Automatic Extraction of Learning Concepts From Exam Query Repositories

Damir Pintar, *Member, IEEE,* Domagoj Begušić, Frano Škopljanac-Mačina, *Member, IEEE,* and Mihaela Vranić, *Member, IEEE*

Original scientific paper

***Abstract*—One of the biggest challenges in the process of establishing modern e-learning systems is figuring out ways to leverage legacy course materials and integrating them in the new information systems. Existing exam query repositories in particular are a very valuable data source, but one which usually lacks enough metadata to help establish relationships between exam questions and corresponding learning concepts whose adoption is being evaluated. In this paper we present the continuation of our research regarding the usage of educational data mining methods able to automatically annotate pre-existing exam queries with information about learning concepts they relate to. In our novel approach we leverage both textual and visual information contained in the queries. By combining the power of natural language processing which focuses on the text of the question, and annotated data extracted from figures accompanying the questions, we are able to further refine our classification methods and achieve noticeably improved results. By identifying learning concepts more accurately we further facilitate automatic creation of exams as well as even better insight into learning concept adoption. Our approach is again applied on data gathered from a large scale university course, and the results were validated in consultation with educational domain experts.**

***Index Terms*—educational data mining, exam queries, learning concepts, classification, e-learning.**

I. Introduction

THE term ”educational data mining” (EDM) denotes an emerging interdisciplinary research field concerned with developing methods for exploring the specific and diverse data encountered in the field of education. Its goals are to provide better insight into the learning process, identify the properties of the learning environment, improve educational outcomes and explain various educational phenomena [2]. Information systems used in the educational domain commonly store large amounts of data from various sources, using different formats and pertaining to different levels of granularity. Data mining methods usually require specific adaptation before they can be applied to a particular problem encountered in the educational domain.

One of the ongoing efforts in further improving the existing e-learning infrastructure and learning process in general is figuring out ways to leverage existing sources and data in ways beyond what they were initially created for. Historical exam results may hide information about deficiencies in structures of the courses, imbalances in exam difficulties or irregularities in testing environments. Submitted homework assignments may offer insight into students learning habits, behaviour towards deadlines etc. Finally, various study materials may be explicitly connected to the learning concepts they are related to even if they do not carry such metadata information within them.

Our main topic of interests were legacy repositories of exam questions which were not annotated with metadata concerning learning concepts they were evaluating. This scenario is pretty common when an educational system is migrating from classical learning to e-learning and existing materials are being transferred to the digital domain; while text materials are being digitized and stored additional effort needs to be made to enrich these sources with enough metadata to make their usage and integration in the e-learning system as streamlined as possible. Automating this process to a certain extent can be of extraordinary help to the teaching staff who can then focus on higher-level issues of improving the learning process.

In our previous research [1], we have focused on text questions from a legacy exam query repository and leveraged natural language processing to create a model which would automatically connect exam questions with the most likely learning concepts they were related to. The questions were first carefully annotated by a group of domain experts, which allowed supervised learning methods access to an adequately large representative sample to learn from. The experts were given a list of *primary* and *secondary* concepts, with primary concept usually being related to the overall course lecture theme, and secondary to the exact learning concept being evaluated. In regards to previously published research results in [1], which focused solely on the natural language processing and leveraging only textual information contained in the exam queries, we have significantly broadened the scope of our research evaluating and comparing both the predictive power of textual information and the visual information contained in the figures accompanying the exam queries. Also, we have revisited our pre-processing steps and introduced a few changes which helped improve our previous results. Additionally, we explore the idea of introducing automatic translation as a dataset preprocessing step, which may help to decouple the results from the language they are originally written in

Manuscript received August 29, 2018; revised October 19, 2018. Date of publication October 26, 2018. The associate editor Prof. Nikola Rožić has been coordinating the review of this manuscript and approved it for publication.

Authors are with the University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia. Emails:{damir.pintar, domagoj.begusic, frano.skopljanac-macina, mihaela.vranic}@fer.hr

Digital Object Identifier (DOI): 10.24138/jcomss.v14i4.605

and allow further generalization of our approach, making it applicable to exam queries written in any language for which appropriate automatic translation interfaces are available.

The rest of the paper is structured as follows: Section II will provide information about related work. Section III discusses our methodology, which involves describing the process of preparing the input data and choosing appropriate methods. Section IV shows application of these methods and provides discussion on gained results. Finally, in Section V we conclude and briefly discuss our plans for further research on this topic.

II. Related Work

Recently, the number of EDM references is growing, as the researchers in this field are actively exploring different ways to analyse educational data stored in e-learning systems (learning materials, detailed student activity, performance and assessment data). Usually, well-known data mining techniques and methods are used, e.g. classification, clustering, prediction, social network analysis (SNA), text-mining, process mining and relationship mining. EDM is still primarily used in academic community, as its main goal is the improvement of learning, teaching and educational processes [2].

