Journal of Modern Technology and Engineering
Vol.8, No.1, 2023, pp.39-53

MULTI-LABELING STRATEGY BASED ON A HEURISTIC TEXT SEGMENTATION FOR EDUCATIONAL CONTENT

Selcan Kayahan[1*], **Korhan Günel**[2], **Urfat Nuriyev**[3,4]

[1]Department of Computer Science, Institute of Science, Ege University, İzmir, Türkiye
[2]Department of Mathematics, Faculty of Science, Aydın Adnan Menderes University, Aydın, Türkiye
[3]Department of Mathematics, Faculty of Science, Ege University, İzmir, Türkiye
[4]Institutte of Control Systems, Baku, Azerbaijan

Abstract. In this study, it is aimed to extract the learning concepts from the educational contents, to segment the context into some overlapped text blocks that have semantic integrity, and to label the paragraphs with in the text blocks with multiple learning concepts. The study uses the History of Art book taught in schools affiliated to the Republic of Turkey Ministry of National Education. Natural language processing and heuristic clustering techniques are applied on the book and it is aimed to determine which learning concepts are associated with each paragraph of the document. For this purpose, feature vectors representing the parsed text blocks are extracted and the Particle Swarm Optimization clustering technique is applied after applying Principal Component Analysis on the feature vectors. In addition, the segmented text blocks matched with the learning concepts presented by an expert in the book to make a performance analysis of the proposed system. Then, the weighted mean squared error is calculated by comparing expert opinions and system outputs. The obtained results give hope about educational content can be decomposed into text blocks labeled with more than one learning concept.

Keywords: heuristic text segmentation, multi-labeling, particle swarm optimization, educational technology.
AMS Subject Classification: 68T50, 68T01.
Corresponding author: Selcan, Kayahan, Ege University, Department of Computer Science, Institute of Science, Izmir, Türkiye, e-mail: *selcan.kayahan@gmail.com*
Received: 12 December 2022; Revised: 15 February 2023; Accepted: 4 March 2023; Published: 30 April 2023.

1 Introduction

Concept is a general design that includes common features of objects or events and gathers them under a common name (Turkish Language Association, 2022). Concepts can also be expressed as basic mental formations that people need in their lives. Objects, events, facts are distinguished and understood in the human mind through concepts. A person's each experience means the restructuring of concepts, thus scenes, schemas and networks of meaning (Bozkurt, 2003). When the educational documents (textbooks) taught in schools are examined, it is seen that a "concept list" is given for each unit or for the whole book. The purpose of giving this list is to ensure that the learner knows the concepts related to that unit or chapter before starting to work on the unit or chapter. Thus, the learner understands that the subject he/she will start learning is about "what" in general. In short, the learner becomes aware of what he/she needs to learn in advance. This situation helps the learner to reconstruct the scenes, schemas and networks of meaning in his/her mind and facilitates the learning of the subject. Therefore, it is important that the concepts given in the textbooks are given completely, accurately and in a way that covers the whole subject. With the proposed model in this study, using the text segmentation approach, a single training document is separated into text blocks and these text blocks are automatically labeled with more than one learning concept. For this purpose, a minimal set of learning concepts within the educational content was determined by using heuristic/metaheuristic optimization techniques, machine learning and statistical language models. Then, the document was divided into overlapping text blocks and each paragraph of the text was labeled with teaching concepts through these text blocks.

Looking at the study from a different point of view, it can be interpreted that the proposed model allows the indexing of text blocks within a single document, based on the logic of indexing documents by search engines. Seen from a higher perspective, a customized search engine model for educational documents is proposed. For example, the learner wants to learn about the “derivative of the closed function” in mathematics. Instead of directing the learner to the document containing the “derivative” topic, the developed model will be able to list all the text blocks in which the specified concept and related sub-concepts (for example, “chain rule”) are explained.

The rapidly increasing amount of data in the digitalized world has also shown its effect in the field of education. In educational sciences, analyzing data collected from instructors and learners is one of the basic steps. With this analysis, learners can be guided and road maps can be presented so that they can continue their learning processes effectively. Baker (2016) emphasizes that there are not enough studies proving that artificial intelligence systems are valid and reliable and that it is necessary to conduct studies in this field. In a study conducted by Sommerhoff et al. (2018) in the USA, educational technologies were defined as “contexts that facilitate the analysis of real-world problems”.

Ouyang & Jiao (2021) proposed three different paradigms for the appropriate use of artificial intelligence techniques in educational technologies. Their proposed idea advocates the hypothesis that artificial intelligence is used as a tool to represent knowledge and cognitive learning models, and that learners are users of these services. In their second approach, learners take part in training the system in cooperation with artificial intelligence and thus artificial intelligence works as a training support system. The last thought is that artificial intelligence is defined as a tool managed by learners and used by learners to improve their own level. Zhai et al (2021) presented a literature review in which they presented a hierarchical model for artificial intelligence methods to be implemented in the field of education. However,Nabiyev & Erümit (2020) state that studies to determine the technical and pedagogical effects of artificial intelligence on educational environments are not yet sufficient. An opposite idea is found in the study of Lacity & Willcocks (2017). It has been stated by some researchers that whether artificial intelligence studies in the field of education will replace teachers in the future.

Recently, ChatGPT, which has become very popular, can be one of the candidate AI application for integrating big data analysis and artificial intelligence issues into the educational technology tools. Indeed, ChatGPT is a natural language processing model based on the GPT-3.5 architecture developed by OpenAI. It has been trained on large text corpora. In this way, it supports features such as multilingual support, semantics and grammar conformity. Natural language processing techniques in GPT-3.5 include word distribution, focusing on specific word groups, word root detection, sentence structure analysis and word meaning comparison. We asked ChatGPT the following question about our study: “Can you label paragraphs with more than one concept in a Turkish text?” Its answer was:

> “Yes, tagging paragraphs in Turkish text with multiple concepts is possible. For instance, a paragraph can contain both history and culture concepts or both economic and political concepts. The crucial point is to select appropriate concepts based on the content of the paragraphs.
>
> As an example, let’s consider the following paragraph: In the last period of the Ottoman Empire, economic crisis and political instability had affected the country. While the majority of the people living in poverty, minority groups

were maintaining their privileged positions."

