READERBENCH – AUTOMATED FEEDBACK GENERATION FOR ESSAYS IN ROMANIAN

Irina TOMA[1], Andreea-Madalina MARICA[2], Mihai DASCALU[3], Stefan TRAUSAN-MATU[4]

The development of online environments has transformed written communication into one of the most frequently used types of interactions between individuals; this effect has increased even more during the COVID-19 pandemic, which imposed physical distancing restrictions. As writing is a key skill in everyday activities, it is important for people to have strong skills and to be capable to communicate their thoughts and beliefs in a structured form. This paper introduces automated scoring and feedback mechanisms for Romanian, derived from an online collection of freely available essays, and integrated in the ReaderBench platform. Several regression models are evaluated in terms of essay scoring accuracy, out of which Gradient Boosting Regression was selected based on its performance (R^2 = .42, MAE = 1.10 on a 10-point scale). The feedback mechanisms provide suggestions for improving the quality of writings based on several rules, which in turn rely on the textual complexity indices computed by the ReaderBench framework, together with meaningful components generated from a Principal Component Analysis.

Keywords: Natural Language Processing, Automated Essay Feedback, Automated Essay Grading, ReaderBench framework

1. Introductions

People use daily written texts to connect with each other, when interacting at their jobs, when communicating with friends, or while expressing ideas and emotions through social media posts, blogs, or articles. Good communication skills are important, as information should be transmitted in a complete and coherent manner, using appropriate language, thus ensuring that the receiver quickly and appropriately understands the message. In contrast, poor communication skills generate frustration and misunderstanding, both on the

[1] PhD student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: irina_toma@rocketmail.com

[2] Bachelors student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: andreea.marica@stud.acs.upb.ro

[3] Prof., Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: mihai.dascalu@upb.ro

[4] Prof., Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: stefan.trausan@cs.pub.ro

sender's and receiver's side, as the message can be incomplete or delivered in an ambiguous manner [1].

Communication abilities, focused especially on writing skills, are taught in schools immediately after children learn how to write. The Romanian curriculum [2] emphasizes the importance of communication starting with primary school, where students develop their abilities to transmit simple messages in a given context, to write short texts expressing their emotions and opinions on a subject, or to extract information from texts. In lower secondary school, the complexity of the tasks increases, as students learn to summarize texts, adapt the transmitted information to the targeted audience, and debate on simple statements. In higher secondary school, the subjects are more elaborated, students learn how to write different types of text (e.g., descriptions, argumentative essays, structured and non-structured essays, news, advertisements), and perform critical literature reviews. During each education cycle, students undertake written exams that asses the levels of their communication skills. The amount of information students accumulate in a relatively short amount of time can become overwhelming. Without proper guidance and support, they might experience anxiety, low motivation, and poor overall performance [3]. Looking from the tutor's perspective, the volume of work increases considerably, as they must evaluate the writings of a large number of students, a process which is both time and resource consuming. Moreover, the evaluation is subjective, as different teachers may score essays differently [4, 5].

A solution to the above-mentioned problem consists of implementing Automated Essay Scoring (AES) systems, which standardize the essay evaluation process, reduces subjectivity, and shortens the evaluation period. Several such systems have been implemented based on Natural Language Processing (NLP) techniques and statistical models, and most of these systems are available for the English language [6]. The first AES system was created in 1966 under the name Project Essay Grade (PEG). Its purpose was to ensure more effective essay evaluations [7]. The score prediction was based on a statistical evaluation of the surface linguistic features of a text such as mean word length, number of words, number of punctuation marks, number of different parts of speech [8]. The system does not use NLP techniques for essay analysis and does not evaluate text semantics. More advanced systems evaluate text features at lexical, semantic, syntactic and discourse levels [9] and provide a score for these texts. E-rater [10] uses statistical approaches, together with NLP techniques, and is structured into five modules. The first three modules extract features from a given text related to syntactic variety, organization of ideas, and vocabulary usage. The fourth module weights the previously identified features, while the fifth module computes the final scores [11]. IntelliMetric [12] uses a combination of statistical methods, NLP techniques and machine learning for essay scoring. The system evaluates over 300 text-related features which are grouped into five Latent Semantic Dimensions [13]: focus and unity, cohesion and consistency, organization and structure, sentence structure, and mechanics and conventions.

