Efficient Detection of Potential for Research Collaborations and Educational Innovation through Semantic Textual Similarity (Short Paper)

Asaf Salman
The Hebrew University of Jerusalem
asaf.salman@mail.huji.ac.il

איתור יעיל של פוטנציאל לשיתופי פעולה מחקריים וחדשנות חינוכית באמצעות דמיון טקסטואלי סמנטי (מאמר קצר)

אסף סלמן
האוניברסיטה העברית בירושלים
asaf.salman@mail.huji.ac.il

Abstract

Educational innovation is the driving engine behind new possibilities for the next generations. While variations of the same innovative progress may manifest throughout history, their ability to transform into a game-changer is still limited. However, if we could efficiently and quickly detect educational studies that deal with a similar challenge, theory, or technology, then we can better collaborate for deepening our knowledge. As the whole is greater than the sum of its parts, such an outcome will probably promote educational innovations that are not exclusive to specific settings. This study demonstrates how it is possible to harness Semantic Textual Similarity (STS), a Natural Language Processing (NLP) task, for such purpose. A dataset consisting of 14,217 journal articles in English that focus on educational technology was mined from ERIC online database. The articles were respectively clustered into 4 equal periods, as their range covers the last 44 years (1978-2021). Each one of these clusters was divided again into 6 sub-clusters denoting a target Education (Early Childhood, Primary, Secondary, Post-Secondary, Special, and Teachers). The articles' abstracts (within the same sub-cluster) were compared against each other by computing a corresponding Cosine Similarity Score (CSS), based on applying a pre-trained sentence-transformers model. The findings indicate that: (a) The average CSSs of the articles can reflect a scope denoting (lack of) potential educational innovation - from studies that resemble each other to studies that (significantly) differ; (b) STS can be utilized as a state-of-the-art technique for efficient detection of potential productive collaborations between researchers who share the same educational interests.

Keywords: Educational Innovation, Collaboration, NLP, Semantic Textual Similarity.

Introduction

Educational innovation is a renewal in educational ideas, practices, methods, or objects that is deliberately attempted to improve our ability to solve educational problems and achieve educational goals (Mutaqin, 2021). However, implementing, adopting, and scaling up an educational innovation is a difficult and uncertain process due to both the educational system that resists change, and the complex organizational procedures involved in it (Berman & McLaughlin, 1976; Cohen & Ball, 2007; Varpio et al., 2012).

Advancements in information and communication technologies (ICT) enable us to look for new ways to promote educational innovation (Okoye et al., 2020; Karma et al., 2021), particularly through *Data Mining* (Garay et al., 2016) and *Machine Learning* techniques (Guns & Rousseau, 2014) to reveal the synergy between different research topics in order to construct a (scientific research) network of potential interdisciplinary collaborations within and across academic institutions. Although such techniques usually still require labeled data, they are more efficient and accurate compared to traditional discovery based on advanced search queries and (pre-)defined keywords or terms, which due to different levels of granularity may suffer from being too abstract or specific (Garay et al., 2016). For instance, as applicable in the search engine of the Education Resources Information Center (ERIC) (ERIC, n.d.).

In this study, I seek to demonstrate how we can practically promote the potential for research collaborations and educational innovation through harnessing *Semantic Textual Similarity* (STS), a *Natural Language Processing* (NLP) Task. This state-of-the-art technique neither relies on annotation nor is limited to (pre-)defined keywords, thereby enabling to fully automate the process, alongside replacing the binary classification (hit/miss) with a much richer and sensitive capture of context. Regarding Education, STS has already been proven to be beneficial for short answer grading (Sultan et al., 2016), cross-lingual plagiarism detection (Glavaš et al., 2018), and comparison of academic courses' materials for recommendations system (Seidel et al., 2020).

Methods

Data Collection

A dataset consisting of 14,217 journal articles in English that focus on educational technology was mined from ERIC online database, using its application programming interface (API). Each extracted record included the article's Title, Abstract, Author(s), Year of Publication, and Descriptors (terms that are assigned to every record in ERIC and reflect the subjects specified in the content). Even though it was possible to extract journal articles from a variety of educational study fields, this dataset focuses on innovation that is related to educational technology, and thus was limited to leading journal articles in this field (N=43). The range of articles covers the last 44 years (1978-2021).