In [3], authors introduce a text-mining framework based on a phrase graph model of text documents. They applied this framework to automatically analyse and cluster learning materials according to their main topics. Authors focused on key phrases identified in text documents, rather than on the individual words. This enabled them to tag different text documents with appropriate topic labels. It is important to note that the authors utilized unsupervised learning methods to extract labels of text documents clusters. Conversely, we used supervised learning techniques to classify exam questions, which made validation of our results more reliable.

Valuable scientific reviews of the development of EDM research can be found in [4] and [5]. Castro et al. in [4] identified that the research issues in EDM were dealing with clustering and classification problems in e-learning systems, and to a lesser extent data visualization and prediction. In [5] Romero and Ventura described more recent interesting research issues in EDM. One of the issues is the automated generation of feedback information about learning materials and tests structure to authors and teachers, e.g. multiple choice questions were analysed using hierarchical clustering techniques to discover similarities among learning concepts covered by the answers.

In [6], Chen et al. introduced a text-mining method for automated building of concept maps, which can help with the designing of learning materials in adaptive e-learning systems. The authors tested their method on a set of academic papers from the field of e-learning. Their method automatically built a concept map of the e-learning domain, primarily by analysing the keywords sections of each paper. Contrary to this approach, we analysed full texts of exam questions.

Supraja et al. [7] present a model for automated labelling of course questions. There are some similarities to our approach, but the main difference is in the classification of questions. Their model classifies questions depending on their learning objectives based on simplified Bloom's Taxonomy (three levels of questions: remember, analyse and transfer). Our research goal was focused on linking questions with identified learning concepts from the course syllabus and available learning materials. Furthermore, we classified questions according to their primary and/or secondary learning concepts as we will show in the next section.

In [8] authors created text-mining algorithms for finding (concept, relation, concept) triples in text and for building concept maps. Authors used their model to automatically build concepts maps from lecture slides. These concept maps were evaluated by human experts using different objective factors such as coverage, suitability and accuracy. The validation results were encouraging because the experts graded automatically built concept maps as very good or good. We must note that the authors decided to ignore all the figures in lecture slides and left that for future work. In our approach we aimed to include most important information from the figures as will be elaborated in the following sections.

Authors in [9] used Formal concept analysis (FCA) method to automatically build an ontology of an engineering course exam. This ontology is described by a concept lattice, which is a directed graph of hierarchically ordered nodes. Each node is called a formal concept, and it consists of sets of objects and their shared attributes. Input data must be prepared as a binary matrix, called a formal context, containing all objects from the domain of interest and all the chosen attributes. In their research the exam questions were the objects, and each question was coupled with a set of its attributes. After building an exam ontology the teachers can check it to ensure that the exam covers all the intended learning concepts and to identify groups of similar questions. In the preprocessing stage of the FCA method the authors in [9] have manually labelled all the exam questions. This laborious task could be automated by applying questions' labelling methods presented in our research.

III. Methodology

The methodology used in our research follows and expands on the methodology described in [1]. In comparison with the recent related research efforts, our method combines text-mining approach based on full question texts with the key information obtained from annotated figures that supplement questions. Also, in the proposed method we added an option for automatic translation of the questions' dataset into English. This has allowed us to compare the accuracy of classification for both datasets, and for their combinations with the annotated figures.

Basic overview of the process can be seen on Fig. 1. In the continuing paragraphs this process will be elaborated upon with special attention given to added steps of the process, depicted by the dotted arrows to emphasize that they can be optionally "turned" off to allow easier comparison to the previous approach and evaluation of their corresponding results.