For each scoring systems, an Automated Writing Evaluation (AWE) application was developed, such as Criterion [14] for E-Rater and MYAccess [15] for IntelliMetric. AWEs provide, besides scoring, relevant feedback for the evaluated essays. One of the advantages of using AWE systems [16] consists of providing fast feedback suggestions to improve text quality. Students receive feedback in the same environment where they are writing the essay, allowing them to easily integrate the suggestions and revise their work. Feedback is received in an iterative manner, motivating them to continue the writing process and improve their writing skills. Another advantage is the separation of the feedback on several directions, such as grammar, syntax, and cohesion, allowing users to focus on each individual aspect. Besides feedback and scoring mechanisms, AWEs include tools useful in essay writing, such as spell checkers, dictionaries, and thesauri.

The experiments presented in this paper target the implementation of AES and AWE systems for Romanian integrated in the ReaderBench website. Starting from previous experiments performed by Dascalu, Gifu, and Trausan-Matu [17], and implemented in the ReadME system [18], this paper introduces several improvements: an online corpus of freely available essays, a Principal Component Analysis (PCA) analysis performed on a large collection of documents, experiments with various regression models for AES, as well as an online interface. The ReaderBench framework [19, 20] computes 216 textual complexity indices for Romanian structured in five categories:

- Surface indices – such as the word entropy, mean sentences and words per paragraph, mean words per sentence, mean punctuation marks per paragraph and per sentence, mean commas per paragraph and per sentence;
- Word complexity indices – refer to the difficulty of a word and include indices such as the mean word length, maximum and mean depth of a word in the hypernym tree, mean senses per word;
- Syntactic and morphologic indices – such as the mean and standard deviation of various parts of speech (e.g., prepositions, nouns, verbs, adverbs, and adjectives) and syntactic dependencies, per paragraph and per sentence;
- Semantic cohesion indices – estimate the local (intra-paragraph, between sentences) and global (i.e., inter-paragraph) cohesion using different semantic models, such as word2vec [21], and co-reference chains. The

word2vec model for Romanian uses a vector size of 300 and was trained on the ReadME corpus[3];

- Discourse structure indices – computed using discourse markers and cue phrases.

The AWE system uses a set of rules for providing feedback to improve the overall quality of texts. Each rule is defined by three elements: 1) a benchmark value that represent the reference value established using well-written essays, 2) a threshold value that defines the tolerance interval, and 3) a set of feedback messages triggered for values beyond the tolerance interval. Similarly to the AES system, the AWE system uses the ReaderBench textual indices to provide feedback [22, 23]. Since the number of indices is quite high, the AWE system must group these features into a reduced number of components and provide feedback accordingly; dimensionality reduction is performed using a Principal Component Analysis (PCA) [24]. Besides the PCA, feedback is also provided using the most representative textual complexity indices.

The next sections of the paper are structured as follows. The *Corpus* section describes the steps followed to create two relevant corpora of essays written in Romanian. The follow-up section, *Automated Essay Scoring*, evaluates several regression models proposed for automated scoring., while the section centered on *AWE* details the feedback generation process. The following section describes the user interface, whereas conclusions and future implementations are presented in last section.

2. Corpus

Two corpora were assembled for follow-up experiments: one for training the regression models used by the AES system, and one for calculating the benchmark values required by the AWE system. The first corpus consists of essays available online on two platforms – clopotel.ro[4] and ereferat.org[5]. These websites provide, free of charge, the full text of the essays written in Romanian, together with a mean user score. The essay themes fall into the following categories: Romanian language and literature, history, geography, biology, and psychology (see Table 1). The total number of essays is 3423 and additional filtering was applied to remove essays shorter that three paragraphs or longer than 100 paragraphs; thus, 3148 texts were selected.

