The mathematics approaches in the analysis of mathematics syllabuses similarities among Japanese and Bosnia and Herzegovina universities

The mathematics approaches in the analysis of mathematics syllabuses similarities among Japanese and Bosnia and Herzegovina universities

M Milijevic[1*] and D Preradovic Kulovac[2]

[1]University of Banja Luka, Faculty of Economics, Department of Quantitative Analysis and Informatics, Majke Jugovica 4, 78 000 Banja Luka, Bosnia and Herzegovina
[2]Corvinus University, Doctoral School of of Economics, Business and Informatics, Fovam ter 8, 1093 Budapest, Hungary

*E-mail: mirjana.milijevic@ef.unibl.org

Abstract. The goal of our research is improvement of mathematics curriculum and popularization of mathematics among students of economics in developing countries. We analyze and compare curricula of pure mathematics courses that are taught to university students of faculties of economics in Japan and in Bosnia and Herzegovina. Data set contains math syllabuses in 2021/22 school year from six public universities in Bosnia and Herzegovina and seven from Japan. The text corpus was pre-processed and then the Term Frequency – Inverse Document Frequency algorithm, and Sentence Transformed Multi QA model were applied to build word vectors, find the similarity among Japanese and Bosnia and Herzegovina mathematics syllabuses using cosine similarity approach, and to find the key competences of these two countries mathematics syllabuses using the word cloud. Our results show the following similarity between the curricula: 60.7 percent using TF-IDF and 80.3 percent using Multi QA model. The key competences in the Japanese mathematics course are narrow and focused, in contrast to Bosnia and Herzegovina's.

1. Introduction

Mathematics and economics are strongly related disciplines. Nowadays we can see the growth of mathematical economics as a discipline, where economists use mathematical models for their research. On the other hand, economists with a quantitative background are often employed as consultants to advise business, industry, or government.

During their teaching at Faculty of Economics, University of Banja Luka, the authors of this paper have noticed insufficiently mathematics performance of their students. Besides, it was noticed that students are not motivated to learn mathematics, they do not find it interesting (Figure 1) and important for their future studies and jobs (26.8 percent of students think that mathematics is not needed in their future profession). Those were findings from bachelor students survey distributed via faculty email at beginning of the Spring semester of 2021/22 school year at Faculty of Economics, University of Banja Luka.

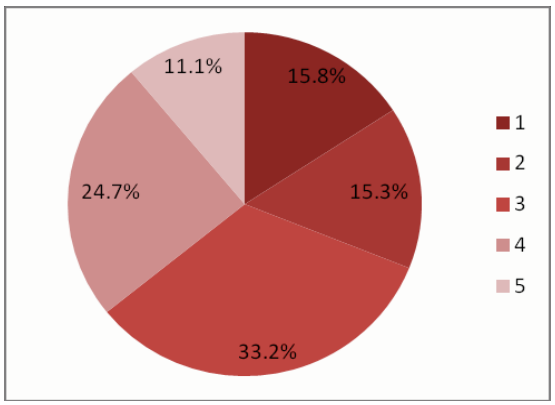

Figure 1. Student responses to the question of whether they found the Mathematics for Economists course interesting on a scale of 1-5, with 1 representing not interesting at all, 5 representing very interesting course, and 3 representing undecided, sample 190 students.

Students cannot relate mathematics course content with the real life problems, especially economics. We think that this is an alarming problem and we want to try to understand the reasons behind it and to give a strategy to overcome it. This paper is a first step toward answering the question "What is the best methodology of teaching mathematics to students of economics that we could use to strengthen their mathematical competences?" To understand the phenomena, we started from curricula, by applying different research approaches to curricula mining.

We can find a lot of research work in this direction. Researchers that we are referring to are mainly professors of economics courses that were facing failure of their students to master the concepts of finance and managerial economics. It is interesting to notice that in 80's, 90's, 2000's and later, we find the same problems when it comes to mathematics teaching to students of economics. Moreover, we can see that this problem persists in different countries too. In [1] it was shown that the performance of students in introductory economics courses is related to basic mathematical knowledge, especially in algebra and arithmetic. Schaffer and Calkins [2] showed that there exists a strong relationship between students' performances in introductory business finance and their mathematical background. Furthermore, in [3] the authors used standard regression methods to quantify students' background and finance class performance. They asked whether poor students' performance in finance comes from students' deficient mathematical skills or from students' lack of finance intuition.

