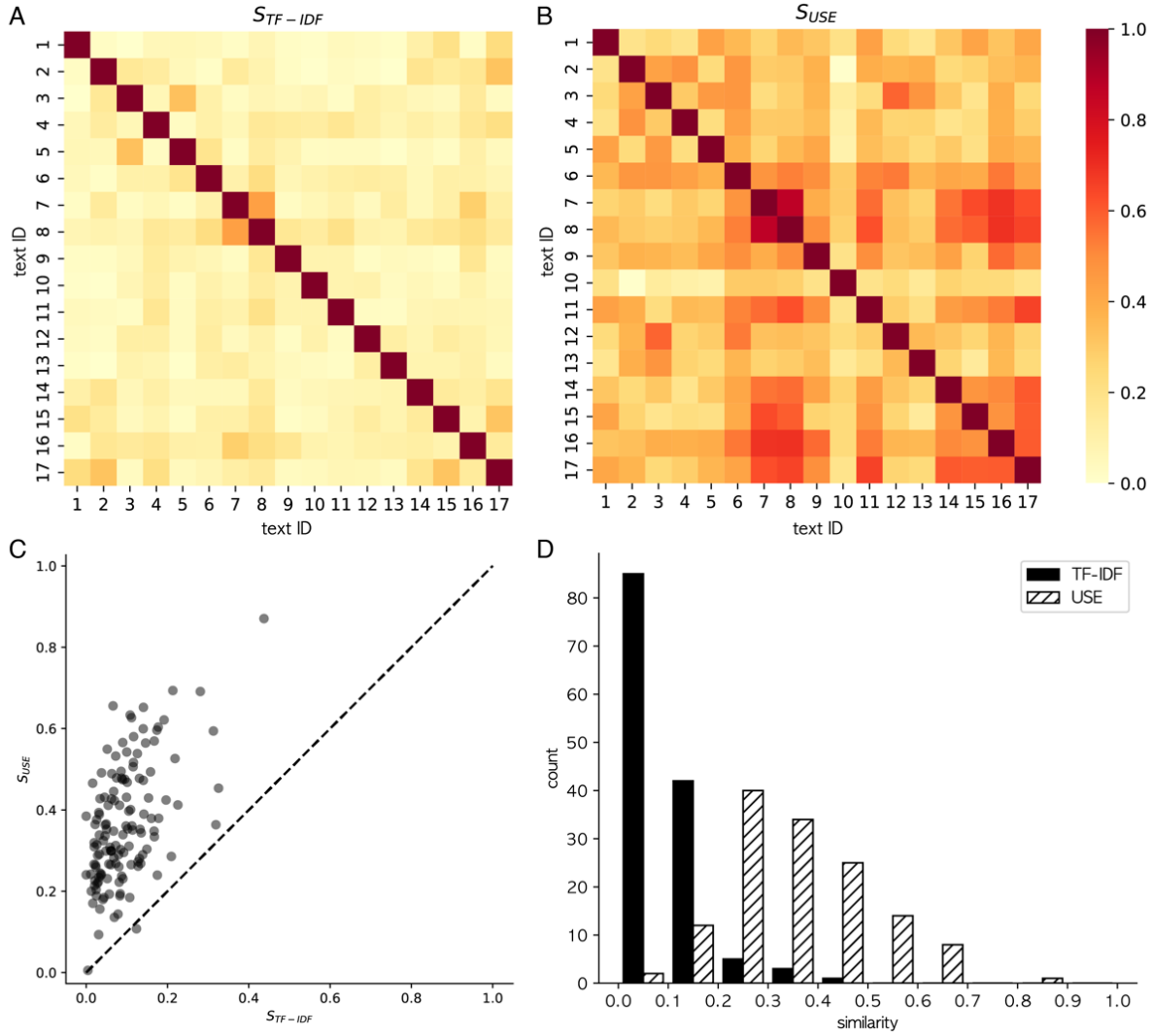

Figure 3. Comparison of the similarity measures based on word frequency (S_{TF-IDF}) and contextualized embedding (S_{USE}). In (**A**,**B**), the similarity values of S_{TF-IDF} and S_{USE} are shown, respectively, for each pair of proposal texts. In (**C**), the similarity values of S_{TF-IDF} and S_{USE} are compared on a scatter plot. In (**D**), the histograms of S_{TF-IDF} (solid) and S_{USE} (dashed) are compared.

However, S_{TF-IDF} values were smaller (mean: 0.09; standard deviation: 0.07) than S_{USE} values (mean: 0.36, standard deviation: 0.14), which was a statistically significant difference ($p < 10^{-9}$, paired t-test).

