

## Automating Transfer Credit Assessment-A Natural Language Processing-Based Approach

Dhivya Chandrasekaran\* and Vijay Mago

Lakehead University, Thunder Bay, Ontario, P7B 5E1, Canada

\*Corresponding Author: Dhivya Chandrasekaran. Email: dchandra@lakeheadu.ca

Received: 13 January 2022; Accepted: 18 March 2022

**Abstract:** Student mobility or academic mobility involves students moving between institutions during their post-secondary education, and one of the challenging tasks in this process is to assess the transfer credits to be offered to the incoming student. In general, this process involves domain experts comparing the learning outcomes of the courses, to decide on offering transfer credits to the incoming students. This manual implementation is not only labor-intensive but also influenced by undue bias and administrative complexity. The proposed research article focuses on identifying a model that exploits the advancements in the field of Natural Language Processing (NLP) to effectively automate this process. Given the unique structure, domain specificity, and complexity of learning outcomes (LOs), a need for designing a tailor-made model arises. The proposed model uses a clustering-inspired methodology based on knowledge-based semantic similarity measures to assess the taxonomic similarity of LOs and a transformer-based semantic similarity model to assess the semantic similarity of the LOs. The similarity between LOs is further aggregated to form course to course similarity. Due to the lack of quality benchmark datasets, a new benchmark dataset containing seven course-to-course similarity measures is proposed. Understanding the inherent need for flexibility in the decision-making process the aggregation part of the model offers tunable parameters to accommodate different levels of leniency. While providing an efficient model to assess the similarity between courses with existing resources, this research work also steers future research attempts to apply NLP in the field of articulation in an ideal direction by highlighting the persisting research gaps.

**Keywords:** Articulation agreements; higher education; natural language processing; semantic similarity

### 1 Introduction

With the significant growth in the enrollment of students in post-secondary institutions and the growing trend of interest in diversifying one's scope of education, there is an increasing demand among the academic community to standardize the process of student mobility. Student mobility is defined



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

as “any academic mobility which takes place within a student’s program of study in post-secondary education [1].” Student mobility could be both international and domestic. While there are various barriers to student mobility, offering transfer credits for students moving from one post-secondary institution to another is considered one of the most critical and labor-intensive tasks [2]. Various rules and regulations are proposed and adopted by institutions across different levels (provincial, national, or international) to assess transfer credits, but most of these methods are time-consuming, subjective, and influenced by undue human bias. The key parameter used in assessing the similarity between programs or courses across institutions is learning outcome (LO), which provides context on the competencies; achieved by students on completion of a respective course or program. To standardize this assessment, LOs are categorized into various levels based on Bloom’s taxonomy. Bloom’s taxonomy proposed by Bloom et al. [3] attempts to classify the learning outcomes into six different categories based on their “complexity and specificity”, namely *knowledge, comprehension, application, analysis, synthesis, and evaluation*.

Semantic similarity, being one of the most researched Natural language processing (NLP) tasks, has seen significant breakthroughs in recent years with the introduction of transformer-based language models. These language models employ attention mechanisms to capture the semantic and contextual meaning of text data and represent them as real-valued vectors, that are aligned in an embedding space such that the angle between these vectors provides the similarity between the text in consideration. In an attempt to reduce the inherent complexity and bias, and exploit the advancements in the field of NLP, we propose a model that determines the similarity between courses; based on the semantic and taxonomic similarity of their learning outcomes. The proposed model

- ascertains taxonomic similarity of LOs based on Bloom’s taxonomy.
- Determines the semantic similarity of LOs using RoBERTa language model.
- provides a flexible aggregation method to determine the overall similarity between courses.

Section 2 discusses the background of the student mobility process and semantic similarity techniques. Section 3 describes in detail the various components of the proposed methodology followed by the discussion of results in Section 4. In Section 5, challenges in automating the process of assessing transfer credit are outlined and the article concludes with the future scope of research in Section 6.

## 2 Background

This section provides a brief overview of the student mobility process across the world and the structural organization of learning outcomes. Various semantic similarity methods developed over the years are discussed in the final sub-section thus providing an insight into the necessary concepts to understand the importance of the research and the choices made to develop the proposed methodology.

### 2.1 Student Mobility

The movement of students across institutions for higher education has been in existence for decades in the form of international student exchange programs, lateral transfers, and so on. Governments across the world follow different measures to standardize the process to ensure transparency and equity for students. According to the Organization for Economic Co-operation and Development (OECD) indicators<sup>1</sup>, there are approximately 5.3 million internationally mobile students. Mobile students include both international students who cross borders to pursue education and foreign

