Big Data in Education

Shannamae Yang-Tay
School of Physical and Mathematical Sciences

Asst Prof Farhan Ali
National Institute of Education

Abstract - Papers published in reputable journals tend to be relevant to the current time. Therefore to learn about the values and ideologies in Education across time, we can analyse the papers published in each period. Given that millions of papers are published yearly, it is unrealistic to read all papers. Instead, we will use Natural Language Processing and Machine Learning to process this Big data. In our project, we aim to observe the changing trends and ideologies in Education by analyzing language used in over 1 million articles published from 1980 to 2020. We form networks with the tokenised keywords from the titles and abstracts of each paper and cluster them using k-means clustering. We created graphs for 5 periods of interest. In these graphs, VosViewer calculates the distance between nodes based on their correlation and we draw edges between nodes if their co-occurance hits specified thresholds. Furthermore, popular terms can be selectively labeled and the node sizes reflect how often they occur. Given these features, the network graphs produced are useful for instantly comparing values across time periods. For example, we can track topics in Education like Technology and Childcare by observing the node size and distance between the nodes “technology” and “childcare” across the graphs.

Keywords – Big Data, Networks, Natural Language Processing, Education

1 INTRODUCTION

As education is an ever growing and changing field, it is impossible to physically read and appreciate all published papers. Furthermore, if we had to choose only a few papers to review, many would likely favour papers written by renown authors. This creates a disproportionate valuation of papers published and offers a skewed perspective of the current climate in education. Our project aims to solve this by using Natural Language Processing to analyse all papers published. We assume that the title, abstract and key words of each article is an accurate representation of the papers’ contents thus extract terms from these sections of each article. Every article can be indiscriminately processed and used to construct maps and graphs that visualise the trends and links between terms. This paper will explain our methodology, and interpretation of the resulting term maps and network graphs. An Edge is a line drawn between 2 nodes. Edge strength will refer to the number of articles where the two neighbouring nodes co-occur.

2 LITERATURE REVIEW

There have been previous papers that utilized Natural Language Processing to process articles and papers. Flis and Van Eck a network visualizing software called VosViewer to convert thousands of articles from psychology journals into density maps and term maps. Their intention was to track the development of the study of psychology and visualize the upcoming trends. To make the analysis easier, they split their data into smaller time periods and creates a set of graph for each topic category in psychology. (Flis & Van Eck, 2018)

Emmanuelle Logette took a slightly different approach to processing thousands of papers for meaning. He mined a literature data base relevant to COVID-19 and extracted terms to create network graphs. Using minimum spanning trees and shortest path calculations, he discovered that in many perspectives such as protein, cell compartments and organ systems, higher glucose levels was linked to greater severity of COVID-19. (Logette et al, 2021)

3 METHODOLOGY

3.1 DATA AND TERM IDENTIFICATION

We used 1MILSED, a data set of 1,035,065 publications compiled from Scopus, Unpaywall, and PlumX. To visualize the changes across decades,

we separated the 1MILSED data into 5 decades, 1980-1989, 1990-1999, 2000-2009, 2010-2019 and 2020.

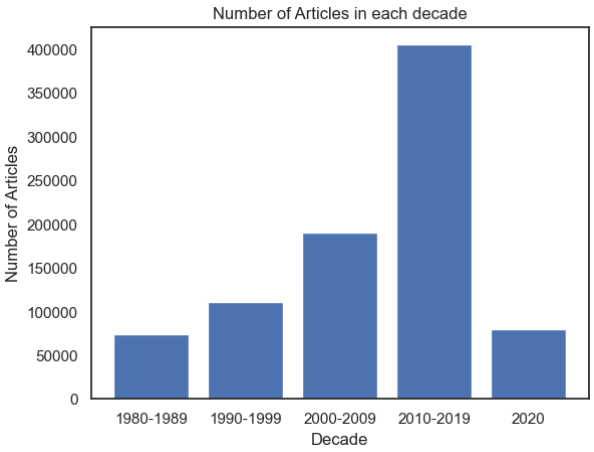

Figure 1: Number of Articles in each decade

Decade 5 only includes 2020 because 2020 has a comparable number of articles as the other decades. We want to ensure that any change across decades is not due to the appearance or disappearance of publishers. Thus, we only took articles from journals that appear in all decades. We focused on the movement and interactions of terms across decades so only terms that appear in all decades are included in our graphs. After this filtering, we have 304 terms in each decade.

3.2 Generating Coordinates

Network visualizing software, VosViewer, calculates the coordinates of terms. Term frequency and document frequency is used to calculate the relevance of terms as well as the distance from each other in a network. We used the following parameters in VosViewer to generate out coordinates and relevance scores: full counting, thesaurus, min occurrence 10, 100% relevancy, LinLog(2-1).

