Facilitating learning outcome assessment– development of new datasets and analysis of pre-trained language models

Akriti Jindal
Lakehead University
Thunder Bay, Canada
ajindal@lakeheadu.ca

Kaylin Kainulainen
Lakehead University
Thunder Bay, Canada
kkainula@lakeheadu.ca

Andrew Fisher
Saint Mary's University
Nova Scotia, Canada
andrew.fisher@smu.ca

Vijay Mago
York University
Toronto, Canada
vmago@yorku.ca

Abstract

Student mobility reflects academic transfer from one postsecondary institution to another and facilitates students' educational goals of obtaining multiple credentials and/or advanced training in their field. This process often relies on transfer credit assessment, based on the similarity between learning outcomes, to determine what knowledge and skills were obtained at the sending institution as well as what knowledge and skills need to still be acquired at the receiving institution. As human evaluation can be both a challenging and time-consuming process, algorithms based on natural language processing can be a reliable tool for assessing transfer credit. In this article, we propose two novel datasets in the fields of Anatomy and Computer Science. Our aim is to probe the similarity between learning outcomes utilising pre-trained embedding models and compare their performance to human-annotated results. We found that ALBERT, MPNeT and DistilRoBERTa demonstrated the best ability to predict the similarity between pairs of learning outcomes. However, Davinci - a GPT-3 model which is expected to predict better results - is only able to provide a good qualitative explanation and not an accurate similarity score. The codes and datasets are available at https://github.com/JAkriti/New-Dataset-and-Performance-of-Embedding-Models.

1 Introduction

Student mobility refers to the movement - or, "transfer" – of students from one post-secondary institution (i.e., college or university) to another. Students might choose to transfer for any number of reasons; common motivating factors include the opportunity to obtain both advanced training and multiple credentials in order to increase the number of future employment options. Additionally, students whose high school grades do not allow them to enter their program or institution of choice might instead enroll first in an institution with less stringent admission requirements. Obtainment of the initial post-secondary credential (e.g., diploma) can then facilitate transfer into the desired credential (e.g., degree) (Lang and Lopes, 2014), particularly when both are within related fields (e.g., Computer Programming diploma and Computer Science degree).

Transferring within similar fields of study often means that there is overlap in topics and/or courses required for both credentials; therefore, in order to effectively recognize students' previous learning, receiving institutions are often required to assess "transfer credit." Although numerous factors might influence this assessment, learning outcomes are considered a particularly valuable tool in the process (Arnold et al., 2020a). Learning outcomes are the measurable objectives defined at the end of an assignment, class, course, or program (Davis, 2009) and indicate the skill or knowledge level that can be expected from a student who has successfully completed the task in question. When a student transfers between institutions, the receiving institution typically reviews course learning outcomes from the previous institution to determine whether they align with the learning outcomes of comparable courses offered at the receiving institution. Generally, program coordinators or other domain experts (e.g., teaching faculty) are the trusted authority designated to determine whether credit is warranted; however, human evaluation can be a complex and challenging task (Fallon, 2015).

The process of assessing transfer credit can be facilitated through the use of Natural Language Processing (NLP) based semantic similarity algorithms. NLP has wide applicability, with the main challenge of measuring textual semantic similarity (Chandrasekaran and Mago, 2021b; Majumder et al., 2016). In the past few decades, there has been rich advancement in defining various measures for similarity between words, short texts, and sentences (Corley and Mihalcea, 2005; Ramage

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 38–47
September 11–12, 2023. ©2023 Association for Computational Linguistics

et al., 2009). Word-embeddings have emerged as a well-known technique that represents text in the form of a real-valued vector that reasonably captures the syntactic and semantic resemblance between them (Turian et al., 2010; Mikolov et al., 2013). Transformer-based pre-trained language models trained on large text corpora have successfully emerged to be paradigmatic models for building vector-based representations of texts (Vaswani et al., 2017). These models have applications in numerous fields such as text summarization (Mohamed and Oussalah, 2019), question/answering (Bordes et al., 2014; Lopez-Gazpio et al., 2017), sentiment analysis (Zhao et al., 2016), and sentence prediction, among others.

