International Journal of Engineering & Technology, 7 (2.4) (2018) 109-116

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



Using unsupervised machine learning to model tax practice learning theory

Alfred Howard Miller *

Higher Colleges of Technology
**Corresponding author E-mail:*

Abstract

The aim of this study was to utilize unsupervised machine learning framework to explore a dataset comprised of assessed output by Bachelors of Business, Taxation learners over four successive semesters. The researcher sought to motivate deployment of an evidence-supported, data-driven approach to understand the scope of student learning from a bachelor's degree in business class taxation class, as a tool for accreditation reporting purposes. Outcomes from the data analysis identified four factors; two related to tax and two related to learning. These factors are, tax theory, and tax practice, along with practical learning and theoretical learning. Research motivated a grounded theory paradigm that explained taxation class learner's scope of acquired knowledge. The resulting four factor model is a result of the study. The emergent paradigm further explains accounting student's readiness for career success upon graduation and provides a novel way to meet outcomes reporting requirements mandated by programmatic business accreditors such as required by the Accreditation Council for Business Schools and Programs (ACBSP).

Keywords: *Big Data Analysis; Computer Pattern Recognition; Taxation Learning Outcomes; Unsupervised Machine Learning; Tax Practice Learning Theory.*

1. Introduction

The knowledge gap being addressed in this study was how to develop alternative trend data that is less subjective and independent of grades, and GPA? A second thrust was to assess the career readiness of graduates of Higher Colleges of Technology, Fujairah Women's College [1]. Taxation class is offered in Semester 7 or 8 of the fourth year of the Bachelor degree. Value added tax was introduced in UAE for January 2018, and may be a sign of the introduction of future, additional tax regimes and modalities [2]. The study method stressed the objectivity of machine learning, to take precedence over methods with greater subjective human input.

The Accreditation Council for Business Schools and Programs (ACBSP), Standard 4 requires schools to report metric trend data measurement. Of prime importance is how they are meeting student learning outcomes. ACBSP does not accept grades or GPA as measurements requirement as these are too subjective. ACBSP instead seeks to motivate member institutions to find novel ways measure student learning. Furthermore, measurement of student learning outcomes is a Council for Higher Education Accreditation (CHEA) public reporting requirement [3].

The deployment of KH Coder software, for data mining/ context analysis, provided an avenue to identify what concepts the students were actualizing in their final reflections [4]. The process graphically modeled, actual taxation learning, that the students acquired. The researcher then explored patterns in the data to assess, what if any of the taxation student learning outcomes, were being met.

Thus, this study utilized the emergent theory research approach. The KH Coder coupled with the Stanford 'part of speech' (POS) tagger, was a scalable data mining application, suitable also for a big data deployment. The approach is a valuable toolkit for program representatives to analyze learning outcomes, and career readiness of graduates [4]. A second goal was to assess career readiness of accounting graduates in the area of taxation.

2. Literature review of previous studies

The machine learning process was used to evaluate data collected from end of term student reflections produced by learners at the completion of all course work, but prior to their attempting the final exam. Measurement of students' acquisition of an understanding of the taxation learning outcomes was applied to data collected from four successive semesters from 2015 through Summer 2017 of Taxation class. Content analysis was used to measure acquisition of key taxation concepts in Bachelor of Science in Accounting program. The course of study is the tax code of the United States. The students are predominantly female Emiratis, United Arab Emirates (UAE) nationals. The UAE enjoys a reputation historically as being a country with little to no taxation. However, this situation is changing and the UAE is set to introduce a 5% VAT or value added tax by 2018 [2].

One of the early advocates of analyzing the corpus or body of language was Susan Hunston [5]. "The main argument in favour of using a corpus is that it is a more reliable guide to language use than native speaker intuition is" (p. 20). Hunston [5] nevertheless valued intuition stating: "[Intuition] is an essential tool for extrapolating important generalizations from a mass of specific information in a corpus" (p. 22). The method is therefore a mixture a computer analysis and researcher intuition.

Stoykova [6] applied a search and retrieval approach to electronic versions of educational text corpus. The researcher used a statistical method to extend the keyword search, and was seeking fre-

Copyright © 2018 Alfred Howard Miller. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

quency distributions, collocation candidates, and to identify grammatical and semantic relations. The keyword search results, demonstrated lexical, grammatical, syntactic and/or semantic constraints. These constraints governed how words were used and combined. Word combinations were explored in their related semantic context. General keywords extracted were semantically relevant.

A group of researchers from Australia [7] sought to reduce subjectivity assessing student earning outcomes, via grading, through reduction of the human factors. Human variation was proposed to be reduced by a process called calibration. The main thrust of their method was to identify variation between graders in 1) areas of application skills and judgement, 2) workshops to narrow these gaps, and 3) if such workshops increased grader confidence in assessment. With the unsupervised machine learning method proposed for this study, human factors are divorced from the process and data analysis is objective, except to the degree necessary to interpret and construct the initial evaluation model. Other authors approached disparity in assessment from the standpoint of intellectual ability, procedural understanding and student motivation [8], [9]. None of these studies, which focused on accounting, used machine-based content analysis.

A previous study from this KH Coder stream of research, compared a section of N = 25 HR management learners to the first group of Taxation learners analyzed which was about the same size. Using a duplicated methodology, a clear-cut divergence was noted between topics reflected upon HR and Accounting learners. This is the expected outcome, and one which was statistically supported [10]. A following study in this same stream, an earlier iteration of the current study, triangulated evidence from the fact-based identification of credential acceptance and stakeholder satisfaction [11].

Anzai, and Matsuzawa [12] used the content analysis methods of word frequency and co-occurrence networks to explore differences in the mission statements of pre-World War II and post-World War II Japanese universities. Their key finding was a greater focus on research in the pre-war university model compared to a greater teaching focus in the post-conflict iterations. A moderating factor was the evolution of post-war universities through privatization to what is known as national university corporations and a greater focus on profitability.

