VILNIUS TECH
Faculty of Fundamental Sciences

Machine Learning-Based ChatGPT Usage Detection in Open-Ended Question Answers

Birutė Pliuskuvienė, Urtė Radvilaitė, Pavel Stefanovič, Simona Ramanauskaitė

Abstract. Today, chatGPT is one of the most widely used large language models for different tasks. The open access to the possibilities of the chatGPT leads to problems in the educational area because many students start to use it to solve various knowledge assessment tasks: homework, tests, midterms, and exams. In such cases, students start cheating instead of trying to understand the study material. In this research, the suitability of traditional machine learning algorithms to detect the usage of chatGPT in the answers text of the open-ended questions was examined. The dataset was collected using the midterm exam answers of the VILNIUS TECH bachelor's students. The experimental investigation has been performed by dividing the dataset into three (student answer, chatGPT answer, rephrased chatGPT answer) and two (student answer, non-student answer) classes separately. The various combinations of text preprocessing have been taken into account, and in this way, the influence of preprocessing options has been analyzed for chatGPT usage detection models' accuracy.

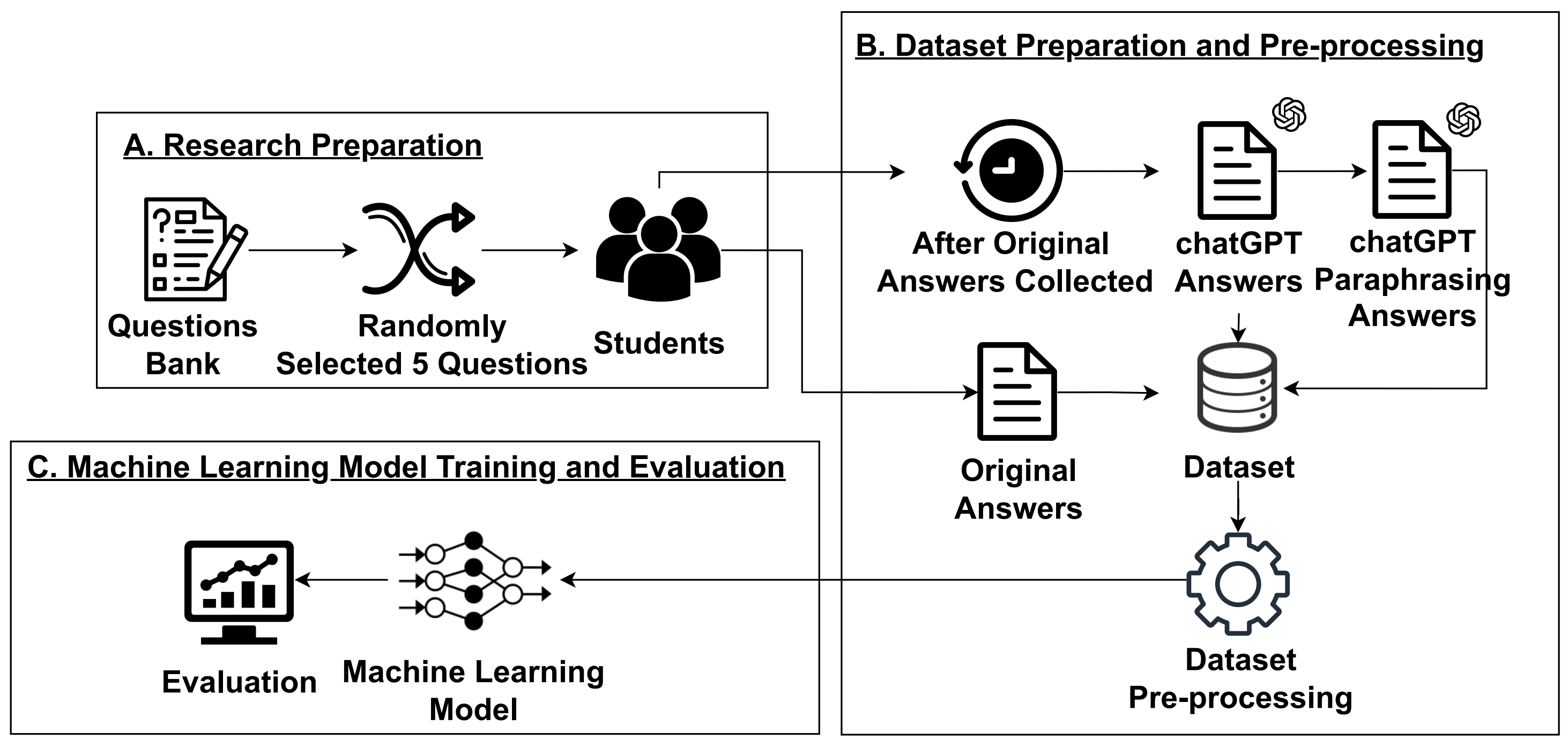

B. Dataset Preparation and Pre-processing
A. Research Preparation
Questions Bank
Randomly Selected 5 Questions
Students
After Original Answers Collected
chatGPT Answers
chatGPT Paraphrasing Answers
Original Answers
Dataset
C. Machine Learning Model Training and Evaluation
Evaluation
Machine Learning Model
Dataset Pre-processing

Figure 1. The principal scheme of the performed research

