CAMARINES SUR
POLYTECHNIC COLLEGES
COLLEGE *of*
COMPUTER STUDIES

CSAC 224 - Machine Learning

# Automated Quiz Statement Classifier Using Machine Learning and PDF Processing

**C-jay Lavapie**
College of Computer Studies
Camarines Sur Polytechnic Colleges
`c-lavapie@my.cspc.edu.ph`

## Abstract

With the advent of the digital age, the need for automated quiz generation has also gained importance among educators and learning platforms. Manually choosing and classifying statements is the conventional approach to creating quizzes, which is usually time-consuming and prone to errors. This paper introduces the Automated Quiz Statement Classifier, a machine learning and natural language processing (NLP) based system that extracts and classifies content worthy of quizzes from PDF files. Through the use of TF-IDF vectorization and logistic regression and support vector machine (SVM) model training on a corpus exceeding 46,000 labeled sentences from various educational sources, the system successfully identifies statements appropriate for use in quizzes. The prototype, developed with Flask and PyMuPDF, allows users to input educational PDFs and obtain output with labeled quiz-relevant statements. Metrics of evaluation like accuracy, precision, and F1-score show excellent model performance, sanctioning its use for automated reviewer and assessment content creation. Future improvements involve increasing the dataset using manually selected samples and adding image detection to create flashcards. The research proves the capability of AI-based tools to automate and enhance educational content development.

## 1 Introduction

At present, in the digital age, automated quiz generation has fast emerged as a critical tool for educators, researchers, and the providers of e-learning platforms. To create quizzes via traditional methods, a selection of appropriate statements and their grouping into types of questions needs to be done manually, which is a lengthy and error prone process. Machine learning techniques can help to automate quiz statement classification to improve efficiency and accuracy to address the challenge mentioned.

The Automated Quiz Statement Classifier presented in this study leverages machine learning and PDF processing to extract, analyze and classify rows from PDF into quiz worthy questions. The system integrates natural language processing (NLP), classification models, and can identify key concepts,

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

distinguish types of questions, and filter out the relevant content. Among others, the approach is addressed towards improving quiz generation in such a way that manual intervention is reduced but the quality of automatically generated questions remains high.

The proposed system takes several stages, such as text extraction from PDF documentation, preprocessing extracted content, feature extraction and classification by machine learning models. The model can train the classifier on diverse educational texts to understand factual statements, definitions, multiple choice questions, and open ended questions etc. Not only this, based on the direct involvement of teachers, this automation streamlines the process of creation of quizzes and also provides an efficient solution for creating assessments at scale.

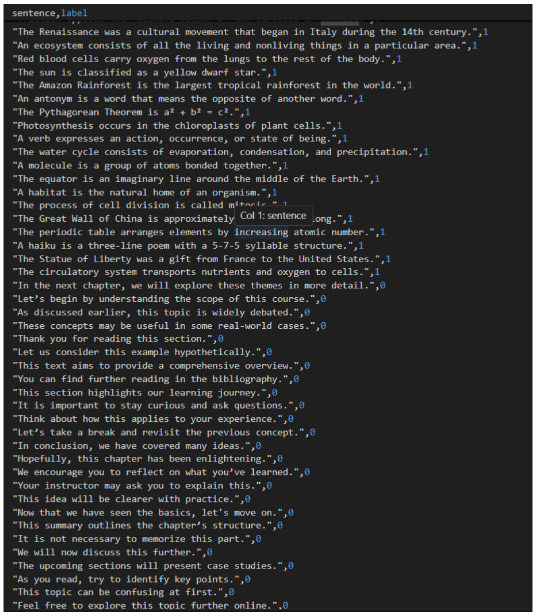
## 2 Related work

This section of the paper focuses on related work that may help this project come into fruition. Various papers published on journals.plos.org, nbci, and cambridge are further studies in order for the readers to grasps how this project is created.