Minami and Ohura [13] used KH Coder for content analysis of 35 student survey responses. They focused on 20 key responses to Question 11, "How Student's Attitude Influences on Learning Achievement? An Analysis of Attitude-Representing Words Appearing in Looking-Back Evaluation Texts". The author's explored a potential link between learner's attitude and achievement performance. The authors characterized their study as Education Data-mining (EDD) missed with Knowledge Discovery Data-mining (KDD). Low performers used language and words similar to high performers yet they were poorly implemented while the middle performance group had identifiable word differences. Their approach analyzing a limited number of student responses drew a parallel to the method deployed to analyze content—reflections in the case report on the SCIL model. Dissimilarities were use of a questionnaire, instead of discussion points and class focus on IT—information retrieval [13].

Authors based in Kuwait, reviewed 148 studies for a 2013 meta-analysis of educational data mining. They concluded that improved progress toward meeting learning outcomes was more likely. They also forecasted that data-driven methods such as modeling, clustering, prediction, relationship mining and distillation for human interpretation would lead to improvements in teaching and setting of policies [14].

Data deployment in accreditation is evident in Baldrige Criteria 4.1a, tracking overall organizational performance; WASC Standard 2, quantifying meeting of educational objectives; NCATE Standard 2, implementation of quality assessment systems; and ACBSP Standard 4, measurement of outcomes assessment and student learning outcomes. For example, for ACBSP, some applications would be tracking student information, mid-point assessments, end of program assessments, program review, along with stakeholder oriented data, such as alumni satisfaction and employer satisfaction [15].

3. Literature review of proposed method

Computerized analysis methods are gaining in popularity due to high data volumes. Trochim [16] traces the advent of theory-driven approaches, application to program evaluation and creation of theoretical frameworks aiding conceptualization methodology [16]. Studies today can draw upon techniques borrowed from search engine optimization and search engine marketing. A highly capable technology is MatLabs from Mathworks. In fact, computational thinking in business education is a boundary spanning technology when machine learning and numerical analysis capabilities of MATLAB are deployed outside of engineering sciences [17].

Vesanto & Alhoniemi [18] advocated the self-organizing map as being an excellent tool in exploratory data mining. They further explained the dendrogram supported nature of clustering using the self-organizing map (SOM). Furthermore, hierarchical cluster analysis uses dendrograms to specify clusters and as a technology, can reinforce the SOM.

According to Berinato [19], data visualization has become a must-have skill for all managers. This is because a visual abstraction is often times the only way to process the volume and velocity of data that arrives for processing. Furthermore, decision-making increasingly relies upon the ability to make-sense from and interpret of this voluminous amount of what is also known as big-data. Due to open source programs, the internet and proprietary, yet affordable tools, visualization is becoming widely accessible. Access to tools without the deeper understanding of their application can result in producing charts that are inadequate or ineffective.

Berinato [19] proposed that data managers and decision makers ask two questions. "Is the information conceptual or data-driven? and is the statement about the topic declarative or exploratory?" [19] (Page 1). Berinato's decision model identifies which of four types of visualization goals will be most effective, namely: "idea illustration, idea generation, visual discovery, or general data visualization" (Page 1) [19]. The implication of what has been stated by Berinato is guidance to the researcher conducting a study that is both quantitative and qualitative in nature [19].

The method to conduct data visualization with multidimensional scaling was explained by Buja, Swayne, Littman, Dean, Hofmann, and Chen [20]. Tamura [21], has used word analysis, co-occurrence analysis, and categorization, to examine a shared socio-cultural experience, from an internet discussion forum. Tamura set word analysis limits and reported high frequency words appearing 40 or more times. Using co-occurrence networks and the Jaccard coefficient, Tamura [20] intuitively created categories.

The Jaccard coefficient is a word frequency algorithm. It divides the frequency of word intersection by the union of word appearance. For example, if the frequency of word a is 4, and frequency of word b is 3, then the frequency of words a and b is 2. As stated as a formula; $2 / (4 + 3 - 2) = 0.4$ [21]. Matsuo, & Ishizuka 2004, p. 2). Tamura [20] reported that computer coding allowed a researcher to objectively handle large amounts of data, and qualitative information could be easily represented numerically which served to decontextualize the data. Tamura [20] considered decontextualization to be an advantage via removing human factors from a crucial portion of data analysis.

The Handbook of Business Discourse was a useful reference for motivating content analysis in the context of business studies. Systematic replicable methods were described for working with large data sets of text, apportioning words into categories, motivated by uniformly applied computer coding rules. Intuition as well as objective means can be used to process the output. The approach has commonality with other data analysis methods, as a means to identify patterns, and trends [5].

Content analysis methods are best applied when triangulated with other methods [22]. Yu, Jannasch-Pennell, and Di Gangi, [23] concurred. They stated that a computer aided quantitative approach to text-mining of natural language, was reliable and consistent, being compatible with grounded theory, provided the researcher remained neutral, as categories emerged from the data. Moreover, they justified content analysis because it could be used to validate evidence [16]. Trochim noted how widespread and interdisciplinary concept mapping methods had become [16].

KH-Coder [3], has been accepted in the U.S. Court of Law [24-25.] Text mining and content analysis has been used in over 600 studies, including studies from the Middle East and North Africa, such the mobile learning study of [25]. In the tradition of Vygotsky, text mining and content analysis can be used to scaffold collaboration of proficient students helping weaker ones, through adoption of pedagogic processes of performance improvement to construct socio-cultural theories based on student responses from activities [26] "Ultimately Algorithms are a set of instructions followed by computers to solve problems" said Flowerdew [27] However, O'Neil [28], challenges big data, and says rogue algorithms are often with too small of sample sizes with only a few dozen students.