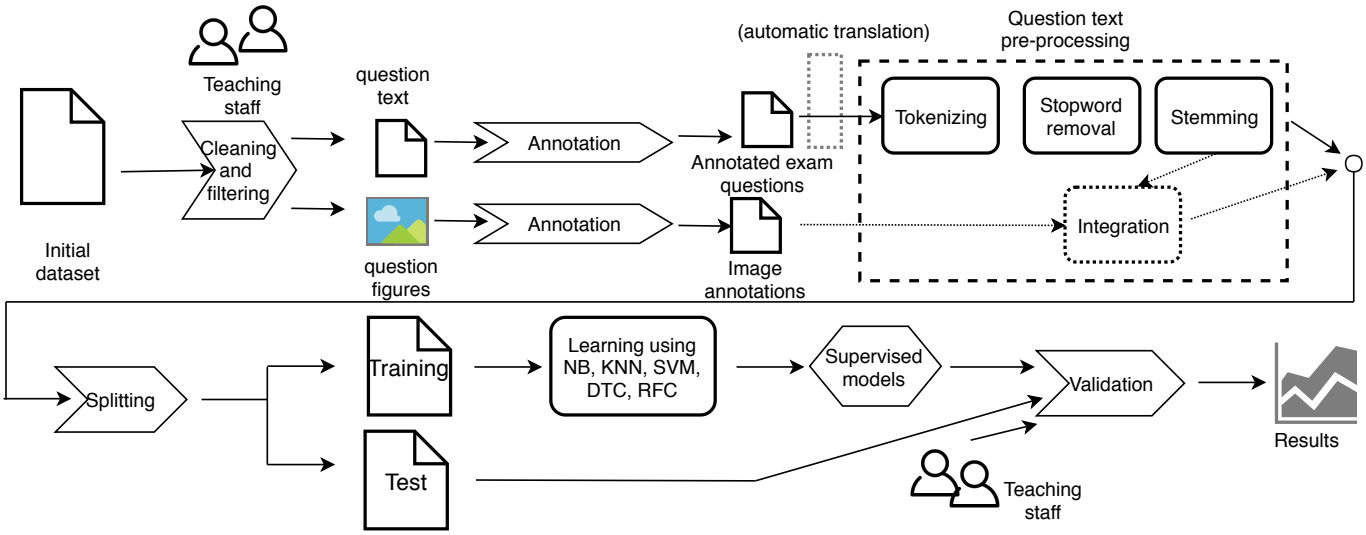


Fig. 1. Overview of the process

A. Dataset Description

The detailed dataset description was already provided in [1], so here we will briefly summarize just its basic properties. Dataset initially contained 3914 questions, with each question having the following 10 attributes:

- **ID** - unique identifier
- **variantID** - ID of source question (for variants)
- **text** - text of the question in UTF-8 format
- **answerA (B, C, D, E)** - answer texts
- **correctAnswer** - correct answer code
- **pictureFile** - filename reference if the question contains a figure, NULL otherwise

As it will be described further, our approach was initially focused on question texts as the primary source of data for automatic learning concept extraction. However, due to the nature of the course being analysed, accompanying figures often hold very valuable information, sometimes being crucial in identifying the learning concept the question is related to. For example, the text may only concisely refer to the measure which needs to be calculated, while the picture provides the entire context related to this measure and in turn the learning concept being evaluated. This fact has motivated us to further explore the value of textual information compared to the visual information contained in exam queries, and to devise a hybrid approach for identifying learning concepts which could leverage both types of the information and achieve higher accuracy compared to models which rely only on one of these types.

Taking this facts in mind, our input data was expanded to include 1661 figures (in various formats) which were referenced by 2407 questions (figures were occasionally recycled by individual questions). Most figures were a scheme of an electrical circuit using the IEC 61346 standard.

B. Cleaning, Filtering and Annotation

In [1] we have outlined the steps undertaken to transform questions from various formats (plaintext, HTML, markdown etc.) and converted them to a uniform format which was a necessary pre-requisite of having a high-quality exam question repository. After the text was cleaned using customized Python scripts. In the filtering phase, we have conditionally removed variant questions, meaning that we didn't use them in training, but were kept for the training set, since it was expected that the teaching staff will tend to use old questions when coming up with new ones, so the validation process needs to account for that.

Annotation was done in cooperation with domain experts, who were given a list of 20 *primary concepts* and 115 *secondary concepts* which related to learning concepts typically being evaluated on exams. This was done through a form which randomly chose a question and asked of users to pick the primary and secondary concept which he/she feels the question is related to. Fig. 2 shows bar charts of primary and secondary concepts in the filtered dataset. Each bar represents a frequency of an identified primary/secondary concept. Since it would be impractical to put concept labels on the x-axis directly, it instead depicts numeric identifiers of concepts (once they have been ordered by ascending frequency). One noted characteristics of this dataset is a severe imbalance of the concepts.

Since part of our research included estimating the effectiveness of automatic translation when it comes to identifying the learning concepts, we created a separate dataset which contained question texts which were automatically translated into English language using Google Cloud Translation API [18]. Since our intention was evaluating automatic translation as a pre-processing step without additional human intervention, the translated text was left as-is, regardless of the fact that certain inadequacies were apparent even by casual perusal of gained translations. It is therefore expected that certain information loss has occurred which we assume will be ultimately reflected as at least a slight deterioration of the final results.