We can tag this paragraph with both 'economy' and 'politics' concepts. Additionally, the 'culture' concept is also present in the paragraph, as the last period of the Ottoman Empire is an essential period for Turkish culture."

With this origin, in this study, it is emphasized how the proposed machine learning and multi-labeled text segmentation approach can be used in educational technologies. In machine learning, document classification and labeling is not considered a new research topic and it is possible to come across many studies in the literature. However, multi-labeling differs from classical classification approaches in many ways.

Burkhardt & Kramer (2018) stated that multi-label classification problems can be solved by machine learning methods, which are divided into two classes: problem transformation approaches and adaptive algorithm approaches. Problem transformation approaches are based on a very simple basis with the logic of labeling data with more than one classifier. In adaptive algorithm approaches, a single classifier works with an online learning approach, constantly updating the arbitrary parameter values needed for the algorithm to work according to the data.

Moyano et al. (2018) presented a literature review article for multi-labeled classifiers. A similar study consists of a literature review for multi-labeling approaches in unbalanced datasets (Tarekegn et al., 2021). In both studies, the differences and emerging difficulties between labeling data with a single label and multi-labeling approaches were highlighted. Tarekegn et al. (2021) divided multi labeling approaches into four different groups as resampling, adaptable classification, ensemble and cost-sensitive methods. For the performance analysis of the methods, the criteria presented in the literature were examined under three main headings as sample, label and ranking-based criteria.

Liu et al. (2005) proposed a multi-labeling approach in their sentiment analysis study on microblog posts, but they achieved an accuracy rate of 0.344. In addition, Tarekegn et al. (2021) evaluated the Hamming loss score, subset accuracy score, example-based F1 score, micro and macro F1 scores, average precision value, coverage, and an error values, which were also mentioned in their literature review, as performance criteria.

Kumar et al. (2018) proposed a multi-label classification approach based on hierarchical placement logic in their approach, which they named MLC-HMF, and they experienced this approach with ready-made data sets created in different fields. What should be emphasized in the study is that the accuracy rate obtained for the data set created especially in the field of education is limited to 0.294 ± 0.021.

Lee et al. (2019) developed a memetic search algorithm to identify distinctive features for document classification. In the study, the authors classified the documents into one or more categories from eleven different categories, such as Art, Computer, Education, based on the differences in tag frequencies. They compared their proposed approach with seven different filter and spiral-based feature selection approaches. As a performance criterion of the method, the accuracy scores were calculated separately for each category. It is seen that the accuracy rates obtained for different categories in the study vary between 0.1322 ± 0.0082 and 0.5345 ± 0.0082 values.

Yang & Liu (2019) developed a method for multi-labeled text classification based on parallel encoding and serial decoding approach. The model they developed actually consists of a combination of convolutional neural networks and an encoder that extracts local neighborhood information and global interaction information from the source text. This model was tested on three different datasets and micro-F1 scores of 0.893, 0.725 and 0.825 were obtained for each dataset, respectively. These values appear to be quite reasonable for document classifiers applying multi-labeling.

In a similar study, Aljedani et al. (2020) carried out a hierarchical approach to multi-labeling and labeling on Arabic texts. With their suggestions optimizing the parameters of the HOMER (Hierarchy Of Multilabel classifiER) algorithm, which uses more than one classifier, they have achieved an accuracy rate of 0.758 and an average micro-F1 score of 0.853 for multi-labeling.

As can be easily observed in all these studies, the success rates of multi-labeling approaches are quite low compared to classical single-label classification and clustering approaches and are still not at a sufficient level. In addition, no study has been found in the literature in which multi- labeled classifiers have been applied to educational technologies.

However, almost all of the document classification studies were carried out on a collection of documents gathered under a corpus. Natural language processing studies carried out on a single document are almost nonexistent. In the study, the idea of going one step beyond the classical document classification and labeling the learning concepts automatically extracted from the content of a single document and the text blocks in the document and thus labeling the document with more than one concept is unique. Content analysis on text blocks in a document is relatively more difficult than content analysis on the entire document. The main reason for this is that the number of words contained in text blocks is much less. To solve this problem, the idea was developed to split the document into blocks of text with common paragraphs. Such an approach has not been encountered in the literature. In this context, it is thought that the study will scientifically add a unique value to the literature in the fields of natural language processing and machine learning.

Natural Language Processing is a sub-branch of artificial intelligence. It includes methods and techniques based on the processing of texts written in natural language. Text segmentation is one of the study areas of natural language processing. Text segmentation is an important step in making sense of texts (Li et al., 2022b). In other words, text segmentation ensures that the boundaries between adjacent sections on certain topics are determined, thus revealing the semantic integrity or difference within a large text (Pak & Teh, 2017). Li et al. (2022b) stated that text segmentation has become a fundamental study field in natural language processing, which is handled at different levels of granularity.

There are many reasons why document segmentation is useful for text analysis. One of the main reasons is that they are smaller and more consistent than entire documents. Another reason is that each partition is used as analysis and access units (Oh et al., 2007). There are also studies that apply text segmentation in semantic analysis. Lattisi et al. (2022) proposed a method using semantics in text segmentation. Hoon & Wei (2017) used a technique in which information is organized into segments called values to improve the search algorithm.

On the other handDuan et al. (2012) applied text segmentation in the field of view mining to determine the direction in which user opinions tend to and to identify related features. Text segmentation also forms the basis of applications for automatically annotating and summarizing text within document content. Adding annotations to a document in natural language processing is usually done by tagging or coloring the important parts of the text with metadata. This is usually done at the word or sentence level.

By dividing the documents into small subsections, the text segmentation methods that enable the determination of the relations between these subsections can be used to determine the semantic similarities and differences on the educational materials. As a matter of fact, Qingrong et al. (2016) conducted a text segmentation study on Chinese poems in order to determine whether text segmentation facilitates the cognitive learning process and its effect on rhythmic reading speed.

This study is based on the logic of tagging each block separately by simply segmenting the educational documents into blocks of text in overlapping paragraphs. Thus, it is ensured that the paragraphs in each text block are labeled with one or more learning concepts.