Measurement of student learning outcomes can be analyzed using trend data normally reported for quality assurance reports and content analysis techniques such as KH Coder. KH Coder [4] permits coding-based text analysis and allows the researcher to deploy a range of techniques which include; word frequency analysis, hierarchical cluster analysis, co-occurrence network, multidimensional scaling and self-organizing map. While KH Coder is open source and user friendly, MatLab a proprietary program available at the college from August 2017 onwards offers greater computational and graphics capability. Both technologies permit user-coding and a user friendly interface. Time required to run statistical procedures can be expected to take anywhere from a several minutes for factor analysis to several hours for the iterative function of the self-organizing maps.

4. Method

Fujairah Women's College (FWC) is the largest Federal institution of higher education in the UAE with 23,000 students enrolled in one of 17 campuses located the country. Emirati nationals in the Accounting major take the BUS 4163 Taxation class course in Semester 7 or 8 during Year 4. Major assessments include an early formative exam worth 15% of the final grade, a midterm of 25%, along with a comprehensive final exam worth 30%. There is also a group project-based assessment, which is submitted prior to the final exam and it is valued at 30%.

The group project requires students to produce an applied interpretation a range of business case situations specific to the range of structures recognized under the US tax code and to explore the implications of a value added tax. The project concludes with a 750-word student reflection, learners comment upon on what they learned in taxation class and how they dealt with the challenges of the taxation course, referencing not only the difficulty of the material but also negotiating the group project [1]. This research study used a data mining-based, content analysis process deployed through KH-Coder to analyze the student reflections.

a) Preparation, Reliability and Validity of Data

The 86 reflections from Academic Year 2015 – 2017, Taxation class were assembled into a single plain text document (.txt). HTML code was used to separate each learners' reflection. Sample sizes were N = 23 (2015), N = 28 (2016), N = 9 (2017, Fall Semester) and N = 26 (2017, Summer 1), for a collective value of N = 86. In this case, a collective analysis motivated the identification of the breadth and uniformity in the content taught over the three years for which the four semesters data had been collected. Four periods of data are accepted as being sufficient to established a trend for accreditation reporting purposes [4].

Qualitative reliability was established as follows. There was only one researcher who coded the data and the process was re-visited, rechecking the coding each time new data was added. Data was added as another semester was completed. This ensured consistency. Repetitious words such as course or assignment titles and section headings provided by the instructor as well as irrelevant details such as student names, student numbers, and characters incomprehensible to the computer program were excised from the document. Next, a machine pre-processor function verifies that the data is properly scrubbed and ready to be read into the content analysis program. The preprocessor function of KH Coder is used to verify the readiness of the compiled data set for quantitatively analysis.

Before KH Coder was run, the software package, Stanford POS tagger (part-of-speech) was uploaded to provide KH Coder a textual frame of reference and specifications of how to proceed with the analysis [24]. Text-mined data from 86 reflections over four semesters were evaluated using a range of KH Coder [4] techniques. The methods included; word frequency analysis, hierarchical cluster analysis, co-occurrence network, multidimensional scaling and self-organizing map. [4]

b) Word Frequency Analysis

Academic Years (AY) 2015, 2016 and 2017 were compared through an analysis of word frequency arranged in the order of high to low. Higher-frequency words or keywords (KWs) reveal general themes from the taxation class final reflections. To limit data overload, KH Coder was set to extract 60 words for each analysis [9]. The part-of- speech (POS) tagger enabled KH Coder to render a frequency-based output by order of high to low frequency in Excel. By specifying POS tagger in the options menu, KH Coder identified the top frequency output of words by the categories of noun, proper noun, adjective, adverb, and verb [29].

c) Hierarchical Cluster Analysis

The cluster method used KH Coder's [Export Document-Word Matrix] which enabled the finding and analysis of word combinations with similar appearance patterns. Results are displayed in a family tree-like dendrogram, which were color-coded by group and connected by tree-branches. Familial relationships could be discerned visually [29].

d) Co-occurrence Network

The displays of co-occurrence networks used circles of varying size to represent keywords extracted from the reflective statements. Line thickness connecting these circles were demonstrative of the strength of these proximity relationship between associated keywords. The strength of association was represented by the thickness of lines connecting the words while the closer proximity, the more likely the words were to co-occur. The structure motivated a visual examination ascertaining the relationship of each keyword to the others and enabled a broader interpretation of the taxation reflections [29].

e) Multi-Dimensional Scaling

The method of Higuchi's [29] enabled extracted word relationships to be displayed in scatter plots in one, two or three dimensions. The [Export Document-Word Matrix] was a tool to find word combinations with similar appearance patterns. Kruskal graphing was the default mode, with Classical and Sammon modes available if the document became too crowded with overlapping words. Jaccard, Euclid and Cosine were available as a distance estimator options. Cluster analysis allows up to twelve user-specified clusters to be displayed; each with a different color. A two dimensional bubble plot is created which may assist in interpretation of groups.

f) Self-organizing Map

The mapping method used the KH Coder produced [Export Document-Word Matrix] coupled with the Euclidean distance calculation. As the map grew more complex, plotting more key terminology, a larger volume of text and a greater number of hexagonal coordinates; it can take over an hour to run on an i7 computer. The map can be color-coded whereas light blue (close proximity to its neighbors on the map) changes to white, which changes to shades of pink, which are most distant from its neighbors [29].