In this direction, this paper aims to propose two novel datasets consisting of course learning outcomes in postsecondary education. We determined the complexity of the outcomes (sentences) through readability analysis. We also implemented various embedding models to scrutinize the similarity between pairs of sentences and compared the models' performance with human-annotated results. Among different models, we found that ALBERT, MPNET and DistilRoBERTa demonstrated the best ability to predict the similarity between pairs of learning outcomes. However, Davinci - a GPT-3 model which is expected to predict better results - is only able to provide a good qualitative explanation and not an accurate similarity score.

2 Context and Motivation

2.1 Learning Outcomes in Postsecondary Education

Learning outcomes are "clearly defined and measurable statements of learning that reflect the scope and depth of performance; what a learner is expected to know, understand and be able to demonstrate after completion of a process of learning" (Lennon et al., 2014, p. 47). Within postsecondary education, outcomes are foundational for both developing curriculum and demonstrating quality assurance (Arnold et al., 2020a; Lennon, 2015). Transfer credit assessment increasingly relies on learning outcomes as a means of evaluating similarity between courses and credentials offered by different postsecondary institutions (Arnold et al., 2020a; Fallon, 2015), with outcomes sometimes being viewed as a "currency" that students can exchange between institutions in order to avoid repeating previous learning (Young et al., 2017). Effectively assessing transfer credit is an important process when considering that the amount of credit received can correspond to increase in academic performance as well as influence academic workload and time to completion for obtaining a postsecondary credential (Gerhardt and Masakure, 2016).

2.2 The Challenge of Assessing Learning Outcomes

Learning outcomes have the potential to establish a common language for communicating student learning and achievement across contexts (Arnold et al., 2020b); however, the overall process of assessing transfer credit tends to be both resource- and time-intensive (Arnold et al., 2020a). Additionally, course comparisons can differ substantially across institutions, and might (or might not) incorporate numerous other considerations related to content, evaluation, and grading (Arnold et al., 2020a). This subjectivity can be detrimental for students and institutions alike (Tortola et al., 2020), with the lack of consistency in standards and processes presenting a notable barrier. A recommendation to address this concern is the implementation of policies and practices that facilitate consistent decision-making, for example by documenting previous assessments (Wheelahan et al., 2016). An additional consideration is the presence of common assumptions regarding the nature and quality of education offered at different types of institutions (e.g., colleges and universities) (Arnold et al., 2020b), which could influence transfer credit decisions. Again, establishing some means of consistency that eliminates such potential bias could facilitate a more accurate and effective assessment process.

3 Methodology

3.1 New Dataset Development

We developed two novel datasets consisting of learning outcomes related to two content areas, namely (1) human anatomy and (2) operating systems. To create each dataset, we first accessed relevant course outlines from postsecondary institutions in Ontario, Canada. All of the outlines were publicly available and could be accessed via the institutions' websites without log-in credentials or other permissions. Next, we extracted the learning outcomes from each course outline and organized them by field (i.e., human anatomy and operating systems), institution (e.g., Institution A, Institution B, etc.), course (e.g., ANAT 101, BIOL 102, etc.), and topic (e.g., Digestive System, Muscular System, etc.). In some instances, we modified the general sentence structure of a learning outcome to either reduce“wordiness”, delete redundant information, and/or separate information pertaining to multiple topics. For example, an outcome that included two topic areas, such as “Explain the structure and function of the muscular and skeletal systems,” would become two separate outcomes (e.g., “Explain the structure and function of the muscular system; Explain the structure and function of the skeletal system”). The resulting datasets consisted of 28 (anatomy) and 59 (operating systems) unique learning outcomes (sentences) representing the knowledge and skills that would be expected of students who successfully completed the respective courses.

To create sentence pairs for analysis, learning outcomes from each dataset were paired together so that (1) both similar and dissimilar pairs were represented uniformly (i.e., by creating both inter- and intra-topic pairings) and (2) no learning outcomes were repeated more than twice. A total of 28 and 45 sentence pairs were analyzed for the anatomy and operating systems datasets, respectively. The datasets are available at https://github.com/JAkriti/New-Dataset-and-Performance-of-Embedding-Models.