Automated question classification and answer systems have improved education environments, content retrieval and data analysis in different fields due to their advancement. This works explores the different methodologies such as machine learning and natural language processing, so as to identify what questions can be classified best, so as to answer them in the least amount of time. Zylich et al. [1] describe a question-answering framework that automate question retrieval and provide on demand relevant course materials to answer commonly asked student questions thereby reducing teacher effort in online learning contexts. Kim et al. [2] also use machine learning to classify mathematical test questions based on difficulty and the benefits from adaptive learning are pointed out for educational users. Supervised machine learning is shown how to perform classification tasks in open ended survey data, thereby allowing for improved qualitative data analysis and interpretation by Haensch et al. [3]. In addition, a detailed survey on question classification methods [4] explains thoroughly the application of different ways to improve intelligent question answering systems that can improve the accuracy as well as the speed by which the response is generated. Taqi and Ali [5] investigated different automatic question classification models and illustrate the value of different machine learning processes for improving the effectiveness of question classification. Other studies [6] describe how spatial analysis tools can assist with development of intelligent question answering and decision making processes. Research in medical and healthcare related domains including tasks involving the automatic classification of clinical questions [7] confirms that these systems are capable of being applied outside of the education domain to retrieve relevant medical information [8] demonstrates how these systems can be applied beyond education, aiding in the retrieval of relevant medical information. Expanding upon these insights, the study "Automatic Question Classifier" [9] investigates techniques like natural language processing and machine learning classifiers to categorize questions effectively. Moreover, the article "Automatic Question Classifiers: A Systematic Review" [10] provides a comprehensive analysis of various automatic question classification systems, highlighting their methodologies and applications, Expanding on these insights, the study Automatic Question Classifier [9] investigates the integration of NLP and machine learning classifiers to improve categorization. Furthermore, Automatic Question Classifiers: A Systematic Review [10] provides an in-depth analysis of various classification systems, highlighting their methodologies and applications. Overall, these studies show the growing influence of automation, artificial intelligence and machine learning in question classification and answering systems, as well as in learning environments, in improving assessment accuracy and in more efficient information retrieval in different areas.

### 2.1 The Dataset

Our dataset contains 46,138 labeled sentences gathered from diverse sources of education including lecture slides, PDF textbooks, teacher notes, and transcripts of presentations and then generated by AI to create similar sentences in different fields. The materials cover a range of different academic fields including computer science, engineering, accountancy, and mathematics. Every sentence is marked with respect to whether it is "quiz-worthy," i.e., appropriate for application in an exam or quiz scenario. To ready the data, we used a number of preprocessing operations. Sentences were lowercased, special characters and punctuation removed, and tokenized. We also eliminated frequent



```
sentence,label
"The Renaissance was a cultural movement that began in Italy during the 14th century.",1
"An ecosystem consists of all the living and nonliving things in a particular area.",1
"Red blood cells carry oxygen from the lungs to the rest of the body.",1
"The sun is classified as a yellow dwarf star.",1
"The Amazon Rainforest is the largest tropical rainforest in the world.",1
"An antonym is a word that means the opposite of another word.",1
"The Pythagorean Theorem is a² + b² = c².",1
"Photosynthesis occurs in the chloroplasts of plant cells.",1
"A verb expresses an action, occurrence, or state of being.",1
"The water cycle consists of evaporation, condensation, and precipitation.",1
"A molecule is a group of atoms bonded together.",1
"The equator is an imaginary line around the middle of the Earth.",1
"A habitat is the natural home of an organism.",1
"The process of cell division is called mitosis.",1
"The Great Wall of China is approximately [illegible] long.",1
"The periodic table arranges elements by increasing atomic number.",1
"A haiku is a three-line poem with a 5-7-5 syllable structure.",1
"The Statue of Liberty was a gift from France to the United States.",1
"The circulatory system transports nutrients and oxygen to cells.",1
"In the next chapter, we will explore these themes in more detail.",0
"Let’s begin by understanding the scope of this course.",0
"As discussed earlier, this topic is widely debated.",0
"These concepts may be useful in some real-world cases.",0
"Thank you for reading this section.",0
"Let us consider this example hypothetically.",0
"This text aims to provide a comprehensive overview.",0
"You can find further reading in the bibliography.",0
"This section highlights our learning journey.",0
"It is important to stay curious and ask questions.",0
"Think about how this applies to your experience.",0
"Let’s take a break and revisit the previous concept.",0
"In conclusion, we have covered many ideas.",0
"Hopefully, this chapter has been enlightening.",0
"We encourage you to reflect on what you’ve learned.",0
"Your instructor may ask you to explain this.",0
"This idea will be clearer with practice.",0
"Now that we have seen the basics, let's move on.",0
"This summary outlines the chapter’s structure.",0
"It is not necessary to memorize this part.",0
"We will now discuss this further.",0
"The upcoming sections will present case studies.",0
"As you read, try to identify key points.",0
"This topic can be confusing at first.",0
"Feel free to explore this topic further online.",0
```

Figure 1: The Dataset

stop words by utilizing NLTK’s stopword corpus. Lemmatization and stemming were considered but eventually not used since they took a toll on classification performance in our initial experiments.