4.1. Analysis

a) Findings

The research results are relevant to the objectives. Word frequency analysis indicated a 70% similarity across the top 60 words and that five out ten of the top words matched across all four terms. These results are in line with the statistical analysis of variance (ANOVA) reported in [30] at 7th QS Maple, for a presentation with this global ranking body. The result obtained using the statistical test failed to reject the null hypothesis, that population mean differences among taxation class were not statistically significant. In contrast, a mean difference between accounting word frequency and that of a test group of HR learners revealed the means of both populations were not equal.

b) Co-occurrence Networks

These tools were used to produce data driven graphics to evaluate centrality and communality for frequently occurring words from the accounting learner data. To analyze centrality, the eigenvector method measuring the influence of a given node in network was utilized [31]. For the eigenvector centrality, co-occurrence network shown below on the left, the pink color of the words Tax, taxation and income, and to a lesser extent business and type indicated high centrality of these constructs. Colors closer to white such as government, pay, and learn were of medium centrality. Whereas, constructs depicted in blue, were of low centrality. Word networks are based on the proximity relationship with other words in the same network and are representative of unique constructs. Constructs below the dashed line in the below left Diagram 1 are taxation related and constructs above the dashed line are related to applied learning. Analysis is based on the approach of [29].

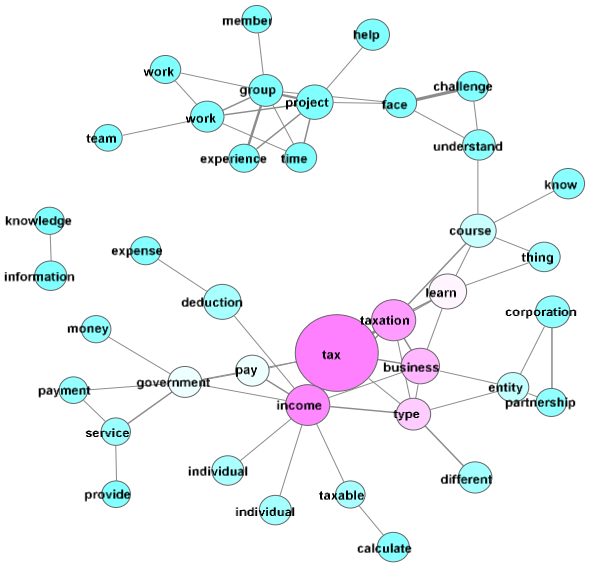


Fig. 1: Centrality Co-Occurrence Network [4].

The communality network shown in Figure 2 which follows, revealed that there were five distinct tax-related construct groupings; these are known as communities. These communities shown in Figure 2 outlined in a solid black outline and are colored as red, beige, purple, green and pink. Three learning related constructs were identified as being isolated from the tax-related concepts, just as they were with the centrality network. The learning constructs are identified by dashed oval and colored aqua, blue and gray. Expense and deduction along with knowledge and information represent distinct construct pairs only aligned with each other. Corporation, entity and partnership forms a similar grouping of three constructs [29].

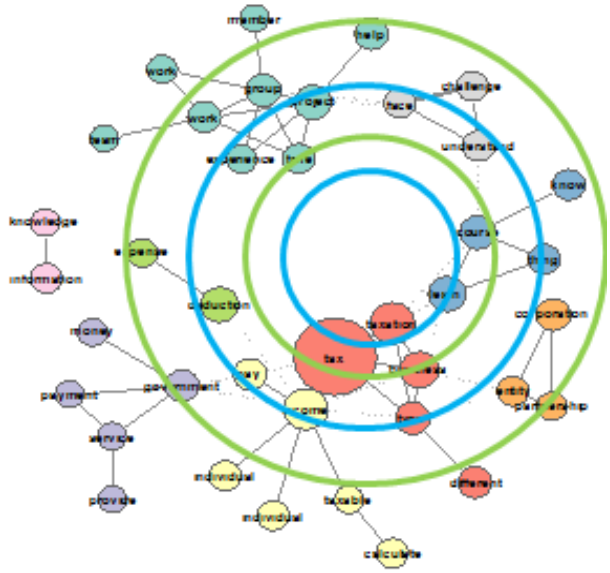


Fig. 2: Community Co-Occurrence Network [4].

c) Correspondence Analysis

This analysis was performed, and the output is shown in Figure 3. The correspondence analysis is interpreted by seeking to identify and understand distinctive characteristics. Distinctive characteristics are more distant from the origin (x, y) = 0. Non-characteristic terminology and ideas are proximally located to where the ordinate and abscissa approach zero, i.e. the origin (0, 0). The most distinctive interpretations are more distant and located beyond two nominal measurements of the x axis and beyond three nominal measurements along the y axis. Along Dimension 2, the constructs, taxpayer, gross and filing were distant along a positive measurement of Dimension 2 (ordinate) while the construct of Tax was distant from these constructs via a negative measurement. With respect to differentiation along Dimension 1 (abscissa), both tax-related constructs were distant from the negative location of new and its associated terminology, which was more related to applied learning.

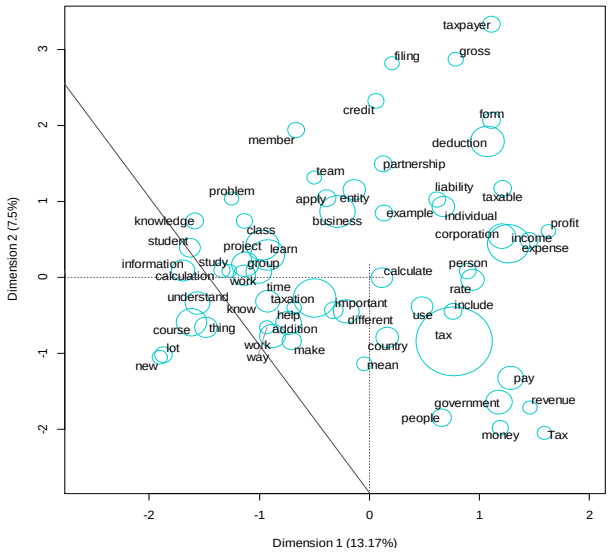


Fig. 3: Correspondence Analysis, Interpretation [32].