A preliminary statistical analysis was performed on the selected essays to better understand their structure. The metrics computed for each essay were: the

[3] https://github.com/aleris/ReadME-RoTex-Corpus-Builder

[4] http://referat.clopotel.ro

[5] https://www.ereferat.org

mean words per sentence, mean sentences per paragraph, mean words per paragraph, and the corresponding standard deviation for each metric. The mean words per sentence is between 5 and 70, 96.56% of the sentences having less than 40 words. The mean number of words per sentence is lower than 10 words for 12.87% of the corpus, between 10 and 15 words for 38.34% of the corpus, between 15 and 30 for 43.55% of the corpus and above 30 words for 5.24% of the corpus. High values of standard deviation relate to text organization, as part of the essays contain headings or lists. Additional follow-up filtering should consider removing these essays from the corpus.

Table 1

Essays divided per category

Category	**# essays**
Romanian language and literature	1,644
Biology	391
Geography	388
History	567
Psychology	158
Total	**3,148**

The paragraphs from the essays contain between 1 and 11 sentences, the majority being short paragraphs, with 2-4 sentences. Mean words per paragraph is between 10 and 250 words, with 48.57% of the texts having 40, up to 70 words per paragraph. Acording to the studies performed by Nicki Stanton [25], coherent paragraphs elaborate on only one main idea and the ideal length of the paragraphs is between 8 and 10 lines to ensure lizibility. While accounting for previous considerations, the documents from the corpus are situated at the inferiour bound of these values.

The essay scores represent a mean score given by users who read and used the specific text; individual essay values were rounded to the nearest integer. The distribution of the scores is bimodal, with two peaks (see Fig. 1), having most of the essays covering high (8 - 10) or medium scores (5 - 6). The small number of essays scored below 5 represents an drawback in creating an AES that covers the entire range of possible scores; all follow-up experiments consider only essays in a normalized scale with essay between 5 and 10.

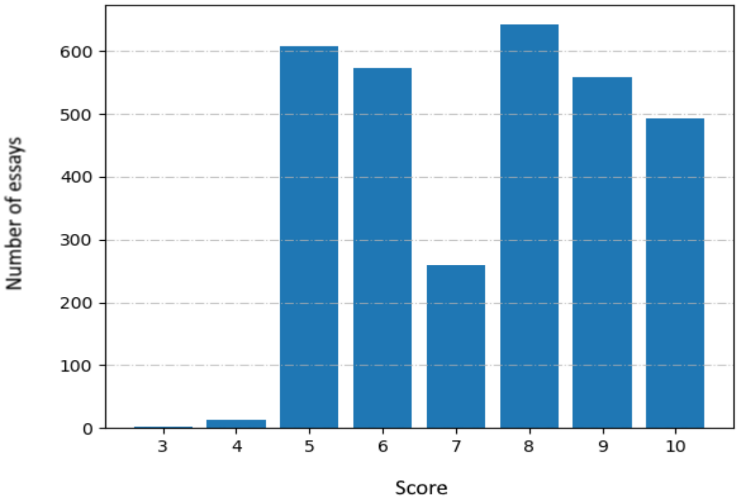


Fig. 1. The distribution of essay scores across the corpus.

The second corpus was required to perform the PCA and we relied on the ReadME corpus, a collection of freely available plain texts extracted from several online resources. The ReadME corpus is extensive, surpassing 1 billion tokens and covering 62% of the total words available in the Explanatory Romanian Dictionary (DEX). For this experiment, we considered subsections of the ReadME corpus, particularly texts from the following categories: science – the texts are extracted from the digital library of the Bucharest University of Economic Studies[6], literature – texts are represented by paragraphs extracted from books without copyright, and journalism – a series of news published in Gazeta de Cluj newspaper[7]. The identified texts were afterwards segmented in smaller parts (25 lines, up to 800 words), thus resulting in 60,595 reference texts for AWE.