To extract features, we employed TFIDF (Term Frequency-Inverse Document Frequency) to numerically represent the sentences. This was able to pick up both the frequency and saliency of words within the dataset. We employed the TfidfVectorizer in scikit-learn with a max_features parameter of 5,000, unigrams and bigrams, and the filtering of terms that were too infrequent and too frequent. We also tested other feature representations such as word2vec and Doc2Vec, but TFIDF was more suitable for our structured text.

Normalization was internal to TF-IDF vectors, and we did not require extra scaling. In certain implementations of the model, we experimented with adding extra features such as sentence length, occurrence of certain keywords (e.g., "define", "explain"), and part-of-speech tag histograms, but our final model employed only TF-IDF vectors since they struck a balance between simplicity and performance.

The data was divided into 50% training, and 50% testing. It was obtained from open-access resources such as MIT OpenCourseWare, Wikibooks, Project Gutenberg, and educational datasets on Kaggle. Following is a small portion of our data:

> "Define the time complexity of the merge sort algorithm." → Quiz-worthy
>
> "Sorting is a common operation in software engineering." → Not quiz-worthy
>
> "Describe accounting principles applicable under accrual-based reporting." → Quiz-worthy

This data was critical to training and testing our classifier to extract meaningful content automatically from educational texts.

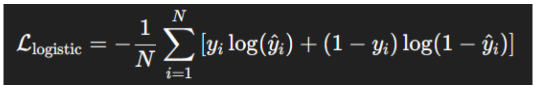
## 3 Methods

This section of the papers describes the step-by-step process that were implemented in creating this project. It will focus on training the model, and logistic regression. Supervised Learning was used in the dataset feeding it to the model. Vectorization was also used

### 3.1 Training and Learning Algorithm

Training consisted of fitting two classification models: logistic regression and a support vector machine (SVM). We selected these models on the basis of interpretability and their strong performance for binary text classification. We employed scikit-learn's implementation for both, with a TF-IDF vectorized representation of sentences as input.

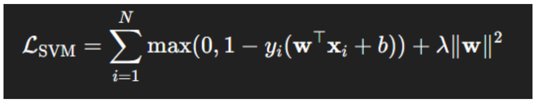
The logistic regression model employs a sigmoid activation function to output the probability of a sentence being quiz-worthy. The task is to minimize the binary cross-entropy loss, which is expressed as:



$$\mathcal{L}_{\text{logistic}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Figure 2: Binary cross-entropy loss used in logistic regression

The SVM, on the other hand, was trained using a linear kernel and optimized with the hinge loss function:



$$\mathcal{L}_{\text{SVM}} = \sum_{i=1}^{N}\max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) + \lambda\|\mathbf{w}\|^2$$

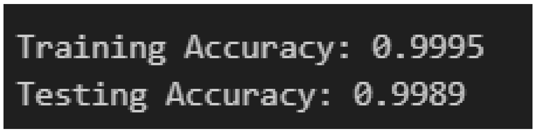
Figure 3: Hinge loss function used in SVM

Training was performed on 50% of the dataset, while 50% for testing. Hyperparameters (e.g., regularization strength, kernel type) were tuned using grid search and evaluated using F1-score and accuracy.

## 4 Experiments/Results/Discussion

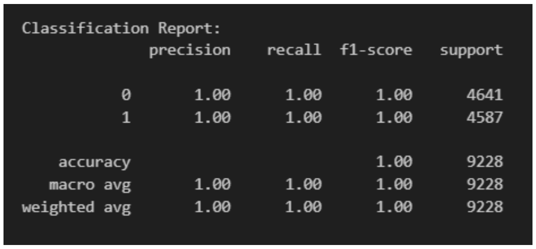
Following the procedures outlined in the earlier section, here are the results from training the logistic regression and support vector machine (SVM) classifiers for predicting quiz-worthy sentences. The performance of the models in acquiring useful patterns from the data is evaluated based on validation set evaluation, considering measures like accuracy, precision, recall, and F1-score. This is supplemented with an analysis of the training process and some comments on model convergence and classification behavior. The strengths and weaknesses of each of the models are discussed further in this section and their potential for real-world use such as automated reviewer generation is mentioned.