Constructs identified by Correspondence Analysis:

- Tax Practice construct = Taxpayer, gross & filing -> Credit, form, deduction, profit
- Tax Philosophy construct = Tax -> Revenue, money, people, pay, government
- Applied Learning construct = New -> Lot, course, thing, understand, information, calculation, student, knowledge, problem, member

d) Multidimensional Scaling

Multidimensional scaling motivated a visualization of the data set through display of individual outputs across nonlinear reduction [31]. Figure 4 depicts output graphed across two dimensions based on similarity. Dimension 1, the horizontal axis, shows the applied learning construct, consisting of terminology largely dissimilar from the taxation constructs. The applied learning construct numerically ranges from -.02 to -.06 along Dimension 1, while tax-related constructs can be found from about -0.1 on up to +0.6. The solid line separates applied learning from tax related text output. Dimension 2 at the strongly positive end of the dimension illustrates a group of constructs which are colored in red; these more general tax philosophy-based constructs, while the remaining dimension from +0.2 to -0.5 depicts the terminology of two sets of constructs which are of more practical applications. The group of turquoise colored tax constructs are more central while those that are beige color, are largely opposite of the applied learning group of constructs along Dimension 1.

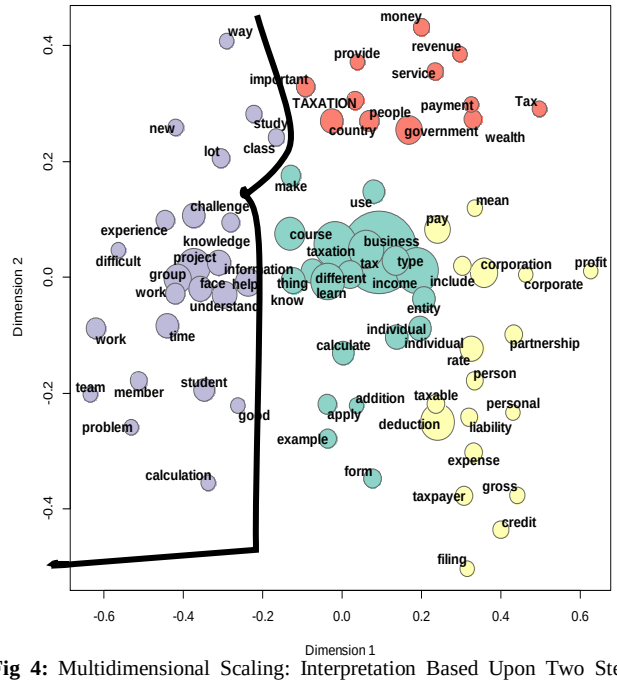


Fig 4: Multidimensional Scaling: Interpretation Based Upon Two Step Approach Of Higuchi [29], Buja Et Al. [20] &, Hofman, & Chen [32]. Heavy Line Separates Tax Topics From Learning Related Factors.

e) Hierarchical Cluster Analysis.

Interpretation of the hierarchical cluster analysis revealed four construct groupings, comprised of pairwise groupings that ultimate yielded two groups of sub-constructs in three cases and in one case three groups of sub-constructs. Each group of sub-constructs had a range of from four to eleven top words.

The first construct noted consisted of words focused on practical learning. There were two key groups of sub-constructs. Standout words for each applied learning construct included, project, group, work, time, student, experience, member and team understand and information in one case and understand, challenge, face, difficult, problem and calculation in the other case. The practical learning group of constructs are depicted in Figure 5 below.

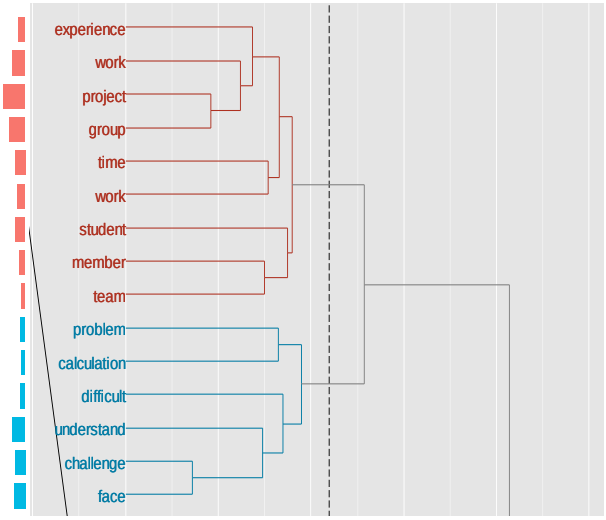


Fig. 5: Hierarchical Analysis: Practical Learning Constructs, [4].

The next grouping of constructs relates two groups of sub-constructs that portrayed students' understanding of the basic philosophy of taxation and these are shown in Figure 6. The most significant words of the tax philosophy construct were government, wealth, revenue, provide, payment and service. income tax, government and country in the case of the first group and country, use, important, people, form, apply, study, money
class, tax and taxation for the second group. The second group of sub-constructs had eleven high frequency words, which equaled another group in total and together they equaled being the largest grouping of sub-constructs.

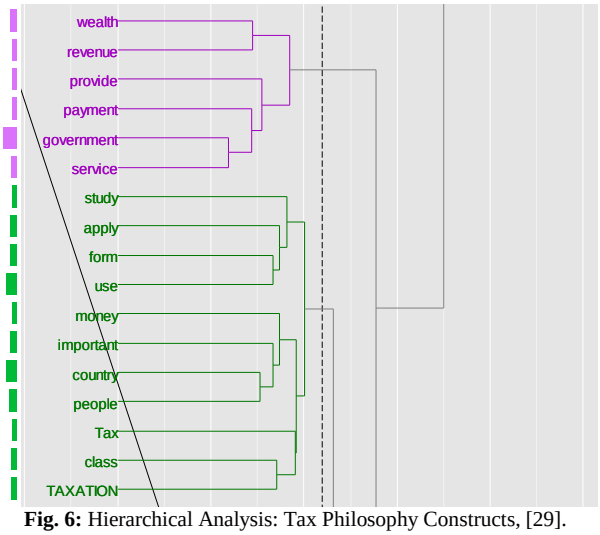


Fig. 6: Hierarchical Analysis: Tax Philosophy Constructs, [29].

