# Comparison of Supervised Machine Learning Methods for Automated Assessment of Student's Responses to Dichotomously Scored Items in Financial Literacy Test

**Conference Paper** · December 2021

CITATIONS
0

READS
4

**1 author:**

Nikita Kolachev
National Research University Higher School of Economics
**6** PUBLICATIONS **3** CITATIONS

SEE PROFILE

# Comparison of Supervised Machine Learning Methods for Automated Assessment of Student's Responses to Dichotomously Scored Items in Financial Literacy Test

Nikita I. Kolachev[1]

[1]HSE University, 20 Myasnitskaya Ulitsa, Moscow, 101000, Russia

**Abstract**

The article is devoted to comparing methods of automatic checking of dichotomously scored items in the financial literacy test. Such approaches to natural language processing as "bag-of-words", n-grams (within word boundaries), n-grams (across the whole response), and the fastText pre-trained embeddings were analyzed. The logistic regression was used to classify students' answers. The analysis was conducted on the data of ninth-graders from one of the Russian regions. As a result, it was concluded that the "bag of words" is not suitable for automated evaluation of responses, and it is better to utilize the n-grams method with vectorization over the whole student's response.

**Keywords**

financial literacy, supervised machine learning, natural language processing, logistic regression, dichotomous items, open-ended questions

## 1. Introduction

Financial literacy is considered an essential skill that helps achieve economic and financial stability and development in a country [1]. The standard definition of financial literacy is as follows: it is "the knowledge and understanding of financial concepts and risks, and the skills, motivation, and confidence to apply such knowledge and understanding to make effective decisions across a range of financial contexts, to improve the financial well-being of individuals and society, and to enable participation in economic life" [1, p. 128].

The growth of the citizens' welfare is formed through their dynamic financial behavior based on the use of savings and insurance instruments [2]. The formation of such behavior requires a sufficiently high level of financial literacy, which serves as a basis for the interaction of citizens with various financial institutions, their conscious use of products of the banking and insurance sector, the formation of a pension provision strategy [3].

Lack of financial literacy does not allow citizens to effectively plan their budget, make decisions in the field of personal or family finances, focused on the long term. The lack of basic

*MACSPro'21: Conference on Modeling and Analysis of Complex Systems and Processes, December 16–18, 2021, Moscow, Russia*

nkolachev@hse.ru (N. I. Kolachev)

0000-0002-3214-6675 (N. I. Kolachev)



knowledge and skills prevents individuals from consciously choosing and evaluating products and services, increasing the risk of becoming a victim of fraud by potentially unscrupulous financial market participants.

One of the conditions for solving the problem of knowledge and skills deficit is the formation of essential competencies in the field of financial literacy among school-age children. Special attention is paid to them because people form habits and behaviors at a young age, learning from parents and others around them, which indicates the importance of early interventions to help shape beneficial actions and attitudes [1]. The school makes a significant contribution to developing children's basic competencies in financial literacy. Already in primary school, children can begin to get acquainted with the basic concepts that reveal the sources of their family's budget and ways to use it effectively. As the child grows older, both the system of these concepts and the complexity of the tasks they face in managing personal and family finances expand.

For understanding students' level of financial literacy, the latter should be measured appropriately. One of the key requirements for educational measurement is its objectivity. The latter is usually twofold: as specific objectivity, which means that comparisons between individuals should be independent of instruments used [4, 5] and as independence of test results from raters/assessors. The second objectivity applies only to tests with open-ended questions. While specific objectivity is generally ensured by using appropriate probabilistic models (e.g., Rasch models), the other one is achieved by implementing and carrying out a variety of procedures: training and "blinding" of raters, studying the consistency of raters' scores, engaging a third party to assess an examinee's answer in case of raters' discrepancies. All those things are associated with material and time costs.

Quickly, cheaply, and quite objectively, it is possible to measure some characteristics of a person using a multiple-choice test. However, such tests cannot constantly assess the full range of competencies, the complexity of thinking, and reasoning ability [6]. This is why large-scale international educational studies (e.g., PISA, TIMSS, PIRLS) use open-ended questions in their measurement tools.

In financial literacy tests, the problem of objectivity doubles because this area of assessment is not tied to a school subject, so raters from different disciplines can check students' answers: mathematics, civics, economics, geography, etc. Evaluation of responses by various subject experts can distort student results, even with careful preparation and training of raters. Moreover, automated assessment of freely constructed responses reduces the cost of school monitoring and helps decrease teachers' workloads.

Given the above, an attempt was made to find an algorithm that would best classify students' answers into correct (1 point) and incorrect (0 points).

## 2. Method

### 2.1. Participants

Four thousand nine hundred and seventy-eight students of ninth grade from 58 schools of one of the Russian regions participated in the study. The sample was gender-balanced, and the

respondents' modal age was 15 years. Thirty-seven percent of the sample were from rural schools.