To create comparable maps, we orientated all maps to have the term “child” at 45 degrees from the x-axis in the top right quadrant.

4 TERM MAPS

We track where the top 50 global terms lie across the decades relative to each other. The closer they are, the more often they appear in the same article together. VosViewer generates a relevance score for each term. We selected 50 terms with the highest relevance scores and names this list the Global Top 50 terms. These Global Top 50 terms are labeled in the term maps for each decade. We can track the distance between nodes to understand how specific terms become more or less closely linked.

The node size corresponds to the link weight. Here, link weight refers to how many terms the particular term occurs with. The larger the node, the more terms it co-occurs with. We used the cubic function to scale the link weights. So the size of the node is calculates by cubing the number of terms that term co-occurs with.

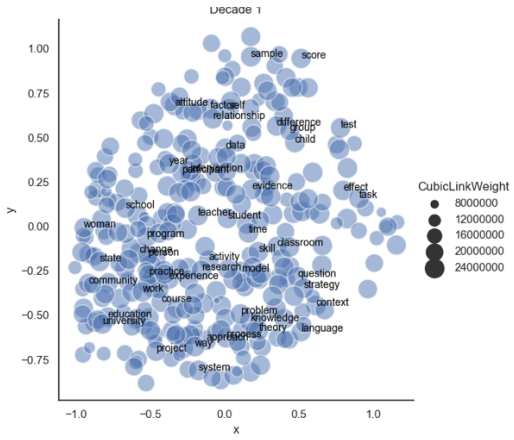

Figure 2: Decade 1 Term Map

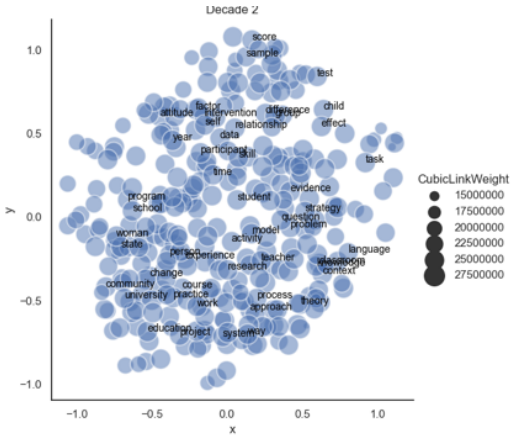

Figure 3: Decade 2 Term Map

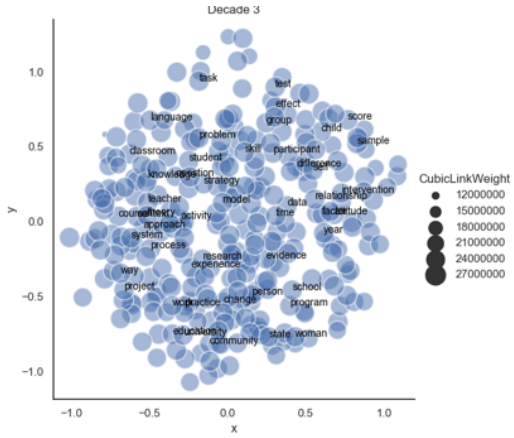

Figure 4: Decade 3 Term Map

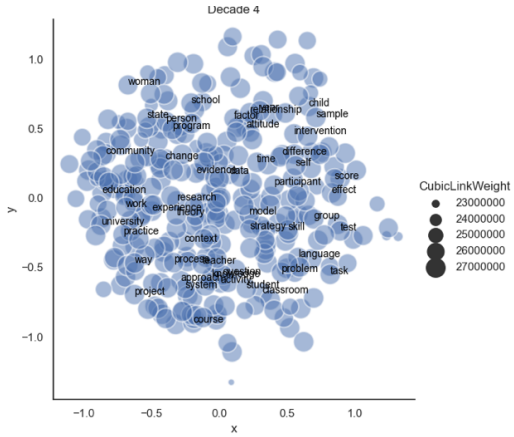

Figure 5: Decade 4 Term Map

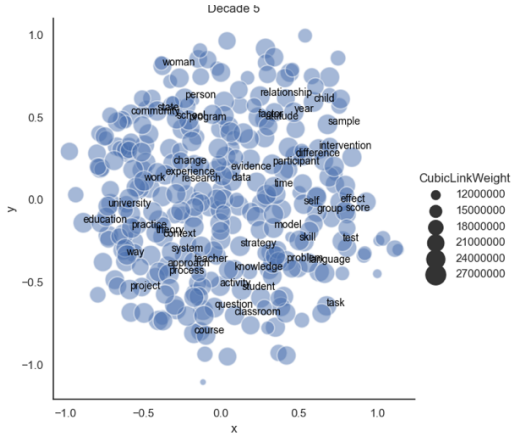

Figure 6: Decade 5 Term Map

5 NETWORK GRAPHS

We calculate the top 50 highest occurring terms in from 1980 to 2020 and call this the Global Top 50 terms. We set the threshold at 30 so that edges are only drawn if the two terms co-occur in over 30 articles. Edges are formed between two nodes if they occur in the same article. To sparsify the edges, we set a threshold for the minimum co-occurences. In this paper, we have set the minimum co-occurrence to 101 so that edges are only formed if the terms occur in over 100 articles. This allows us to visualise the close links to any selected term across the decades. For simplicity, if there is an edge between 2 nodes, we will say they are closely linked.

5.1 TOPIC: MATHEMATICS

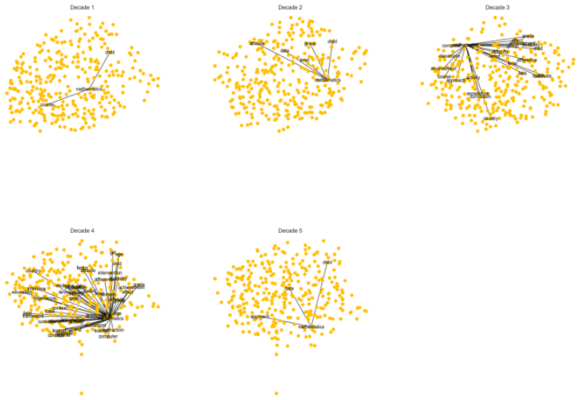

Figure 7: Terms with over 100 co-occurrences to “mathematics”

“mathematics” is only closely linked to “course” and “child” in Decade 1. In Decades 2, 3 and 4, “mathematics” becomes closely linked to more holistic terms like “attitude”, “level”, “disability”, “language” and “interest”. While Decade 1 displayed superficial understandings and priorities in teaching children math, Decades 2, 3 and 4 display more understanding of the difficulties and obstacles in teaching children math. The links to “interest”, “language” and “ disability” shows a heavier emphasis on understanding how to make math education more catered to the individial and those with disabilities.