Data Analysis

The dataset was preprocessed and analyzed in three phases. First, based on integrating relevant Descriptors in each article, I applied a computerized (automatic) Multiclass Classification of the articles into 1 of 6 synthesized labels (themes) denoting the target Education (Early Childhood, Primary, Secondary, Post-Secondary, Special, and Teachers). After that, the articles were double-clustered: [1] By their Publication Year to 1 of 4 clusters denoting Range of Years

(1978-1988, 1989-1999, 2000-2010, 2011-2021), as delimiting equal periods of at least a decade is a reasonable long enough time for prominent and feasible technology-oriented educational innovations to manifest; [2] Each one of these clusters was divided again into 6 sub-clusters denoting a target Education.

Second, a pre-trained sentence-transformers model (all-distilroberta-v1) was applied to estimate whether the content of a pair of articles' abstracts (within the same sub-cluster) is semantically close. This model maps sentences and paragraphs to a 768 dimensional dense vector space and can be used for clustering or semantic search by computing a *Cosine Similarity Score* (CSS) to the similarity of two texts (Reimers & Gurevych, 2019). The CSS ranges from 0 (when the content of the two texts are completely different) up to 1 (when the content is identical in terms of meaning). In addition, an efficient method of *Paraphrase Mining* was utilized during the comparison process (Reimers & Gurevych, 2019). After assigning a corresponding CSS for each pair of abstracts, I generated an *Average Cosine Similarity Score* (Avg_CSS) from all the CSSs of the same article (record). As a result, a reliable measure of each article's similarity (or divergence) degree to its relatives was obtained.

Third, I visualized the processed dataset through both a barplot (see Figure 1) and a swarmplot (see Figure 2) in order to gain insights, as it allows to represent and inspect the distribution of several data attributes systematically and efficiently.

Findings

The sampled dataset is relatively random due to uneven distribution of articles within and across clusters. Moreover, there is a gradual increase in the number of articles in each (sub-)cluster over the periods, which could be attributed to publication and accessibility issues. It is evident that Post-Secondary Education has attracted most of the academic attention, while studies regarding Teachers' Education have significantly increased in 2000-2021.

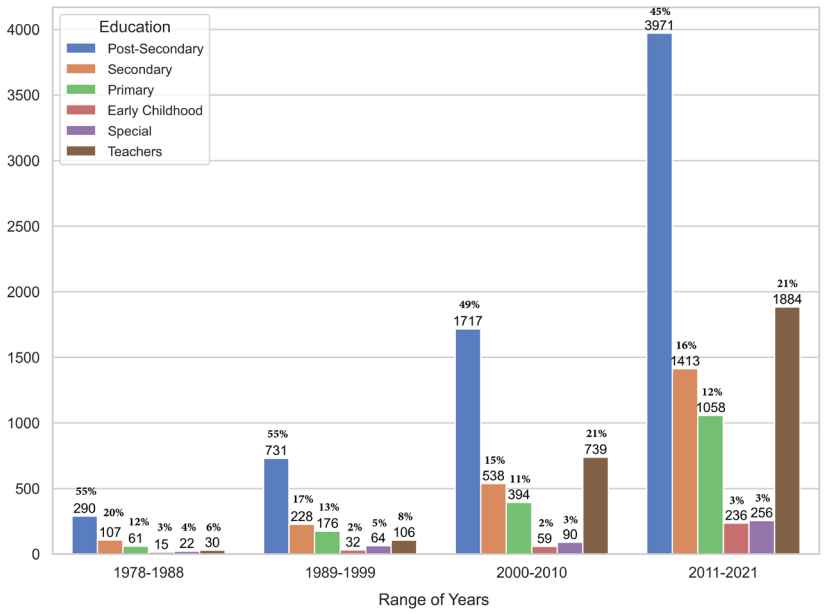

Figure 1. The distribution of articles in (sub-)clusters according to the following attributes: [1] Range of Years, and [2] Education.

