

Framing automatic grading techniques for open-ended questionnaires responses. A short survey

Gabriella Casalino¹, Barbara Cafarelli², Emiliano del Gobbo³, Lara Fontanella⁴, Luca Grilli², Alfonso Guarino³, Pierpaolo Limone², Daniele Schicchi⁵ and Davide Taibi⁵

¹University of Bari, Department of Computer Science “Aldo Moro”, Piazza Umberto I, 70121 – Bari (BA), Italy

³University of Foggia, Department of Humanities, Via Arpi, 176, 71121 – Foggia (FG), Italy

²University of Foggia, Department of Economics, Management and Territory, Via da Zara, 11, 71121 – Foggia (FG), Italy

⁴University “G. d’Annunzio” of Chieti-Pescara, Department of Legal and Social Sciences, Via dei Vestini, 31, 66100 – Chieti (CH), Italy

⁵National Research Council of Italy, Institute for Educational Technology, Via Ugo la Malfa, 153, 90146 – Palermo (PA), Italy

Abstract

The assessment of students’ performances is one of the essential components of teaching activities, and it poses different challenges to teachers and instructors, especially when considering the grading of responses to open-ended questions (i.e., short-answers or essays). Open-ended tasks allow a more in-depth assessment of students’ learning levels, but their evaluation and grading are time-consuming and prone to subjective bias. For these reasons, automatic grading techniques have been studied for a long time, focusing mainly on short-answers rather than long essays. Given the growing popularity of Massive Online Open Courses and the shifting from physical to virtual classrooms environments due to the Covid-19 pandemic, the adoption of questionnaires for evaluating learning performances has rapidly increased. Hence, it is of particular interest to analyze the recent effort of researchers in the development of techniques designed to grade students’ responses to open-ended questions. In our work, we consider a systematic literature review focusing on automatic grading of open-ended written assignments. The study encompasses 488 articles published from 1984 to 2021 and aims at understanding the research trends and the techniques to tackle essay automatic grading. Lastly, inferences and recommendations are given for future works in the Learning Analytics field.

Keywords

systematic review, automatic grading, open-ended questions, Learning Analytics


Proceedings of the Second Workshop on Technology Enhanced Learning Environments for Blended Education (teleXbe2021), October 5–6, 2021, Foggia, Italy

✉ gabriella.casalino@uniba.it (G. Casalino); barbara.cafarelli@unifg.it (B. Cafarelli); emiliano.delgobbo@unifg.it (E. del Gobbo); lara.fontanella@unich.it (L. Fontanella); luca.grilli@unifg.it (L. Grilli); alfonso.guarino@unifg.it (A. Guarino); pierpaolo.limone@unifg.it (P. Limone); daniele.schicchi@itd.cnr.it (D. Schicchi); davide.taibi@itd.cnr.it (D. Taibi)

ORCID: 0000-0003-0713-2260 (G. Casalino); 0000-0003-1088-7306 (E. del Gobbo); 0000-0003-0931-2054 (L. Grilli); 0000-0002-9055-9689 (A. Guarino); 0000-0003-0154-2736 (D. Schicchi); 0000-0002-0785-6771 (D. Taibi)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

Grading students' written assignments is a crucial task for teachers at all levels of education. Yet, it is one of the toughest and burdensome tasks, subject to several challenges. Questionnaires can, in general, be closed-ended and open-ended. Closed-ended questions give students a set of previously written options to answer. Therefore, they are more suitable for automatic corrections and more likely to provide an objective evaluation. There already exists ample and consolidated literature on methodologies and techniques that can be used to evaluate such questionnaires to assess students' performance and specific skills. In this context, Item response theory (IRT) provides a valuable and theoretically well-founded framework for educational measurement [1]. In Massive Online Open Courses (MOOCs), IRT models have been used to avoid cheating on graded tests [2, 3] to improve peer assignment in peer assessment activities [4, 5], and also to build an adaptive learning module for a conversational agent to support learners [6]. Despite the benefit of closed-ended questions, not all the questions can be formulated in a closed-ended fashion. On the other hand, open-ended questions, requiring a text answer in a natural or formal language, allow a more in-depth assessment of students' capabilities and learning performances. However, their evaluation is time-consuming, requires a lot of concentration, and is more likely to be affected by graders' subjectivity. To overcome these issues, automatic grading methodologies for short textual answers have been studied for more than a decade. Different techniques have been used for implementing auto assessment and building the assessor module for intelligent tutoring systems. Many early works on automatic grading are mainly based on the similarity between students' answers and a reference answer. These methods perform well for questions that have a single or a minimal number of correct answers. However, some open-ended questions ask students to express their reasoning, precluding the construction of a reference answer.