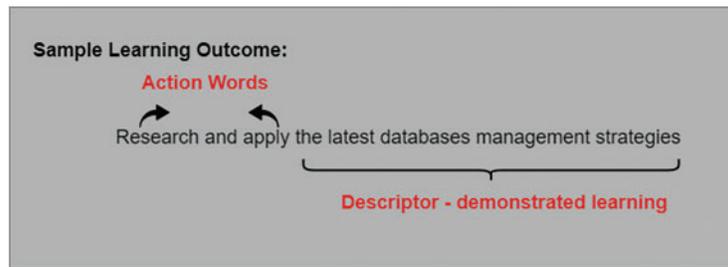
---

<sup>1</sup><https://www.oecd-ilibrary.org/content/component/17d19cd9-en>

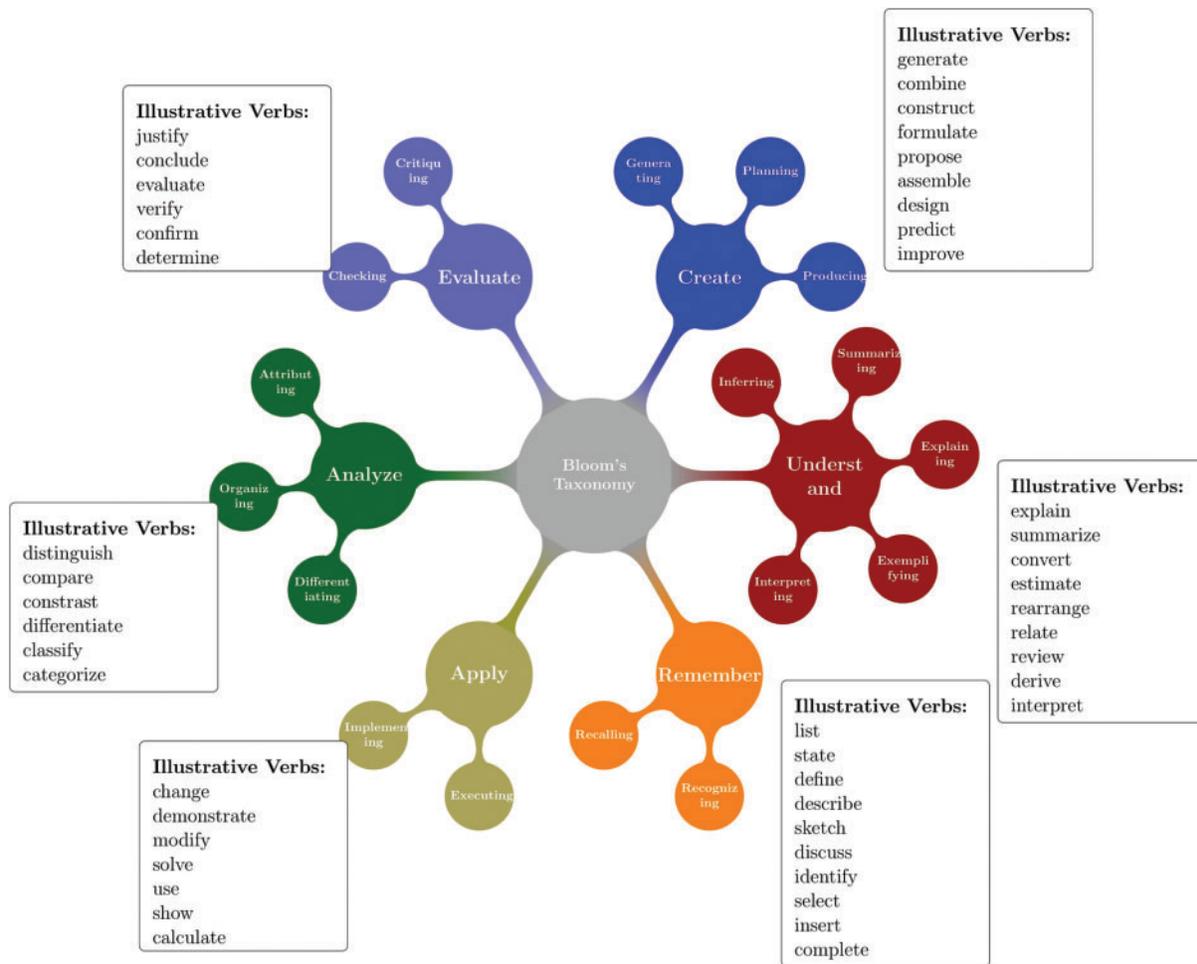
students who are from a different nationality than the country of the host institution. Mobile students face a wide range of barriers both academic and non-academic. Academic barriers include the lack of necessary qualifications and non-transferability of credits, while non-academic barriers may include social, cultural, financial, and psychological barriers. Governments across the world have taken various measures to reduce these barriers to enable a smooth transition for students. The Bologna process formed as a result of the Bologna declaration of 1999, provides guidance for the European Higher Education Area comprising 48 countries in the standardization of higher education and credit evaluation. In the United Kingdom, institutions like Southern England Consortium for Credit Accumulation and Transfer (SEEC) and Northern Universities Consortium for Credit Accumulation and Transfer (NUCCAT) oversee the collaboration between universities to allocate academic credits which are treated as currency awarded to students on completion of requirements. Canada offers provincial supervision of articulation agreements between institutions, with the provinces of British Columbia and Alberta leading and the provinces of Ontario and Saskatchewan following yet way behind. The Ontario Council on Articulation and Transfer (ONCAT) carries out funded research to explore venues to increase the agreements between universities and colleges in Ontario. The credit transfer system in the United States is decentralized and often carried out by non-profit organizations designated for this specific purpose. In Australia, the eight most prominent universities established the Go8 credit transfer agreement to offer credit to students who move between these institutions. While there are various such governances on a national level, most international credit evaluations are carried out in a need-based manner. In addition to being an academic barrier, credit evaluation also has a direct impact on one of the most important non-academic barriers-the financial barrier. Hence, all agencies offer special attention to make this process fair and accessible to the population of mobile students worldwide.

## **2.2 Learning Outcomes**

Credit evaluation is carried out by domain experts in the receiving institution by analyzing the learning outcomes of the courses the students have completed in their previous institution. Learning outcomes are often statements with two distinct components namely the action words and the descriptor. The descriptor part provides the knowledge the students have learned in each course or program and the action words provide the level of competency achieved for each specific knowledge item. An example of learning outcomes for a computer programming course is provided in [Fig. 1](#). The taxonomy for educational objectives was developed by Bloom et al. [3] and later revised by Anderson et al. [4]. The six levels of the original taxonomy are *knowledge, comprehension, application, analysis, synthesis, and evaluation*. In the revised version, two of these levels were interchanged and three levels were renamed to provide better context to the level of the acquired knowledge. Hence the levels of the revised taxonomy are “*Remember, Understand, Apply, Analyze, Evaluate, and Create.*” Each level encompasses sub-levels of more concrete knowledge items, and these are provided in [Fig. 2](#). In order to estimate the similarity between learning outcomes, it is important to understand their structure and organization.



**Figure 1:** Learning outcome



**Figure 2:** Bloom's taxonomy and corresponding illustrative verbs

### 2.3 Semantic Similarity

The semantic textual similarity (STS) is defined as the similarity in the meaning of text data in consideration. Various semantic similarity methods proposed over the past two decades can be broadly classified as knowledge-based and corpus-based methods [5,6]. Knowledge-based methods rely on

ontologies like WordNet<sup>2</sup>, DBpedia<sup>3</sup>, BabelNet<sup>4</sup>, etc. The ontologies are conceptualized as graphs where the words represent the nodes grouped hierarchically, and the edges represent the semantic relationship between the words. Rada et al. [7], followed a straightforward approach and introduced the *path* measure in which the semantic similarity is inversely proportional to the number of edges between the two words. However, this method ignored the taxonomical information offered by the underlying ontologies. Wu et al. [8] proposed the *wup* measure that measured the semantic similarity in terms of the least common subsumer (LCS). Given two words, LCS is defined as the common parent they share in the taxonomy. Leacock et al. [9] proposed the *lch* measure by extending the *path* measure to incorporate the depth of the taxonomy to calculate semantic similarity in Tab. 1. Other knowledge-based approaches include feature-based semantic similarity methods, that calculate similarity using the features of the words like their dictionary definition, grammatical position, etc., and information content-based methods that measure semantic similarity using the level of information conveyed by the words when appearing in a context.