“mathematics” is closely linked to “child” in all decades which makes it clear that the Education research community has consistently acknowledged that mathematics is important to

childrens' education despite the change in methodology and ideals.

In Decade 5, there are few edges however, this may be because this decade only consists on articles from 2020. Despite the sharp drop in close links in Decade 5, it still reflects the shift in focus from course work in Decade 1 to approaches in teaching children math in Decade 5.

5.2 TOPIC: SUCCESS

Figure 8: Terms with over 100 co-occurrences to "success"

Figure 8 shows the shift in the perception of "success" where the focus shifts from the results to the process. In Decades 1 and 2, "success" is closely linked to result orientated terms like "program" and "problem. On the other hand, in Decades 3 and 4, "success" is closely linked to process oriented terms like "strategy", "approach", "model", "process", "participant" and "challenge". We can also see a clear shift in the type of success the research and education community is interested in. In Decade 1, "success" is associated to individuals like "child" and "student". In the following Decades, we see that more value is places on success as a community and institution. In Decades 2, 3, 4, "success" is closely linked to more collaborative terms like "group", "community", "participant" and "institution".

5.3 TOPIC: TECHNOLOGY

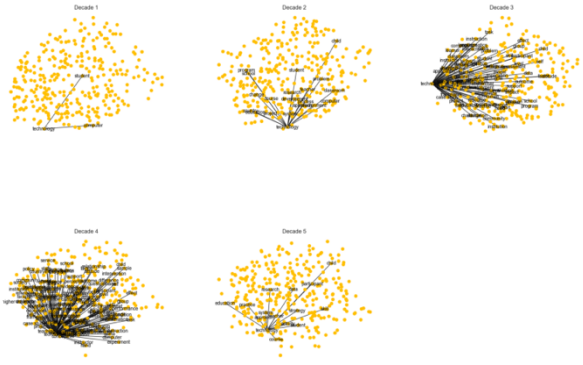

Figure 9: Terms with over 100 co-occurrences to "technology"

Figure 9 shows that technology was still unexplored by the Research community in the first decade. The only close links to "technology" are "student" and "computer". In Decade 2, "technology" is closely linked to "research", "approach", "development" which indicates a strong interest in technology in and its development in Education. Decade 3, shows more wide-spread and varies use of technology in education, with words like "community", "curriculum", "language", "mathematics", and "literacy". Decade 4 shows that there is a greater interest in the use and effects of technology in society with strong links to the words, "communication", "implication" ,"life", "literature, "role" and "relationship". Comparing Decade 1 and Decade 5, it is clear that the research community and educators in decade 5 place a higher value on technology in the field of education. By looking at the close links in each decade, we observe how the interest in technology has changed and the understanding of its value in education has deepened.