Firstly, we want to see if an improvement of mathematics curriculum could lead to better understanding of the subject. We analyzed and compared curricula of Mathematics for Economics that is taught to students of economics in Japan, as a representative of developed country and in Bosnia and Herzegovina, as a representative of developing country, using mathematics approaches in text mining.

The field of text mining encompasses various ways of analyzing text, distinguishing between plain and structured text mining approaches. According to Agrawal [4], plain text mining techniques include document and information retrieval, text categorization, similarity, and text summarization. In our research, we used similarity analysis and text summarization. Similarity analysis aims to determine the similarity and/or distance between documents. It can be determined using predefined terms. Text summarization is a text mining technique that provides results suitable for human interpretation and suitable for visualization because they are based on graph methods, the most commonly used being cosine similarity with TF-IDF [5].

This paper is divided into four parts: Introduction, where we present the rationale for the research and the focus area, Methodology, where we present the research design and the models we used. In the third part, we summarize the data sources and their structure. In the last part of the paper, Results and Discussion, we report our calculations and provide directions for further research as well as the limitations of this paper.

2. Methodology

Curricula collected from universities in Bosnia and Herzegovina were translated into English (without source list, i.e. literature for students) and loaded into a .txt file. All syllabi from Japanese universities are summarized in a second .txt file.

Preprocessing and modeling of the data is done in Python in the following order: we lowercased the text, removed punctuation marks and stop words. Stop words were removed using the Gensim unsupervised learning library, which was extended to include additional defined stop words: week, university, and university names.

The text corpus was then tokenized and lemmatized using the Natural Language Toolkit (NLTK) library, which is highly rated among other tokenization models in the field. Tokenization in text mining involves breaking the text into parts, such as words or phrases. These parts are called tokens [6], [7].

We preferred lemmatization to obtain the base of words (dictionary form) with different endings, asthis approach gives slightly better results compared to stemming [8].

In the following steps, we vectorized the word corpus using the TF-IDF approach. TF-IDF stands for term frequency-inverse document frequency and it calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. It has many uses. For example, in [4], TF-IDF is used for solving of the query retrieval problem. For a collection of documents D, a word w and a document $d \in D$, we calculate

$$w_d = f_{w,d} \log \frac{|D|}{f_{w,D}} \tag{1}$$

where $f_{w,d}$ denotes the number of times w appears in d, |D| is the size of D, and $f_{w,D}$ is the number of documents including the word w, [9]. TF (term frequency) measures how many times a term appears in a document. For example, if we have a document D containing 4000 words, and the word “united” appears 15 times. In this case, the term frequency of the word “United” is equal to

$$\frac{15}{4000} = 0.00375 \tag{2}$$

On the other hand, IDF (inverse document frequency) assigns lower weight to frequent words in a document. For example, the word “for” appears in a document 1000 times but it is not significant in the analysis. Also, IDF assigns greater weight to infrequent words. For example, if we have 5 documents and the word “country” appears in 2 of them, IDF is calculated as

$$IDF = \log\left(\frac{5}{2}\right) \tag{3}$$

To summarize, if a word has greater occurrence in documents, it will have higher term frequency, and the less occurred word will have higher importance (IDF). TF-IDF is a product of TF and IDF.

The vectors created using the TFIDF approach are used in the following step to calculate the cosine similarity between the mathematics curricula of the faculty of economics in Japan and Bosnia and Herzegovina.

Another vectors approach was using the Sentence Transformer Multi QA model "multi-qa-MiniLM-L6-cos-v1". The Multi QA model used here is a model pre-trained in various domains and is suitable for use with cosine similarity analysis. It generates a 384 dimensional dense vector space of length 1. For more details, see Wang, [10], [11].

Cosine similarity: In text analysis, we represent each document by a vector. After that, we measure an angle between each pair of vectors and calculate its cosine, which is measure of similarity between documents. It is known that cosine of an angle α between vectors a and b is defined as

$$\cos\alpha = \frac{a \bullet b}{|a||b|} \tag{4}$$

where ● denotes dot product, and | | denotes the length of a vector. For more details on cosine similarity we refer to [12].