There are learning concepts in written materials such as educational documents, textbooks, and resource books, which have a lot of use, especially at the k-12 level. Exacting learning concepts within a specific learning field is often a difficult, controversial, time-consuming and insignificant process even for an expert in this field Günel et al. (2016). Analysing the documents with data processing methods helps to make a semantic analysis of the resources used in education, to determine what the documents mean, and to determine the relationships between the teaching concepts and the general resource. Text segmentation is a method that divides documents into small subsections and enables the relationships between these subsections to be seen. Text segmentation methods can determine semantic similarities and differences on educational materials.

In this study, by analyzing the books prepared by the Ministry of National Education and shared as open access, it is aimed to label the text blocks in the education document with the learning concepts regardless of the learning area for learners. For this purpose, the content was analyzed using natural language processing, machine learning and heuristic or non-heuristic optimization techniques.

2 Methodology

The information in a document consists mainly of the semantics of words (Dinçer & Karaoğlan, 2003). In vector-based methods, which are frequently used in the field of information retrieval, texts are represented by the words they contain and/or vectors consisting of different features of the text (Meadow et al., 2000; Bilgin, 2019).

For most individuals, it is quite difficult to determine where to start working on a subject they are interested in. In order to achieve the learning objectives, a general roadmap is needed under the supervision of an expert. However, as the subject becomes more specialized in the individual learning process, it becomes increasingly difficult to create this roadmap based on the individual's personal knowledge level and to find an expert who will create this map. Zubrinik et al. (2012) suggested that information retrieval and management systems can assist in the creation of a learning plan, that is, roadmaps, by acting as an expert. If a person learns the concepts and the relationships between them by using this roadmap, he/she will also reach the learning goals.

In the study, it was aimed to identify the learning concepts from the unstructured raw data text- containing learning materials and to label the text with teaching concepts by segmenting it.

Preprocessing: In the study, all educational contents are considered as word strings. In the preprocessing stage, all words in the documents were first converted to lowercase letters, then multiple spaces and line jump characters were removed from the documents. The documents are also free of all mathematical formulas, numbers, variables and symbols. Special characters and punctuation marks such as $\#, \%, \&$ and \$ were then removed from the documents. At the end of the preprocessing, the all contents are transformed into a word sequence with only one space between consecutive words.

Stemming: While stemming, which is one of the main problems in the fields of information retrieval and natural language processing, can be done more easily in languages such as English, whose morphological structure is not very complex, it is very difficult in agglutinative languages such as Turkish because of the semantic and structural changes in the suffix sentence. In the study, syllable-based stemming approach was used to determine the longest stem. The Turkish spelling algorithm first determines the position of the vowels, and if there is no vowel, it separates the word into syllables from the index in which the first vowel is found. After the spelling process, it has been tried to find the longest body of a word that has a suffix, taking into account the states of the noun (presence, separation, denotation), the spelling of the word with its suffixes and its softening. After each suffix removed from the word, it was checked whether the rest of the word was in the dictionary by using the binary search algorithm.

In the next step, we propose a heuristic approach to segmentation of preprocessed documents. In the first step of the approach, as seen in Figure 1, the text is separated into sub-text blocks containing common paragraphs. In Figure 1, it is visually shown that overlapping sub-text blocks are removed from the document by quoting from the Art History book taught in schools

Sanat, insanoğlunun tarih öncesi çağlardan bu yana kendisini ifade etmekte kullandığı, yaşlanma ve hayata ait en yaratıcı ve en etkin yollardan biridir. Duygu ve düşüncelerin temellendirdiği bir içeriğe ve biçime sahip olan sanat, doğrudan yaşam ve hayatla ilgili olduğundan tüm insanlara hitap eder. Bu nedenle kalıcı ve evrenseldir. Gücü toplumsal sınırları aşar. Tarih boyunca neyin sanat ve sanat eseri olarak adlandırılacağına dair fikirler sürekli değişmiş, zaman içinde sanatın anlamına değişik kısıtlamalar getirilip farklı sanat tanımları yapılmıştır.

Thomas Munro (Tomas Munro), doğada ve sanatta güzeli konu edinen estetiği ön plana çıkararak sanatı "Sanat doyurucu estetik yaşantılar oluşturmak amacıyla dürtüler yaratma becerisidir. Sanat, güzel ile uğraşır." diye tanımlamıştır. Clive Bell (Kliv Bel) ise uyumu ön plana çıkararak "Sanat çizgi, şekil ve renk ilişkilerinin kendi aralarındaki kombinasyonudur." derken R. G. Collingwood (Kalingvud), sanatın temel olarak "Duyguların yaratıcı ifadesi veya dışavurumu" olduğunu söylemiştir. Tolstoy (Tolstoy) ise "İnsanın bir zamanlar yaşamış olduğu duyguyu, kendinde canlandırdıktan sonra, aynı duyguyu başkalarının da hissedebilmesi için hareket, ses, çizgi, renk veya kelimelerle belirlenen biçimlerle ifade etme ihtiyacından sanat ortaya çıkmıştır." der.

Sanatı; duygu ve düşünceleri etkileme gücünden hareketle "İnsanların, tabiat karşısındaki duygu ve düşüncelerini çizgi, renk, biçim, ses, söz ve ritim gibi unsurlarla güzel ve etkili bir biçimde, kişisel bir üslupla ifade etme çabasıdır." diye tanımlayabiliriz.

Sanatın tanımı yapılırken estetik (güzellik ile güzelliği araştıran felsefe dalı) kavramı ön plana çıkarılmıştır. Fakat bu kavramlar her dönemin, bölgenin, toplumun ve bireyin bakış açısına göre değişen göreceli kavramlardır. Evrensellik ise sanatın en temel niteliğini ortaya koymaktadır. Çünkü sanatın her türlü kişisel veya yerel nitelendirmeyi ve sınıflandırmayı aşan bir yapısı vardır. Sanatçı eserlerinde içinde yaşadığı toplumun kültürünü ve dünyaya bakış açısını yansıtırken başka toplumların kültürlerinden de etkilenir ve eseri ile başka insanları ve toplumları etkiler.