3.2 Complex Sentence Dataset (Chandrasekaran and Mago, 2021a)

Recently, a dataset comprising 52 sentence pairs related to definitions of Computer Science terminology was developed and analyzed. The authors conduct readability analysis anticipating that their dataset exhibits a low readability index. This claims that their dataset is more complex in comparison to two benchmark datasets (Sentences Involving Compositional Knowledge “SICK”(Marelli et al., 2014) and Semantic Text Similarity “STS” (Shao, 2017)). Their main objective is to show how the increase in complexity of sentences leads to a significant decrease in the performance of embedding models.

3.3 Readability Analysis

The readability score is a metric defined to measure the complexity of a sentence and deliberate the grade level of education required for a person to understand the piece of text. Depending on the complexity of learning outcomes, it is important to comprehend how reasonably the embedding models perform to evaluate the similarity scores between them. The indices used to determine the readability scores of the sentences in the proposed datasets are – a) Flesch-Kincaid Grade Level (Coleman and Liau, 1975), b) Coleman-Liau Index (Kincaid et al., 1975), c) Automated readability Index (Kincaid et al., 1975), d) Linsear Write and e) Gunning fog index (Gunning et al., 1952).

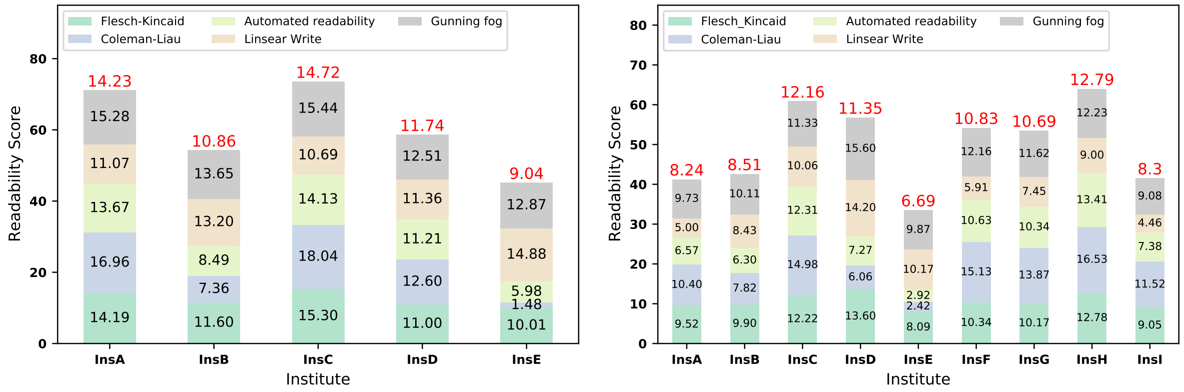
The readability scores of learning outcomes from each institute (e.g., Institute A, Institute B, etc.) are evaluated using the above indices. The aggregate of all these indices provides an overall readability score of each institute as highlighted in Figure 1 (for each dataset). For example, an average score of 11.74 shows that a reader needs a qualification of grade 11 to understand the text. Therefore, following this notation we observe that a reader requires education of collegiate level and above to understand the Anatomy sentences, and knowledge of grade 12 and above for Computer Science sentences.

3.4 Annotation

To develop a basis for comparing the performance of embedding models, the proposed datasets are each manually evaluated by three human respondents with relevant contextual expertise. The Anatomy dataset is evaluated by two graduated scholars and one graduate student in Kinesiology. The Computer Science dataset is evaluated by three thesis-based Master’s students. The annotators have been made aware of the applicability of this work. Each sentence pair is annotated on a scale of 0 to 9, where 0 (9) represents completely dissimilar (similar) sentences. To affirm the competency of these human ratings, we computed inter-rater agreement using *Krippendorff’s alpha coefficient* represented as α, where data with a coefficient value between $0.667 < \alpha < 0.8$ is considered reliable (Krippendorff, 2011; Hayes and Krippendorff, 2007). For the Anatomy dataset, $\alpha = 0.71$, and for the Computer Science dataset, $\alpha = 0.68$ which indicates that the annotation is reliable.