First, the most prominent trend is the growing similarity (high Avg_CSS) among Post-Secondary Education articles over the periods. Thus, it may suggest that they are relatively addressing less and less diverse issues. Second, from 2011 to 2021, Teachers' Education articles become more similar to each other compared to previous periods. Third, suspectable articles that could denote potential educational innovations are likely to be found at the beginning of the range (Avg_CSS $\leq$ .25), as their content has significantly diverged from the mainstream. Especially, when the standard terminology in the field was intentionally not used, as it may indicate about innovative theory, approach, or pioneering definitions worthy to explore. Nevertheless, integrating other variables is definitely essential for more robust detection. Fourth, the results demonstrate that utilizing STS encapsulates the ability to cluster educational studies with similar content. Finally, let alone computing the CSS for each pair of abstracts enables to extract the highest among them, as indications for potential productive collaborations between researchers who share the same educational interests (see Table 1).

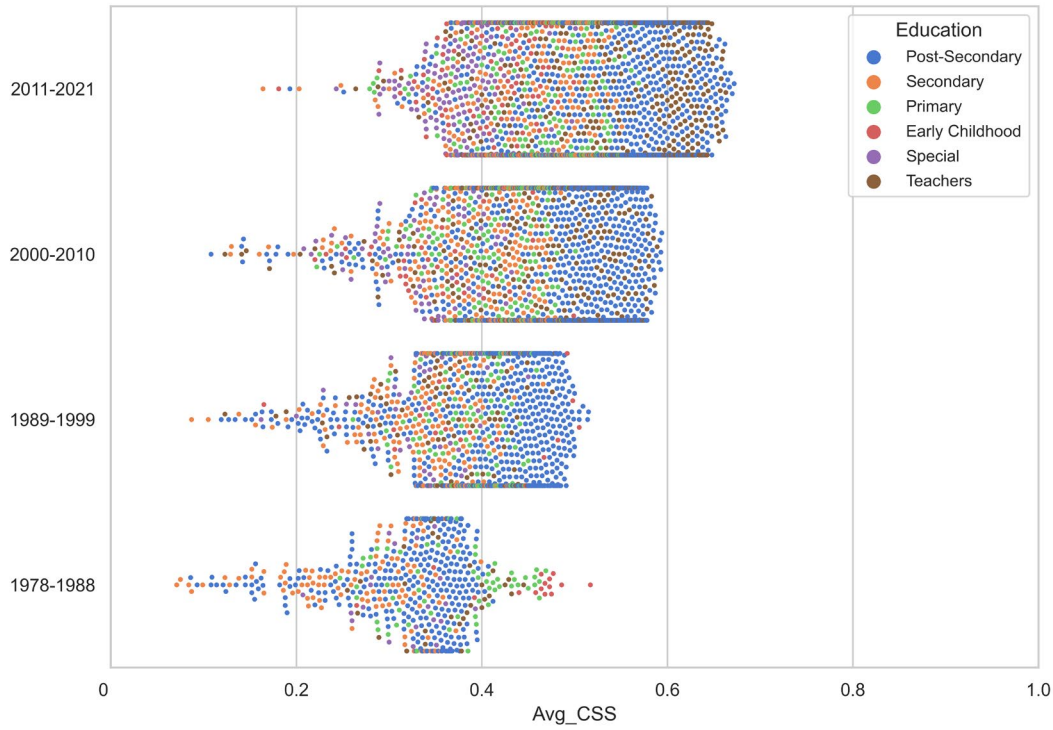

Figure 2. The distribution of articles by integrative allocation and clustering according to the following attributes: [1] Range of Years, [2] Education, and [3] Average Cosine Similarity Score (Avg_CSS). The y-axis denotes the four periodic clusters, while the x-axis denotes the Avg_CSS range (0-1). Each article is represented by one dot that is positioned according to the article's corresponding (sub-)clusters and Avg_CSS, while its color denotes the target Education.