Sanat eserleri, sanatçının kendine özgü duygu, düşünce, heyecan, hayal gücü ve yeteneklerini yansıttığı gibi diğer insanların da duygu, düşünce ve hayal dünyasını geliştiren eserlerdir. Sanat eserlerinin en önemli özellikleri özgün, tek ve evrensel olmasıdır. Daha önce yapılmış olan bir eserin taklit edilmesiyle ortaya çıkan ürün ne kadar güzel olsa da sanat eseri olarak kabul edilmez. Çünkü sanat eseri sanatçının gördüğü, kavradığı ve gerçeklik olarak kabul ettiği, ruhunda hissettiği duyguların çeşitli şekilde anlatılmasıdır. Sanatçı; herkesten farklı hisseden, düşünen, yorumlayan ve bunları eserlerine farklı şekilde yansıtan kişidir.

Figure 1: Parsing text blocks with common paragraph

affiliated to the Ministry of National Education of Turkey. The idea of creating overlapping text blocks is unique in terms of revealing the relationships between text blocks and facilitating multi-labeling.

Each text block formed after this process on a single document is considered as a separate text file and feature extraction is applied for these text blocks. It is obvious that text blocks contain much fewer words compared to the document. For this reason, the bag of words model is created by using monograms and bigrams while extracting the feature vectors that characterize the text blocks. One of the research problems that arise at this point is to determine the size of the bag of words model. Considering that each text block contains a different number of monograms, it is clear that the size of the bag of words can be as much as the number of words of the text blocks containing the minimum number of monograms. However, the size of the bag of words can be decisive for classification or clustering. After the feature vectors of the text blocks were extracted, dimensionality reduction was applied using Principal Component Analysis.

The steps of this phase continue as follows:

***n*-gram extraction:** After the all contents were preprocessed, n-terms for $1 \leq n \leq 6$ were extracted. The n-term is a sequence of n consecutive words. In order to create the attributes of each n-term, firstly, the frequency of the n-terms in the text should be extracted. In order to do this quickly, the starting and ending positions of n consecutive words in the text are stored in two separate strings. Then, while n-terms are sorted alphabetically with the quicksort algorithm, they are sorted simultaneously in sequences showing the starting and ending positions of the terms. The time complexity of the quicksort algorithm is $O(m.\log(m))$ in the best case and $O(m^2)$ in the worst case, m specifying the number of terms to be sorted. Thus, the same terms are ordered one after the other and it is easy to determine whether two consecutive terms are the same using only the differences between the start and end positions. If the lengths of the terms are the same, whether two consecutive terms are the same is decided by using the Boyer-Moore text matching algorithm (by scanning character by character from right to left). Since the texts to be compared are of the same length, if m is the number of characters in the word, the time complexity of this operation will be $O(1)$ in the best case and $O(m)$ in the worst case. If two consecutive words are the same, their frequency is increased by 1 and the next two consecutive n-terms are compared. If the words are not the same, the first n-terms are added to the word pool with their frequency.

According to the n-gram model, the monograms and bigrams in the whole document were extracted. The n-gram model is a probabilistic language model and works according to the Equation 1 (Manning et al., 2009; Günel et al., 2016).

$$P(w_i|w_{i-(n-1)} \cdots w_{i-1}) = dfracC(w_{i-(n-1)} \cdots w_{i-1}w_i)C(w_{i-(n-1)} \cdots w_{i-1}) \tag{1}$$

n-gram filtering: One of the important problems with the extraction of n-terms as candidate learning concepts is whether the first or last word of the n-word sequence is a stop word. If so, this n-gram should be ignored. Because learning concepts in educational content cannot start or end with a single word. For example, while the expressions “swarm intelligence” and “particle speed” are learning concepts in the field of “Global Optimization”, the word sequences ”swarm intelligence with” and ”particle speed and” are not learning concepts.

Stop words are words that do not make sense on their own in the document and have low discrimination value when the text as a whole is taken into account. However, although they are displayed with high frequency in the document, they only have syntactic functions in the sentence, they do not contribute much to the information extraction systems in terms of semantics. For these reasons, they are often overlooked. There are different approaches in the literature for the removing stop words. In classical methods, words are compared with words in the stop word pool. In approaches based on Zip's law, expressions consisting of a single word with a high frequency in the document are seen as a stop word and deleted if the reverse document frequency is also low. In Mutual Information approaches, the problem is considered as a classification problem and single words with low mutual information value are considered as stop words. In the term-based random sampling approach, stop words are tried to be determined on randomly selected text phrases by using the Kullback-Leibler separation measure. There are different techniques based on these four approaches in the literature. In the study, two of these methods were hybridized to determine the stop words. After eliminating the n-terms whose first or last word is a stop word, the spelling check phase was started (Kumova & Karaoğlan, 2017). At this stage, the statistical approach proposed by Aşliyan et al. (2007) was used. The stated approach is based on the probability of successive Turkish letters and the detection of misspelled words.Aşliyan et al. (2007) determined the misspelled words with an accuracy rate of 97% by using the probability of succession of 3-letter sequences. In this study, n-word sequences that could not pass the spelling check were removed from being a candidate learning concept.

Generating feature vectors: In the study, some of the features used in natural language processing applications have been changed appropriately for the learning concept extraction problem. The attributes used are briefly summarized as follows:

- n: word count in candidate learning concept t_j, $t_j = w_1w_2 \dots w_n$
- Frequency: The number of times that t_j, which is a candidate learning concept, is occurred in the educational content, $f(t_j)$
- Term Frequency: The ratio of the n-gram frequency of the t_j word string to the total frequency of all n-grams of the training content, $tf(t_j)$
- Text Block Frequency: t_j number of text blocks containing the word string, $df(t_j)$.