3. Automated Essay Scoring

Several regression models were trained using the textual complexity indices from ReaderBench which were applied on the first developed corpus, the AES corpus: Linear Regression, Lasso Regression, Ridge Regression, Gradient Boosting Regression, Random Forest Regression, and Multilayer Perceptron. Model hyperparameters were chosen using grid search to maximize cross validation performance. The alpha hyperparameter for Lasso Regression was set at 0.01, while a value of 100 was used for Ridge Regression. Several hyperparameters, including tree-specific or boosting parameters, were set for the Gradient Boosting Regression: learning rate – 0.1, number of estimators – 100, the maximum depth of an individual estimator – 2, the minimum number of

[6] http://www.biblioteca-digitala.ase.ro/biblioteca/model/index2
[7] https://gazetadecluj.ro/stiri/stiri-cluj/

observations which are required by a node to be considered for splitting – 2, and a least squares cost function. The following hyperparameters were considered for the Random Forest Regressor: the maximum height of a decision tree – 100, the number of features taken into account for splitting a node – 3, the number of examples needed for splitting a node – 10 and the number of estimators – 1000. The following parameters were selected for MLP: the activation function for the hidden layer was set as Relu, a regularization parameter of 0.01, and a learning rate step of 0.1.

From the total number of essays in the corpus, a static split was performed: 80% of all entries were used for training the models and 20% for testing. Three metrics were calculated for each model: mean squared error (MSE), R^2 score, and mean absolute error (MAE). The results are displayed in Table 2. Gradient Boosting Regression was the most predictive model that explained $R^2 = .42$ variance; the model also exhibited the lowest value for mean squared error (MSE = 1.87) and mean absolute error (MAE = 1.1). Thus, this model was selected for follow-up analyses of essay scoring.

Table 2

Regression models evaluation

Model	**MSE**	**R^2 score**	**MAE**
Linear Regression	2.22	.32	1.21
Lasso Regression	2.21	.32	1.21
Ridge Regression	2.22	.32	1.23
Gradient Boosting Regression	1.87	.42	1.10
Random Forest Regression	2.07	.36	1.20
Multilayer Perceptron	3.02	.07	1.22

4. Automated Writing Evaluation

Personalized feedback is computed based on the textual complexity indices provided by ReaderBench and several representative principal components [24]. The PCA groups similar indices into several orthogonal components; however, a set of preliminary statistical filters were applied. As a first step, the indices that deliver a small volume of linguistic information were removed. In these category fall indices that have values of 0 or -1 for at least 80% of all entries. Also, texts that have at least 10% of the values for their indices classified as outliers were removed from the dataset. An index is considered outlier if the difference between its value and the mean value of the index calculated for all the AWE corpus exceeds the double value of the standard deviation. Considering these conditions, 74 indices and 294 documents were eliminated from the analysis. The next step consists of analyzing the statistical distributions of the indices; all indices that have the skewness value between than -4 and 4 or the kurtosis distribution between -10 and 10 were kept in the analysis [26], while the others were eliminated. After this preprocessing step, the total number of indices eliminated was 77.

The last preprocessing step uses the Pearson correlation coefficient to eliminate multicollinear indices from the normalized dataset; values higher than .9 denote extremely similar indices. The filtering algorithm consists of building an undirected graph, where the nodes are represented by the textual complexity indices and the vertices reflect the correlation relation between the nodes, if the value exceeds the imposed threshold. A greedy algorithm is applied on top of this graph. The nodes are sorted in descending order based on their rank; thus, nodes with most vertices are parsed first. The process is iterative: the most connected node is removed and alongside all corresponding edges. After this step, 20 indices were removed.

The PCA was applied on top of the remaining set of 45 indices. A total of four components were identified as being the most representative for the current analysis. The components explained a total variance of .693, contained more than two indices per component, and each component explained at least .01. A threshold value of .4 was set for the eigenvalues (i.e., loading factors) to exclude the less important indices from the components. Table 3 presents the textual complexity indices included in the components, together with their loading factors and the cumulative component variances. The four components reflect the following:

- C1: Lexical diversity – this component contains indices referring to mean values of different parts of speech, dependencies, as well as discourse elements;
- C2: Word complexity – the second component reflects the diversity of concepts. The textual complexity indices grouped in the component are related to word entropy and relations in the hypernym tree;
- C3: Local cohesion – this component is focused on local cohesion, containing indices that refer to the cohesion between paragraphs and word entropy;
- C4: Global cohesion – the last component is centered on global cohesion and contains the following indices: cohesion between the first paragraph and the middle section and transition cohesion between last sentence of current paragraph and the next paragraph.