3.5 Web Interface

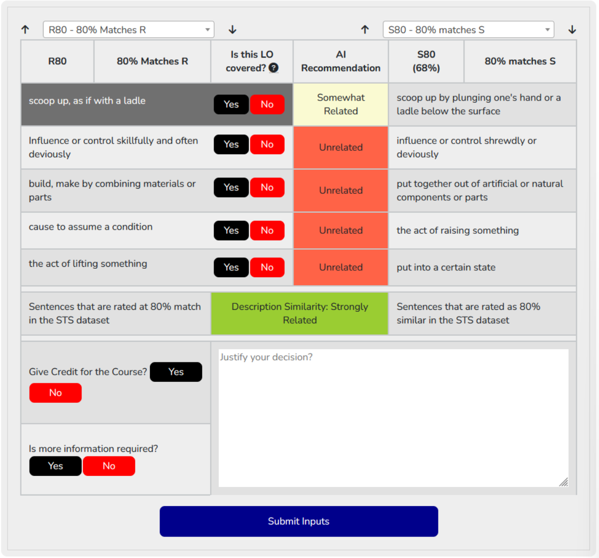
To ensure that the implementation of pre-tarined embedding models is successful in assisting transfer credit assessment, a web interface is developed to streamline the process. This begins by prompting users to upload new programs to the website



(a) Anatomy Dataset

(b) Computer Science Dataset

Figure 1: Readability analysis of learning outcomes from different institutes (denoted as InsA, InsB, and so on) for (a) Anatomy (b) Computer Science dataset using five different indices (values indicated in black). The aggregate scores are highlighted in red on each of the stacked bar graph.



↑ R80 - 80% Matches R ▾ ↓ ↑ S80 - 80% matches S ▾ ↓

R80	80% Matches R	Is this LO covered?	AI Recommendation	S80 (68%)	80% matches S
scoop up, as if with a ladle		Yes No	Somewhat Related	scoop up by plunging one's hand or a ladle below the surface	
Influence or control skillfully and often deviously		Yes No	Unrelated	influence or control shrewdly or deviously	
build, make by combining materials or parts		Yes No	Unrelated	put together out of artificial or natural components or parts	
cause to assume a condition		Yes No	Unrelated	the act of raising something	
the act of lifting something		Yes No	Unrelated	put into a certain state	

Sentences that are rated at 80% match in the STS dataset | Description Similarity: Strongly Related | Sentences that are rated as 80% similar in the STS dataset

Give Credit for the Course? Yes No

Justify your decision?

Is more information required? Yes No

Submit Inputs

Figure 2: An example of the web interface where learning outcome comparisons can be observed

that contain information about the courses and their expected learning outcomes. Once an institute identifies a program they would like to transfer credit to, a comparative analysis is performed where a natural language processing algorithm is used to determine the semantic similarity between each course.

From these results, members of the receiving institute are able to access the screen shown in Figure 2 where they can observe suggestions from the algorithm for each learning outcome comparison before making their own decisions. After each user has provided input, the owner of the analysis can then observe the overall consensus before making a final recommendation on the transfer credit and generating a report to show the outcome.

Model	Version
BERT$_{\text{base}}$ (Devlin et al., 2018)	*bert-base-nli-mean-tokens*
BERT$_{\text{Large}}$ (Devlin et al., 2018)	*bert-large-nli-mean-tokens*
RoBERTa$_{\text{base}}$ (Liu et al., 2019)	*roberta-base-nli-mean-tokens*
RoBERTa$_{\text{Large}}$ (Liu et al., 2019)	*nli-roberta-large*
ALBERT (Lan et al., 2019)	*paraphrase-albert-small-v2*
DistilRoBERTa (Sanh et al., 2019)	*all-distilroberta-v1*
DistilRoBERTa (Sanh et al., 2019)	*nli-distilroberta-base-v2*
MPNeT (Song et al., 2020)	*all-mpnet-base-v2*
GPT-3 (Brown et al., 2020)	Davinci OpenAI

Table 1: Pre-trained embedding models used to generate sentence embeddings.