### 4.1 Training and Classification Report



```
Training Accuracy: 0.9995
Testing Accuracy: 0.9989
```

Figure 4: Testing and Training Accuracy



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4641
           1       1.00      1.00      1.00      4587

    accuracy                           1.00      9228
   macro avg       1.00      1.00      1.00      9228
weighted avg       1.00      1.00      1.00      9228
```

Figure 5: Classification Report

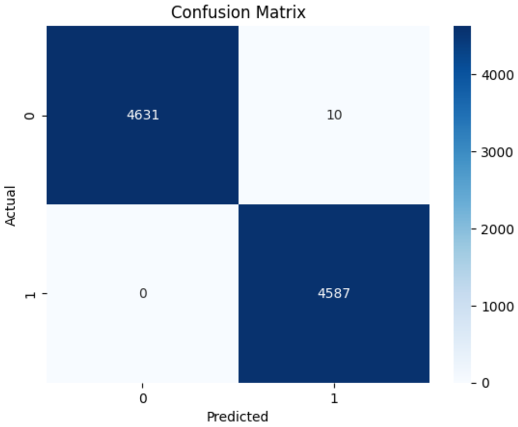
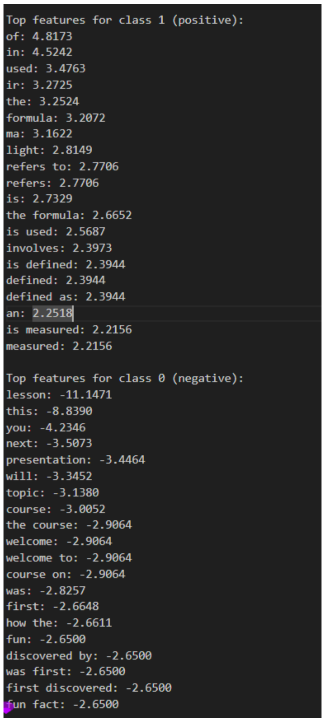


Figure 6: Confusion Matrix



```
Top features for class 1 (positive):
of: 4.8173
in: 4.5242
used: 3.4763
ir: 3.2725
the: 3.2524
formula: 3.2072
ma: 3.1622
light: 2.8149
refers to: 2.7706
refers: 2.7706
is: 2.7329
the formula: 2.6652
is used: 2.5687
involves: 2.3973
is defined: 2.3944
defined: 2.3944
defined as: 2.3944
an: 2.2518
is measured: 2.2156
measured: 2.2156