## 2.2. Instrument

The financial literacy test was constructed based on the PISA framework for financial literacy [1]. The primary element of the instrument is a testlet built on a specific story (situation) of social life. Testlets are a form of item development based on common stimuli (texts, diagrams, tables, etc.). The items were constructed on the following principles:

- *comprehensiveness* (items contain different formats of information presentation and aimed at different areas of financial literacy assessment following the PISA research approaches);
- *the presence of a problem* (items include questions that involve solving a particular financial problem, including taking into account alternative financial possibilities);
- *contextuality* (items are based on real-life situations);
- *personal involvement* (the issues offered in the items are relevant to students);
- *levels of difficulty* (items were developed taking into account mastery of social practices of using financial products and services by students and forms of their cognitive activity);
- *competence* (items allow assessing the following range of competencies: identify financial information, analyze information in a financial context, evaluate financial issues, and apply financial knowledge and understanding).

The whole test pool consisted of 9 testlets comprised of 6 ordinary dichotomous or polytomous items (questions). The testlets were assigned to 4 test variants. The test design is presented in Figure 1. The test was administered based on the incomplete test design when students did not perform all items but interacted only with some of them (in our case, 18 items which are 33% of the whole test). The respondents' results were equated based on the common items approach using probabilistic (item response theory, IRT) models. The last testlet in three test variants served as a link.

The test demonstrated appropriate reliability (Rasch reliability = .71). Twenty-one questions were open-ended.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variant 1 | Testlet 1 | Testlet 2 | Testlet 3 | | | | | | |
| Variant 2 | | | Testlet 3 | Testlet 4 | Testlet 5 | | | | |
| Variant 3 | | | | | Testlet 5 | Testlet 6 | Testlet 7 | | |
| Variant 4 | | | | | | | Testlet 7 | Testlet 8 | Testlet 9 |

**Figure 1:** The incomplete test design

## 2.3. Data collection

The test took 40 minutes. Students performed the test in class settings. For several days after the testing procedure, there was a check of open-ended questions by raters from among students' teachers.

The raters were trained to score the students' answers. The training explained the approaches of the international PISA study, which formed the basis of the evaluation system for the proposed financial literacy assessment tool. Items were predominantly checked by two raters, which allowed the study of discrepancies between their scores.

After expert work, a data set with the actual students' answers and scores was formed. For purposes of clarity, we selected only dichotomously scored items (0 = wrong answer, 1 = correct answer). We used 11 145 student responses to 17 open-ended dichotomously scored questions for machine learning algorithms. Most of the respondents' answers (69%) received 1 point.

## 2.4. Data analysis plan

This paper compared the following approaches to working with natural language (text) in order to predict students' results: "bag of words", the n-grams method (creating n-grams within word boundaries and across the whole response), and pre-trained embeddings. "Bag of words" is an approach to text tokenization that transforms texts into unitary sparse vectors, making of words features that are appropriate for further analysis. The n-grams approach is also a tokenization technique that extracts an n-length sequence of characters or words and converts them into machine-reading vectors. For instance, if we want to excerpt two words or symbols, it would be a bigram. Researchers show that in Russian texts, it makes sense to use n-grams consisting of letters rather than words [7]. The n-grams might be extracted within word boundaries or even across the answer. Another popular approach is pre-trained word embeddings. They are vectorized representations of words capturing semantic and syntactic information. Words closer in meaning are expected to have more relative vector coordinates. Pre-trained word embeddings are retrieved from predefined fixed-sized text corpora. In contrast to the other two approaches, the pre-trained embeddings are the words with known vector coordinates.

Logistic regression was used to classify students' answers. Logistic regression is a method of binary classification based on information from predictors. This method was chosen over the others because it can be better understood by all agents of the educational process: stakeholders, principles, teachers, parents, and even students. Moreover, the results of the implementation of this method are more interpretable due to its low complexity. So, the dependent variable is the student's score. The features obtained from the above-mentioned natural language processing techniques are independent variables.

Metrics of classification quality were the area under the error curve (ROC AUC), as the classification design is not balanced, and recall because it is most relevant in the case of student assessment. In addition, we included the log loss metric since it is a more selective loss function amongst others [8]. Text preprocessing only had the conversion of students' responses to lowercase letters.

The learning procedure took place using the Python programming language (version 3) [9] with the scikit-learn package (version 0.24.2) [10]. The random split of the sample into a training and a test part was in the proportion of 70% to 30%. The values of quality metrics were calculated either on the training sample, using three cross-validation samples, or on the test one. Of all possible configurations of the n-grams method, the best quality was demonstrated by extracting 2 to 6 characters when vectorizing within word boundaries and from 2 to 8 characters when vectorizing throughout a student's response. These are the configurations of the n-grams

method shown in the tables below. Also, we included pre-trained embeddings taken from the "fastText" library. This choice is made since "fastText" is one of the Russian language's most efficient vector word representations [11].