USE was more informative than TF-IDF for comparing text proposals. As shown in Figure 3D, the histogram of S_{TF-IDF} was skewed toward zero, with a median of 0.08, indicating that most of the S_{TF-IDF} values were near zero, as highlighted by the small off-diagonal values present in Figure 3A. In contrast, the histogram of S_{USE} was more symmetrical (Figure 3D), with a median of 0.35, close to its mean (0.36). Thus, this wider range of S_{USE} allows one to distinguish semantically different texts from others.

3.2. NLP vs. Human Evaluation

Figure 4A shows the histogram of the similarity scores calculated by the human evaluators (S_{human}). The values of S_{human} were centered around the mean of 3.17 and covered most of the maximum range from 1 to 5 with a standard deviation of 0.83.

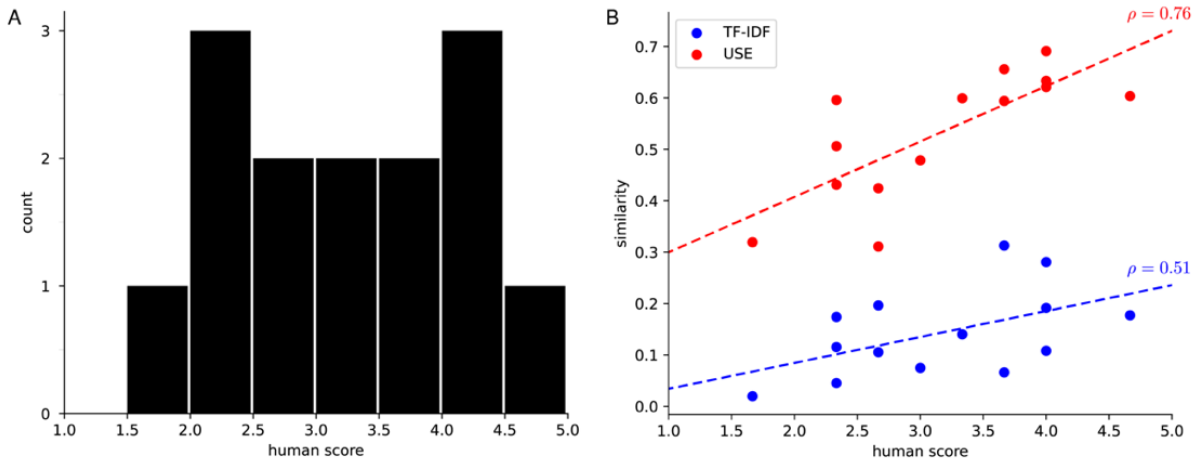

Figure 4. Histogram of similarity scores calculated by the human evaluators (**A**) and their comparison to NLP-based similarity measures (**B**).

The human scores were positively correlated with the scores obtained using the NLP algorithms. Figure 4B shows $S_{\mathrm{TF-IDF}}$ and S_{USE} as functions of S_{human}. Both $S_{\mathrm{TF-IDF}}$ and S_{USE} tended to increase as S_{human} increased. However, S_{USE} was more strongly correlated with S_{human} than $S_{\mathrm{TF-IDF}}$. The Pearson correlation coefficient between S_{USE} and S_{human} was 0.76, whereas it was 0.51 between $S_{\mathrm{TF-IDF}}$ and S_{human}.

3.3. Clustering Results

Figure 5 shows the dendrograms of the hierarchical clustering results based on TF-IDF (A) and USE (B). In these dendrograms, the horizontal lines denote cluster merging and vertical lines signify the trace of recursive merging, starting from individual samples at the bottom.