Key competencies are represented using the word cloud, a way of summarizing the text that shows the correlation between font size and frequency of each word and provides insight into the themes of the text [13].

We have compiled the curricula of mathematics at the faculties of economics of six state-funded universities in Bosnia and Herzegovina. Data were collected from the universities' official websites if the curricula were publicly available or through written email requests if they were not. The curricula were collected in February and March 2022 in the local language.

All curricula are at the bachelor's level; they are compulsory subjects with a similar structure: Number of teaching hours per teacher and per teaching assistant, ECTS credits, course requirements, course topics per week, and course objectives or core competencies.

There are eight publicly funded universities in Bosnia and Herzegovina. We managed to obtain data from six of them: University of Banja Luka, Bihać, East Sarajevo, Mostar, Sarajevo, and University of Zenica.

The data from the Japanese universities include Kyoto, Nagoya, Osaka, Tohoku, Kyushu, Tokyo and Hokkaido Universities and are available in the form of subject lists of mathematics courses in English at undergraduate level. All data are structured in the form of spreadsheets. Data were collected between January and March 2022 through written email inquiries.

3. Results and discussion

After applying the preprocessing steps explained in the methodology above, the similarity between the Japanese and Bosnian mathematical curricula is calculated using the Multi QA model of the Python Library Sentence Transformer, resulting in 80.3 percent of similarity.

The second vectorization approach, the TF-IDF algorithm, gives us vectors of different sizes, where we had 204 elements for the Japanese document and 286 elements for the Bosnian curricula. To calculate the cosine similarity, we had to make the vectors the same size. To get the same size of the vectors, we recreated two vectors with the counted word keys and their frequencies. After unifying these two vectors, we extracted the values of each vector. In this way, we loaded vectors with 0 values for each "empty" place. TF-IDF Model showed 60.7 percent similarity between two documents.

The first limitation of this study concerns the comparison of the models. We are aware that it is not possible to compare these two approaches, and because of this limitation, we would like to point the further direction of the research to the inclusion of additional approaches to calculating similarity using more mathematics curricula of the different countries. Then, we can compare the extent to which these curricula are comparable with each other in each approach.

The key competencies of the Japanese mathematics curricula for bachelor students of economics are shown in a word cloud in Figure 1. The Japanese mathematics curricula are dominated by the terms linear, matrix, and function. These words occur most frequently in the main text. They are followed by the words vector, equation, economics, and method.

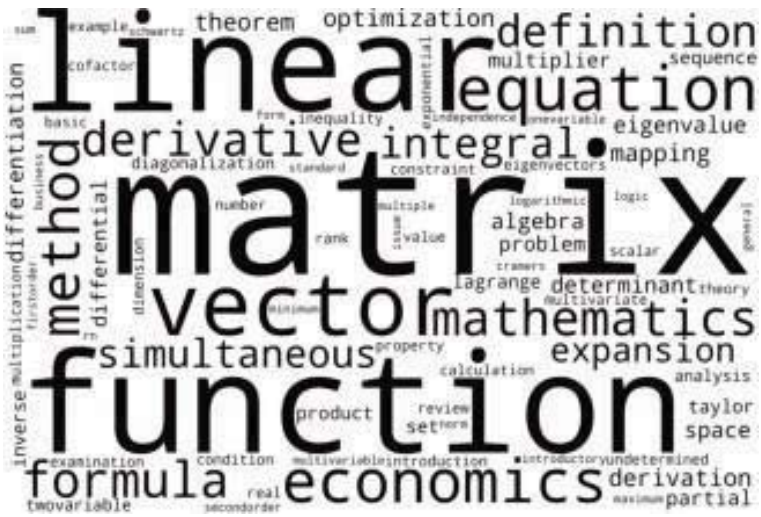

Figure 2. The key competencies in the mathematics bachelor curricula of students of economics in Japan

The most important keyword in mathematics curricula in Bosnia and Herzegovina is function. Unlike in Japan, almost equal importance is given to other competencies such as differential, variable, integral, equation, matrix and real.