The purpose of the feedback generation component is to provide relevant suggestions for improving the quality of a text. Therefore, a set of rules was introduced based on the 45 indices selected from ReaderBench, and the 4 components generated by the PCA. For each textual complexity index an upper and a lower bound were set, together with several feedback messages. The interval bounds were calculated as the mean index value minus/plus 2.5 times the standard deviation. For PCA, the minimum and maximum threshold values were calculated in relation to the AES corpus.

Table 3

PCA loading factors (M – mean; SD - standard deviation)

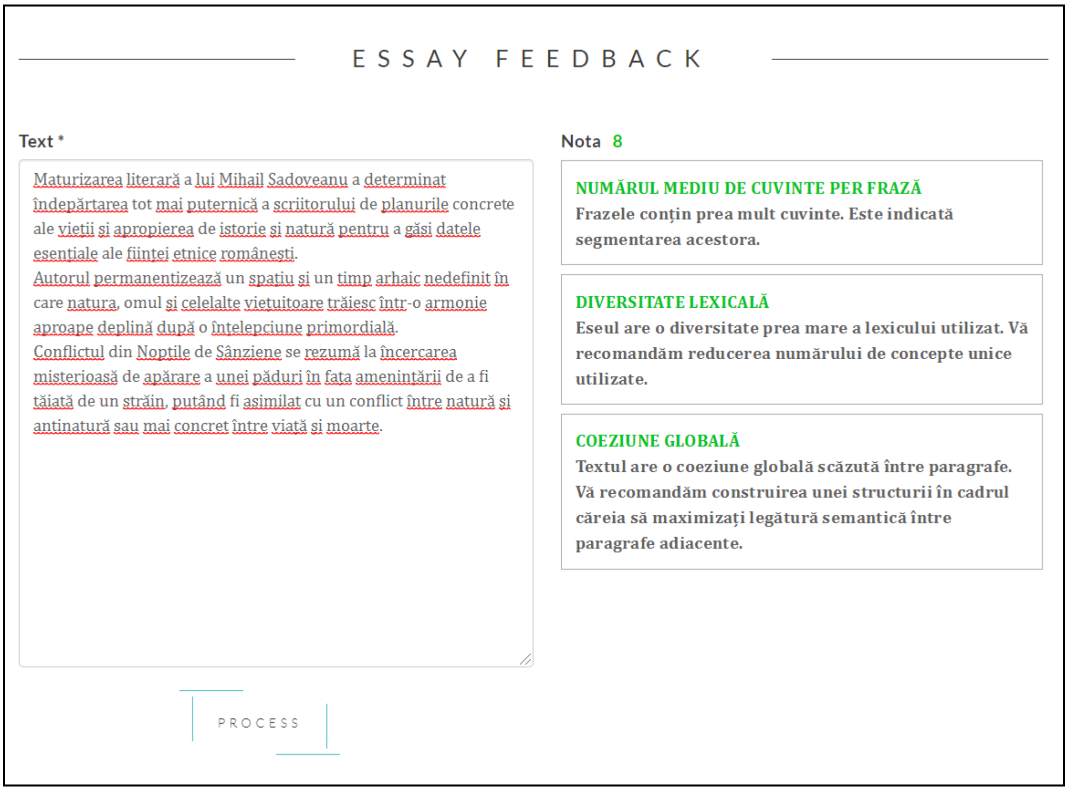
Name	**C1**	**C2**	**C3**	**C4**
M verbs per sentence	.939			
M oblique nominals per sentence	.884			
M coordinating connectives per sentence	.876			
M indefinite pronouns per sentence	.874			
M subject-object dependencies per sentence	.865			
M adverbs per sentence	.860			
M adverbial clauses per sentence	.854			
M nominal subjects per sentence	.844			
M fixed multiword expressions per sentence	.834			
M words per sentence	.828			
M determiners per sentence	.822			
M markers pe sentence	.817			
M interrogative pronouns per sentence	.804			
M coordinating conjunctions per sentence	.803			
M pronouns, third person, per sentence	.795			
M simple subordinators per sentence	.782			
M adjectival clauses per sentence	.777			
M reason and purpose connectors per sentence	.764			
M causal complements per sentence	.759			
M punctuation marks per sentence	.757			
M contrast connectors per sentence	.738			
M unique pronouns per sentence	.714			
M opposite connectors per sentence	.703			
M copulas per sentence	.681			
M indirect objects per sentence	.626			
M auxiliary dependencies per sentence	.607			
M open clausal complements per sentence	.551			
M adjacent cohesion between paragraphs	.540			
M character distance between words and their lemmas	.512			
M nouns per sentence		.889		
M word entropy at document level		.791		
M character entropy at word level		.670		
M maximum word depth in hypernym tree from root		.663		