In the light of the above, it is of particular interest for the research community focused on Learning Analytics [7, 8, 9, 10, 11, 12] to find out new methods and tools for supporting teachers in the assessment of students' learning performances through open-ended questions. Moreover, with the Covid-19 pandemic outbreak, we have witnessed a questionnaires deluge. The shifting of learning contexts from physical to virtual classrooms has made the evaluation of the students more difficult: very often, during these years characterized by home confinement, teachers have struggled to see students' faces [13], therefore to deeply understand their engagement, participation and lessons comprehension level [14, 15, 16]. For this reason, questionnaires as well as online tests have proliferated as tools to evaluate step-by-step students' performances.

Besides, in recent years, we have witnessed an ever-increasing availability of Massive Open Online Courses (MOOCs), e.g. Coursera¹, EduOpen². Through such MOOCs, students can acquire - via self-learning - a wide variety of competences and abilities. At the end of the lessons, they usually find questionnaires which are often closed-ended precisely due to the challenges posed by open-ended responses previously highlighted. In this perspective, automatic techniques for grading open-ended questionnaires can provide new potentials for intelligent tutoring [17, 18] throughout these online courses.

¹<https://www.coursera.org/>

²<http://learn.eduopen.org/>

Goals of automatic grading of questionnaires responses There are numerous benefits to be obtained from automatic grading in general, automatic grading of natural language responses, and automatic grading of open-ended questionnaires responses. These are themed around summative assessment (for providing grades), formative assessment (to give feedback), and effectiveness. Concerning summative assessment, the demands of large class sizes and assessment practices [19] require efficient and cost-effective solutions. In addition, humans make mistakes when grading, and consistency is needed when the inter-rater agreement is imperfect as result of fatigue, bias, or ordering effects [20]. Another benefit is that the idea of automatic grading in itself may promote the formalization of assessment criteria when not performed otherwise [21]. One must also consider the immediacy that automatic grading systems can provide, where test-takers would otherwise need to wait for human markers to complete the grading [22]. Concerning formative assessment, automatic grading is interesting for broader applications such as e-learning and intelligent tutoring systems. Finally, as for effectiveness, automatic grading is becoming very competitive with human grading for the assessment of open-ended questionnaires responses (both short-answers [23] and essays [24]).

Previous literature There are plenty of research contributions on automatic grading techniques for open-ended responses. Yet, there are very few surveys and reviews. In [25], the authors surveyed automatic grading techniques for short-answers, analysing 80 papers published between 1996 and 2014. They mainly focused on the advancement of methods and approaches. The authors found that statistical methods were the most used to tackle automatic grading, and Natural Language Processing techniques were widely adopted for extracting lexical, morphological, semantic and syntactic features from data. Moreover, they observed that this body of works was emerging, still there were barriers in the advancement of the research due to the impossibility of publishing the datasets employed for privacy reasons. In [26], the authors focused on the different software that have been adopted over the years to tackle such a problem, limiting the search to initiatives published in 2018/2019. They found that open-ended question grading software can be divided into two big groups. One group uses statistical approaches dealing with false-positive answers. While they are generally lowering the workload of creating questions, their primary disadvantage is low feedback: most such questions are incapable of hinting feedback and answer-until-correct feedback because providing such feedback may lead learners to false-positively graded answers, exposing the system vulnerability and lowering its reputation among the learners. The second group includes questioning systems likely to produce false-negative results; they generally perform better at providing hints and answer-until-correct feedback. Their challenges are the higher workload of creating questions, the necessity to account for every possible correct answer by a teacher, and lower error detection. The optimal choice for creating automatized e-learning courses are template-based open-ended question systems like those in [27, 28, 29] which allow answer-until-correct feedback and can find and report various types of errors. Such systems require more staff time to create questions but less staff time to manage the learning process in the courses once they are run.

With respect to previous literature, this article aims at depicting the state-of-the-art and the evolution over the years of automatic grading techniques for open-ended questionnaires responses. In particular, we perform a systematic mapping of the advancement in this research

field.

Highlights The primary contributions of this work are:

- An up-to-date overview on automatic grading techniques for open-ended questionnaires' responses;
- To encompass 488 papers published between 1984 and 2021 through the adoption of an automatic review tool;
- Inferences and recommendations are given.

Structure of the paper This article is organized as follows: Section 2 provides basic notions on questionnaires and techniques for analyzing natural language text; Section 3 details the methodology adopted to carry out this survey; Section 4 is devoted to present the results obtained and the main findings; and, lastly, Section 5 concludes with final remarks and recommendation for researchers.

2. Background

In this section, we dwell on the different types of questionnaires (Section 2.1), and on automatic text analysis (Section 2.2).

2.1. Questionnaires

One of the most popular forms of automated assignments is quizzes. The major advantage of using quizzes is that answers can be graded without teacher's intervention, so the learner can have immediate feedback about his/her learning level at any time. However, quizzes have significant disadvantages:

- (a) the possibility of guessing the correct answers,
- (b) insufficient feedback.