The final part of the data preparation phase involved deciding how to handle figures accompanying a significant portion

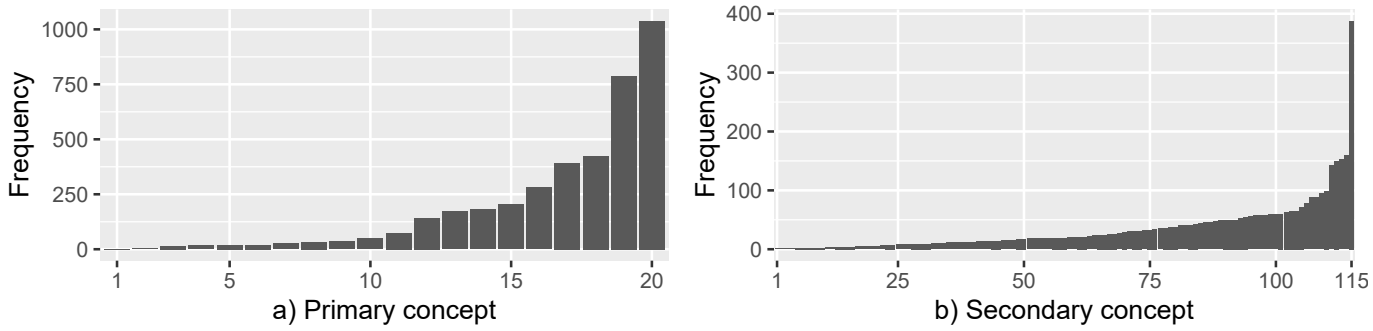


Fig. 2. Concept frequency bar charts (x-axis depicts numeric identifiers of primary/secondary concepts after they have been arranged by ascending frequency)

of exam queries. As stated, the queries referenced figures from a second repository containing 1661 images in GIF and JPG format. These images originally came from various sources and didn't follow a common and consistent standard of depicting similar concepts so the fonts, resolution, symbols, sizes etc. varied greatly.

To convert these images into a format usable for supervised learning methods, we decided to identify a number of image elements which are visually distinctive and connected to the learning concepts being identified. Ultimately we ended up with 16 new attributes, all of them being logical flags informing whether a certain element is present on the figure or not. All attributes are named *iTERM*, where *i* stands for 'indicator' and *iTERM* is a shorthand name for the relevant element whose existence is being described by the attribute. The final list of these attributes is (certain attributes are grouped for the sake of conciseness):

- **iCirc** - indicator whether the figure represents an electrical circuit
- **iRes, iCap, iInd** - presence of a resistor, capacitor or inductor, respectively
- **iResAd, iIndDot** - presence of an adaptable resistor or coils with shared magnetic fields
- **iSourceDCU, iSourceDCI, iSourceAC** - presence of various types of electrical sources
- **iDotCharge** - presence of one or more dot charges
- **iAmpermeter, iVoltmeter, iWattmeter** - presence of various measuring instruments
- **iTriphase** - presence of a triphase electrical circuit
- **iGraph** - indicator whether the picture shows a graph of any sorts
- **iPhoto** - indicator whether the picture is a photograph

These terms are obviously closely related to the course domain, but the approach can easily be generalized for other courses which strongly rely on non-textual information to evaluate learning concepts.

One important aspect of the newly devised attributes is that they do not require expert knowledge of the domain to be identified, but also that only brief visual inspection of the figure is enough to identify these elements. This allowed for manual annotation of the figures to be feasible. Still, a relatively large number of figures motivated us to devise custom-made annotation software to further expedite the process. Fig. 3

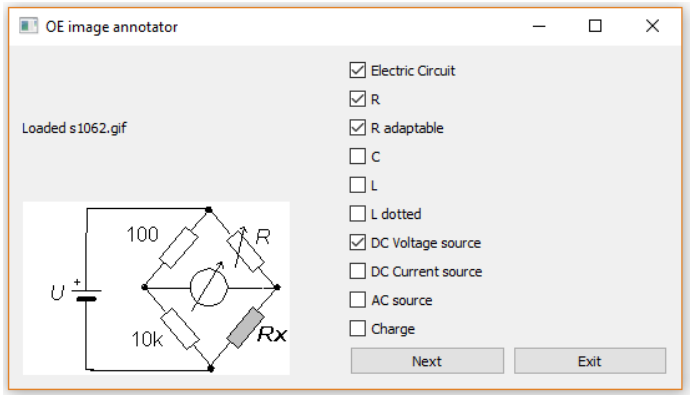


Fig. 3. Custom image annotator

shows a screenshot of this custom-made annotator. Ultimately we ended up with a tabular structure containing identifier of all 1661 figures and binary flags pertaining to the above attributes.

This annotated dataset holds an additional value. Our intention is to use our predictive model for future queries, but without the need to manually annotate the accompanying figures. This dataset will hopefully allow us to train a neural network (or multiple networks) which will be able to automatically identify the presence of the above elements. Due to the nature of the domain and these elements being visually distinctive, we are fairly confident that this approach will be implemented and validated successfully in our future work.

C. Pre-processing Exam Questions

The pre-processing steps contained the usual natural language processing steps were undertaken: removal of interpunction, sentence segmentation and tokenization, stopword removal and stemming. Tokenization was facilitated using the stopword repository [13] (expanded with a few additions which were noticed and deemed uninformative concerning learning concepts). These words were then stemmed using a stemmer based on information available at [14] and API from [15].