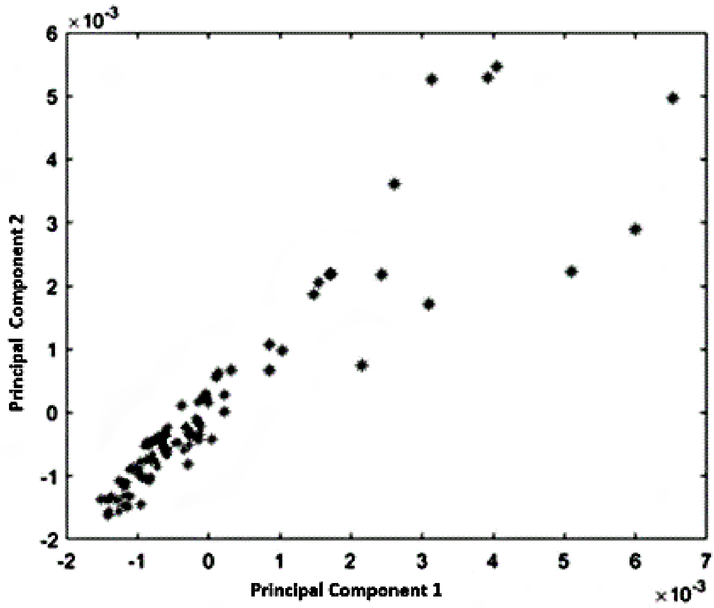

Figure 2: Positioning of text blocks on the plane after PCA process for History of Art History textbook

- Inverse Text Block Frequency: In this study, inverse document frequency value, which is one of the indispensable attributes in document classification and approaches to clusters, has been reinterpreted as Inverse Text Block Frequency since it is studied on a single document. This value is calculated by the equation $itbf(t_j) = \log\left(\frac{N}{df(tj)}\right)$ where N is the number of all text blocks in the document d.

- Text Block Frequency - Inverse Text Block Frequency: The similar approach specified in the previous attribute was used for the term frequency - inverse document frequency attribute, and the text block frequency - inverse text block frequency values were obtained. This value is calculated as $tbf - itbf(t_j) = tbf(t_j) \times itbf(t_j)$ for each t_j word string.

Normalization: In the study, min-max normalization was used to the feature vectors of each candidate learning concept defined with the above-mentioned feature values. Thus, it is ensured that the attribute values take values in the range of [0,1].In addition, the minimum number of candidate learning concepts included in each text block was calculated and the feature vectors of the text blocks were created with the bag of words approach by using the feature vectors of the calculated number of word sequences.

Dimensionality Reduction: Dimensional reduction with Principal Component Analysis (PCA) was applied on the normalized feature vectors of the paragraphs (Pearson, 1901). With this method, the most basic features of the dataset are determined and the dataset is reduced.

PCA approach was applied on the feature vectors of 92 segmented text blocks in the document used in the study. The result of determining the principal components and positioning the text blocks on the plane is shown in Figure 2.

As can be seen in Figure 2, the text blocks are located in the search space with an almost linear distribution. In addition, considering that most of the data is very close to each other, it is observed that it is not easy to decompose the data into clusters. In the next stage of the study, unsupervised clustering approach was used since it was aimed to label independent of the domain (document content).

Clustering: In this study, instead of the classification approach, the clustering approach is preferred to automate the labeling process unattended. Clustering is an unsupervised machine learning method. The reason why clustering approaches are preferred in the study can be

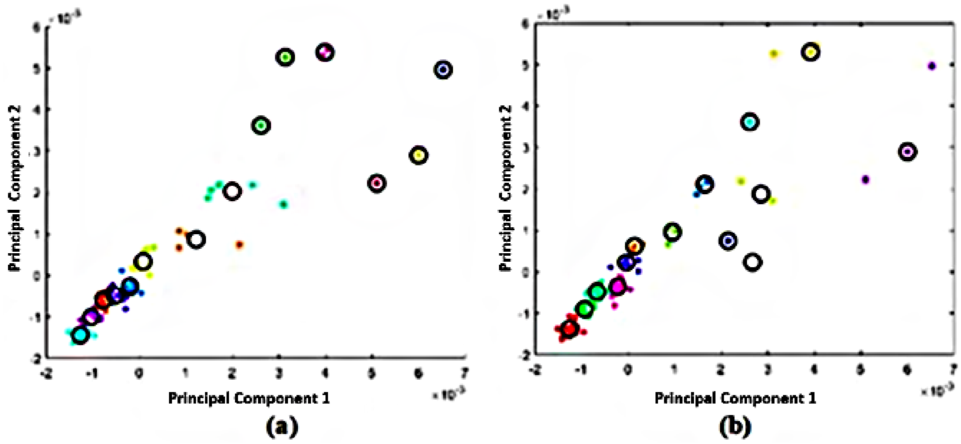

Figure 3: Clustering of text blocks with (a) k-means (b) PSO for the History of Art textbook

explained with a well-known sentence of Yann Le Cun describing supervised and unsupervised learning: If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake (Geron, 2019).

In this study, using the normalized data, text blocks that are heuristically parsed with the Particle Swarm Optimization method, which is one of the techniques based on the swarm intelligence approach, are clustered. PSO is a meta-heuristic optimization method introduced (Kennedy & Eberhart, 1995). In the PSO approach, all particles are initially randomly located in the search space. In the swarm, the particles quickly gather around the particle with the best position, using their own cognitive information and the information they obtained from the swarm.

In this study, the position vectors of the particles indicate the coordinates of the cluster centers. After the document used in the study was parsed into 92 different text blocks, each text block was described with a feature vector. Considering that the text will be labeled with 14 different teaching concepts, it is aimed to gather 92 text blocks under 14 clusters. For this purpose, in the first step of PSO, a swarm of particles is created to randomly distribute in the search space, which is limited by the range of minimum and maximum values of the feature vectors components. The position vector of each particle in the swarm is defined to contain the coordinates of the 14 candidate cluster centers. The particle swarm contains a total of 50 particle with dimensions of 14×2. At each step of the algorithm, the distances of the feature vectors of the text blocks to the centers are calculated. Therefore, a distance matrix of type 92×14 is created for 92 paragraphs and 14 clusters.

Then, the cluster center to which each text block is closest is determined and the text block is included in this cluster. After this process for each particle, the total distance of the feature vectors of all text blocks from the cluster centers to which they are included, and the cost of the particle are determined. After the costs of the particles are determined, the particle with the lowest cost in the swarm is determined. After this stage, the positions of the particles are updated in each iteration with the classical PSO approach. When the maximum number of iterations is reached, it can be observed that all particles gather around the particle with the best position in the swarm, that is, the optimal position of the cluster centers. As a result, this location is used to determine cluster centers. Text blocks are automatically labeled using this particle's position vector.