**Table 1:** Formula of the semantic similarity distance measures taken into consideration for determining verb similarity

S. no.	Semantic similarity measure	Formula
1	$sim_{path}(t_1, t_2)$	$\frac{1}{1 + min\_len(t_1, t_2)}$
2	$sim_{wup}(t_1, t_2)$	$\frac{2depth(t_{lcs})}{depth(t_1) + depth(t_2)}$
3	$sim_{lch}(t_1, t_2)$	$-\log\left(\frac{minlength(t_1, t_2)}{2 \times maxdepth}\right)$

Corpus-based semantic similarity methods construct numerical representations of data called embeddings using large text corpora. Traditional methods like Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) used one-hot encoding techniques or word counts to generate embeddings. These methods ignored the polysemy of text data and suffered due to data sparsity. Mikolov et al. [10] used a simple neural network with one hidden layer to generate word-embeddings when used in simple mathematical formulations, produced results that were closely related to human understanding. Pennington et al. [11] used word co-occurrence matrices and dimension reduction techniques like PCA to generate embeddings. The introduction of transformer-based language models, which produced state-of-the-art results in a wide range of NLP tasks, resulted in a breakthrough in semantic similarity analysis as well. Vaswani et al. [12] proposed the transformer architecture, which used stacks of encoders and decoders with multiple attention heads for machine translation tasks. Devlin et al. [13] used this architecture to introduce the first transformer-based language model, the Bidirectional Encoder Representations from Transformers (BERT) that generated contextualized word embeddings. BERT models were pre-trained on large text corpora and further fine-tuned to a specific task. Various versions of BERT were subsequently released namely, RoBERTa [14]-trained on a larger corpus over longer periods of time, ALBERT [15]-a lite version achieved using parameter reduction techniques, BioBERT [16]-trained on a corpus of biomedical text, SciBERT [17]-trained on

<sup>2</sup><https://wordnet.princeton.edu/>

<sup>3</sup><https://www.dbpedia.org/>

<sup>4</sup><https://babelnet.org/>

a corpus of scientific articles, and TweetBERT [18]-trained on a corpus of tweets. Other transformer-based language models like T5 [19] GPT [20], GPT-2 [21], GPT-3 [22], etc., use the same transformer architecture with significantly larger corpora and an increased number of parameters. Though these models achieve state-of-the-art results the computational requirements render them challenging to implement in real-time tasks [23]. Some of the other drawbacks of these transformer-based models include the lack interpretability and the amplification of societal bias that exists in the training data [24–26].

### 3 Methodology

This section describes in detail the three modules of the proposed methodology namely,

- Pass 1: Taxonomic similarity
- Pass 2: Semantic similarity
- Pass 3: Aggregation

Given the learning outcomes of the courses in comparison, Pass 1 generates a *taxonomic\_similarity\_grid*, and Pass 2 generates a *semantic\_similarity\_grid* where the rows and columns are populated with the learning outcomes, and the cells are populated with the taxonomic similarity value and the semantic similarity value. These two grids are further passed on to Pass 3 where the final similarity between learning outcomes is assessed factoring in both the similarity values and further aggregated to arrive at the course level similarity. The architecture of the proposed model is presented in Fig. 3.

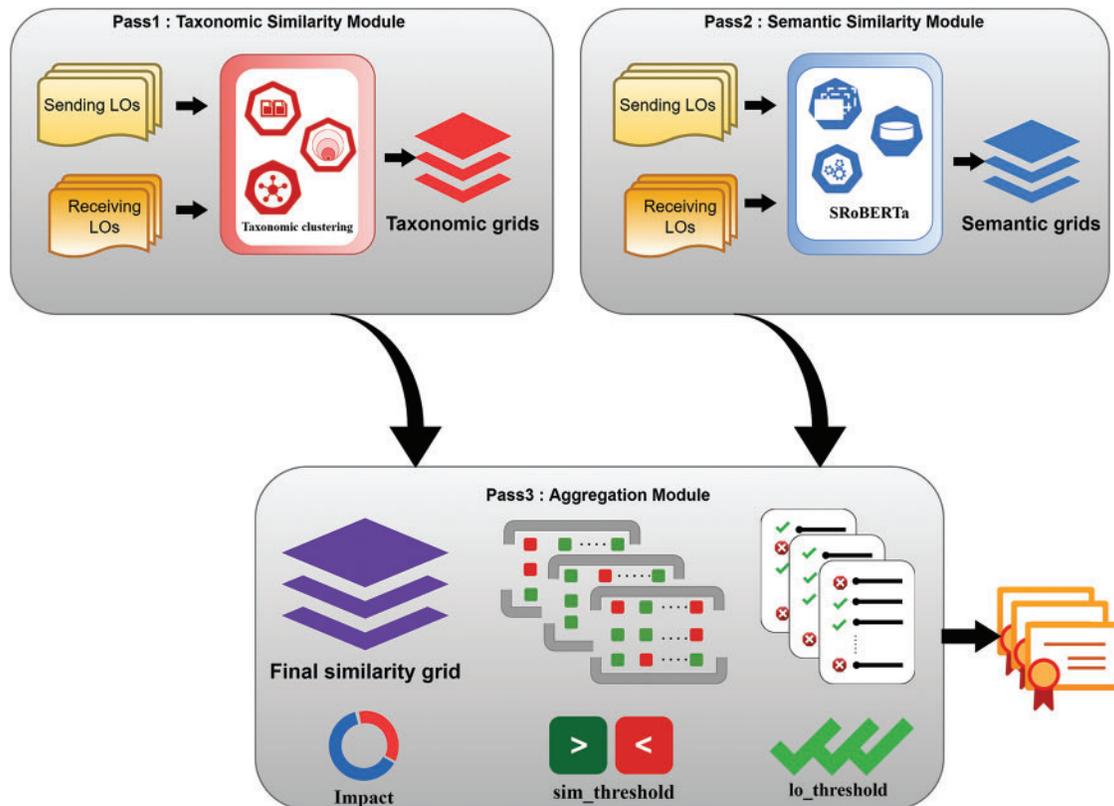


Figure 3: Proposed model architecture

### 3.1 Pass 1: Taxonomic Similarity

A clustering-inspired methodology is proposed to determine the taxonomic similarity between learning outcomes. Six different clusters, one for each level in Bloom's taxonomy are initialized with a list of illustrative verbs that best describe the cognitive competency achieved, specifically in the field of engineering [27], as shown in Fig. 2. In this pass, the verbs in the learning outcomes are identified using spaCy pos tagger<sup>5</sup> and WordNet synsets [28], and then these verbs are used to determine the cluster to which the learning outcomes are most aligned with. While encountering verbs already available in the list, a straightforward approach is followed, and the learning outcomes are assigned to the respective cluster. However, for learning outcomes with new verbs, an optimal cluster is determined based on the semantic similarity between the new verb and the verbs in the existing clusters. The best measure to calculate this similarity is determined as a result of the comparative analysis carried out between various knowledge-based and corpus-based semantic similarity measures.