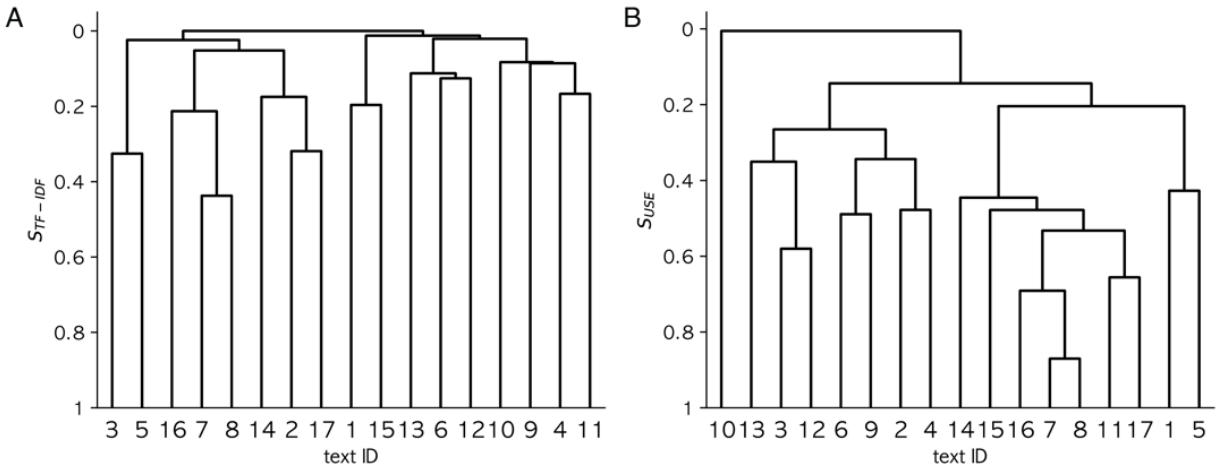

Figure 5. Dendrograms of clustering results using TF-IDF (**A**) and USE (**B**).

TF-IDF and USE produced considerably distinct clustering results, except for a few texts. Texts 7 and 8 were the most similar in terms of TF-IDF and USE and were clustered to the same group by the first merge. This group comprising 7 and 8 was later merged with text 16. After this point, the two clustering results deviate from each other, resulting in very different groups.

4. Discussion

TF-IDF and USE produced noticeably different numerical representations of the text proposals. $S_{\mathrm{TF-IDF}}$ and S_{USE} were modestly correlated with a Pearson correlation coefficient

of 0.54. This relatively low correlation indicates that the two NLP-based quantities capture different aspects of the texts. Considering the human evaluation results, S_{USE} is more correlated with S_{human} than $S_{\mathrm{TF-IDF}}$ is. This difference is discussed in more detail as follows.

TF-IDF produced low similarity values for most of the text pairs. This was surprising because the same project topic (JetBot-based autonomous cars) for the proposals was given to the students. The fixed topic may have decreased the diversity of vocabularies and produced higher $S_{\mathrm{TF-IDF}}$ values because TF-IDF is only based on the frequencies of words. In contrast, the TF-IDF-based measure regarded most of the proposals as distinct ($S_{\mathrm{TF-IDF}} \approx 0$). Thus, even for the same topic, the TF-IDF-based similarity measure could distinguish differences in the proposals. However, $S_{\mathrm{TF-IDF}}$ is limited as a similarity measure because it produced a limited range of values and lacked information regarding the level of similarity for post processing, such as clustering.

In contrast, USE produced a wider range of similarity values than those of TF-IDF on the same dataset (Figure 3D). This higher sensitivity of S_{USE} originates from the *contextualized* embedding in the transformer encoder structure. For instance, the same word in different proposals may coincide with different motivations or approaches. These semantic differences are undetectable by TF-IDF, which counts only the occurrences of the word. However, the contextual meaning of a word and its relationships with other words in a sentence can be captured by USE. This additional information can be beneficial during clustering (Figure 5).

USE results were closer to the human evaluations than TF-IDF in comparing proposal pairs. This difference may be due to humans considering the occurrence of words and their contextual meanings and structures when comparing texts [31–33]. The transformer structure adopted in USE was designed to mimic how human readers pay attention to particular words in a text. We argue that the attention-based contextualization of USE produced more human-like similarity measures.

The hierarchical clustering results are interpreted considering the existing studies on the ideal group size (4–7) as follows. Using the dendrograms presented in Figure 5, most similar texts are sequentially combined into a group until the group sizes range from four to seven. With TF-IDF (Figure 5A), it is natural to combine the most similar text pair (texts 7 and 8) into a group and add text 16 into the group to obtain a group size of three. As the group size is rather small, it would be reasonable to combine this group with another group of texts 2, 14, and 17 in the same hierarchy, resulting in a group of texts 2, 7, 8, 14, 16, and 17, sized (six) within the ideal range. Among the remaining texts, texts 3 and 5 are the most similar pairs and naturally form a group. However, at this point, it is unclear whether to combine this group of size two with the previous one of size six. The process is straightforward for the remaining texts. Combining texts 1 and 15 and merging them with texts 6, 12, and 13 forms a group of size five. The remaining texts 4, 9, 10, and 11 form another group. Similarly, the hierarchical clustering based on USE (Figure 5B) results in a group of texts 7, 8, 16, 11, 17, 15, and 14 (by the order of merging) and another group of texts 3, 12, 13, 6, 9, 2, and 4 (by the order of merging). Then, the remining texts 1, 5, and 10 form another group.