Table 1. An example of pair of abstracts with a high *Cosine Similarity Score*.

Abstract's Text	Author(s)	Year of Publication	Education	CSS
This study of 902 boys and 828 girls in secondary school shows that gender differences in computer experience have a direct relationship to computer attitudes. Data gathered support the hypothesis that male students have more computer experience than female students and found boys showed more positive attitudes toward computers than girls.	Shashaani, Lily *	1994	Secondary	0.824
A study on gender and computer use surveyed 773 ninth-grade students from Tokyo (Japan) and from Stockholm (Sweden). Regardless of country, males reported higher scores of aptitude with and enjoyment of computers than females did. Overall, boys also showed more positive attitudes toward mathematics and sciences; girls consistently reported languages as their favorite subjects.	Makrakis, Vasilios & Sawada, Toshio **	1996	Secondary	

Note.

* Shashaani, L. (1994). Gender-differences in computer experience and its influence on computer attitudes. *Journal of Educational Computing Research*, 11(4), 347-367.

** Makrakis, V., & Sawada, T. (1996). Gender, Computers and Other School Subjects among Japanese and Swedish Students. *Computers & Education*, 26(4), 225-31.

Conclusion

This study shed light on the practical benefits derived from STS in efficient detection of potential for research collaborations and educational innovation. Utilizing such a technique is not my own initiative or original thought, as it is already used for many tasks in other research fields (e.g., Medicine, Computer Science, Psychology, etc.). Moreover, although STS was applied on a dataset with unique characteristics and limitations, its relevance and scalability are just a matter of sufficient data or required adjustments suitable for exploring potential

innovations in a specific subject. Exploiting STS could also promote collaborations that will deepen our knowledge on a variety of educational issues, which in turn will probably advance inclusive educational innovations that cross the boundaries of place, culture, and socio-economic status.

References

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. In *The Educational Forum* (Vol. 40, No. 3, pp. 345-370). Taylor & Francis Group.

Cohen, D. K., & Ball, D. L. (2007). Educational innovation and the problem of scale. *Scale up in Education: Ideas in Principle*, *1*, 19–36.

ERIC. (n.d.). *How does the ERIC search work?*. Retrieved November 26, 2021, from https://eric.ed.gov/?advanced

Garay, Y., Akbar, M., & Gates, A. Q. (2016, June). Towards identifying potential research collaborations from scientific research networks using scholarly data. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (pp. 217-218). IEEE.

Glavaš, G., Franco-Salvador, M., Ponzetto, S. P., & Rosso, P. (2018). A resource-light method for cross-lingual semantic textual similarity. *Knowledge-based systems*, *143*, 1–9.

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, *101*(2), 1461–1473.

Karma, I., Darma, I. K., & Santiana, I. (2021). Blended Learning is an Educational Innovation and Solution During the COVID-19 Pandemic. *International Research Journal of Engineering, IT & Scientific Research*.

Mutaqin, E. J. (2021, April). Educational Innovation. In *International Conference on Elementary Education* (Vol. 3, No. 1, pp. 163–171).

Okoye, K., Nganji, J. T., & Hosseini, S. (2020). Learning analytics for educational innovation: A systematic mapping study of early indicators and success factors. *International Journal of Computer Information Systems and Industrial Management Applications*, *12*, 138–154.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Seidel, N., Rieger, C. M., & Walle, T. (2020). Semantic Textual Similarity of Course Materials at a Distance-Learning University. In T. Price, P. Brusilovsky, S. I-Han Hsiao, K. Koedinger, & Y. Shi (Eds*.), Proceedings of 4th Educational Data Mining in Computer Science Education (CSEDM) Workshop co-located with the 13th Educational Data Mining Conference (EDM): Vol. 2734.*

Sultan, M. A., Salazar, C., & Sumner, T. (2016, June). Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1070–1075).

Varpio, L., Bell, R., Hollingworth, G., Jalali, A., Haidet, P., Levine, R., & Regehr, G. (2012). Is transferring an educational innovation actually a process of transformation?. *Advances in Health Sciences Education*, *17*(3), 357-367.