---

**Algorithm 1:** Best semantic similarity measure for verbs.

---

**Input:**  $D_{verb} [v_1, v_2, s]$  = SimVerb3500 dataset containing verb pairs  $(v_1, v_2)$  and associated semantic similarity benchmark values  $s$

$wv$  = pretrained *word2vec* word embedding model

$gl$  = pretrained *GloVe* word embedding model

$wn$  = *WordNet* lexical knowledge base

**Output:**  $best\_measure$  = the semantic similarity measure that achieves the best correlation to the benchmark values.

1:  $S \leftarrow \emptyset$

2: **procedure** VERB\_SIMILARITY

3:     **for each**  $(v_1, v_2, s) \in D_{verb}$  **do**

4:          $S+ = s$

5:          $w1 \leftarrow wv.embedding(v_1)$

6:          $w2 \leftarrow wv.embedding(v_2)$

7:          $g1 \leftarrow gl.embedding(v_1)$

8:          $g2 \leftarrow gl.embedding(v_2)$

9:          $synsets1[syn_{11} \dots syn_{1i}] \leftarrow wn.get\_synsets(v_1, POS:verb)$  ➤ List of synsets

10:          $synsets2[syn_{21} \dots syn_{2j}] \leftarrow wn.get\_synsets(v_2, POS:verb)$  ➤ List of synsets

11:         **for each**  $syn_{1i} \in synsets1$  **do**

12:             **for each**  $syn_{2j} \in synsets2$  **do**

13:                  $wup\_sim\_list[syn_{1i}][syn_{2j}] \leftarrow cal\_wup(syn_1, syn_2)$  ➤ *wup\_similarity*

14:                  $path\_sim\_list[syn_{1i}][syn_{2j}] \leftarrow cal\_path(syn_1, syn_2)$  ➤ *path\_similarity*

15:                  $lch\_sim\_list[syn_{1i}][syn_{2j}] \leftarrow cal\_lch(syn_1, syn_2)$  ➤ *lch\_similarity*

16:             **end for**

17:         **end for**

18:          $wup\_sim \leftarrow max(wup\_sim\_list[n])$

19:          $path\_sim \leftarrow max(path\_sim\_list[n])$

20:          $lch\_sim \leftarrow max(lch\_sim\_list[n])$

21:          $word2vec\_sim \leftarrow wv.similarity(w1, w2)$

22:          $glove\_sim \leftarrow gl.similarity(g1, g2)$

23:     **end for**

---

(Continued)

<sup>5</sup><https://spacy.io/usage/v3>

**Algorithm 1:** Continued

---

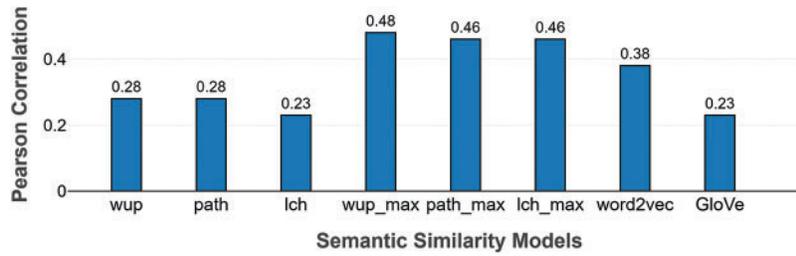
```

24:  measure_list+ = wup_sim, path_sim, lch_sim, word2vec_sim, glove_sim
25:  for each m ∈ measure_list do
26:      cor ← get_pearson_correlation(S, m)
27:  end for
28:  best_measure ← max(cor)
29:  return best_measure
30: end procedure

```

---

Three knowledge-based measures namely *wup*, *lch*, and *path* are measured using WordNet ontology. In this ontology, there are more than one synsets for verbs hence it is necessary to identify the best synset. Given a verb pair, *wup*, *lch*, and *path* select the first synset of the verbs, whereas *wup\_max*, *lch\_max*, and *path\_max* identifies the synset that has the maximum similarity with the verb pairs. Gerz et al. [29] proposed SimVerb-3500 a benchmark dataset consisting of verb pairs with associated similarity values provided by annotators using crowd-sourcing techniques. The performance of six knowledge-based semantic similarity measures and word embeddings models (*word2vec* and *GloVe*) on the SimVerb-3500 is compared and the results are depicted in Fig. 4. This pseudocode is provided in Algorithm 1



**Figure 4:** Pearson’s correlation of various semantic similarity measures on Simverb3500 dataset. *wup* [8], *path* [7], *lch* [9], *word2vec* [10], *GloVe* [11]

The best results are achieved by *wup\_max* measure, hence used in the clustering process. Silhouette width is defined as “the measure of how much more similar a data point is to the points in its own cluster than to points in a neighboring cluster [30].” The silhouette width of a verb  $V_i$  is measured as,

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (1)$$

where,  $a_i$  is the average distance between the verb and the verbs in its own cluster and  $b_i$  is the average distance between the verb and the verbs in its neighboring cluster. The following equations formulate the calculation of these distance measures.

$$a_i = \frac{\sum_{V_j \in C(V_i)} d(V_i, V_j)}{|C(V_i)|} \quad (2)$$

$$b_i = \frac{\sum_{V_k \in C_{neig}(V_i)} d(V_i, V_k)}{|C_{neig}(V_k)|} \quad (3)$$

$V_j$  represent the verbs in the same cluster  $C$  and  $V_k$  represent the verbs in the neighboring cluster  $C_{neig}$ , the function  $d(V_i, V_j)$  and  $d(V_i, V_k)$  calculates the distance between the verbs in this case the *wup\_measure*.