The two main categories of quiz questions are:

- *Closed-ended questions*: a closed-ended question is a question the learner can answer by selecting one of the options provided. They are often easier and take less time to answer, and the answers are more straightforward to analyse: the teacher, creating questions, can give detailed explanations for each wrong choice. Such questions are focused primarily on checking factual knowledge, contain a limited range of possible correct answers, and guide the learner's thoughts. However, closed-ended questions suffer from the disadvantage (a), i.e., the learner can answer correctly by guessing a choice, and they may encourage systematic guessing instead of reasoning for solving the task. Mitigation strategies for such downside exist, e.g., closed-ended questions can be provided with customised feedback for each wrong choice, explaining to learners their errors without the teacher interventions.

- *Open-ended questions*: they require a text answer in a natural or formal language that the learner will provide. Learners cannot guess answers for such questions from their text, forcing them to think about the answers and actually perform the tasks (e.g., maths calculations). However, analyzing free-text answers represent a challenge: the teacher cannot explain every possible mistake (i.e., all the aspects involved in the correctness of a response such as contents, presentation, syntax). The usage of open-ended text questions solves the guessing problem (disadvantage (a), above mentioned) but suffer from the downside (b), i.e., insufficient feedback. Open-ended questions with free-text answers tend to use simple short answers. The reason is that the number of possible correct natural-language answers rise steeply with the answer complexity. The flexibility of natural language plays against the teacher in this case. It is often practically unfeasible listing all semantically equivalent sentences that can be used to answer the question correctly. Variability in shorter natural-language answers and formal-languages answers is also a problem: if the answer contains several independent variable parts, the number of possible correct answers the question's author should provide raises exponentially.

2.2. Automatic text analysis

The autonomous analysis of free-text answers belongs to a research field known as Natural Language Processing (NLP). NLP concerns the study of methodologies that make a machine capable of analyzing natural language autonomously.

Literature shows different types of strategies to approach NLP ranging from statistical to information formal language theory. In this regard, Artificial Intelligence has changed the way to look at the related tasks by leveraging machine learning (ML) and deep learning (DL) methodologies. ML/DL exploits several techniques to infer knowledge examining a massive amount of data representative of a problem. Using such techniques allows the system to emulate the reasoning process of human beings. Their application has led to relevant outcomes in many different related tasks such as machine translation (i.e. converting one natural language into another autonomously) [30], assessment of text complexity (i.e. evaluating the text complexity of a sentence/document autonomously) [31], vocabulary enhancement (i.e. supporting students to improve vocabulary) [32], Social Media Analysis (i.e. analysis of textual data from Social Media) [33, 34], and autonomous essay scoring (i.e. grades to essays written in an educational setting autonomously) [35].

The analysis of natural language through the above methodologies considers converting the natural language to a numerical form describing the input text. Representing a text with numbers is a non-trivial process that should be tackled based on the system's goal. Usually, such representation is either computed apriori (e.g., by preprocessing systems) or created by the system that arranges the input data to optimize its performance to solve a task. ML/DL systems exploit text representation to apply a set of methodologies that vary on the task. For instance, grading an essay automatically through ML/DL systems needs a text representation that considers several aspects: grammar rules, word usage, rules for spelling, punctuation, capitalization, sentence structure variety, and the relevance of the content. After obtaining the numerical description of such features, the system learns how to map the representation of the essay to a small set of grades (e.g., from 1 to 6), acting as an instructor would do.

NLP is an ongoing research field that challenges researchers to develop new, more effective methodologies to overcome issues related to both the input representation and solving tasks methodologies. For example, the representation process might be computationally expensive or infeasible, leading to information loss or inaccurate representations. Choosing the solving task methodology needs to exploit a long process that tests several techniques that work accordingly to the computed representation. Moreover, the most crucial flaw of ML/DL systems is the difficulty to explain the way they make decisions; therefore, a user needs to be supported by external systems to fully understand the rationale of the system behavior.

3. Method

In this section, we provide details on the employed methodology for conducting this short survey.

Our work is loosely based on [36, 37, 38, 39]. The method adopted can be split into the following phases:

1. Planning of the review;
2. Investigation of the research questions;
3. Description of both sources of information and strategies used to collect data;
4. Definition of the selection and exclusion criteria used to filter the studies;
5. Comparison of the selected studies and research questions.

Planning In this step, we identified the research questions, the sources of information, and finally, the method used to select works focusing on automatic grading of open-ended questionnaires' responses. To reduce researcher bias, one of the authors of this work developed the protocol, while all together made discussions and comparisons about results found from the selected articles.

Research questions The definition of the research questions is the most crucial part of any review [38]. To define the research questions of this survey, we have identified and classified the existing literature focusing on automatic grading techniques for open-ended questionnaires' responses. Table 1 describes our research questions.