6 CONCLUSION

By using tokenising the relevant terms of each article and creating network graphs, we are able to easily visualise the linked between words of interest like "technology" without having to sift through all the papers. We can take these graphs a step further by creating accurate topic clusters so that we can instantly see how topic clusters move and grow throughout time.

ACKNOWLEDGMENT

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project. I would also like to acknowledge Vamsi Grandhi and Farhan Ali for creating the 1MILSED dataset

REFERENCES

Flis, I., & van Eck, N. J. (2018). Framing psychology as a discipline (1950–1999): A large-scale term co-occurrence analysis of scientific literature in psychology. *History of Psychology*, *21*(4), 334–362. https://doi.org/10.1037/hop0000067

Logette, E., Lorin, C., Favreau, C., Oshurko, E., Coggan, J. S., Casalegno, F., Sy, M. F., Monney, C., Bertschy, M., Delattre, E., Fonta, P.-A., Krepl, J., Schmidt, S., Keller, D., Kerrien, S., Scantamburlo, E., Kaufmann, A.-K., & Markram, H. (2021). A machine-generated view of the role of blood glucose levels in the severity of COVID-19. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.695139

APPENDIX

Node Labels from section 5.3: Technology

Decade 1: 'computer', 'technology', 'student'

Decade 2: 'approach', 'technology', 'change', 'child', 'classroom', 'computer', 'country', 'course', 'development', 'education', 'environment', 'problem', 'process', 'program', 'project', 'research', 'school', 'student', 'system', 'teacher'

Decade 3: 'activity', 'technology', 'application', 'approach', 'assessment', 'attitude', 'case', 'case study', 'challenge', 'change', 'child', 'class', 'classroom', 'communication', 'community', 'computer', 'context', 'course', 'curriculum', 'data', 'decision', 'development', 'education', 'educator', 'effect', 'effort', 'environment', 'evaluation', 'experience', 'factor', 'field', 'framework', 'goal', 'group', 'impact', 'implication', 'information', 'institution', 'instruction', 'instructor', 'interaction', 'knowledge', 'language', 'learner', 'learning', 'level', 'literacy', 'mathematics', 'model', 'need', 'opportunity', 'order', 'outcome', 'participant', 'person', 'practice', 'problem', 'process', 'program', 'project', 'question', 'research', 'researcher', 'resource', 'school', 'science', 'self', 'skill', 'strategy', 'student', 'support', 'system', 'task', 'teacher', 'teaching', 'technique'

Decade 4: 'activity', 'technology', 'analytic', 'application', 'approach', 'area', 'assessment', 'attitude', 'belief', 'benefit', 'case', 'case study', 'challenge', 'change', 'child', 'class', 'classroom', 'communication', 'community', 'competence', 'computer', 'concept', 'condition', 'content', 'context', 'country', 'course', 'curriculum', 'data', 'decision', 'development', 'difference', 'education', 'educator', 'effect', 'effectiveness', 'effort', 'environment', 'evaluation', 'evidence', 'example', 'experience', 'experiment', 'factor', 'faculty', 'feature', 'feedback', 'field', 'framework', 'game', 'goal', 'group', 'hand', 'higher education', 'idea', 'impact', 'implication', 'information', 'institution', 'instruction', 'instructor', 'instrument', 'interaction', 'interest', 'intervention', 'interview', 'knowledge', 'language', 'learner', 'learning', 'lesson', 'level', 'life', 'literature', 'mathematics', 'medium', 'model', 'need', 'network', 'opportunity', 'order', 'outcome', 'participant', 'participation', 'perception', 'performance', 'person', 'perspective', 'policy', 'practice', 'practitioner', 'problem', 'process', 'program', 'project', 'question', 'questionnaire', 'recommendation', 'relationship', 'research', 'researcher', 'resource', 'response', 'review', 'role', 'sample', 'school', 'science', 'self', 'service', 'simulation', 'situation', 'skill', 'society', 'state', 'strategy', 'student', 'support', 'survey', 'system', 'task', 'teacher', 'teaching', 'team', 'technique'

Decade 5: 'activity', 'technology', 'approach', 'child', 'course', 'data', 'education', 'participant', 'practice', 'research', 'skill', 'strategy', 'student', 'system', 'teacher'