The value of  $S$  ranges from  $-1$  to  $1$  where values closer to  $1$  indicate that the data point is assigned to the correct cluster. Based on this principle, when the model identifies a *new\_verb* in a learning outcome, the silhouette width for each cluster is determined and the learning outcome is assigned to the cluster with maximum silhouette width. For learning outcomes with more than one verb, the verb assigned to the highest taxonomic level determines the final cluster value of the learning outcome in question. Since each cluster represents a corresponding level of competency in Bloom's taxonomy, the final taxonomic similarity between the learning outcomes is measured as,

$$\text{taxonomic\_similarity}(lo1, lo2) = \text{abs}(C_{lo1} - C_{lo2}) \quad (4)$$

where  $C_{lo1}$  and  $C_{lo2}$  represent the cluster ids of the learning outcomes in comparison calculated as,

$$C_{loi} = \max(C(v_1), C(v_2) \dots C(v_n)) \quad (5)$$

where  $n$  is the number of verbs in the learning outcome  $loi$  and  $C(v_n)$  represents the respective cluster of the verb  $v_n$ . For two courses having  $m$  and  $n$  number of learning outcomes, a  $m \times n$  dimensional *taxonomic\_similarity\_grid* is constructed and populated with the respective *taxonomic\_similarity* values. The pseudocode of Pass 1 is provided in Algorithm 2.

---

#### Algorithm 2: Pass\_1: Taxonomic Similarity

---

**Input:** *Slo\_list* = List of  $m$  learning outcomes from the sending institution

*Rlo\_list* = List of  $n$  learning outcomes from the receiving institution.

*Cluster\_list* = {C1, C2, C3, C4, C5, C6} List of six clusters initialized with illustrative verbs, such that C1 = {v1, v2 ... vn} is a list of  $n$  verbs

**Output:** *taxonomic\_similarity\_grid* = A  $m \times n$  grid containing taxonomic similarity values between sending and receiving learning outcomes.

1: **procedure** BUILD\_TAXONOMIC\_SIMILARITY\_GRID

2:   **for each** (*slo*)  $\in$  *Slo\_list* **do**

3:     *slo\_verbs*  $\leftarrow$  *detect\_verbs(slo)*

➤ Find the verbs in the learning outcomes

4:     **for each** *sv*  $\in$  *slo* **do**

5:       **if** *sv*  $\in$  *Cluster\_list* **then**

6:          *B(sv)*  $\leftarrow$  *get\_index(Ci)*

➤ If the verb is among the predefined verbs assign the corresponding cluster id.

7:       **else**

8:          *B(sv)*  $\leftarrow$  *get\_BestCluster(sv)*

➤ Assign the cluster id of the best cluster based on the silhouette width

9:       **end if**

10:     **end for**

11:     *B(slo)*  $\leftarrow$   $\max(B(sv1), \dots B(svn))$

12:   **end for**

13:   ➤ Repeat for learning outcomes from receiving institution

14:   **end procedure**

(Continued)

**Algorithm 2:** Continued

---

```

16: for each (rlo) ∈ Rlo_list do
17:   rlo_verbs ← detect_verbs(rlo)
18:   for each rv ∈ rlo do
19:     if rv ∈ Cluster_list then
20:       B(rv) ← get_index(Ci)
21:     else
22:       B(rv) ← get_BestCluster(rv)
23:     end if
24:   end for
25:   B(rlo) ← max(B(rv1), . . . B(rvn))
26: end for
27:                                     ➤ Calculate taxonomic similarity
28: for each (slo) ∈ Slo_list do
29:   for each (rlo) ∈ Rlo_list do
30:     taxonomic_similarity_grid[slo][rlo] ← abs(B(rlo)−B(slo))
31:   end for
32: end for
33: return taxonomic_similarity_grid
34: end procedure

```

---

**3.2 Pass 2: Semantic Similarity**

Recent transformer-based language models generate contextual word embeddings that are fine-tuned for a specific NLP task, and various researchers have attempted to generate sentence embeddings by averaging the output from the final layer of the language models [31]. However, Reimers et al. [32] established that these techniques yielded poor results in regression and clustering tasks like semantic similarity and proposed the Sentence BERT (SBERT) model to address this issue. SBERT is a modified version of BERT-based language models where a Siamese network and triple network structures are added to the final layer of the BERT network to generate sentence embeddings that capture the semantic properties and thus can be compared using cosine similarity. They further compared the performance of both SBERT and SROBERTa language models in the STS [33] and SICK [34] benchmark datasets. Owing to the fact that the model is specifically designed for semantic similarity tasks, the computational efficiency of the model architecture, and the significant performance achieved, SROBERTa is used in this pass to measure the semantic similarity of the LOs. SROBERTa uses the base RoBERTa-large model with 24 transformer blocks, 1024 hidden layers, 16 attention heads, and 340 M trainable parameters with a final mean pooling layer. The model is trained on the AllNLI dataset which contains 1 million sentence pairs categorized into three classes namely ‘contradiction, entailment, and neutral’, and the training data of the STS benchmark dataset. The semantic similarity between the learning outcomes is determined by the cosine value between the embeddings as,

$$semantic\_similarity(lo1, lo2) = \frac{lo1 \cdot lo2}{\sqrt{\sum_{i=1}^n lo1} \sqrt{\sum_{i=1}^n lo2}} \quad (6)$$

The *semantic\_similarity\_grid* with the dimension of  $m \times n$  is formed and the *semantic\_similarity* values are added to the cells. The output from the two initial passes is fed to the final pass for aggregation. The pseudocode for Pass 2 is provided in Algorithm 3.