Two different clustering approaches, k-means and PSO, were tested in the study. Labeling of the data shown in Figure 3 with the k-means and PSO approach is shown in Figure 4a and Figure 4b. In Figure 3, each color indicates a separate learning concept (cluster), while oval shapes represent cluster centers.

Table 1: Some statistical information about the document

Qualification	Value
Total word count	6213
Stop words count	1196
Paragraph count	91
Monogram count	5017
Singular monogram count	1546
Bigram count	5351
Singular bigram count	4843
Monograms average count per block of text	369.856
Singular monograms average count per block of text	138.747
Monograms average count per paragraph	68.659
Singular monograms average count per paragraph	55.131
Learning concepts count	14
Ratio of learning concepts to singular monograms	0.009
Average learning concept count tagged by the expert and per paragraph	1.626
Text blocks count tagged with a single learning concept by the expert	49
Text blocks count tagged with double learning concept by the expert	27
Text blocks count tagged with three learning concept by the expert	15
Average learning concept count tagged by the system and per paragraph	2.286
Text blocks count tagged with a single learning concept by the expert	12
Text blocks count tagged with double learning concept by the expert	41
Text blocks count tagged with three learning concept by the expert	38

3 Experimental Results

In the study, the Art History History book, which is taught in schools affiliated to the Ministry of National Education of the Republic of Turkey, was used. “Bağımsızlık”, “Kut”, “Toy”, “Töre”, “ülke”, “Divan”, “Saray”, “Reâya”, “İstimalet”, “Ulus Devlet”, are the learning concepts in the Art History book. Some statistical data about the content of the book are presented in Table 1.

After preprocessing the book content, it was parsed into 91 overlapping text blocks and these text blocks were labeled with concepts by the expert. As can be seen in Table 1, after manual labeling, it was observed that the expert using 14 concepts labeled each text block of the book with an average of 1.626 learning concepts. The expert labeled only 15 of the 91 text blocks with three different learning concepts. 27 text blocks were labeled with 2 different learning concepts. It was observed that the remaining 49 text blocks were labeled with only 1 learning concept.

After expert opinions, n-gram extraction, creation of feature vectors, dimension reduction was done with Principal Component Analysis. Then, clustering was applied with the Particle Swarm Optimization (PSO) approach. The main reason for choosing this approach is that it aims to automatically label blocks of text regardless of the field. Particle Swarm Optimization does clustering with unsupervised learning. The text blocks as a result of clustering are labeled as seen in Figure 4.

As can be seen in Figure 4, each paragraph is labeled with at least one and at most three concepts. The dataset produced from the document, including expert opinions, was made available in XML format[1].

For system performance measurement, expert opinions and system outputs were compared. The path followed for the performance measure is briefly summarized below.

If LC is the set of learning concepts obtained from document d containing N paragraphs, for $k = 1, 2, \ldots, N$, the function is defined in Equation 2 for $i, j \in \{1, 2, 3\}$.

$$\chi_k(i,j) = \begin{cases} 0, & \text{if } o_{k,j} = t_{k,i} \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

where $T_k = \{t_{k,i} \in LC \mid t_{k,i} \text{ is } i^{\text{th}} \text{ label determined by the expert for the } k^{\text{th}} \text{ paragraph}\}$, and $O_k = \{o_{k,j} \in LC \mid o_{k,j} \text{ is } j^{\text{th}} \text{ label determined by the system for the } k^{\text{th}} \text{ paragraph}\}$

[1] https://github.com/kgunel/veriseti-Sanat-Tarihi

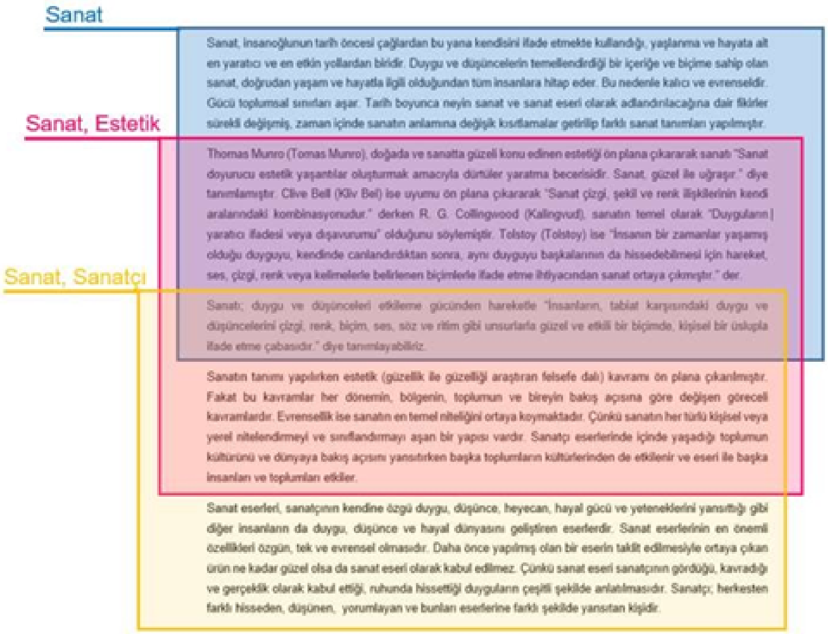

Figure 4: Labeling of text blocks containing common paragraphs. The labels "Sanat", "Sanatçı" and "Estetik" equivalent in English words Art, Artist and Aesthetics respectively.