Name	**C1**	**C2**	**C3**	**C4**
M word depth in hypernym tree from root		.608		
M paths to hypernym tree root based on all word senses		.455		
M named entities – persons		.427		
M 2-gram character entropy			.854	
M syllables per word			.828	
SD of character entropy at word level			.758	
M intra-paragraph cohesion			.531	
M transition cohesion between last sentence of current paragraph and upcoming paragraph, as well as the entire current paragraph and first sentence of the upcoming one				.505
SD word entropy at sentence level				.406
M cohesion between first paragraph and middle section				.403
Explained variance	**.446**	**.111**	**.090**	**.044**

5. User Interface

Both the scoring component and the feedback generation mechanism are exposed to the public in the web interface provided by ReaderBench[8]. The website [27] displays tools for Natural Language Processing, Online Communities Analysis, and Sentiment Analysis available for free. The newly introduced functionality is available under the *Demo* section, in the *Essay Feedback* area. The interface is simple and intuitive, as seen in Fig. 2. The *Essay Feedback* page is divided in two parts. In the left-hand side, the user introduces the text to be processed. A request is afterwards sent to the ReaderBench server to compute the appropriate score and feedback. The result is displayed in the right part of the screen. The first displayed element is the score (ro. “Nota”) computed using the Gradient Boosting Regression model. Under the score, the page displays a scrollable list of suggestions for improving the quality and structure of the text. The suggestions are generated according to the ReaderBench complexity indices and the previously identified principal components. Each suggestion contains a representative title and a short description. For the text in the left side, the following writing suggestions are displayed in the top of the list: Mean words per phrase (ro. “Numărul mediu de cuvinte per frază”), Lexical diversity (ro. “Diversitate lexicală”) and Global cohesion (ro. “Coeziune globală”). The first suggestion is based on the ReaderBench textual complexity index, while the other two are based on the Principal Components.

[8] http://readerbench.com



Maturizarea literară a lui Mihail Sadoveanu a determinat îndepărtarea tot mai puternică a scriitorului de planurile concrete ale vieții și apropierea de istorie și natură pentru a găsi datele esențiale ale ființei etnice românești.
Autorul permanentizează un spațiu și un timp arhaic nedefinit în care natura, omul și celelalte viețuitoare trăiesc într-o armonie aproape deplină după o înțelepciune primordială.
Conflictul din Noptile de Sânziene se rezumă la încercarea misterioasă de apărare a unei păduri în fata amenințării de a fi tăiată de un străin, putând fi asimilat cu un conflict între natură și antinatură sau mai concret între viață și moarte.

Nota 8

NUMĂRUL MEDIU DE CUVINTE PER FRAZĂ
Frazele conțin prea mult cuvinte. Este indicată segmentarea acestora.

DIVERSITATE LEXICALĂ
Eseul are o diversitate prea mare a lexicului utilizat. Vă recomandăm reducerea numărului de concepte unice utilizate.

COEZIUNE GLOBALĂ
Textul are o coeziune globală scăzută între paragrafe. Vă recomandăm construirea unei structurii în cadrul căreia să maximizați legătură semantică între paragrafe adiacente.

Fig. 3. *Essay Feedback* demo page

6. Conclusions

Written communication is the most present form of communication in the digital world. Texts people write should be correct in terms of grammar and syntax, coherent, and tailored according to the targeted audience. Writing skills can be exercised through automatic tools such as AES and AWE systems, that provide fast scoring and feedback suggestions. Moreover, automated systems deliver an objective evaluation that can be incorporated in the initial text in an unlimited number of iterations.