Identifier	Research Question
RQ1	What is current trend on automatic grading techniques for open-ended questionnaires' responses?
RQ2	What are the most used techniques for automatic grading of open-ended questionnaires' responses?

Table 1

List of the research questions (RQ) with specific questions (SQ) addressed in this work.

The articles were searched on Scopus³, the famous abstract and citation database, using the following query:

TITLE-ABS-KEY((*automatic text based grading*) **OR** (*open ended questions automatic grading*) **OR** (*short answers automatic grading*) **OR** (*open ended questions automatic assessment*) **OR** (*short answers automatic assessment*) **OR** (*short answers automatic marking*) **OR** (*open ended questions automatic marking*) **OR** (*short answers automatic scoring*) **OR** (*open ended questions automatic scoring*) **OR** (*short answers machine learning grading*) **OR** (*natural language processing exam grading*) **OR** (*natural language processing exam scoring*) **OR** (*natural language processing exam marking*) **OR** (*natural language processing exam assessment*) **OR** (*nlp exam grading*) **OR** (*nlp exam scoring*) **OR** (*nlp exam marking*) **OR** (*text answers grading*) **OR** (*text answers marking*) **OR** (*essays automatic grading*) **OR** (*essays automatic scoring*) **OR** (*essays automatic marking*) **OR** (*essays automatic assessment*) **OR** (*descriptive answers automatic marking*) **OR** (*descriptive answers automatic grading*) **OR** (*descriptive answers automatic assessment*))

The final query has been built studying the partial queries results and the documents' bibliographical references. The query was run on 12 July 2021 and returned 630 documents.

Selection and exclusion criteria After obtaining the studies, we removed the impurities from the search results. With impurities, we mean the names of conferences correlated to the search keywords that were in the search results. The abstracts of the studies were all inspected to include/exclude the right/not adapt studies from the review. In this phase, we filtered out works that were not relevant keeping only the most representative. In Table 2, we specify the *selection* and *exclusion* criteria adopted for this review.

Selection criteria

Article focused on automatic grading techniques for open-ended questionnaires' responses

Article focused on techniques for analyzing natural language texts with applications to grading

Exclusion criteria

Article without abstract and title written in English language

Article not focused on automatic grading techniques for open-ended questionnaires' responses

Table 2

Selection and *exclusion* criteria for the results found in the databases.

At the end of the *selection and exclusion* step the final dataset consists of 488 documents. The dataset has been exported in the `bibtex` format and analyzed using *Bibliometrix* [40], a R open source package, for performing comprehensive science mapping analysis. This package provides a full set of analytical tools suited for analyzing bibliographical data originating from several abstract databases. The package has also a Graphical User Interface companion to perform analyses directly without coding.

³<https://www.scopus.com/search/form.uri?display=basic#basic>

4. Findings

In this section, we show the results obtained from our systematic mapping of literature concerning automatic grading techniques for open-ended questionnaires' responses.

Table 3 includes key information about the dataset of identified papers. In particular, the collected documents span between 1984 and 2021, and come from 325 sources (journals, books, conferences, etc.).

Timespan	1984:2021
Sources (Journals, Books, etc)	325
Documents	488
Average citations per documents	7.697
Average citations per year per doc	1.018
References	11763
Authors	1231

Table 3

Key information about dataset of identified papers.

4.1. RQ1 – What is current trend on automatic grading techniques for open-ended questionnaires' responses?

Figure 1 identifies the documents published by year. We can observe that prior to 2001 the scientific publication on the topic is negligible, but starting from that time, the production started rising, with an impressive increasing rate in the last 5 years (2016-2021). This goes clearly in step with the introduction and adoption of technologies and computation for learning purposes. Indeed, it is around 2000-2005 that researchers started working on the analysis of students performance through learning management systems [41] such as Moodle⁴ (see, for instance, [42, 43, 44]). The ever-increasing availability of students' data fostered and brought out the need of using automatic techniques for analyzing both closed-ended and open-ended questionnaires responses. Furthermore, the spike we can note on 2020 could be partly explained by Covid-19 pandemic outbreak which changed the way of lecturing at schools constraining teachers to massively exploit technological web-based aids and therefore researchers to deeply and widely studying the topic. Figure 2 depicts – through a lollipop chart – the top-20 most relevant sources ordered by the number of paper published through them on automatic grading of open-ended questionnaires' responses. "Lecture Notes in Computer Science" by Springer⁵ has resulted to be the most relevant container for research initiatives on the topic with 46 papers out of 488. The second place is awarded to the "ACM International Conference Proceeding Series" with 18 articles. These two items represent outliers given they are container of several conference proceedings. Other relevant venues are "Advances in Intelligent Systems and