## 3. Results

Table 1 presents the results of cross-validation on the training sample. Each metric is a mean value of three cross-validation samples. It can be noted that the n-grams method gives better results, expressed in a more accurate classification of students' answers than the "bag-of-words" method. However, the latter demonstrates the lowest log loss value. At the same time, the technique of pre-trained embeddings is not inferior to n-grams. The study of classification quality on the test sample will be conducted for all methods, except for the "bag of words".

**Table 1**
Results of cross-validation of classification quality metrics on the training sample

| Approach | ROC AUC | Recall | Log-loss |
|---|---|---|---|
| Bag-of-words | .719 | .866 | .321 |
| N-grams (within word boundaries) | .756 | .941 | .438 |
| N-grams (across the whole response) | .759 | .940 | .451 |
| Pre-trained embeddings (fastText) | .756 | .941 | .507 |

Table 2 shows the results of cross-validation on the test sample. Each metric is a mean value of three cross-validation samples. It is noticeable that both n-gram methods give close results, with a slight advantage of the n-gram method for the whole answer based on ROC AUC and recall metrics. Nevertheless, according to the area under the error curve, the best classification of students' responses is given by the pre-trained embeddings of the fastText library. At the same time, it has the worst log loss value among the metrics.

**Table 2**
Results of cross-validation of classification quality metrics on the test sample

| Approach | ROC AUC | Recall | Log-loss |
|---|---|---|---|
| N-grams (within word boundaries) | .719 | .969 | .440 |
| N-grams (across the whole response) | .720 | .974 | .451 |
| Pre-trained embeddings (fastText) | .756 | .941 | .504 |

## 4. Discussion

According to the study results, we may conclude that both the n-grams method and the fastText library embeddings for the Russian language can automatically assess students' answers in the financial literacy test. However, preference should be given to the n-grams method with vectorization over the whole student's response because this method has the best recall index; that is, this method more accurately gives the score of 1 to those schoolers who were also given

a score of 1 by the raters. Moreover, it is usually preferable to train models on the targeted data that a researcher possesses rather than implement pre-trained algorithms.

The proposed research cannot be considered definitive. Although the algorithms presented in this study are simple, understandable to all participants of the testing process, further work in this area can be devoted to studying the application of neural networks and other pre-trained embeddings for the automated scoring of students' answers. Also, it might be noted that the current study is dedicated to analyzing dichotomous data. Different algorithms might be found for polytomous items. Another limitation relates to the fact that for fair results, we need reliable and valid data. Before implementing machine learning algorithms, we must be sure that the rater's scores are not biased. In addition, it should be kept in mind that studied algorithms might not be transferred to other areas of assessment such as reading, mathematics, or science literacy. It depends on the degree to which students' answers have more or less standardization. For instance, items in mathematics literacy require more standardized and numerical responses than in reading literacy. Financial literacy items do not need many numerical operations, therefore, its algorithms might be more applicable to reading literacy than to mathematics one. Comparisons of algorithms across domains should be performed in the subsequent research.

## Acknowledgments

The author thanks Elena L. Rutkovskaya and Anastasia V. Polovnikova for developing the financial literacy items and for the opportunity to conduct this research.

## References

[1] OECD, Pisa 2018 assessment and analytical framework, 2019. URL: https://www.oecd-ilibrary.org/docserver/b25efab8-en.pdf?expires=1633161302&id=id&accname=guest&checksum=B9AD01740690F4A82979FB895CF331C8.

[2] A. Lusardi, O. Mitchell, V. Curto, Financial literacy among the young, Journal of Consumer Affairs 44 (2010) 358–380.

[3] A. Lusardi, O. Mitchell, The economic importance of financial literacy: Theory and evidence, Journal of Economic Literature 52 (2014) 5–44.

[4] G. Rasch, On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements, Danish Yearbook of Philosophy 14 (1977) 58–93.

[5] G. Fischer, Applying the principles of specific objectivity and of generalizability to the measurement of change, Psychometrika 52 (1987) 565–587.

[6] O. Liu, J. Rios, M. Heilman, L. Gerard, M. Linn, Validation of automated scoring of science assessments, Journal of Research in Science Teaching 53 (2016) 215–233.

[7] T. Litvinova, O. Litvinova, P. Panicheva, Authorship attribution of russian forum posts with different types of n-gram features, in: Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval, Association for Computing Machinery, 2019, pp. 9–14.

[8] V. Vovk, The fundamental nature of the log loss function, 2015. URL: https://arxiv.org/pdf/1502.06254.pdf.

[9] P. S. Foundation, Python language reference, version 3, 2021. URL: http://www.python.org.
[10] F. Pedregosa, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[11] A. Babii, M. Kazyulina, A. Malafeev, fasttext-based methods for emotion identification in russian internet discourse, in: Proceedings of the 13th ACM Web Science Conference, 2021, pp. 112–119.