The current paper describes the development of an AES and AWE system for Romanian. The work follows the experiments conducted through the ReadME platform and brings several improvements in terms of corpus, PCA analysis and scoring models. The newly features are integrated in the ReaderBench online platform and are available for free to end-users. For the automated score calculation several regression models were evaluated. The metrics used in the evaluation were mean squared error, R^2, and mean absolute error. The model that offers the best results was Gradient Boosting Regression, as it scored the lowest value for mean squared error (1.87) and mean absolute error (1.10) and the highest values for R^2 (.42). Feedback generation is provided using the textual complexity indices from ReaderBench and principal components that group several indices. Each feedback message has a minimum and a maximum threshold determined on a corpus of well-formatted documents from multiple categories (science, literature, and journalism).

Future implementations are scheduled in three phases. The first phase will address the feedback mechanism. The textual complexity indices will be computed at more granular levels, such as paragraph, sentence, in order to offer specific feedback on fine-grained text sequences. The second phase includes enhancements in terms of user experience. The feedback from the user interface will be highlighted in the input text. As an example, the phrases that are not correlated to one another could be underlined in the left side of user interface and the feedback suggestions from the right side could be highlighted. The last phase of the implementation process targets user engagement through gamification. The introduction of progress mechanics, such as points or badges, challenges and stories would keep the user focused on the application and engaged in the learning process. Moreover, users would benefit from having personal statistics on their progress in the platform.

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number TE 70 *PN-III-P1-1.1-TE-2019-2209*, ATES – "Automated Text Evaluation and Simplification" and by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

REFERENCES

[1]. *R. Prabavathi and P.C. Nagasubramani*: "Effective oral and written communication." in *J. Appl. Adv. Res*. vol. **3**, no. S1, pp. 29, 2018.

[2]. *Institutul de Ştiinte ale Educaţiei*: "Programe Şcolare în Vigoare," retrieved from http://programe.ise.ro/Actuale/Programeinvigoare.aspx, at June 1, 2020.

[3]. *N.S.M. Daud, N.M. Daud, and N.L.A. Kassim*: "Second language writing anxiety: Cause or effect?" in *Malaysian journal of ELT research*. vol. **1**, no. 1, pp. 19, 2016.

[4]. *J. Wang, G. Engelhard Jr, K. Raczynski, T. Song, and E.W. Wolfe*: "Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach." in *Assessing Writing*. vol. **33**, pp. 36–47, 2017.

[5]. *M. Meadows and L. Billington*: "A review of the literature on marking reliability." in *London: National Assessment Agency*. 2005.

[6]. *M.A. Hussein, H. Hassan, and M. Nassef*: "Automated language essay scoring systems: A literature review." in *PeerJ Computer Science*. vol. **5**, pp. e208, 2019.

[7]. *L.M. Rudner and P. Gagne*: "An overview of three approaches to scoring written essays by computer." in *Practical Assessment, Research, and Evaluation*. vol. **7**, no. 1, pp. 26, 2000.

[8]. *E.B. Page*: "Project Essay Grade: PEG." M. D. Shermis & J. Burstein (Eds.). pp. 43–54. *Lawrence Erlbaum Associates Publishers* (2003).

[9]. *M. Liu, Y. Li, W. Xu, and L. Liu*: "Automated essay feedback generation and its impact on revision." in *IEEE Transactions on Learning Technologies*. vol. **10**, no. 4, pp. 502–513, 2016.

[10]. *Y. Attali and J. Burstein*: "Automated essay scoring with e-rater V.2.0." Annual Meeting of the International Association for Educational Assessment. pp. 23. *Association for Educational Assessment*, Philadelphia, PA (2004).

[11]. *J. Burstein, C. Leacock, and R. Swartz*: "Automated evaluation of essays and short answers." 2001.