**Algorithm 3:** Pass\_2: Semantic Similarity**Input:**  $Slo\_list$  = List of  $m$  learning outcomes from the sending institution. $Rlo\_list$  = List of  $n$  learning outcomes from the receiving institution. $model$  = Pretrained and fine-tuned SRoBERTa transformer-based semantic similarity model**Output:**  $semantic\_similarity\_grid$  = A  $m \times n$  grid containing semantic similarity values between sending and receiving learning outcomes1: **procedure** BUILD\_SEMANTIC\_SIMILARITY\_GRID2: **for each**  $slo, rlo \in Slo\_listRlo\_list$  **do**3:  $\vec{slo} \leftarrow model.sentence\_vectors(slo)$   $\triangleright$  Get vector representation of learning outcomes4:  $\vec{rlo} \leftarrow model.sentence\_vectors(rlo)$ 5:  $semantic\_similarity\_grid[slo][rlo] \leftarrow \cos(\vec{slo}, \vec{rlo})$   $\triangleright$  Calculate semantic similarity6: **end for**7: **return**  $semantic\_similarity\_grid$ 8: **end procedure****3.3 Pass 3: Aggregation**

The focus of the final pass is to provide flexibility in the aggregation process to enable the decision-making authorities to accommodate the variations in the administrative process across different institutions. Three important tunable parameters are provided to adjust the level of leniency offered by the decision-making authority in providing credits namely,  $impact$ ,  $sim\_threshold$ , and  $lo\_threshold$ . Given the taxonomic similarity grid and the semantic similarity grid from Pass 1 and Pass 2 the  $impact$  parameter determines the percentage of contribution of both the similarities to the overall similarity. The  $sim\_threshold$  defines the value above which two learning outcomes are considered to be similar, and finally, the  $lo\_threshold$  determines the number of learning outcomes that need to be similar in order to consider the courses in comparison to being similar. The higher the value of these three parameters the lesser the leniency in the decision-making process.

The  $final\_similarity\_grid$  is built by aggregating the values from the previous modules, in the ratio determined by the  $impact$  parameter as shown in Fig. 3. The LOs along the rows of the grid belong to the receiving institution's course hence traversing along the rows, the maximum value in the cells is checked against the  $sim\_threshold$  value to determine if the LO in the row is similar to any LO in the columns. The course level similarity is derived by checking if the number of learning outcomes having a similar counterpart meets the  $lo\_threshold$ . The pseudocode for Pass 3 in Algorithm 4.

**Algorithm 4:** Pass\_3: Aggregation**Input:** SSG =  $semantic\_similarity\_grid$ , a  $m \times n$  grid containing semantic similarity values between sending and receiving

learning outcomes

TSG =  $taxonomic\_similarity\_grid$ , a  $m \times n$  grid containing taxonomic similarity values between sending and receiving

learning outcomes

 $\alpha$  = parameter which determines the ratio of two similarity value in the overall similarity $\beta$  = value above which two learning outcomes are considered similar $\gamma$  = number of learning outcomes with similar counterparts

(Continued)

**Algorithm 4:** Continued**Output:**  $TC$  = Final Credit Decision

---

```

1: procedure COURSE_LEVEL_SIMILARITY
2:   for each  $row, col \in SSG$  do
3:     for each  $row, col \in TSG$  do
4:        $final\_sim \leftarrow (\alpha \times SSG[row, col] + (1-\alpha) \times TSG[row, col])$ 
5:       if  $final\_sim \geq \beta$  then
6:          $final\_similarity\_grid[row, column] \leftarrow TRUE$ 
7:       else
8:          $final\_similarity\_grid[row, column] \leftarrow FALSE$ 
9:       end if
10:    end for
11:  end for
12:  for each  $row \in final\_similarity\_grid$  do
13:    if  $Count(TRUE) \geq \gamma$  then
14:       $TC = Yes$ 
15:    else
16:       $TC = No$ 
17:    end if
18:  end for
19:  return  $TC$ 
20: end procedure

```

---

**4 Dataset and Results****4.1 Benchmark Dataset**

One of the major challenges in the given field of research is the absence of benchmark datasets to compare the performance of the proposed system. Although there are existing pathways developed manually, previous research show that most of them are influenced by bias (based on the reputation of institutions, year of study, and so on) and administrative accommodations [35]. To create a benchmark dataset devoid of bias, a survey was conducted among domain experts to analyze and annotate the similarity between courses from two different institutions. A survey with learning outcomes of 7 pairs of courses (sending and receiving) from the computer science domain was distributed among instructors from the department of computer science at a comprehensive research university in Canada. In order to avoid bias, the names of both the institutions were anonymized and explicit instructions were provided to the annotators to assume a neutral position. Although the survey was circulated among 14 professors only 5 responses were received. This lack of responses is mainly attributed to the fact that most faculties are not involved in the transfer pathway development process, and it is carried out widely as an administrative task. The survey questionnaire consisted of questions to mark the similarity between the courses over a scale of 1 to 10 and a binary response ('yes' and 'no') for whether or not credit should be offered to the receiving course. The course pairs were annotated with a final decision value based on the maximum number of 'yes' or 'no' responses received from the annotators. The results of the survey are tabulated in Tab. 2 where the responses 'yes' and 'no' are color coded in 'green' and 'red' respectively. One of the interesting inferences from the survey results is the agreement between the responses in the threshold value of similarity above which the annotators were willing to offer credit. 4 out of 5 annotators offered credit only if they considered the similarity

between the courses falls above 7 on the given scale of 1 to 10. It is also interesting to note that in spite of having no information other than the learning outcomes the annotators differed in their level of leniency which inspired the need to offer flexibility to control leniency in the proposed methodology. For example, from [Tab. 2](#) it is evident that while annotator ‘A2’ has followed a more lenient approach and offered credit for 6 out of 7 courses, annotator ‘A5’ has adopted a stricter approach by offering credit for only 1 out of the 7 courses.

**Table 2:** Survey results for the proposed benchmark dataset

	A1	A2	A3	A4	A5	Final annotation
1	COMP 4121					
2	COMP 5341					
3	COMP 3321					
4	COMP 4321					
5	COMP 3519					
6	COMP 5470					
7	COMP 5471					

## 4.2 Results and Discussion

The results of the proposed model are provided in [Tab. 3](#). In order to provide context to the need for the proposed methodology, the results of the model are compared to the results obtained when only the semantic similarity of the learning outcomes is considered. The proposed model at a neutral setting achieves 85.74% agreement with the human annotation by annotating 6 out of the 7, credit decision correctly, while the semantic similarity model achieves only 54.75% agreement. For the neutral setting of the proposed model, the three parameters in the aggregation pass are set as follows. The *impact* parameter is set at 0.30 meaning that the semantic similarity contributes 70% to the overall similarity and the taxonomic similarity contributes to the remaining 30%. The *sim\_threshold* is set at 0.65, meaning that the overall similarity should be more than 65% in order for the learning outcomes to be similar to each other. The *lo\_threshold* is set at 0.5 which considers that at least half of the available learning outcomes have similar counterparts. In order to demonstrate the options for flexibility, the proposed model is run by modifying the *lo\_threshold* parameter. As shown in [Tab. 4](#), for the lenient setting the *lo\_threshold* parameter is set at 0.33 and 0.66 and the model achieves an agreement of 85% with the most lenient annotator and the strictest annotator, respectively. Similarly, lowering the *impact* parameter makes the model more aligned to being strict and vice versa. However, it is important to understand that increasing the impact of taxonomic similarity to a higher percentage result in determining the overall similarity based on only two or three action words in the learning outcomes. Decreasing the *sim\_threshold* will make the model more lenient and increasing it will make the model stricter. The results of the model at various parameter settings are tabulated in [Tabs. 5 and 6](#) in APPENDIX A. Since the proposed model calculates similarity based on Bloom’s taxonomy and text data of the learning outcomes, the model can be used across universities and departments. This ensures that the model is not only flexible in terms of leniency but also accommodate changing scenarios.