However, in regards to the process described in [1], we have introduced a few tweaks and fixes which helped keep more information in the text of the questions and subsequently slightly improve the results. Most notably, the symbols for

TABLE I
QUESTION PRE-PROCESSING

original question	svi su otpori jednaki r=1 Ω. koliki je napon uab?
preserving the units and symbols	svi su otpori jednaki rr=1 Ω. koliki je napon uab?
stopword removal	otpori jednaki rr=1 Ω. napon uab?
stemming	otpor jednak rr=1 Ω. napon uab?
numbers/special characters removal	otpor jednak rr Ω napon uab

TABLE II
PRE-PROCESSED QUESTIONS AND CORRESPONDING CONCEPTS

text	**concept1**	**concept2**
napon uab prikazan spoj iznos	DC circuits	bridge circuit
odredi snag naponski izvor daj krug ww rr	DC circuits	triangle-star
odredi napon otpornik rr ww	DC circuits	superposition
otpor jednak rr Ω napon uab	DC circuits	Thevenin/Norton

some very important domain concepts (such as U for the voltage and I for the current), which were omnipresent in the question texts, unfortunately clash with stopwords "in" and "and" in the Croatian language. This resulted in those symbols being removed despite them possibly helping with identifying the learning concept. To address this issue, we introduced new, custom symbols for these concepts which prevented their unwanted removal due to them being misidentified as stopwords. Table I shows how the processing of question works on one randomly chosen question, while table II shows an example of three exam question texts with annotated concepts, the way they look after the pre-processing step. The concepts are translated into English, but the pre-processed question text was by necessity left in the original, Croatian language.

For the automatically translated exam questions texts, we had more stemmer tools at our disposal, such as "PorterStememer", "LancasterStemmer", or even "Snowball" - a small string processing language designed for creating stemming algorithms. We opted for "PorterStemmer" [16] due to its popularity and convenient out-of-the-box solutions. As for tokenization and stopword removal, those functionalities are natively integrated in Python's Natural Language Toolkit packages, which makes choosing English language even more convenient.

Image annotations did not require any additional preprocessing due to their binary nature. The part of our research which used textual attributes together with image annotations simply added the binary annotations as additional variables before proceeding with the model-building.

D. Vectorization, Classification and Validation

For detailed description of the vectorization, classification and validation steps, we again refer to [1], while here we will briefly summarize the most important steps and elaborate on the additions introduced by the approach described in this paper.

TABLE III
PRIMARY CONCEPT CLASSIFICATION RESULTS

Method	NB	SVM	KNN	DTC	RFC
Train acc.	84.04%	98.85%	74.52%	99.9%	99.52%
Test acc.	80.23%	92.02%	72.88%	91.54%	90.67%

TABLE IV
SECONDARY CONCEPT CLASSIFICATION RESULTS

Method	NB	SVM	KNN	DTC	RFC
Train acc.	64.62%	97.79%	60.96%	99.52%	98.85%
Test acc.	60.77%	80.58%	57.02%	77.12%	78.56%

Performed vectorization used the *Bag of Words* model which turns words into a categorical variable and calculates each category's frequency. This was deemed preferable to the alternative, *TF-IDF* due to relatively small amount of text in the questions. For the purpose of classifying, a collection of common, most popular classifiers was chosen, due to ease of their use and general availability, making it easy to replicate our approach. These classifiers are:

- Multinomial Naive Bayes
- Linear Support Vector Machine
- Decision Tree Classifier
- Random Forest Classifier

A brief description of these classifiers can be found in [1], while [17] represents one of the best sources for the elaborate explanation with insight into their mathematical and statistical foundations.

All classifiers are trained and tested using cross-validation, with 75% of the dataset used for training, and 25% for testing. Classifier precision was also evaluated on the training data to examine the potential overfitting issues.

In our initial research which focused on pure natural language processing and using only textual information for the classification of learning concepts, we have devised four different flavours of classifying. First we tried to identify only the primary concept. Then we repeated the process, trying to identify only the secondary concept. Finally, we used a hierarchical classifying approach where we first classified the primary concept, and then trained a second classifier to try to identify the secondary concept within the subset of observations with already assigned primary concept.

To keep the results concise, for the automatically translated questions, we will focus only on the best performing classifiers from [1], Support Vector Machine and Random Forest Classifier. For image classification, we will use the same classifiers, and we will compare how images fare on their own compare to using only text, before finally combining all available attributes and training the best-performing classifiers on the entire dataset.