A. Research Preperation. The research has been prepared and conducted with 118 students of the last bachelor's degree course at VILNIUS TECH university. During the midterm exam, five questions (15 questions bank) were randomly given to each student to answer. The students had to answer the given questions by themselves, without additional material. The questions required reveal students' understanding rather than definitions. During the midterm exam, 118 students attended. All answers have been collected (in Lithuanian), evaluated, and labelled in the dataset as original student answers. The example of the question in English: "*In your own words, explain what artificial intelligence is. Give two specific examples of the application of artificial intelligence*."

B. Dataset Preparation and Pre-processing. Students could get additional points by participating in additional testing, where they had to answer five random questions from the question bank, but this time using chatGPT. The two requirements have been given: 1) the student must provide the original answer given by chatGPT; 2) the student must paraphrase chatGPT's answer and provide it. Totally 53 students participated and provided answers in this test. The obtained text was combined and the whole dataset consists of 1120 student answers: 590 original answers (118 students), 265 chatGPT answers (53 students), and 265 paraphrased chatGPT answers (53 students). To perform a deeper analysis of the newly prepared dataset, the dataset was pre-processed using different filters and two datasets were formed (Fig. 2). A dataset has been divided into two separate datasets by labelling dataset items differently. The first dataset has 3 classes - original students' answers, chatGPT answers, and paraphrased chatGPT answers. Another dataset has 2 classes - original students' answers and chatGPT answers with paraphrased chatGPT answers considered in the same class.