Top features for class 0 (negative):
lesson: -11.1471
this: -8.8390
you: -4.2346
next: -3.5073
presentation: -3.4464
will: -3.3452
topic: -3.1380
course: -3.0052
the course: -2.9064
welcome: -2.9064
welcome to: -2.9064
course on: -2.9064
was: -2.8257
first: -2.6648
how the: -2.6611
fun: -2.6500
discovered by: -2.6500
was first: -2.6500
first discovered: -2.6500
fun fact: -2.6500
```

Figure 7: Top Features

We can see that the model has a really high accuracy rate based on this results. After testing, it is avail to consistently filter out sentences that are not quiz worthy. Thus proving that the model is not overfitted.

### 4.2 PDF Extraction

We used PyMuPDF, fpdf, and regular expression for cleaning out sentences. It is proccess using PyMuPDF which extracts the texts from the pdf, and then it sanitize the text for any escape sequences by replacing them with dash and commas. We then clean and split the texts based on what was replaced in the sanitization, filtering out filler header sentences so that it doesnt show in the results. And returns it into a list which is fed into the joblib model and nltk.

### 4.3 Model Application using Flask

We developed a simple prototype so that we can use the model without having to interact with the codes of the program.

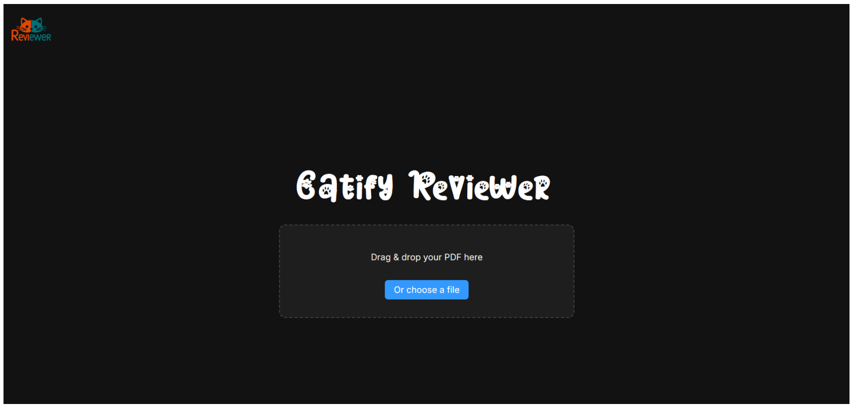


Figure 8: Home Page

To use the app, we will just need to upload a pdf file and the model will process it and will be downloaded as a pdf format for the user to use.

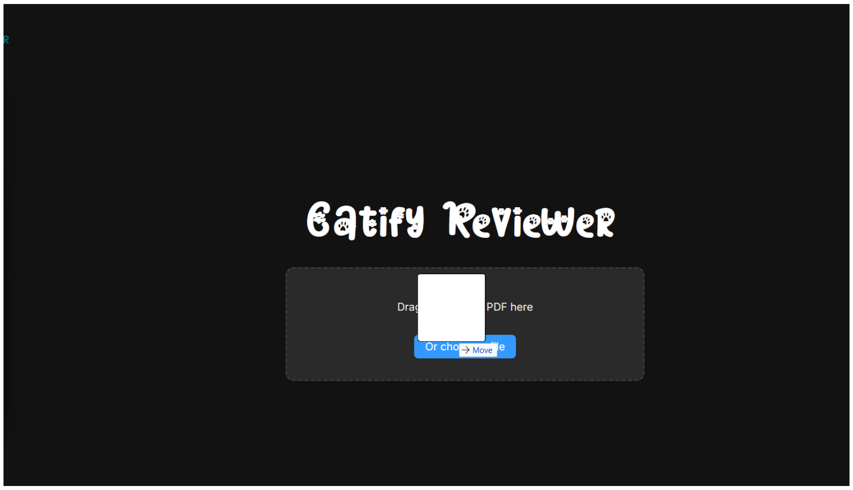
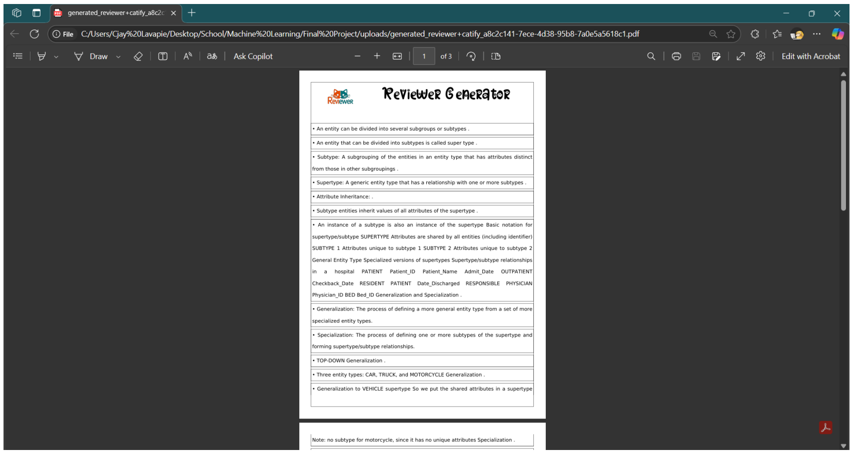


Figure 9: Sample Use


generated_reviewer+catify_a8c2c
File
C:/Users/Cjay%20Lavapie/Desktop/School/Machine%20Learning/Final%20Project/uploads/generated_reviewer+catify_a8c2c141-7ece-4d38-95b8-7a0e5a5618c1.pdf
Draw
Ask Copilot
1
of 3
Edit with Acrobat
Reviewer

Reviewer Generator

• An entity can be divided into several subgroups or subtypes .

• An entity that can be divided into subtypes is called super type .

• Subtype: A subgrouping of the entities in an entity type that has attributes distinct from those in other subgroupings .

• Supertype: A generic entity type that has a relationship with one or more subtypes .

• Attribute Inheritance: .

• Subtype entities inherit values of all attributes of the supertype .

• An instance of a subtype is also an instance of the supertype Basic notation for supertype/subtype SUPERTYPE Attributes are shared by all entities (including identifier) SUBTYPE 1 Attributes unique to subtype 1 SUBTYPE 2 Attributes unique to subtype 2 General Entity Type Specialized versions of supertypes Supertype/subtype relationships in a hospital PATIENT Patient_ID Patient_Name Admit_Date OUTPATIENT Checkback_Date RESIDENT PATIENT Date_Discharged RESPONSIBLE PHYSICIAN Physician_ID BED Bed_ID Generalization and Specialization .

• Generalization: The process of defining a more general entity type from a set of more specialized entity types.

• Specialization: The process of defining one or more subtypes of the supertype and forming supertype/subtype relationships.

• TOP-DOWN Generalization .

• Three entity types: CAR, TRUCK, and MOTORCYCLE Generalization .

• Generalization to VEHICLE supertype So we put the shared attributes in a supertype

Note: no subtype for motorcycle, since it has no unique attributes Specialization .

Figure 10: Sample Output