IV. IMPLEMENTATION AND RESULT DISCUSSION

In this section we will provide the results of our research comparing the between-model accuracy of classification for all used approaches. It must be stated that the results shown here

TABLE V
HIERARCHICAL CLASSIFICATION RESULTS

Method	NB	SVM	KNN	DTC	RFC
Test acc.	59.90%	79.9%	51.06%	79.80%	81.35%

are newly generated compared to those shown in [1], even for the first part of the research which used the equivalent approach of leveraging only textual information in native language. The reason for result fluctuations are due to the tweaked pre-processing steps, described earlier in the paper. Table III shows the results when trying to identify only the primary concept. In concordance with what was shown in [1], Support Vector Machines, Decision Tree Classifier and Random Forest Classifier show the best results. Interestingly, the new pre-processing step has resulted in the boost of Decision Tree Classifier which made it actually take over the second place, previously held by Random Forest Classifier in [1]. Also, as a peculiar side-note, SVM actually performed slightly worse now that certain words, previously mistaken for stopwords, were held in the dataset. The difference is small and statistically insignificant though.

Looking at Table IV we can witness similar relative behaviour between classifiers. Overall performance is weaker, which is expected due to a relatively large number of secondary concepts and noticed imbalance between them. Naive Bayes and kNN classifier in particular performed rather poorly, not only on the test set, but on the training set too.This time around, compared to corresponding table in [1], SVM and especially DTC have achieved a noticeable boost, with a slight rise gained by RFC. Tweaked pre-processing has made DTC results come really close to RFC, with the previous difference of 5% going to just 1%, making DTC a feasible choice.

Finally, in Table V we can see the result of hierarchical classification. Here, only test results are shown, due to peculiarities of the process where the classifiers were trained in a two-step process, only training secondary concept classifiers on a subset of observations with the same primary concept. In this case Random Forest Classifier has beat the Support Vector Machine classifier, although the difference is not necessarily significant. A big issue with hierarchical classifying was relatively small number of samples for a large number of secondary concepts.

So far we can confirm our previous conclusions that the Support Vector Machine and Random Forest Classifier showcase the strongest predictive power, with Decision Tree Classifier possibly being the optimal choice if interpretability is required. Having not only the benefit of an automatically assigned concept, but an actual visualization of the process how the concept was assigned would be very much appreciated by the domain experts, potentially gaining additional insight into the fact how wording of questions may relate to concepts they are evaluating.

Our next task was researching whether the added step of automatically translating the questions before doing the NLP and classifier training steps was a feasible option. While this translation commonly wasn't serviceable for human consumption without additional intervention from domain experts, it offered an advantage of easy automatic integration in the process as an additional pre-processing step, as well as allowing the usage of NLP packages oriented towards English language, regardless of the language the questions originated at. This helps generalize the approach and allow it to be used even in circumstances when no helpful NLP packages exist for the native language the exam queries are written in. Of course, as previously stated, automatic translation of exam queries does not offer the level of quality the professional translation would achieve, especially concerning some domain-specific terms, so a certain loss of information is expected.

TABLE VI
HIERARCHICAL CLASSIFICATION ACCURACY - NATIVE LANGUAGE VS. AUTOMATIC TRANSLATION

Language\Method	SVM	RFC
Croatian	79.9%	81.35%
English (autotranslate)	66.83%	68.56%

TABLE VII
HIERARCHICAL CLASSIFICATION ACCURACY - IMAGE ANNOTATION ATTRIBUTES ONLY

Method	SVM (One vs. Rest)	SVM (One vs. One)	RFC
Accuracy	30.42%	36.81%	37.50%

The final results confirmed our expectations regarding this fact. Table VI shows the result of hierarchical classification applied on both variants of exam questions, and the native language significantly outperforms the automatically translated texts. We expect however that once the professional translation of exam queries become available (which might be the case due to the planned course internationalization in the upcoming years), the same process might achieve significantly better results.

Finally, we have oriented ourselves towards non-textual information and leveraged the previously prepared image annotations. Since the goal of this research was primarily to compare the extent of usefulness of visual information compared to textual information when it comes to classifying the learning concepts (as well as gaining insight in the effectiveness of the combined approach), for this final bit of research we have only used a subset of exam queries which were accompanied with a figure. We have trained predictive models by:

- using only visual information (i.e. image annotations)
- using combined textual and visual information

The results of hierarchical classification using only the annotations as predictive attributes can be seen in Table VII.

We tried two flavours of SVM and RFC, and we can conclude that in their current form image annotations by themselves do not have strong predictive power, at least when it comes to combining concepts in the hierarchical classification. The reason for that can be a significant overlap when it comes to visual elements used by learning concepts, where a very similar figure can be used to evaluate a varied collection of concepts, with the potentially distinguishing element not being present in current collection of annotations.

Finally, we have combined textual and visual information to see how all combined data performs when it comes to classification accuracy. To fully leverage all prepared datasets, we have also repeated the process for the automatically trans-

TABLE VIII
HIERARCHICAL CLASSIFICATION ACCURACY - COMBINED ATTRIBUTES

Language \Method	SVM	RFC
Croatian	83.56%	86.06%
English (autotranslate)	74.33%	75.29%

lated questions, to see whether additional data can offset the apparent loss of information caused by autotranslation. Results can be seen in Table VIII.