The next grouping of constructs appears on the following page as part of Figure 7, and they relate to tax practice. There were three constructs that defined tax practice, namely deduction, taxpayer, expense, example, person, include, credit mean, gross, filing and personal for the first sub-construct. The second construct consisted of corporation, entity, partnership and profit. The third construct included, tax, income, pay, individual, calculate, rate, corporate, and liability. One of the construct with the most individual words was also listed under tax practice where the first group of sub-construct was defined by 11 high frequency words. The tax practice group of construct is shown in Figure 7. Two of the highest frequency word groups appeared in the third set of sub-constructs namely, Taxation and income. These were the highest frequency words of the entire hierarchical analysis, and this group is shown in olive drab.

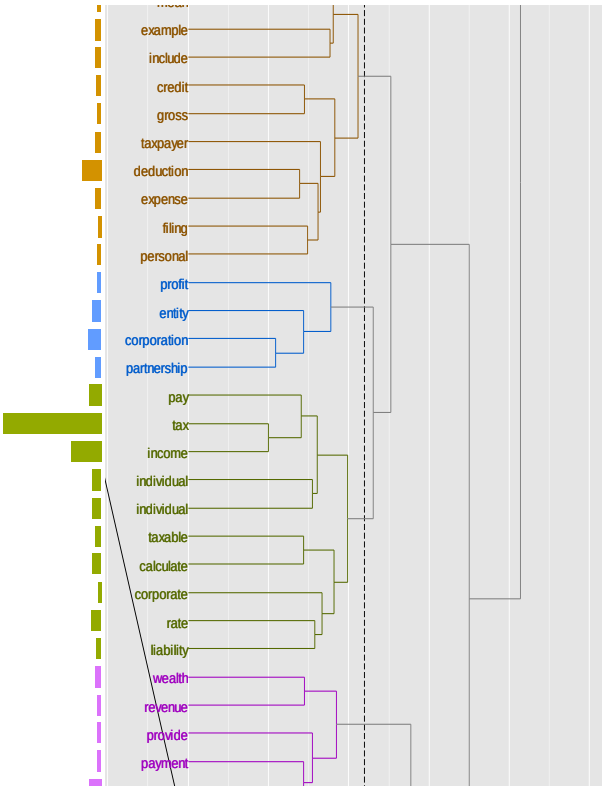


Fig 7: Hierarchical Analysis: Tax Practice Constructs, [4].

The final grouping of constructs appears on the following page as part of Figure 8, and they relate to the theoretical learning dynamic. There were two constructs that defined theoretical learning. One of which, the first, was probably the most prevalent sub-construct overall, going on the strength of the keywords. The first sub-construct included the following key words, namely taxation, learn, business, course, help, know, different, type, thing, and addition. The other sub-construct included the following keywords, information, knowledge, new, make, lot, good, and way.

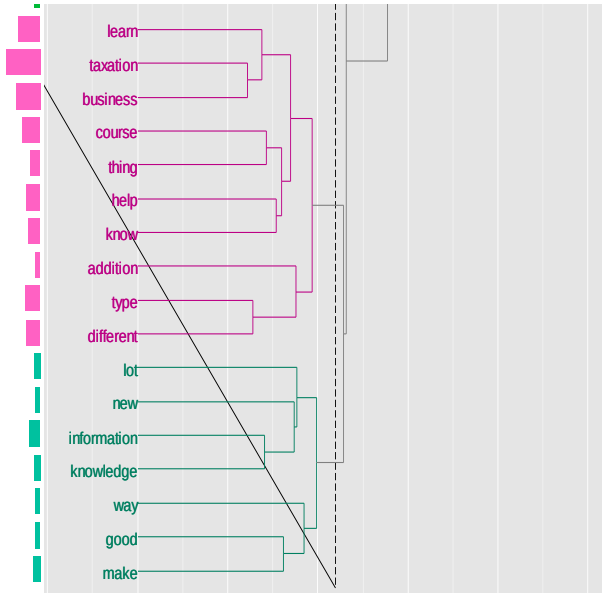
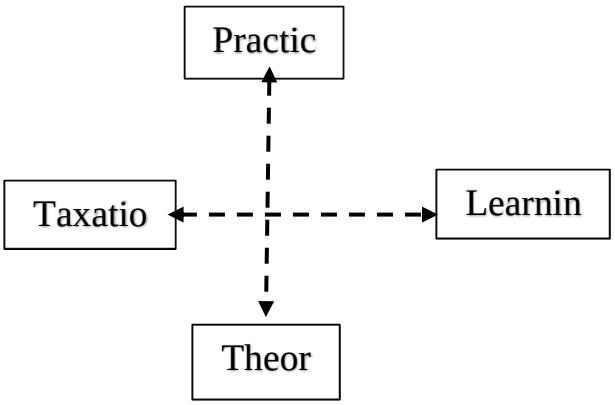


Fig. 8: Hierarchical Analysis: Practical Learning Constructs, [4].

f) Self-organizing Map.

This tool creates an artificial neural network and is constructed using a learning algorithm via unsupervised machine learning [33]. The interpretation of the self-organizing map shown in Figure 9, considers the following factors; pink color asserts a large difference in vectors to the neighboring nodes. Purple colored nodes, followed by green and blue represent the other three constructs.

The cluster method was utilized and iterations were applied for clusters through 3 to 9 achieving a best fit with 4 clusters. The constructs diametrically oppose each other pair opposites and relatively distant. Constructs of contained within the pink zone are representative a theoretical understanding of tax or tax philosophy. The mapped zone shaded green represents a practical understanding or tax practice construct and it bridges between the pink zone of tax philosophy and the applied and theoretical learning construct zones represented respectively by purple and blue color-coded zones. The mapped blue zone is representative of an internalized theoretical understanding, while the purple zone on the map is a categorization of the practical group experience.