In this case the mean square error for the multi-labeling system is calculated by the Equation 3.

$$MSE = \frac{1}{N} \sum_{k=1} N \left(\chi_k(i,j)\right)^2 \tag{3}$$

In addition, a heuristic weighting has been implemented for the labels according to their order of importance. Accordingly, considering that each paragraph is labeled with at least one and at most three concepts, the target labels were indexed with $i = 1, 2, 3$ respectively, and the weights were graded with 0.6, 0.3 and 0.1 values, respectively. In this case, the weighted mean squared error is calculated with the Equation 4

$$wMSE = \frac{1}{N} \sum_{k=1} N \left(0.6\chi_k(1,j) + 0.3\chi_k(2,j) + 0.1\chi_k(3,j)\right)^2 \tag{4}$$

The numbered paragraphs in the study are divided into text blocks consisting of three paragraphs, starting from the first paragraph. Thus, blocks of text are formed heuristically to include paragraphs $1-3, 2-4, 3-5, \ldots$ When the last paragraph of the document is reached, a cyclic text justification approach is used by pretending that the document is going back to the beginning. The basic logic of this approach is based on the idea that similar concepts can be mentioned in the introduction and conclusion sections of the document. Therefore, the final text blocks created include paragraphs 90, 91, and 1, and paragraphs 91, 1, and 2. As a result of the study, mean square errors were obtained as $wMSE = 0.557$ for k-means and $wMSE = 0.527$ for PSO clustering approach. This value can be interpreted as the primary weighted learning concept is clearly determined for each paragraph, and the secondary weighted learning concepts are successfully identified in some of the paragraphs. As it can be understood from the weighted mean square error formula, if no learning concept belonging to a paragraph can be determined, that is, if all the labels made for this paragraph are incorrect, an error of 1 unit is obtained. Therefore, an error of up to N units can be obtained for N paragraphs. This can be seen with the maximum amount of error that can be obtained by precisely labeling each paragraph with three different learning concepts. If some of the paragraphs are labeled with fewer learning concepts, the label weights can be updated so that the maximum error amount to be obtained is still N. In order to meet this condition, it is sufficient that the sum of the weights of the learning concepts grading their relations with the text blocks is 1. The average error amount per paragraph is still 1 unit. Therefore, the error wMSE value should be in the range of [0,1]. As a matter of fact, the 0.5 unit weighted mean square error indicates that one of the three labels for each paragraph can be clearly identified, while a part of the second label agrees with the expert's opinion. As can be seen, in both clustering methods used in the proposed model, the wMSE values were calculated very close to each other. Independent of the clustering approach, similar labeling is important in order to show the consistency of the approach.

4 Discussion and Conclusion

Considering that natural language processing applications have gained importance in almost every field, it will be seen that text processing methods are also needed on educational documents. In particular, it is seen as an important problem area to reveal which subject the materials taught as textbooks "talk" about. Determining the concepts in the books in a complete, correct and way to cover the whole subject will ensure that the education is qualified. In this study, it is aimed to determine how much the chapters in a book taught in high schools are related to which concepts. The fact that the most frequently mentioned words and concepts in the Art History History book, which is used as a data source in the study, are "culture" and "civilization", has brought a certain limitation to the study. The proposed method has also been done previously using a different book. In addition, it is thought to be carried out and tested on different fields and by increasing the variety of documents. At the end of the study, many research problems such as how many paragraphs a text segment should contain, whether the paragraph count can be determined dynamically to be different for each text block, examining the effect of the clustering approach to be chosen on performance, determining the number of learning concepts to be used in labeling the text blocks.

One of the limitations of this study is the application of the proposed approach on a single document. Although the setup of the study was done in this direction, it is planned to develop a method in order to make efficient text segmentation and labeling in documents on different subjects, regardless of the field, in the continuation of the study.

In addition, in the continuation of the study, it is aimed to develop an adaptive self-testing system model with the multi-labeled text segmentation approach and to determine the necessary standards for the integration of the model into educational support systems. It is aimed to update the labeling system in line with the feedback provided by the learners to the system during the self-testing process. In this way, it is aimed to direct the learners to the text segments where the learning concepts introduced in the topic and related sub-learning concepts are found rather than the whole topic presented in the educational document. Thus, in the individual learning process, a visual roadmap will be presented to the learner by descending to the level of learning concepts within the subject rather than the subject.

5 Acknowledgement

The authors would like to thank the anonymous reviewers for their helpful comments to improve the paper.

References

Aljedani, N., Alotaibi, R. & Taileb, M. (2020). HMATC: Hierarchical multi-label Arabic text classification model using machine learning. *Egyptian Informatics Journal*, *22*(3), 225-237. doi: `https://doi.org/10.1016/j.eij.2020.08.004`

Aşliyan, R., Günel, K. & Yakhno, T. *Detecting Misspelled Words in Turkish Text Using Syllable n-gram Frequencies.*, In: Ghosh A., De R.K., Pal S.K. (eds) Pattern *Recognition and Machine Intelligence.* PReMI 2007, Lecture Notes in Computer Science, vol 4815. Springer, Berlin, Heidelberg, 2007. doi: https://doi.org/10.1007/978-3-540-77046-6_68

Baker, R.S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600-614. doi: https://doi.org/10.1007/s40593-016-0105-0

Bilgin, M. (2019). Kelime Vektörü Yöntemlerinin Model Oluşturma Sürelerinin Karşılaştırılması. *Bilişim Teknolojileri Dergisi*, *12*(2), 141-146. doi: https://doi.org/10.17671/gazibtd.472226

Bozkurt, B.Ü. (2018). Kavram, Kavramsallaştırma Yaklaşımları ve Kavram öğretimi Modelleri: Kuramsal Bir Derleme ve Sözcük öğretimi Açısından Bir Değerlendirme, (Concepts, Conceptualization Approaches, and Concept Teaching Models: A Theoretical Review and An Evaluation in Terms of Teaching Vocabulary). *Ankara Üniversitesi Dil Dergisi*, *169*(2), 5-23.

Burkhardt, S. Kramer, S. (2018). Online multi-label dependency topic models for text classification. *Mach Learn*, *107*, 859-886. doi: https://doi.org/10.1007/s10994-017-5689-6

Dinçer, B.T., Karaoğlan, B. *Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish.*, In: Yazıcı, A., Şener, C. (eds) Computer and Information Sciences - ISCIS 2003. ISCIS 2003. Lecture Notes in Computer Science, vol 2869. Springer, Berlin, Heidelberg. 2003. doi: https://doi.org/10.1007/978-3-540-39737-3_31

Duan, D., Qian, W., Pan, S., Shi, L. & Lin, C. VISA: a visual sentiment analysis system. In: *Proceedings 5th International Symposium Visa Information Communicate Interaction*— VINCI'12, ACM: Hangzhou, 2012, pp. 22-28. 2012. doi: https://doi.org/10.1145/2397696.2397700

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'reilly, 2nd Edition.