3.6 NLP Algorithm

Transformer is a neural network architecture that emerged as a breakthrough in NLP (Vaswani et al., 2017). Along with the encoder-decoder structure, self-attention mechanism is the key characteristic of transformers for the algorithms to learn the long-range relationship between words in a sequence. This architecture has surpassed the performance of various traditional networks like convolutional and recurrent neural networks known for language understanding (Mikolov et al., 2011). Furthermore, *Sentence transformer* is a transformer-based model designed to generate a fixed-size dense vector for a sentence of any length (Reimers and Gurevych, 2019). A brief outline of transformer-based models along with their sentence transformer version used in this paper is given in Table 1. The resulting sentence embeddings are then compared using cosine similarity.

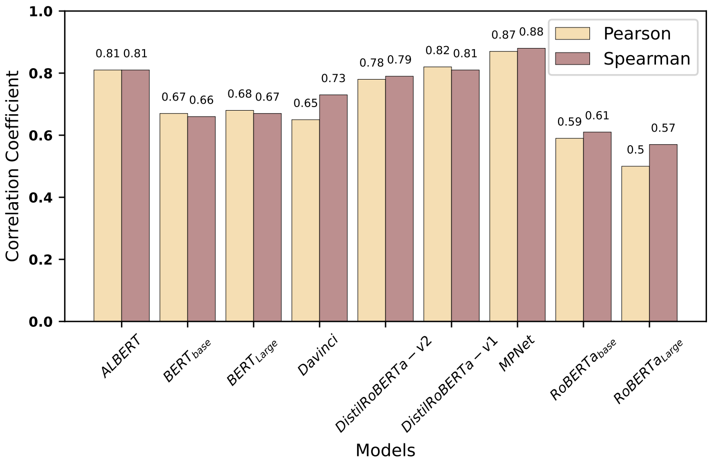


Figure 3: Pearson's and Spearman's correlation coefficient to analyse relationship between similarity values of human annotators and embedding models (Section 3.6) for Anatomy dataset.

S1: Discuss the structural organization and function of the respiratory system and its major organs.

S2: List the parts of the respiratory system and identify their functions.

ALBERT	BERT$_{base}$	BERT$_{Large}$	Davinci	DistilRoBERTa-v1	DistilRoBERTa-v2	MPNeT	RoBERTa$_{base}$	RoBERTa$_{Large}$	Human
0.8288	0.8197	0.9002	0.9098	0.7738	0.7967	0.7706	0.8901	0.8622	0.7407

Table 2: Similarity scores of a sentence pair from Anatomy dataset, evaluated using versions of pre-trained embedding models discussed in Section 3.6. Human ratings are normalized between 0 and 1.

4 Results

This section provides an extensive comparative analysis of various embedding techniques (discussed in Section 3.6) implemented for evaluating the similarity scores of learning outcomes in proposed datasets. To evaluate the relationship between the similarity scores of human annotators and embedding models we employ Pearson's and Spearman's rank coefficients for the Anatomy and Computer Science datasets (including the dataset proposed by (Chandrasekaran and Mago, 2021a) and the dataset proposed in this paper).

4.1 New Proposed datasets

4.1.1 Human Anatomy dataset

The results presented in Figure 3 demonstrate Pearson's and Spearman's correlation coefficients for the proposed Anatomy dataset of learning outcomes. While comparing different pre-trained embedding models we found that MPNeT achieves the best performance with 0.87 Pearson's value and 0.88 Spearman's value. This shows that the scores of MPNeT are very highly correlated to the human annotator's ratings. Moreover, DistilRoBERTa-v1 and ALBERT models show good performance with nearly 0.81 for both Pearson's and Spearman's correlation predicting that the ratings are highly correlated. However, every other model fails to predict a better similarity score. An example of sentence pair specifying how similar the sentences are according to embedding models is highlighted in Table 2, where we clearly observe that MPNeT is closest to human rating.

Example 1- ChatGPT

S1: Apply the basic planes of motion and fundamental movements associated with human movement using proper anatomical terminology.
S2: Describe the structure and function of bones and bony prominence.