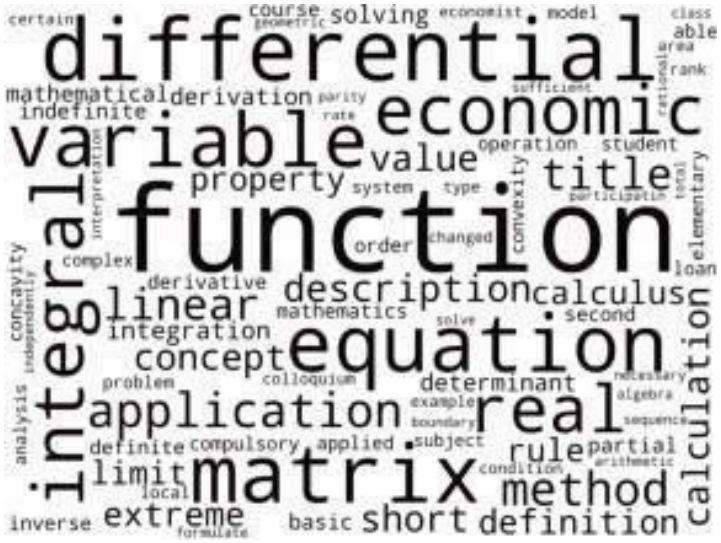

Figure 3. The key competencies in the mathematics bachelor curricula for economics students in Bosnia and Herzegovina

Judging from the word cloud representation of the most frequent words, Japanese mathematics courses for undergraduate economics students focus on a few topics, while the topics in Bosnia and Herzegovina are broader and more scattered. Perhaps here we find part of the reason for student dissatisfaction mentioned at the beginning of this paper.

Our further research proposal is in the direction of experimenting with bigrams and trigrams, as well as determining the minimum word frequency. Another possibility is to look more closely at additional words that can be included in the stop word list to improve results.

References

[1] Ballard C L and Johnson M F 2004 Basic Math Skills and Performance in an Introductory Economics Class, *J. Econ. Educ.* **35**(1) 3-23
[2] Schaffer B F and Calkins D O 1980 An appraisal of prerequisites to business finance, *J. Financ. Educ*. **9**(Fall) 51-55
[3] Ely D P and Hittle L 1990 The impact of math background on performance in managerial economics and basic finance courses, *J. Financ. Educ.* **19**(Fall 1990), 59-61
[4] Agrawal R and Batra M 2013 A Detailed Study on Text Mining Techniques, *Int. J. Soft Comput. Eng. IJSCE* **2**(6) 118-121
[5] Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe E D, Gutierrez J B and Kochut K 2017 Text Summarization Techniques: A Brief Survey, 10.48550/arXiv.1707.02268
[6] Mohan V 2016 Text Mining: Open Source Tokenization Tools: An Analysis, *Advanced Computational Intelligence Journal* **3**(1) 37-47
[7] Hardeniya N 2015 *NLTK essentials: build cool NLP and machine learning applications using NLTK and other Python libraries*, Packt Publ, Birmingham, UK
[8] Balakrishnan V and Lloyd-Yemoh E 2014 Stemming and lemmatization: A comparison of retrieval performances, Proceedings of SCEI Seoul Conferences, 10-11 Apr 2014, Seoul, Korea, pp. 174–179
[9] Salton G and Buckley C 1988 Term-weighting approaches in automatic text retrieval, *Inf. Process. Manag.* **5**(24) 513-523
[10] Wang K, Reimers N and Gurevych I 2021 TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning, 10.48550/arXiv.2104.06979
[11] *** Hugging Face 2022 *Sentence transformers Pre trained models: Multi QA MiniLM L6 dot-v1* https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-dot-v1
[12] Sidorov G, Gelbukh A, Gómez-Adorno H and Pinto D 2014 Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model, *Comput. Sist.* **18**(3) 491-504
[13] Heimerl F, Lohmann S, Lange S and Ertl T 2014 *Word Cloud Explorer: Text Analytics Based on Word Clouds*, 47th Hawaii International Conference on System Sciences, Hawaii, US, Jan. 2014, pp. 1833–1842