Günel, K., Polat, R., & Kurt. M. (2016). Analyzing Learning Concepts in Intelligent Tutoring Systems. *International Arab Journal of Information Technology*, *13*(2), 265-286.

Hoon, G.K., Wei, T.C. (2017). *Flexible facets generation for faceted search.*, In: Vasant, P. & Duy, V.H. (eds) *First EAI International Conference on Computer Science and Engineering.* Penang, Malaysia. 2017. doi: http://dx.doi.org/10.4108/eai.27-2-2017.152348

Kennedy, J., Eberhart, R. (1995). Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, 1995, pp. 1942-1948, vol.4. 1995. doi: http://dx.doi.org/10.1109/ICNN.1995.488968

Kumar, V., Pujari, A.K., Padmanaphan, V., Sahu, S.K. & Kagita, V.R. (2018). Multi-label classification using hierarchical embedding. *Expert System with Application*, *91*, 263-269. doi: https://doi.org/10.1016/j.eswa.2017.09.020

Kumova, S., Karaoğlan, B. (2017). Stop Word Detection as a Binary Classification Problem. *Anadolu University Journal of Science and Technology*, *18*(2), 346-359. doi: https://doi.org/10.18038/aubtda.322136

Lacity, M., Willcocks, L.P. (2020). *Robotic Process Automation and Risk Mitigation: a Definitive Guide.* B Publishing, Ashford, UK.

Lattisi, T., Farina, D. & Ronchetti, M. (2022). Semantic Segmentation of Text Using Deep learning. *Computing and Informatics*, *41*(1), 78–97. doi: https://doi.org/10.31577/cai_2022_1_78

Lee, J., Yu, I., Park, P. & Kim, D.W. (2019). Memetic feature selection for multilabel text categorization using label frequency difference. *Information Sciences*, *485*, 263-280. doi: https://doi.org/10.1016/j.ins.2019.02.021

Li, M., Bai, M. & Lv, Y. (2022a). Text segmentation by integrating hybrid strategy and non-text filtering, *Multimedia Tools and Applications*, *81*, 44505–44522. doi: https://doi.org/10.1007/s11042-022-13029-1

Li, D.J., Chiu, B., Shang, S. & L. Shao, L. (2022b). Neural Text Segmentation and its Application to Sentiment Analysis. *IEEE Transactions On Knowledge And Data Engineering*, *34*(2), 828-842. doi: https://doi.org/10.1109/TKDE.2020.2983360

Liu, L., Kang, J., Yu, J. & Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. *International Conference on Natural Language Processing and Knowledge Engineering*, Wuhan, China, 597-601, doi: https://doi.org/10.1109/NLPKE.2005.1598807

Manning, C., Raghavan, P. & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge University Press.

Meadow, C., Boyce, B. & Kraft, D. (2000). *Text Information Retrieval Systems.* Second ed. Academic Press.

Moyano, J.M., Gibaja, E.L., Ciosand, K.J. & Ventura, S. (2018). Review of ensembles of multi-label classifiers: Models, experimental study and prospects. *Information Fusion*, *44*, 33-45. doi: https://doi.org/10.1016/j.inffus.2017.12.001

Nabiyev, V., Erümit, A.K. (2020). *Eğitimde Yapay Zeka: Kuramdan Uygulamaya.* Ankara: Pegem Akademi Yayıncılık.

Oh, H., Myaeng, S.H. & Jang, M.G. (2007). Semantic passage segmentation based on sentence topics for question answering. *Information Sciences*, *177*(18), 3696–3717. doi: https://doi.org/10.1016/j.ins.2007.02.038

Ouyang, F., Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*(100020), 6 pages. doi: https://doi.org/10.1016/j.caeai.2021.100020

Pak, I., Teh, L.P. (2017). Text Segmentation Techniques: A Critical Review. In: Zelinka, I., Vasant, P., Duy, V. & Dao, T. (eds) *Innovative Computing, Optimization and Its Applications. Studies in Computational Intelligence*, vol 741. Springer, Cham. 2017. doi: ttps://doi.org/10.1007/978-3-319-66984-7_10

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *2*(11), 559–572. doi: https://doi.org/10.1080/14786440109462720

Qingrong, C., Wentao, G. & Scheepers, C. (2016). Effects Of Text Segmentation On Silent Reading Of Chinese Regulated Poems: Evidence From Eye Movements. *The Journal of Chinese Linguistics*, *44*(2), 265-286. doi: https://doi.org/10.1353/jcl.2016.0011

Sommerhoff, D., Szameitat, A., Vogel, F., Lodere, O. & Fischer, F. (2018). What Do We Teach When We Teach the Learning Sciences? A Document Analysis of 75 Graduate Programs. *Journal of the Learning Sciences*, *27*(2), 319-351. doi: https://doi.org/10.1080/10508406.2018.1440353

Tarekegn, A.N., Giacobini, M. & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, *118*, 33-45. doi: https://doi.org/10.1016/j.patcog.2021.107965

Turkish Language Association (Türk Dil Kurumu). Güncel Türkçe Sözlük. Retrieved December 01, 2022, from https://www.tdk.gov.tr

Yang, Z., Liu, G. (2019). Hierarchical Sequence-to-Sequence Model for Multi-Label Text Classification. *IEEE Access*, *7*, 153012-153020. doi: https://doi.org/10.1109/ACCESS.2019.2948855

Zhai, X., Chu, X., Chai, C.S., Jong, M.S.Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 1-18. doi: https://doi.org/10.1155/2021/8812542

Zubrinik, K., Kalpic, D. & Milicevic, M. (2012). The automatic creation of concept maps from documents written using morphologically rich languages. *Expert System with Application*, *39*(16), 12709-12718. doi: https://doi.org/10.1016/j.eswa.2012.04.065