The relatively high correlation between USE and the human evaluation advocates the deep-learning-based text embedding as a similarity measure for group assignments. In general, evaluation by human experts is expensive and time-consuming and remains to be a bottleneck [34]. BLEU [35] has been widely adopted as an efficient proxy for human evaluation, which enables a dramatic improvement in machine translation [36]. This illustrates the importance of the well-established evaluation metric for advancing NLP applications. Similarly, this study demonstrates that the context-based embedding could serve as a semantic measure for proposal-based group assignments.

Furthermore, the proposed method opens the possibility of automated comparison of written proposals and their usage for group assignments. The proposed method comprises two steps—text embedding and clustering, each of which is efficient and well understood. This study shows that such a straightforward framework could substitute human judges

and automate developing NLP-based group assignments. This finding is a first step towards integrating state-of-the-art NLP approaches to class administration.

5. Limitations of the Study

A major limitation of this study is the small number of samples. In this study, we could only collect 17 text proposals from a small class. The small sample size could limit the generality of the conclusions. To evade this limitation, the proposed method has been intentionally designed to be generic. The NLP-based quantities (frequency of words and word embedding) are not fine-tuned to the presented topic or collected texts. Thus, the proposed method is applicable to more general participants focusing on different project topics.

In addition, the text similarity measure used in this study needs to be validated in future studies. In this study, the cosine similarity measure (Equation (2)) was chosen following the literature on TF-IDF and deep-learning-based word embeddings. Other similarity measures and algorithms should be investigated with more texts.

6. Conclusions and Future Research Directions

In this study, we explored the feasibility of assigning project teams by analyzing written proposals using NLP and machine learning techniques. The text proposals were encoded into numerical representations based on term frequency (TF-IDF) or contextualized embedding (USE). These numerical representations were used to cluster similar proposals.

Conclusions are drawn by answering the research questions as follows. First, TF-IDF and USE produce rather distinct numerical representations of the text proposals, resulting in different similarity values for the same set of texts. Second, the similarity measure based on the contextualized embedding (USE) was closer to human evaluation results than that based on the term frequency (TF-IDF). Third, the former was more effective for producing more fine-grained similarity values than the latter. Thus, the clustering result with USE illustrates potential benefits for forming project teams based on student knowledge and interests described in their proposals.

Our future research will focus on generalizing the proposed framework to more complex cases. To test the generality of our method, we aim to collect more data containing diverse topics and scaling up the text size. It is significant to investigate the sensitivity to presented topics or participant characteristics. Another future research direction is adapting the proposed method to educational settings. For instance, NLP-based tools can be used to aid instructors in providing feedback and guidance on term projects and encourage interactions within and between groups. We are eager to further explore NLP applications for enhancing group-learning outcomes.