As expected, the results confirmed the hypothesis that using all available data may result in the best gained classification accuracy. Even the autotranslated questions showcased a noticeable boost. RFC slightly outperformed SVM, which means that it would be our algorithm of choice for this type of predictive modelling. We also strongly suggest considering DTC, especially for already mentioned benefits of interpretability.

It may be interesting to explore misclassified examples to gain knowledge about cases which confused our final classifier. This may give us insight into what additional attributes we may use to further enhance the predictive model. For example, question ID 3116 got assigned the "various AC circuits" secondary concept, even though the correct one was "triangle-star transformation". This in particular is not even a misclassification per se, because the question is in fact related to AC circuits - however, it was designed to test the triangle-star transformation method specifically. This wasn't implicitly referenced in the question but could have only been deduced from looking at the accompanying figure - which unfortunately lacked any specific annotations connected with triangle-star transformation. This example was useful to us since it immediately led us to improve our annotation process to include the "bridge connection" annotation, which is easy to visually distinguish yet is very common with questions concerning triangle-star transformation (since it effectively removes the problematic "bridge" from the circuit). Another example is question ID 3024 which correctly predicted the secondary concept of "mutual inductance" but managed to misclassify the primary concept, guessing "AC circuits" instead of "electromagnetism". Again, by leveraging domain knowledge, it is readily apparent this is really not a "true" misclassification since the mutual inductance does pertain to AC circuits, since the concept is first introduced in the lecture about electromagnetism only to then be subsequently applied in lecture about AC circuits. The conclusion that can be gained from this example is that it might not be enough to try and improve our classifier, but also to focus a bit more on our concept definitions and structure, try to remove overlap as much as possible and introduce a more concise granularity levels between concept layers.

At the end of this section we have to address the notion of reproducibility. Since used dataset is a repository of exam questions which will be actively used on the course it relates to in the upcoming years, and since questions themselves are intellectual property of the teaching staff who authored them, we are currently unable to provide fully open access to it. However, we are willing to provide the dataset or its subsets exclusively for research purposes if certain data protection procedures are respected. For such requests, please contact the paper authors directly via provided contact information.

V. CONCLUSION AND FUTURE WORK

This paper showcased an approach which can allow teachers to better leverage their repositories of test and exercise questions by automatically extracting learning concepts from text questions. Both textual and visual information was leveraged and the final results have shown that a relatively high level of accuracy can be achieved, which makes the proposed approach feasible for implementation. However, it must be stated that for supervised learning methods to be effective, the pre-processing step of manual annotation is a necessary requirement. To facilitate this step, it is highly recommended that custom software annotator solutions be devised, which will both expedite the process and ensure higher quality annotations.

Further improvements of the proposed model are already being undertaken. Since the question answers (currently absent from the model) can often implicitly reveal the learning concept being evaluated through the measuring unit of offered value, the next step of our research will integrate the answers as additional predictive attributes for the model. Furthermore, since our results have shown the question figures are often extremely informative when it comes to the relationship between the exam question and the learning concept, we will also focus our research towards automatic interpretation and annotation of figures. This can have added benefit besides identifying learning concepts tied with the figures, since an automatic image annotator can be used to increase the usefulness of e-learning search engines, or help organize large image databases so they can be more readily integrated into new e-learning solutions. Manual process of annotating figures can be very arduous, even with adequate custom-made tool support, so efforts will be made to relegate this task to specially trained neural networks.

Finally, our immediate efforts will definitely go in the direction of devising further improvements for our learning models. Additionally, we hope that our models can be integrated in a larger e-learning infrastructure where they could offer continual and dynamic support for easier creation of new exams, on-line testing and recommendations of learning materials based on interactively evaluated results.

ACKNOWLEDGEMENT

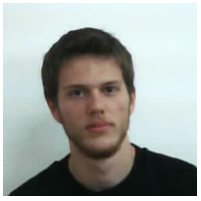
The research team would like to thank Croatian Science Foundation (Hrvatska zaklada za znanost - www.hrzz.hr). The work has been fully supported by Croatian Science Foundation under the project *UIP-2014-09-2051 eduMINE - Leveraging data mining methods and open technologies for enhancement of the e-learning infrastructure*.

REFERENCES

[1] D. Begušić, D. Pintar, F. Škopljanac-Mačina, M. Vranić, Annotating Exam Questions Through Automatic Learning Concept Classification, in Proceedings of the 26th International Conference on Software, Telecommunications and Computer Networks - SoftCOM, Split-Supetar, Croatia, 2018.