There is not a strong similarity between the two sentences. S1 is about applying knowledge of human movement and anatomy, while S2 is about describing the structure and function of bones. They both involve anatomy, but they are not closely related in terms of content.

Davinci: 0.8201 | Human score: 0.6667

On the other hand, where Davinci model is anticipated to provide better results, we observed that though this model is not able to predict a reliable similarity score, it provides a piece of genuine qualitative information as shown above. This sentence pair is expected not to be similar as they are related to two different aspects of human anatomy. However, Davinci provides a high similarity score 0.8201 which is far away from the human similarity score, although it reflects a good explanation when explored on ChatGPT.

4.1.2 Computer Science dataset

For the Computer Science dataset proposed in this paper, the performance of embedding models is displayed in Figure 4. Also, an example of sentence pair with similarity values between them from different models is given in Table 3. Clearly, among the various embedding models, DistilRoBERTa-v2 shows the highest value for Pearson's and Spearman's coefficient displaying a very good correlation between the two ratings. However, we observed that the annotated values for this dataset fail to follow the normal distribution. Therefore, it is not reasonable to certainly rely on Pearson's test to judge the correlation between human evaluation and embedding models. Furthermore, Davinci- the most capable GPT-3 model fails to determine a good correlation. However, this model provides a justifiable interpretation of the similarity between the two sentences as shown in the example below.

> Example 2- ChatGPT
>
> S1- Design and install operating system to experience its functioning.
> S2- Configure Package repositories and package management.
>
> *There is some similarity between the two sentences. Both S1 and S2 involve working with operating systems and installing or configuring different components or features. However, the specific tasks described in the sentences are not closely related. S1 involves designing and installing an operating system to experience its functioning, while S2 involves configuring package repositories and package management.*
>
> Davinci: 0.8192 | Human score: 0.0222

Here the explanation is true and acceptable, even though the Davinci similarity score is very high i.e. 0.8192 in comparison to human rating.

4.2 Complex Sentence dataset(Chandrasekaran and Mago, 2021a)

In this paper, the authors show how the embedding models do not perform well to predict the similarity values of complex sentence dataset proposed by them. We try to analyze these sentence pairs using the versions of models highlighted in Section 3.6. Notably, we found that ALBERT model executes best performance with 0.83 Pearson's coefficient and 0.77 Spearman's value as shown in Figure 5. Furthermore, DistilRoBERTa-v2, DistilRoBERTa-v1, and MPNeT have high Pearson's (Spearman's) value of 0.81(0.78), 0.81(0.78), and 0.79(0.77) respectively which outperforms the results of the models presented in (Chandrasekaran and Mago, 2021a). Table 4 shows the results of embedding models for a particular sentence pair from the dataset where ALBERT performs the best in comparison to human ratings. However, the Davinci model which is expected to predict better results is only able to provide a good explanation and not a valid score. For example, the subjective description of the similarity between a chosen sentence pair from the dataset is very well elaborated by ChatGPT, as shown below. Whereas, this model predicts a high similarity score of 0.7639 which shows the sentences have a good similarity which is not actually true as observed from the human score.

> Example 3- ChatGPT
>
> S1: A procedure for solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation.
> S2: Various methods or formats for organizing data in a computer
>
> *There is not a strong similarity between the two sentences. S1 is describing an algorithm, which is a specific type of procedure used to solve a mathematical problem. S2 is describing ways of organizing data in a computer. While both sentences involve some type of procedure or method, they are not closely related in terms of content.*
>
> Davinci: 0.7639 | Human score: 0.1200