Author Contributions: Conceptualization: W.K. and Y.Y.; methodology: W.K. and Y.Y.; software: Y.Y.; validation: W.K. and Y.Y.; formal analysis: W.K. and Y.Y.; investigation: W.K. and Y.Y.; resources: W.K. and Y.Y.; data curation: W.K. and Y.Y.; writing/original draft preparation: W.K. and Y.Y.; writing/review and editing: W.K. and Y.Y.; visualization: Y.Y.; supervision: Y.Y.; project administration: Y.Y.; funding acquisition: Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1G1A1011136).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Newell, C.; Bain, A. *Team-Based Collaboration in Higher Education Learning and Teaching: A Review of the Literature*; Springer: Berlin/Heidelberg, Germany, 2018.
2. Michaelsen, L.K.; Watson, W.E.; Craigin, J.; Fink, D. Team learning: A potential solution to the problem of large classes. *Organ. Behav. Teach. J.* **1982**, *7*, 21–33. [CrossRef]
3. Capelli, P.; Rogovsky, N. New work systems and skill requirements. *Int. Labour Rev.* **1994**, *133*, 205–220.
4. Fiechtner, S.B.; Davis, E.A. Why some groups fail: A survey of students' experiences with learning groups. *J. Manag. Educ.* **1984**, *9*, 58–73. [CrossRef]
5. Baer, J. Grouping and achievement in cooperative learning. *Coll. Teach.* **2003**, *51*, 169–175. [CrossRef]
6. Monson, R. Groups that work: Student achievement in group research projects and effects on individual learning. *Teach. Sociol.* **2017**, *45*, 240–251. [CrossRef]
7. Schneider, R.M.; Krajcik, J.; Marx, R.W.; Soloway, E. Performance of students in project-based science classrooms on a national measure of science achievement. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.* **2002**, *39*, 410–422. [CrossRef]
8. Krajcik, J.S.; Czerniak, C.M. *Teaching Science in Elementary and Middle School: A Project-Based Learning Approach*; Routledge: London, UK, 2018.
9. Dutson, A.J.; Todd, R.H.; Magleby, S.P.; Sorensen, C.D. A review of literature on teaching engineering design through project-oriented capstone courses. *J. Eng. Educ.* **1997**, *86*, 17–28. [CrossRef]
10. Howe, S.; Rosenbauer, L.; Poulos, S. The 2015 Capstone Design Survey Results: Current Practices and Changes over Time. *Int. J. Eng. Educ.* **2017**, *33*, 1393.
11. Pembridge, J.J.; Paretti, M.C. Characterizing capstone design teaching: A functional taxonomy. *J. Eng. Educ.* **2019**, *108*, 197–219. [CrossRef]
12. Paretti, M.C. Teaching communication in capstone design: The role of the instructor in situated learning. *J. Eng. Educ.* **2008**, *97*, 491–503. [CrossRef]
13. Ford, J.D.; Teare, S.W. The right answer is communication when capstone engineering courses drive the questions. *J. STEM Educ.* **2006**, *7*, 5–12.
14. Marin, J.A.; Armstrong, J.E., Jr.; Kays, J.L. Elements of an optimal capstone design experience. *J. Eng. Educ.* **1999**, *88*, 19–22. [CrossRef]
15. Meyer, D.G. Capstone design outcome assessment: Instruments for quantitative evaluation. In Proceedings of the Frontiers in Education 35th Annual Conference, Indianopolis, IN, USA, 19–22 October 2005; pp. F4D7–F4D11.
16. Hotaling, N.; Fasse, B.B.; Bost, L.F.; Hermann, C.D.; Forest, C.R. A quantitative analysis of the effects of a multidisciplinary engineering capstone design course. *J. Eng. Educ.* **2012**, *101*, 630–656. [CrossRef]
17. Hayes, J.; Flower, L. Identifying the organization of writing processes. In *Cognitive Processes in Writing*; Gregg, L., Steinberg, E., Eds.; Erlbaum: Hillsdale, NJ, USA, 1980; pp. 3–30.
18. Hayes, J. A new framework for understanding cognition and affect in writing. In *The Science of Writing: Theories, Methods, Individual Differences, and Applications*; Levy, M., Ransdell, S., Eds.; Erlbaum: Mahwah, NJ, USA, 1996; pp. 1–27.
19. Berninger, V.W.; Winn, W.D. Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In *Handbook of Writing Research*; MacArthur, C., Graham, S., Fitzgerald, J., Eds.; Guilford Press: New York, NY, USA, 2006; pp. 96–114.
20. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Processing Manag.* **2003**, *39*, 45–65. [CrossRef]
21. Cer, D.; Yang, Y.; Kong, S.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. Universal Sentence Encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 31 October–4 November 2018; pp. 169–174.
22. Yang, Y.; Cer, D.; Ahmad, A.; Guo, M.; Law, J.; Constant, N.; Abrego, G.H.; Yuan, S.; Tar, C.; Sung, Y.; et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5 July 2000; pp. 87–94.
23. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2009.
24. Open-Source Korean Text Processor. Available online: https://github.com/open-korean-text/open-korean-text (accessed on 12 May 2022).
25. Choi, M.; Hur, J.; Jang, M.-G. Constructing Korean lexical concept network for encyclopedia question-answering system. In Proceedings of the 30th Annual Conference of IEEE Industrial Electronics Society, Busan, Korea, 2–6 November 2004; Volume 3, pp. 3115–3119.
26. Yun, H.; Sim, G.; Seok, J. Stock Prices Prediction using the Title of Newspaper Articles with Korean Natural Language Processing. In Proceedings of the International Conference on Artificial Intelligence in Information and Communication, Okinawa, Japan, 11–12 February 2019; pp. 19–21.
27. Kudo, T.; Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 31 October–4 November 2018; pp. 66–71.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 1–11.

29. TensorFlow Hub. Available online: https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3 (accessed on 19 April 2022).
30. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 632–642.
31. Stanovich, K.E.; West, R.F. Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Mem. Cogn.* **1979**, *7*, 77–85. [CrossRef]
32. Michael, H.; Frank, K. Modeling Human Reading with Neural Attention. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 85–95.
33. Zheng, Y.; Mao, J.; Liu, Y.; Ye, Z.; Zhang, M.; Ma, S. Human behavior inspired machine reading comprehension. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 425–434.
34. Kreutzer, J.; Caswell, I.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A.; Subramani, N.; Sokolov, A.; Sikasote, C.; et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 50–72. [CrossRef]
35. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
36. Marie, B.; Fujita, A.; Rubino, R. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; Volume 1, pp. 7297–7306.