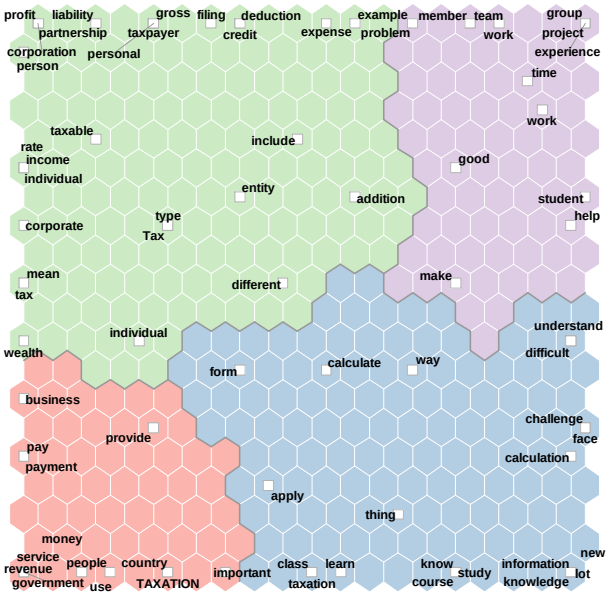


Fig. 9: Self-Organizing Map, [4].

5. Conclusion

The research results are relevant to the objectives. Computer aided text-mining of natural language, was a reliable and consistent quantitative approach; compatible with grounded theory, to validate evidence gathered for content analysis [16]. The researcher maintained a neutral stance when conducting interpretation of the data, so as to correctly identify relevant emergent properties toward development of the grounded theory. Hunston [6] an early advocate of text analysis, found using corpora more reliable than intuition, but that intuition was necessary for "extrapolating important generalizations" p. 22. Currently new data is being evaluated that was collected for 24 current additional learners over Fall 2017, with more data to be collected Spring and Summer semesters of 2018. New data helps diminish the weakness of the small sample size and motivates potential further validation and testing, to advance the proposed model moving forward.

The dendrogram produced using hierarchical cluster analysis supported the self-organizing map produced by the unsupervised ma-

chine learning process. The emergent grounded theory from this study, is a factor model describing the reflected experience of Year 4 Bachelor's Degree Emirati learners from their taxation class at Fujairah Women's College. The model advances four factors, theoretical learning, practical learning, tax philosophy and tax practice. The face validity of these factors is supported by the terminology resent in student reflections submitted with the final project. Emirati learning taxation is important because the UAE is introducing a more comprehensive taxation regime, with a 5% VAT tax being introduced in 2018. The model explains Emirati national motivation to not only learn taxation in an applied sense, but to also prepare themselves from a practical standpoint to aspire to tax-related workplace roles such as being auditors, tax preparers and accountants. See Figure 10 for the proposed grounded theory Taxation Learning Model for this study.

The tools available in KH Coder [4] enable a neutral machine-driven quantitative analysis, of factual evidence. The machine learning output can be interpreted intuitively for validation learning outcomes. The qualitative assessment of the quantitative outcomes enables schools to monitor the program and report effectively to accreditation authorities, through initial, reaffirmation and quality assurance reports, and to their stakeholders, meeting public outcomes reporting requirements [4].

Computer-aided content analysis yielded distinction of two learning factors, not previously revealed by this stream of research. These are the practical learning, and theoretical learning factors emergent from a reassessment of the integrated learning factor reported in two previous studies.

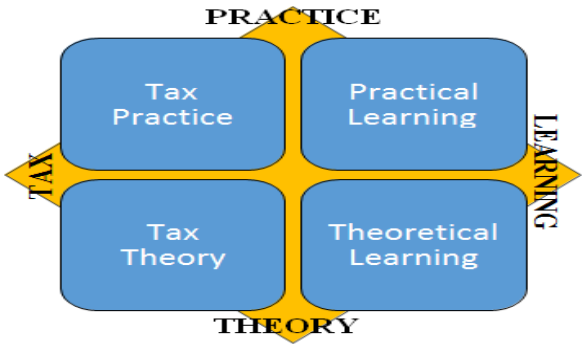


Fig. 10: Theoretical Model: The Tax Practice Learning Theory.

References

[1] Higher Colleges of Technology (2016). Retrieved from http://www.hct.ac.ae/ (2017).

[2] United Arab Emirates Ministry of Finance (2017). VAT, retrieved from https://www.mof.gov.ae/En/budget/Pages/VATQuestions.aspx

[3] ACBSP (2018). Baccalaureate-Graduate Degree Accreditation, Retrieved Januray 29, 2018 from http://www.acbsp.org/?page=baccalaureate

[4] KH Coder, (2017). Open source software, Higuchi, Koichi, Ritsumeikan University, Japan. Available at http://khc.sourceforge.net/.

[5] Hunston, S. (2010). Corpora in Applied Linguistics, Cambridge University Press.

[6] Stoykova, V. (2017). Extracting Academic Subjects Semantic Relations Using Collocations, EAI Endorsed Transactions on Energy Web and Information Technologies 17(14). DOI: 10.4108/eai.4-10-2017.153161

[7] O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B, & Watty, K. (2015). Does calibration reduce variability in theassessment of accounting learning outcomes? Assessment & Evaluation in Higher Education. DOI: 10.1080/02602938.2015.1008398

[8] Andartari A., Susanti, s., & Andriani, V. (2013). Effect of Intellectual Capabilities (IQ) and Learning Motivation at the Results of Accounting Subject on SMA Labschool Rawamangun. DOI10.21009/JPEB.001.1.1

[9] Xiong, Y., Zhou, H. & Ogilby, S. M. (2014). Investigation of the Effects of Cognitive Elaboration on Accounting Learning Outcomes, Journal of Education and Learning; 3(4); DOI: 10.5539/jel.v3n4p1

[10] Miller, A. H. (2016). Computer-Aided Content Analysis of the Corpus of Business Discourse: A Comparison of Accounting and HR Learners, NETs 2016 Osaka Japan July 25, 2015.