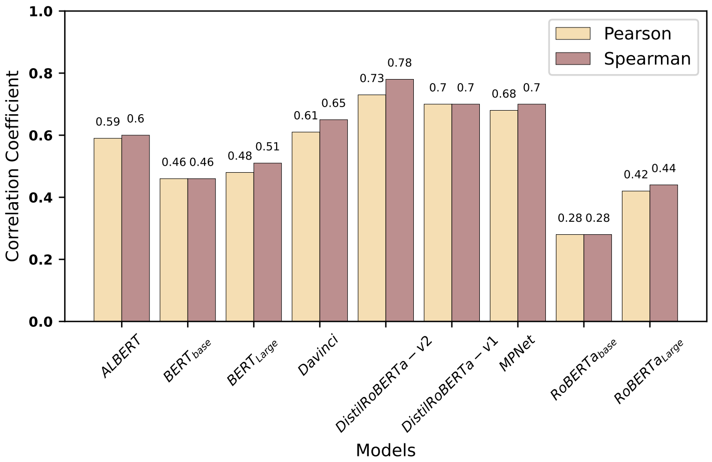


Figure 4: Pearson's and Spearman's correlation coefficient to analyse relationship between similarity values of human annotators and embedding models (Section 3.6) for proposed Computer Science dataset.

S1- Manage securely remote systems.

S2- Maintain a Unix workstation and set it up as a network client.

ALBERT	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{Large}}$	Davinci	DistilRoBERTa-v1	DistilRoBERTa-v2	MPNeT	$\text{RoBERTa}_{\text{base}}$	$\text{RoBERTa}_{\text{Large}}$	Human
0.2695	0.5247	0.5274	0.8145	0.4241	0.5708	0.4246	0.4882	0.4653	0.6667

Table 3: Similarity scores of a sentence pair from proposed Computer Science dataset, evaluated using versions of pre-trained embedding models discussed in Section 3.6. Human ratings are normalized between 0 and 1.

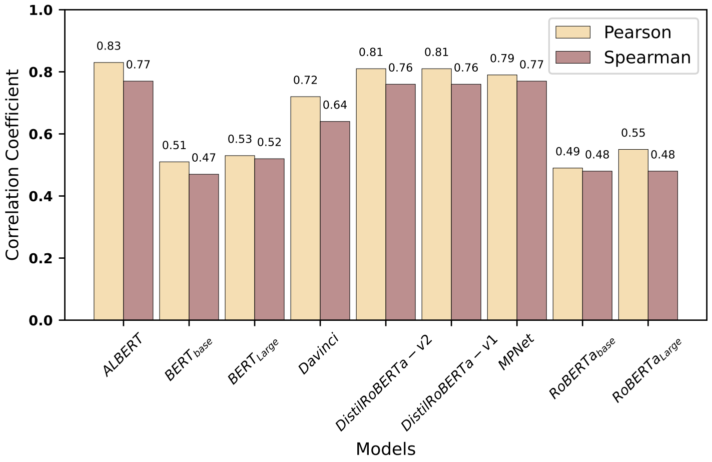


Figure 5: Pearson's and Spearman's correlation coefficient to analyse relationship between similarity values of human annotators and embedding models (Section 3.6) for complex sentence dataset.

S1- A procedure for solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation.

S2- Various methods or formats for organizing data in a computer.

ALBERT	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{Large}}$	Davinci	DistilRoBERTa-v1	DistilRoBERTa-v2	MPNeT	$\text{RoBERTa}_{\text{base}}$	$\text{RoBERTa}_{\text{Large}}$	Human
0.1300	0.4967	0.5656	0.7639	0.1623	0.3924	0.2001	0.6812	0.4069	0.0667

Table 4: Similarity scores of a sentence pair from Complex sentence dataset, evaluated using versions of pre-trained embedding models discussed in Section 3.6. Human ratings are normalized between 0 and 1.

5 Conclusion

Transfer credit assessment usually consists of course comparisons via the evaluation of learning outcomes, which represent an important tool for assessment but are also subject to potential inconsistencies and bias. Therefore, an automated system to assess transfer credit based on learning outcomes across institutes can facilitate the process by providing a reliable and consistent measure of similarity. Over the years there has been a rich advancement in the era of large language models to measure semantic similarity between texts. Various pre-trained embedding models have been developed to represent text for algorithms to understand and compare semantic similarity. In this paper, we aim to propose two novel datasets of learning outcomes for courses in Human Anatomy and Computer Science operating systems and perform an analysis using embedding models to assist in transfer credit assessment. We found that versions of ALBERT, MPNeT and DistilRoBERTa outperform Davinci (a GPT-3 model) that only provides a good qualitative interpretation of the similarity between pairs of sentences. Application of these models within the context of transfer credit assessment can contribute to greater efficiency and consistency when determining learning outcome similarity.