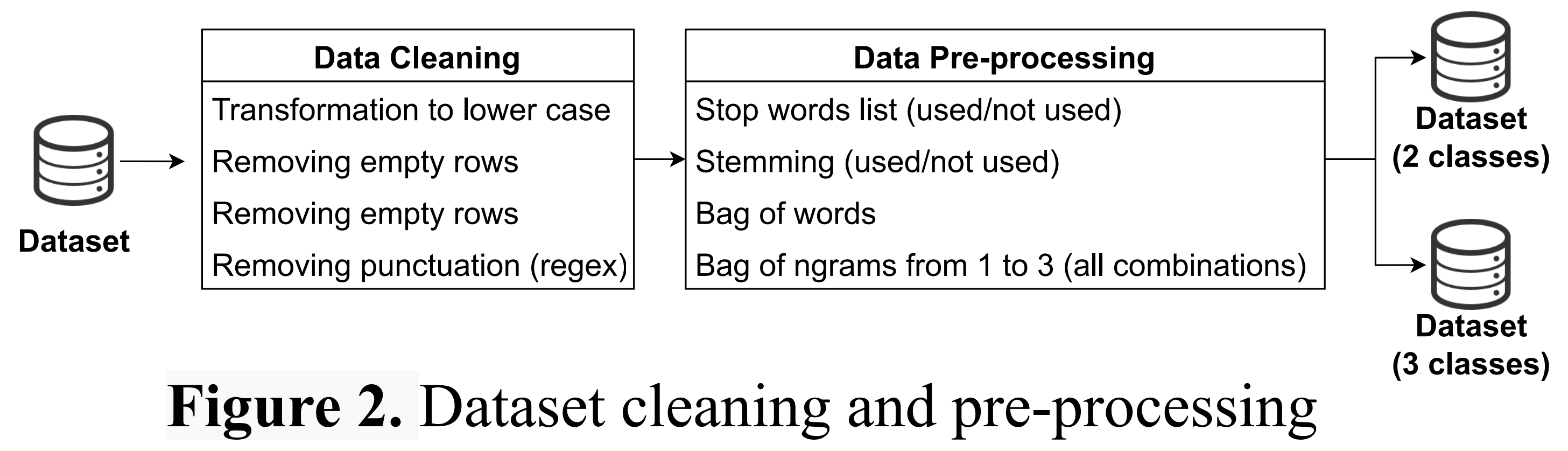
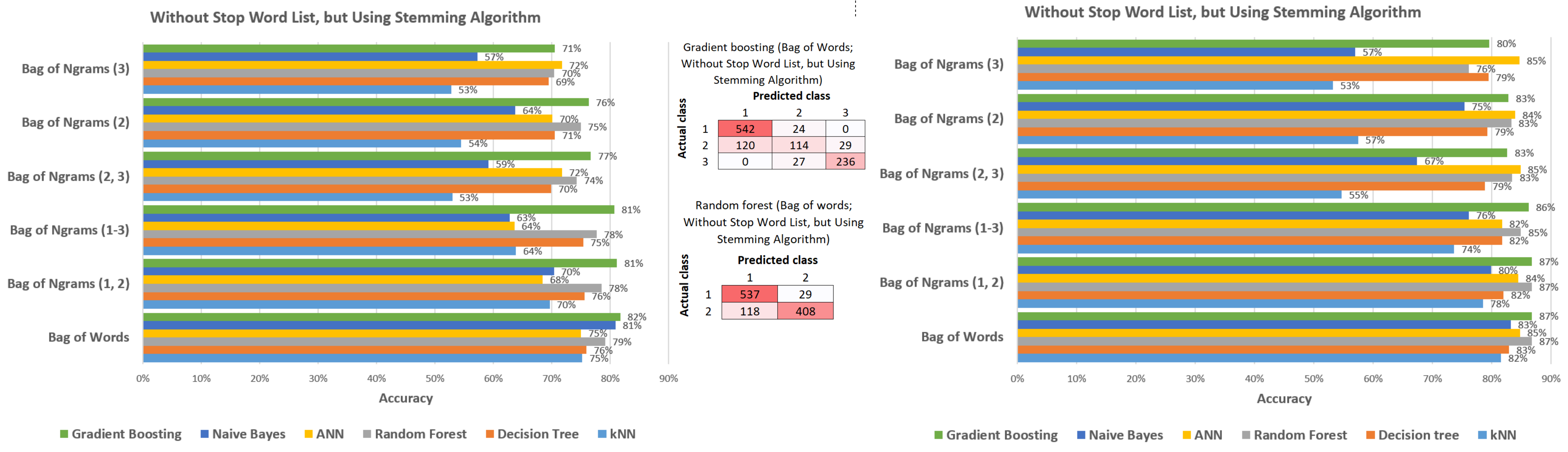


Figure 2. Dataset cleaning and pre-processing

C. Machine Learning Model Training and Evaluation. Several traditional classification algorithms have been chosen: artificial neural networks (ANN), k-nearest neighbours (kNN), decision trees, random forests, gradient boosting trees, and Naive Bayes. To train and test each machine learning model, stratified k-fold cross-validation was used, where. To estimate the performance of each model, four measures were calculated: accuracy, recall, precision, and the F1 score. To summarise the experimental investigation results, only the accuracy results are presented in the diagrams.



Gradient boosting (Bag of Words; Without Stop Word List, but Using Stemming Algorithm)

Actual class	Predicted class 1	2	3
1	542	24	0
2	120	114	29
3	0	27	236

Random forest (Bag of words; Without Stop Word List, but Using Stemming Algorithm)

Actual class	Predicted class 1	2
1	537	29
2	118	408

Figure 3. Summarized results of performed experimental investigation

Conclusion. The presented research methodology and dataset-building strategy integrate different cases of students' answers: individual answers, chatGPT-generated answers and student-rephrased chatGPT answers. With this method, one can detect plagiarism at different levels and complexity - a full copy or a rephrased version. At the same time on its basis developed models are more adaptable to teacher and education system needs, considering whether the situation allows for rephrased answers or not. Considering the 3 class model results combined with 2 class results (accepting both chatGPT based cases into one case), the 3 class model is more acceptable since it achieves the same accuracy as separately built 2 class models (in both cases 87% accuracy is achieved).