[11] Miller, A. H. (2017). Preparing Students for Career Success in Accounting: The SCIL-based Model with a Focus on Content Analysis, Transnational Journal of Business, Retrieved from: http://www.acbsp.org/members/group.aspx?id=143359

[12] Anzai, S., & Matsuzawa, C. (2013). Missions of the Japanese National University Corporations in the 21st Century: Content analysis of mission statements - Academic Journal of Interdisciplinary Studies, 2013 - mcser.org

[13] Minami, T., & Ohura, Y. (2015). How Student's Attitude Influences on Learning Achievement? An Analysis of Attitude Representing Words Appearing in Looking Back Evaluation Texts, International Journal of Database Theory

[14] In addition, Application, 8(2), 129-144. Retrieved from http://dx.doi.org/10.14257/ijdta.2015.8.2.13

[15] AlShammari, I. A., Aldhafiri, M. D., & Al-Shammari, Z. (2013). A meta-analysis of educational data mining on improvements in learning outcomes. College Student Journal, 47(2), 326-333.

[16] DePape, J., Lockard, N. & Laramy, R. (2007). Using Accreditation Self-Study Results to Better Understand Student from Recruit through Alumnus. The Center for Teaching and Learning, Preparing Facilitators of Learning for a Diverse World. Take the Credit. Retrieved from http://www.cair.org/wp-content/uploads/sites/474/2015/07/

[17] Trochim, W. M., (2016). Hindsight is 20/20: Reflections on the Evolution of Concept Mapping. Evaluation and Program Planning. DOI: 10.1016/j.evalprogplan.2016.08.009

[18] Gross, S., Kim, M., Schlosser, J., Mohtadi, C. Lluch, D., & Schneider, D. (2014). Fostering computational thinking in engineering education: Challenges, examples, and best practices. 2014 IEEE Global Engineering Education Conference (EDUCON), pp 450 - 459. DOI: 10.1109/EDUCON.2014.6826132

[19] Vesanto, J. & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 11(3). Retrieved from http://ftp.it.murdoch.edu.au/units/ICT219/Papers%20for%20transfer/papers%20on%20Clustering/Clustering%20SOM.

[20] Berinato (2016). Visualizations That Really Work. Harvard Business Review, Retrieved from https://hbr.org/2016/06/visualizations-that-really-work Laramy.pdf

[21] Buja, A., Swayne D. F., Littman M. L., Dean, N., Tamura, T. (2011). Application of text-mining methodology to sociological analysis of internet text in Japan. Retrieved from http://www.cajs.tsukuba.ac.jp/monograph/articles/01_201103/cajs01_201103_077-097.pdf

[22] Matsuo, Y, & Ishizuka, M. (2004). Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools, 13(1) pp 157–169. Retrieved from https://www.researchgate.net/profile/Mitsuru_Ishizuka/publication/2572200_Keyword_Extraction_from_a_Single_Document_using_Word_Cooccurrence_Statistical_Information/links/02e7e522976acdaa9e000000.pdf

[23] Bargiela-Chiappini, F., Nickerson, C., & Plancken, B. (2008) Business Discourse. Retrieved from http://www.palgraveconnect.com/pc/doifinder/10.1057/9780230627710

[24] Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. The Qualitative Report, 16(3), 730-744. Available from http://nsuworks.nova.edu/tqr/vol16/iss3/6/

[25] Pelet, J-E, Khan. J., Papadopoulou, P., & Bernardin, E. (2014). M Learning: Exploring the Use of Mobile Devices and Social Media, in (Ed) Baporikar, N. (2014). Handbook of Research on Higher Education in the MENA Region: Policy and Practice, 261-296. Hershey, PA: IGI Global.

[26] Posner, R., (2012) Opinion, United States of America, Plaintiff-Appellee, v. Deanna L. Costello, Defendant-Appellant, No. 11-291 U.S. Court of Appeals Seventh Circuit Court.

[27] Smith, G. (2014). Data and Intuition, The Conglomerate, Retrieved from http://www.theconglomerate.org/corpuslinguistics/

[28] Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation* International Journal of Corpus Linguistics, 14(3), 393-417. DOI:10.1075/ijcl.14.3.05flo

[29] O'Neil, C. (2016). 'Rogue Algorithms' and the dark side of big data. Knowledge@Wharton. Wharton, University of Pennsylvania. Available from http://knowledge.wharton.upenn.edu/article/rogue algorithms-dark-side-big-data/?utm_ source=kw_newsletter&utm_medium=email&utm_campai n=2016-09-22

[30] Higuchi. K. (2016). A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial using Anne of Green Gables (Part I) Ritsumeikan Social Sciences Review, p. 77-91. Retrieved from http://www.ritsumei.ac.jp/file.jsp?id=325881

[31] Miller, A. H. (May 2017). Assessing work readiness in accounting graduates via the SCIL-based model. Seventh QS_MAPLE Conference Dubai, UAE, May 01- 04, 2017. Retrieved from: http://www.qsmaple.org/7thqsmaple/

[32] Borg, I, & Groenen, P. (2005). Modern multidimensional scaling: Theory and applications (2nd Ed.). New York, NY: Springer-Verlag.

[33] Hofman, H., & Chen, L. (2008). Data Visualization with Multidimensional Scaling. Journal of Computational and Graphical Statistics, 17(2), 444-472. Doi: 10.1198%2F106186008X318440 [24]. Austin, D. (2017).

[34] Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics. 43 (1): 59–69. Doi: 10.1007/bf00337288.