[2] C. Romero and S. Ventura, Data mining in education, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, pp. 12-27, Jan. 2013, DOI: 10.1002/widm.1075.
[3] K. Hammouda and M. Kamel, Data Mining in E-Learning, pp. 374-404. London: Springer London, 2007, DOI: 10.1007/978-1-84628-758-91_3.
[4] F. Castro, A. Vellido, A. Nebot, and F. Mugica, Applying Data Mining Techniques to eLearning Problems, pp. 183-221. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, DOI: 10.1007/978-3-540-71974-8_8.
[5] C. Romero and S. Ventura, Educational data mining: A review of the state of the art, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, pp. 601-618, Nov 2010, DOI: 10.1109/TSMCC.2010.2053532.
[6] N.-S. Chen, Kinshuk, C.-W. Wei, and H.-J. Chen, Mining e-learning domain concept map from academic articles, Computers & Education, vol. 50, no. 3, pp. 1009-1021, 2008, DOI: 10.1016/j.compedu.2006.10.001.
[7] S. Supraja, K. Hartman, S. Tatinati, and A. W. H. Khong, Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes, in Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017, Wuhan, Hubei, China, June 25-28, 2017 (X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, eds.), International Educational Data Mining Society (IEDMS), 2017.
[8] T. Atapattu, K. Falkner, and N. Falkner, A comprehensive text analysis of lecture slides to generate concept maps, Computers & Education, vol. 115, pp. 96-113, 2017, DOI: 10.1016/j.compedu.2017.08.001.
[9] F. Skopljanac-Macina and B. Blaskovic, Formal concept analysis overview and applications, Procedia Engineering, vol. 69, pp. 1258-1267, 2014. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013, DOI: 10.1016/j.proeng.2014.03.117.
[10] E. Baralis and L. Cagliero, Highlighter: Automatic Highlighting of Electronic Learning Documents, in IEEE Transactions on Emerging Topics in Computing, vol. 6, no. 1, pp. 7-19, 1 Jan.-March 2018, DOI: 10.1109/TETC.2017.2681655.
[11] P. Chen, Y. Lu, V. W. Zheng, X. Chen and B. Yang, KnowEdu: A System to Construct Knowledge Graph for Education, in IEEE Access, vol. 6, pp. 31553-31563, 2018, DOI: 10.1109/ACCESS.2018.2839607.
[12] P. Janardhanan, L. Heena and F. Sabika, Effectiveness of Support Vector Machines in Medical Data mining, in Journal of Communications Software and Systems, vol. 11, no. 1, pp. 25-30, 2015, DOI: 10.24138/jcomss.v11i1.114.
[13] TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatian stopword repository. Available: http://www.zemris.fer.hr/~jan/amn/download/hrstop-1.2, [Accessed: 18-May-2018]
[14] N. Ljubešić, D. Boras, O.Kubelka, Retrieving Information in Croatian: Building a Simple and efficient rule-based stemmer, Proceedings of INFuture 2007., pp 313-320
[15] Natural Language Processing group, Faculty of Humanities and Social Sciences, University of Zagreb, Available: http://nlp.ffzg.hr/, [Accessed: 18-May-2018]
[16] The Porter Stemming Algorithm, https://tartarus.org/martin/PorterStemmer/, [Accessed: 25-August-2018]
[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001., DOI:10.1007/978-0-387-84858-7
[18] Google Cloud Translation API, https://cloud.google.com/translate/docs/, [Accessed: 20-August-2018]

Damir Pintar is an assistant professor at the University of Zagreb, Faculty of Electrical Engineering and Computing. His interests include data management, machine learning, data mining and Big Data technologies, with a particular focus on predictive modelling and practical implementations of data science principles in industry and academia. He is actively involved in a number of data science-related projects in which he acts both as a researcher and as the project leader. He is a member of IEEE as well as the Centre of Research Excellence for Data Science and Advanced Cooperative Systems.



Domagoj Begušić is a Bachelor of Science in Computing and a graduate student at University of Zagreb, Faculty of Electrical Engineering and Computing. His interests include data science, natural language processing and integration of data mining methods for the improvement of e-learning systems.

Frano Škopljanac-Mačina is a laboratory manager and a PhD student at University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electrical Engineering Fundamentals and Measurements. His research work is focused on using formal methods and automated assessment techniques in designing advanced adaptive e-learning systems. He is a member of IEEE since 2012.

Mihaela Vranić is an assistant professor at University of Zagreb, Faculty of Electrical Engineering and Computing. She is involved in a number of science and industry projects as researcher or project leader. She is the author of a number of scientific papers in the area of her scientific interest, which include electronic business, database, data warehousing, business intelligence and data mining. She is also a member of program committees and reviewer at several conferences and journals in her field of interest. From the academic year 2018/19 onwards she serves as Vice Dean of Education at University of Zagreb, Faculty of Electrical Engineering and Computing. She is a member of IEEE and the Centre of Research Excellence for Data Science and Advanced Cooperative Systems.