6 Limitations

Due to the complexity measures (readability analysis) requiring a minimum of 100 words, some of the smaller learning outcome sets require padding. To try and minimize the effect this will have on the results, we append the word "a" until the set can be measured. Furthermore, the datasets involve learning outcomes from the same courses being offered at different years of class. Therefore, while conducting human annotation, the comparison among learning outcomes is not consistent, which leads to a low inter-rater agreement among human values for both datasets. While utilizing the pre-trained embedding models, due to fewer number of sentences in the dataset we were not able to pre-train the models. This reflects a need to further enhance the dataset.

Acknowledgment

The authors would like to thank the Research Team, DaTALab, Lakehead University, for their support. They would also like to thank Lakehead University, CASES, and the Ontario Council for Articulation and Transfer, without their support this research would not have been possible.

References

Christine Arnold, Mary Wilson, Jean Bridge, and Mary Catherine Lennon. 2020a. Learning outcomes, academic credit and student mobility.

Christine Arnold, Mary Wilson, Michael Potter, and Leesa Wheelahan. 2020b. Shifting paradigms in postsecondary education: Historical, conceptual, and theoretical frameworks governing outcomes-based approaches to credit transfer. *Learning Outcomes, Academic Credit and Student Mobility*, 201:199.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dhivya Chandrasekaran and Vijay Mago. 2021a. Comparative analysis of word embeddings in assessing semantic similarity of complex sentences. *IEEE Access*, 9:166395–166408.

Dhivya Chandrasekaran and Vijay Mago. 2021b. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Courtney D Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 13–18.

Barbara Gross Davis. 2009. *Tools for teaching*. John Wiley & Sons.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nicole Fallon. 2015. Leaning outcomes in credit transfer: A key tool for innovation in student mobility.

Kris Gerhardt and Oliver Masakure. 2016. Postsecondary student mobility from college to university: Academic performance of students. *Canadian Journal of Higher Education*, 46(2):78–91.

Robert Gunning et al. 1952. Technique of clear writing.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Daniel Lang and Valerie Lopes. 2014. Deciding to transfer: A study of college to university choice. *College Quarterly*, 17(3):1.

Mary Catharine Lennon. 2015. Incremental steps towards a competency-based post-secondary education system in ontario.

Mary Catherine Lennon, Brian Frank, James Humphreys, Rhonda Lenton, Kirsten Madsen, Abdelwahab Omri, and Roderick Turner. 2014. *Tuning: Identifying and measuring sector-based learning outcomes in postsecondary education*. Higher Education Quality Council of Ontario Toronto.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199.

Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. 2016. Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, 20(4):647–665.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černockỳ. 2011. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 196–201. IEEE.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Muhidin Mohamed and Mourad Oussalah. 2019. Srl-esa-textsum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management*, 56(4):1356–1372.

Daniel Ramage, Anna N Rafferty, and Christopher D Manning. 2009. Random walks for text semantic similarity. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing (TextGraphs-4)*, pages 23–31.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yang Shao. 2017. Hcti at semeval-2017 task 1: Use convolutional neural network to evaluate semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 130–133.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Elisa Tortola, Christine Arnold, and Zanele Myles. 2020. Foundations for learning outcomes and credit transfer. *Learning Outcomes, Academic Credit and Student Mobility*, 201:19.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Leesa Wheelahan, Gavin Moodie, Mary Catherine Lennon, Amanda Brijmohan, and Eric Lavigne. 2016. Student mobility in ontario: A framework and decision making tool for building better pathways. *Centre for the Study of Canadian and International Higher Education, OISE-University of Toronto: Toronto, ON, Canada*.

Stacey Young, PG Piché, and Glen A Jones. 2017. Two towers of transformation: The compatibility of policy goals of differentiation and student mobility. *Toronto: Center for the Study of Canadian and International Higher Education, OISE-University of Toronto*.

Jun Zhao, Kang Liu, and Liheng Xu. 2016. Sentiment analysis: mining opinions, sentiments, and emotions.