[12]. *L.M. Rudner, V. Garcia, and C. Welch*: "An evaluation of IntelliMetric^{TM} essay scoring system." in *The Journal of Technology, Learning and Assessment*. vol. **4**, no. 4, 2006.

[13]. *S. Dikli*: "An overview of automated scoring of essays." in *Journal of Technology, Learning, and Assessment*. vol. **5**, no. 1, 2006.

[14]. *J. Burstein, M. Chodorow, and C. Leacock*: "Automated essay evaluation: The Criterion online writing service." in *AI Magazine*. vol. **25**, no. 3, pp. 27–36, 2004.

[15]. *H.C. Chou, M. Moslehpour, and C.-Y. Yang*: "My access and writing error corrections of EFL college pre-intermediate students." in *International Journal of Education*. vol. **8**, no. 1, pp. 144–161, 2016.

[16]. *R. Haswell*: "The Complexities of Responding to Student Writing; Or, Looking for Shortcuts via the Road of Excess." in *Across the Disciplines*. vol. **3**, 2006.

[17]. *M. Dascalu, D. Gifu, and S. Trausan-Matu*: "What Makes your Writing Style Unique? Significant Differences Between Two Famous Romanian Orators." In: Nguyen, N.-T., Manolopoulos, Y., Iliadis, L., and Trawinski, B. (eds.) 8th Int. Conf. on Computational Collective Intelligence (ICCCI 2016). pp. 143–152. *Springer*, Halkidiki, Greece (2016).

[18]. *I. Toma, T.-M. Cotet, M. Dascalu, and S. Trausan-Matu*: "ReadME – Your Personal Writing Assistant." , 2019.

[19]. *M. Dascalu, P. Dessus, S. Trausan-Matu, M. Bianco, and A. Nardy*: "ReaderBench, an environment for analyzing text complexity and reading strategies." In: Lane, H.C., Yacef, K., Mostow, J., and Pavlik, P. (eds.) 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013). pp. 379–388. *Springer*, Memphis, USA (2013).

[20]. *M. Dascalu, G. Gutu, S. Ruseti, I.C. Paraschiv, P. Dessus, D.S. McNamara, S. Crossley, and S. Trausan-Matu*: "ReaderBench: A Multi-Lingual Framework for Analyzing Text Complexity." In: Lavoué, E., Drachsler, H., Verbert, K., Broisin, J., and Pérez-Sanagustín, M. (eds.) 12th European Conference on Technology Enhanced Learning (EC-TEL 2017). pp. 495–499. *Springer*, Tallinn, Estonia (2017).

[21]. *Y. Goldberg and O. Levy*: "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." in *arXiv preprint arXiv:1402.3722*. 2014.

[22]. *M. Dascalu, S. Trausan-Matu, D.S. McNamara, and P. Dessus*: "ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism." in *International Journal of Computer-Supported Collaborative Learning*. vol. **10**, no. 4, pp. 395–423, 2015.

[23]. *M.-D. Sirbu, R. Botarleanu, M. Dascalu, S. Crossley, and S. Trausan-Matu*: "ReadME – Enhancing Automated Writing Evaluation." 18th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2018). pp. 281–285. *Springer*, Varna, Bulgaria (2018).

[24]. *S. Karamizadeh, S.M. Abdullah, A.A. Manaf, M. Zamani, and A. Hooman*: "An overview of principal component analysis." in *Journal of Signal and Information Processing*. vol. **4**,

no. 3B, pp. 173, 2013.

[25]. *N. Stanton*: “Constructing effective paragraphs.” What Do You Mean,‘Communication’? pp. 281–292. *Springer* (1986).

[26]. *D.L. Jackson, J.A. Gillaspy Jr, and R. Purc-Stephenson*: “Reporting practices in confirmatory factor analysis: an overview and some recommendations.” in *Psychological methods*. vol. **14**, no. 1, pp. 6, 2009.

[27]. *G. Gutu-Robu, M.D. Sirbu, I.C. Paraschiv, M. Dascalu, P. Dessus, and S. Trausan-Matu*: “Liftoff - ReaderBench introduces new online functionalities.” in *Romanian Journal of Human - Computer Interaction*. vol. **11**, no. 1, pp. 76–91, 2018.