⁴<https://moodle.org>

⁵<https://www.springer.com/gp/computer-science/lncs>

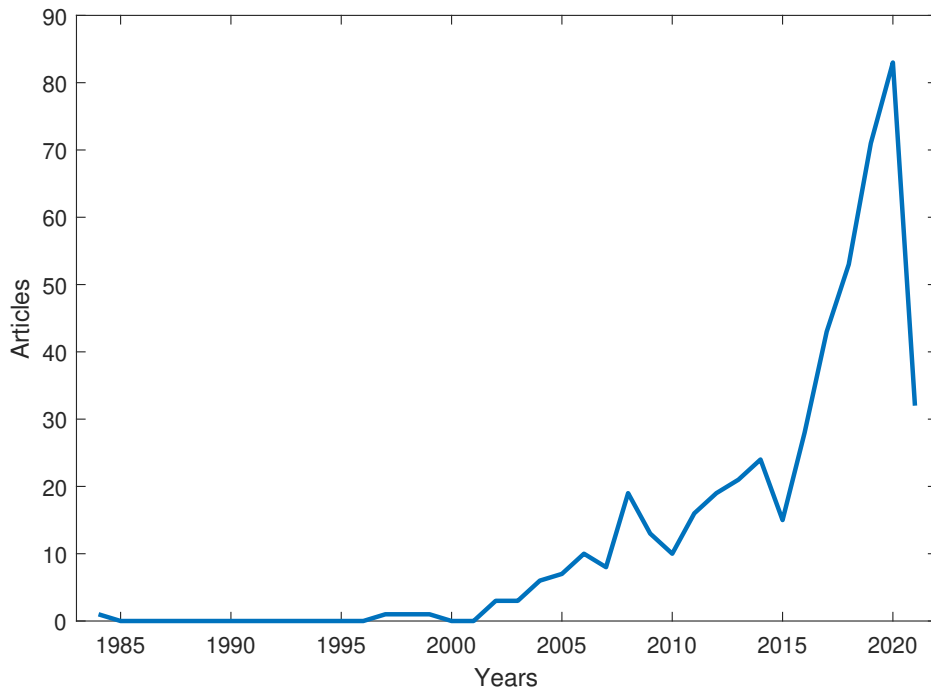


Figure 1: Annual scientific production.

Computing” by Springer⁶ with 10 papers, and “Communications in Computers and Information Science” by Springer⁷ with 8 articles. Further papers have been found in a comparable amount (1 to 6) on very wide set of sources (see Table 3). Figure 3 shows the most cited documents in the dataset tracked by Scopus. The first outlying result – with 258 citations – is the paper by M. Richardson et al. [45] which is focused on machine comprehension of texts, particularly interesting for the entire Natural Language Processing research area. The second one [46], with 148 citations, is a research properly focused on using semantic similarity techniques for automatic short-answer grading. L.S. Larkey’s article [47] (third place with 142 citations) is one of the oldest ones we have found. The author jointly used Linear Regression and clustering methods to classify texts. Overall, the most cited articles tackle lexical and latent semantic analysis (e.g., [48, 49, 50, 51]), using also Latent Dirichlet Allocation (e.g., [52]); others use machine learning (e.g., [53, 54, 55, 56]), while R. Siddiqi et al. [57] present a short-answer marking system exploiting structure matching, i.e., matching a prespecified structure, developed via a purpose-built structure editor. Lastly, interestingly we have found only one comprehensive literature review (e.g., [25]).

Figure 4 depicts the most active authors, by number of published documents, in this area. We can observe that the number of documents for each author is not really high, conversely to what

⁶<https://www.springer.com/series/11156>

⁷<https://www.springer.com/series/7899>

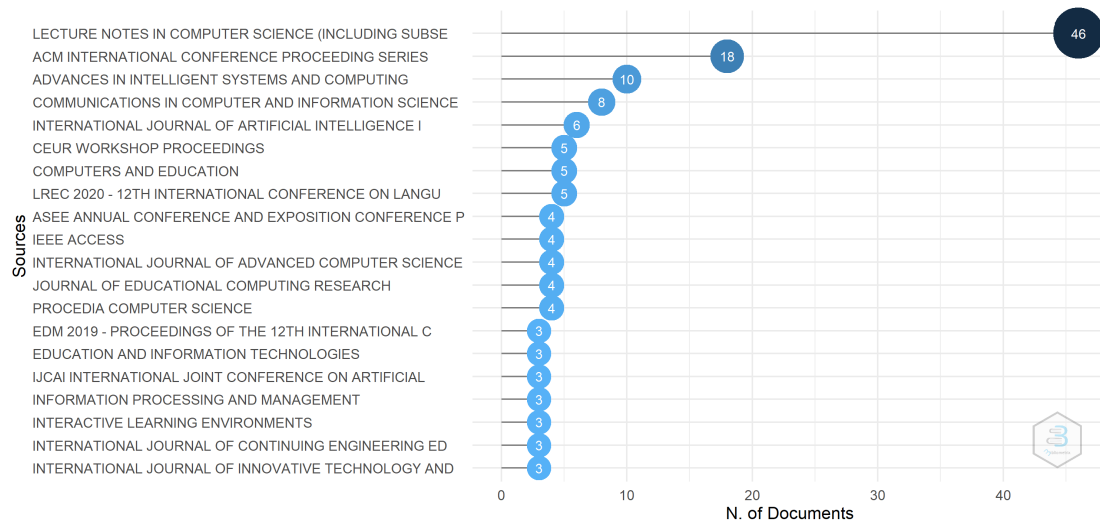


Figure 2: Most relevant sources for identified papers.