While the output still has some little bit of inconsistencies when filtering out sentences. It was able to fully extract all the sentences that are pertaining to a subject or quiz-worthy.

## 5 Conclusion/Future Work

The project was for me a success given the limitation on resources that I had. I was able to achieve my desired results realistically. Logistic Regression was the model to use and Vectorization was the most common out of all the Related Works that I have gathered. If I were to improve upon this. Due to the limitation in my resources, if given the chance, I would like to gather enough resources and build upon that for my database instead of having AI generated database so that it can be more concise when it comes to filtering out sentences when classifying them. I would manually extract sentences found in the pdfs if given. And I would like to add the feature of having to detect images included in the pdf and have a flashcard output instead of a pdf file.

## References

[1] Chen, H.-Y., Shih, P.-C. and Wang, Y. (January 9, 2025) *Exploration of designing an automatic classifier for questions containing code snippets-A case study of oracle SQL certification exam questions*, PLOS ONE. Available at: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0309050 (Accessed: 22 March 2025).

[2] Mulla, N. and Gharpure, P. (2023) *Automatic question generation: A review of methodologies, datasets, Evaluation Metrics, and applications, Progress in Artificial Intelligence.* Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9886210/ (Accessed: 22 March 2025).

[3] Sun, H. et al. (2023) *Question classification for intelligent question answering: A comprehensive survey*, MDPI. Available at: https://www.mdpi.com/2220-9964/12/10/415 (Accessed: 22 March 2025).

[4] V. A. Silva, I. I. Bittencourt, and J. C. Maldonado, *"Automatic question classifiers: A systematic review,"* IEEE Trans. Learn. Technol., vol. 12, no. 4, pp. 485–502, Oct./Dec. 2019.

[5] Taqi, Mustafa Kadhim and Rosmah Ali. *"Automatic question classification models for computer programming examination: A systematic literature review." Journal of theoretical and applied information technology* 93 (2016): 360-374.

[6] Zylich, B. et al. (2020) *Exploring automated question answering methods for teaching assistance, Artificial Intelligence in Education: 21st International Conference*, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I. Edited by I.I. Bittencourt et al. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7334161/ (Accessed: 22 March 2025).

[7] Haensch, A.-C. et al. (2022) *The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis, Frontiers in big data.* Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9403118/ (Accessed: 22 March 2025).

[8] Kim, G.I., Kim, S. and Jang, B. (2023) *Classification of mathematical test questions using machine learning on datasets of Learning Management System questions*, PloS one. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10584180/ (Accessed: 22 March 2025).

[9] Pan, X. et al. (2020) *A survey on deep learning for named entity recognition in natural language processing, ACM Computing Surveys* (CSUR). Available at: https://dl.acm.org/doi/10.1145/3386252 (Accessed: 22 March 2025).

[10] Mitkov, R. and Ha, L. Q. (2003) *Computer-aided generation of multiple-choice tests, Natural Language Engineering.* Available at: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/computeraided-generation-of-multiplechoice-tests/ (Accessed: 22 March 2025).