**Table 3:** Performance comparison of the proposed model with human annotation

S. no.	Human	Semantic similarity	Proposed methodology-neutral
1	COMP 4121	COMP 4121	COMP 4121
2	COMP 5341	COMP 5341	COMP 5341
3	COMP 3321	COMP 3321	COMP 3321
4	COMP 4321	COMP 4321	COMP 4321
5	COMP 3519	COMP 3519	COMP 3519
6	COMP 5470	COMP 5470	COMP 5470
7	COMP 5471	COMP 5471	COMP 5471
Agreement out of 7		4	6

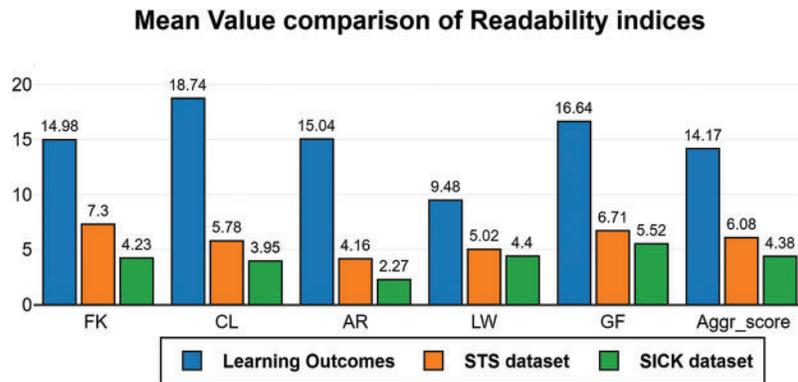
**Table 4:** Performance of variations of proposed models with specific annotators

S. no.	Most lenient annotator	Proposed model-lenient	Most strict annotator	Proposed model-strict
1	COMP 4121	COMP 4121	COMP 4121	COMP 4121
2	COMP 5341	COMP 5341	COMP 5341	COMP 5341
3	COMP 3321	COMP 3321	COMP 3321	COMP 3321
4	COMP 4321	COMP 4321	COMP 4321	COMP 4321
5	COMP 3519	COMP 3519	COMP 3519	COMP 3519
6	COMP 5470	COMP 5470	COMP 5470	COMP 5470
7	COMP 5471	COMP 5471	COMP 5471	COMP 5471
Agreement out of 7		6	6	6

## 5 Challenges and Future Works

One of the important limitations of this research is the fewer number of data points in the benchmark proposed which is attributed to the limited availability of quality learning outcomes, limited response from domain expert annotators, and cost. It is pertinent to understand that there are unique challenges to be addressed in the attempt to automate articulations to overcome these limitations. Although the existing transformer-based models achieve near-perfect results in benchmark datasets, a thorough understanding of these datasets brings to light one of their major shortcomings. Rogers et al. [36] conclude in their survey with a clear statement on the limitations of the benchmark datasets, “As with any optimization method, if there is a shortcut in the data, we have no reason to expect BERT to not learn it. But harder datasets that cannot be resolved with shallow heuristics are unlikely to emerge if their development is not as valued as modeling work.” Chandrasekaran et al. [37] established the significant drop in performance of these models with the increase in complexity of sentences. The comparison of the complexity of the learning outcomes in the proposed benchmark

dataset to the sentences in the STS dataset is shown in Fig. 5, which clearly indicates that learning outcomes are more complex.



**Figure 5:** Comparison of the readability indices of the learning outcomes used in the proposed dataset and the existing benchmark datasets (STS and SICK dataset)

Also, the introduction of domain-specific BERT models shows clear indications that though the transformer models are trained using significantly large corpora with millions of words in their vocabularies, a domain-specific corpus is required to achieve better results in domain-specific tasks. Learning outcomes are not only complex sentences but also contain domain-specific terminologies from various domains. Identifying these research gaps in semantic similarity methods is essential to contextualize and focus future research on addressing them. In addition to the technological challenges, it is also important to understand the challenges faced in the field of articulation. One of the approaches to enhance the performance of existing semantic similarity models is to train them with a large dataset of learning outcomes with annotated similarity values. However, learning outcomes are often considered to be intellectual properties of the instructors and are not publicly available. While almost all universities focus on building quality learning outcomes, most learning outcomes are either vague or don't follow the structural requirements of learning outcomes [38]. Even if a significant amount of learning outcomes is collected, annotation of their similarity requires expertise in subject matter and understanding of the articulation process. This annotation process is considerably more expensive than the annotation of English sentences as well. For example, one of the popular crowdsourcing platforms charges \$0.04 for annotators with no specification and \$0.65 for an annotator with at least a master's degree. Furthermore, the selection of annotators with the required expertise needs manual scrutiny using preliminary questionnaires and surveys which makes the process time-consuming. Finally, articulation agreements are developed across different departments, and it is imminent to provide a clear understanding of the model and its limitations to encourage automation of the process. The proposed model allows transparency and flexibility in the assessment of credit transfer and future research will focus on addressing these limitations by adding more course-to-course comparisons to the benchmark, developing domain-specific corpora, tuning the semantic similarity models with the aid of datasets, and also on ways to improve the generation of learning outcomes through automation.

## 6 Conclusion

The assessment of transfer credit in the process of student mobility is considered to be one of the most crucial and time-consuming tasks, and across the globe, various steps have been taken to mitigate

this process. With significant research and advancements in the field of natural language processing, this research article attempts to automate the articulation process by measuring the semantic and taxonomic similarity between learning outcomes. The proposed model uses a recent transformer-based language model to measure the semantic similarity and a clustering-inspired methodology is proposed to measure the taxonomic similarity of the learning outcomes. The model also comprises a flexible aggregation module to aggregate the similarity between learning outcomes to course-level learning outcomes. A benchmark dataset is built by conducting a survey among academicians to annotate the similarity and transfer credit decisions between courses from two different institutions. The results of the proposed model are compared with those of the benchmark dataset at different settings of leniency. The article also identifies the technical and domain-specific challenges that should be addressed in the field of automating articulation.