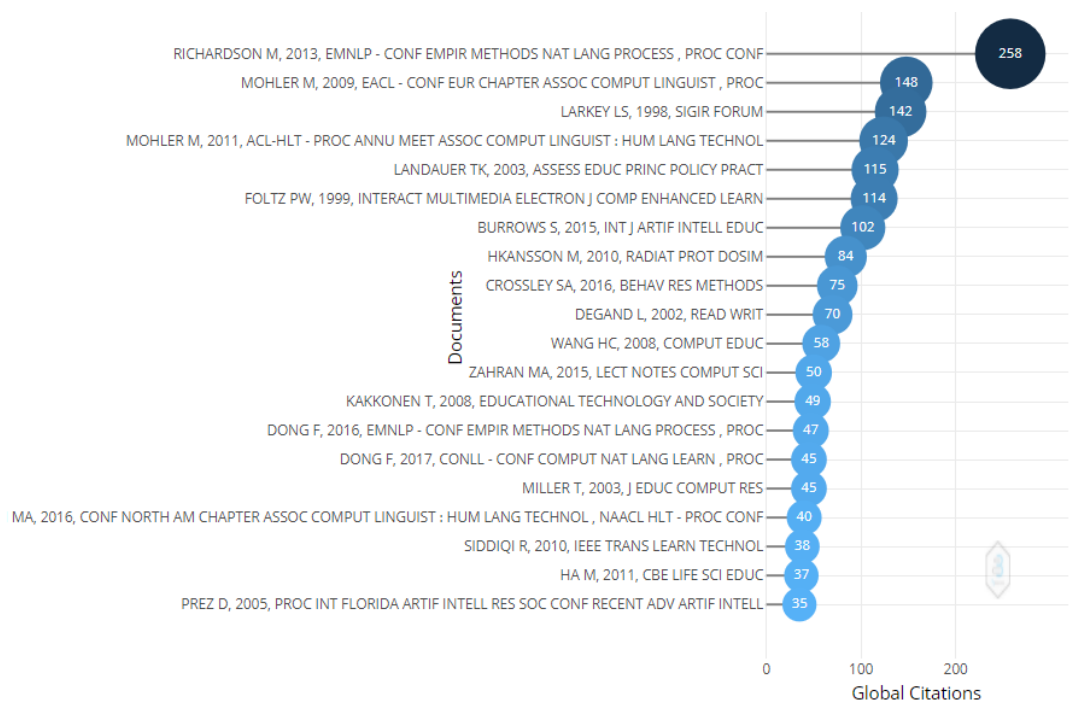


Figure 3: Most global cited documents.

we could expect due to the flourishing interest on automatic analysis of students' responses to open-ended questionnaires. Figure 5 shows the authors' publications over time for the top-20 authors (i.e., authors with more articles) in the dataset. The y-axis lists the authors, while the

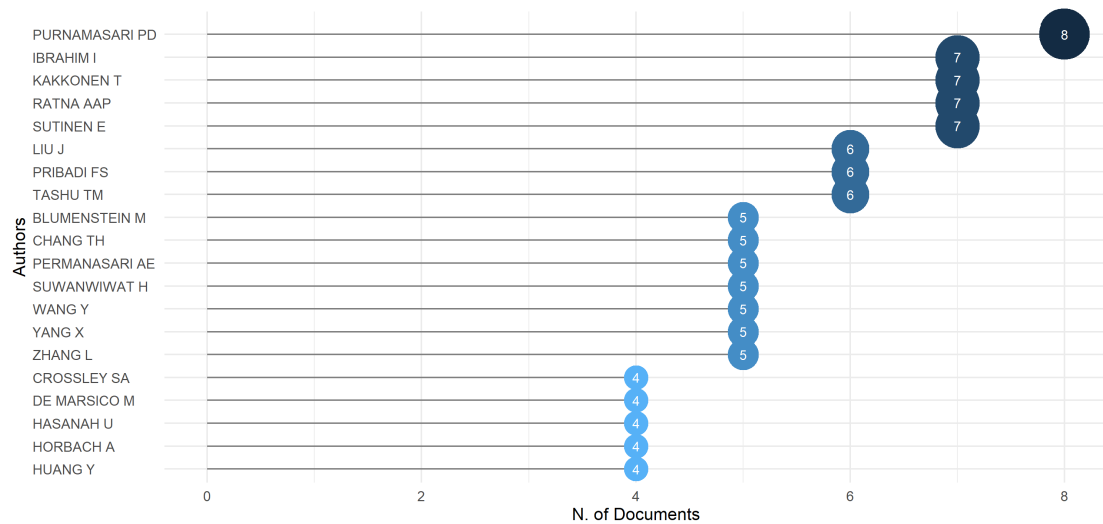


Figure 4: Most relevant authors by the number of published articles.

x-axis represents the years from 2004 to 2021. The red-colored line represents the authors' active period timeline. The blue-colored circles indicate the number of articles published in a specific year (e.g., S.A. Crossley has published one article in 2016). The blue intensity is proportional to the total citations per year of the document published in that year (e.g., the document published in 2016 by Crossley et al. has received 12.5 citations per year).

We observe that top-20 authors' publications are recent: 16 of them have been active on the topic mostly in the last five years, according to the recent research framework development.

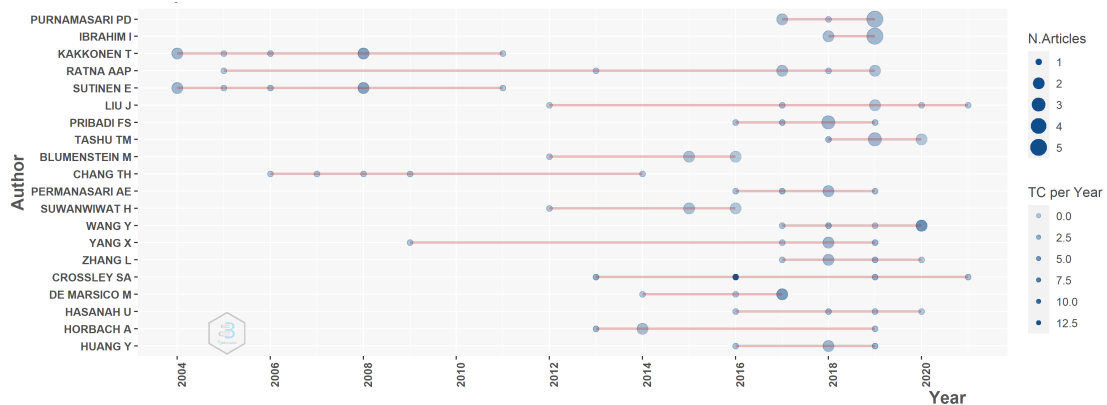


Figure 5: Top-20 authors' production over time.

Figure 6 depicts – through a heatmap – the distribution on the articles by authors' country. The heatmap adopts colors from grey to dark blue. Grey is used for countries with no entries in our dataset. Blues are for countries with at least one contribution in the dataset. The darker the country color, the more articles authors coming from it have contributed to publish on the

topic. United States, China, India, Germany, United Kingdom, and Indonesia are the countries with largest number of articles. In addition, we can observe the collaborations between authors of two different countries c_i, c_j by means of grey-colored links whose thickness is in direct proportion with the number of papers written by authors coming from c_i and c_j . The vast majority of collaborations occur between United States and China, China and Japan, China and Australia.

Country Collaboration Map

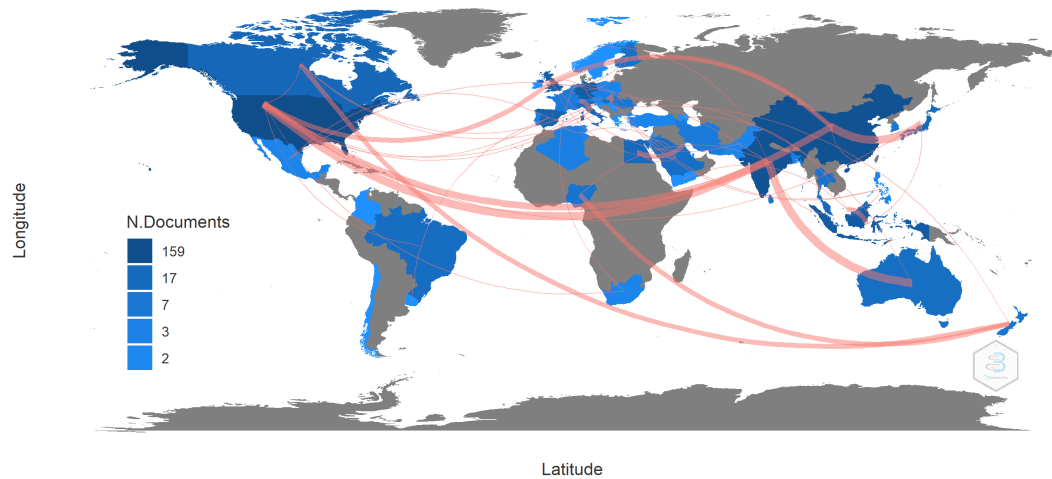


Figure 6: Selected articles for each country and collaborations among countries.