**Acknowledgement:** The authors would like to extend our gratitude to the research team in the DaTALab at Lakehead University for their support, Arunim Garg for his feedback and revisions on this publication. We 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.

**Funding Statement:** This research was supported by Ontario Council on Articulation and Transfer.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] S. Junor and A. Usher, "Student mobility & credit transfer: A national and global survey," *Educational Policy Institute (NJI)*, pp. 1–53, 2008.
- [2] A. Heppner, A. Pawar, D. Kivi and V. Mago, "Automating articulation: Applying natural language processing to post-secondary credit transfer," *IEEE Access*, vol. 7, pp. 48295–48306, 2019.
- [3] B. S. Bloom and D. R. Krathwohl, "Taxonomy of Educational Objectives: The Classification of Educational Goals," in *Handbook I: Cognitive Domain*, New York: David McKay Co., Inc., 1956.
- [4] L. W. Anderson and L. A. Sosniak, in *Bloom's Taxonomy*, Chicago, IL: Univ. Chicago Press, 1994.
- [5] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity-A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–37, 2021.
- [6] M. Kulmanov, F. Z. Smali, X. Gao and R. Hoehndorf, "Semantic similarity and machine learning with ontologies," *Briefings in Bioinformatics*, vol. 22, no. 4, pp. bbaaa199, 2021.
- [7] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, man, and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.
- [8] Z. Wu and M. Palmer, "Verbs, semantics and lexical selection," in *Proc. of the 32nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, Las Cruces, New Mexico, USA, pp. 133–138, 1994.
- [9] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," *WordNet: An Electronic Lexical Database*, vol. 49, no. 2, pp. 265–283, 1998.
- [10] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.
- [11] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543, 2014.

- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, “Attention is all you need,” in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Long Beach California USA, pp. 6000–6010, 2017.
- [13] J. Devlin, M. -W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186, 2019.
- [14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, 2019.
- [15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma *et al.*, “ALBERT: A lite BERT for self-supervised learning of language representations,” in *Int. Conf. on Learning Representations*, Virtual, pp. 1–17, 2020.
- [16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim *et al.*, “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [17] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3606–3611, 2019.
- [18] M. M. A. Qudar and V. Mago, “TweetBERT: A pretrained language representation model for twitter text analysis,” arXiv:2010.11091, 2020.
- [19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2021.
- [20] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.
- [21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.
- [22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan *et al.*, “Language models are few-shot learners,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
- [23] E. Strubell, A. Ganesh and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3645–3650, 2019.
- [24] L. Floridi and M. Chiriatti, “GPT-3: Its nature, scope, limits, and consequences,” *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.
- [25] S. Prasanna, A. Rogers and A. Rumshisky, “When BERT plays the lottery, All tickets Are winning,” in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Virtual, pp. 3208–3229, 2020.
- [26] A. Rogers, “Changing the world by changing the data,” in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, Virtual, pp. 2182–2194, 2021.
- [27] A. J. Swart, “Evaluation of final examination papers in engineering: A case study using bloom’s taxonomy,” *IEEE Transactions on Education*, vol. 53, no. 2, pp. 257–264, 2009.
- [28] G. A. Miller, “Wordnet: A lexical database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [29] D. Gerz, I. Vulic, F. Hill, R. Reichart, and A. Korhonen, “Simverb-3500: A large-scale evaluation set of verb similarity,” in *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*, Austin, Texas, pp. 2173–2182, 2016.
- [30] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Int. Conf. on Learning Representations*, Virtual, pp. 1–43, 2020.

- [32] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP/IJCNLP)*, Hongkong China, pp. 3973–3983, 2019.
- [33] Y. Shao, "Hcti at semeval-2017 task 1: Use convolutional neural network to evaluate semantic textual similarity," in *Proc. of the 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, pp. 130–133, 2017.
- [34] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi *et al.*, "A SICK cure for the evaluation of compositional distributional semantic models," in *Proc. of the Ninth Int. Conf. on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, European Language Resources Association (ELRA), pp. 216–223, 2014.
- [35] J. L. Taylor and D. Jain, "The multiple dimensions of transfer: Examining the transfer function in American higher education," *Community College Review*, vol. 45, no. 4, pp. 273–293, 2017.
- [36] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.
- [37] D. Chandrasekaran and V. Mago, "Comparative analysis of word embeddings in assessing semantic similarity of complex sentences," in *IEEE Access*, vol. 9, pp. 166395–166408, 2021. <https://doi.org/10.1109/ACCESS.2021.3135807>.
- [38] K. Schoepp, "The state of course learning outcomes at leading universities," *Studies in Higher Education*, vol. 44, no. 4, pp. 615–627, 2019.

## Appendix A.

**Table 5:** Performance of the proposed model on varying *sim\_threshold* parameter

S. no.	Human annotation	sim_threshold = 60	sim_threshold = 65	sim_threshold = 70
1	COMP 4121	COMP 4121	COMP 4121	COMP 4121
2	COMP 5341	COMP 5341	COMP 5341	COMP 5341
3	COMP 3321	COMP 3321	COMP 3321	COMP 3321
4	COMP 4321	COMP 4321	COMP 4321	COMP 4321
5	COMP 3519	COMP 3519	COMP 3519	COMP 3519
6	COMP 5470	COMP 5470	COMP 5470	COMP 5470
7	COMP 5471	COMP 5471	COMP 5471	COMP 5471
Agreement out of 7		4	6	4

**Table 6:** Performance of the proposed model on varying *impact* parameter

S. no.	Human annotation	impact = 20	impact = 30	impact = 40
1	COMP 4121	COMP 4121	COMP 4121	COMP 4121
2	COMP 5341	COMP 5341	COMP 5341	COMP 5341
3	COMP 3321	COMP 3321	COMP 3321	COMP 3321
4	COMP 4321	COMP 4321	COMP 4321	COMP 4321
5	COMP 3519	COMP 3519	COMP 3519	COMP 3519
6	COMP 5470	COMP 5470	COMP 5470	COMP 5470
7	COMP 5471	COMP 5471	COMP 5471	COMP 5471
Agreement out of 7		4	6	4