4.2. RQ2 – What are the most used techniques for automatic grading of open-ended questionnaires’ responses?

Figure 7 shows the evolution of topics over the years. Topics are listed on the y -axis, years on the x -axis. The blue-colored line indicates the crucial period when the topic t has been tackled by the scientific publications. The blue circle size changes according to the number of papers published in the specific year concerning t . As an example, the topic “semantic similarity” has been tackled primarily from 2019, while “latent semantic analysis” has been carried since 2009. With a further inspection of these results, and the application of a method inspired by [58], we noticed that the early academic production was focused on methods like *Latent Semantic Analysis* (applied mainly until 2017) and generically on *Natural Language Processing*, the last developments see rising interest towards applications of *Deep Learning* to the automatic grading of open-ended questionnaires. This is coherent with the general trend in this category of promising techniques. The chance to have more easily access to powerful CPUs and GPUs for computation, combined with the availability of pre-trained deep learning architectures ready to work with minimal implementation efforts (see, e.g. [59, 60, 61] for the case of Convolutional Neural Networks).

Figure 8a displays the primary authors’ keyword used to define their academic publications.

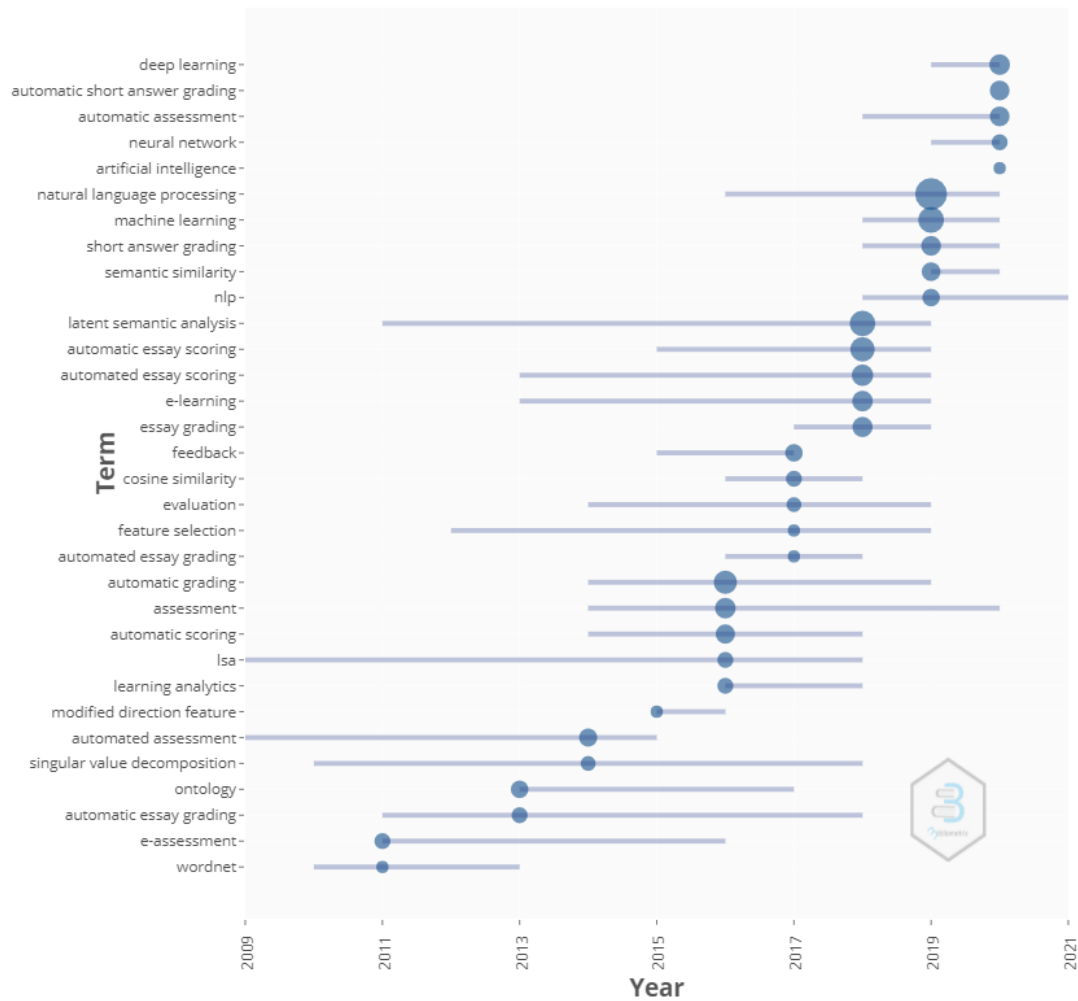
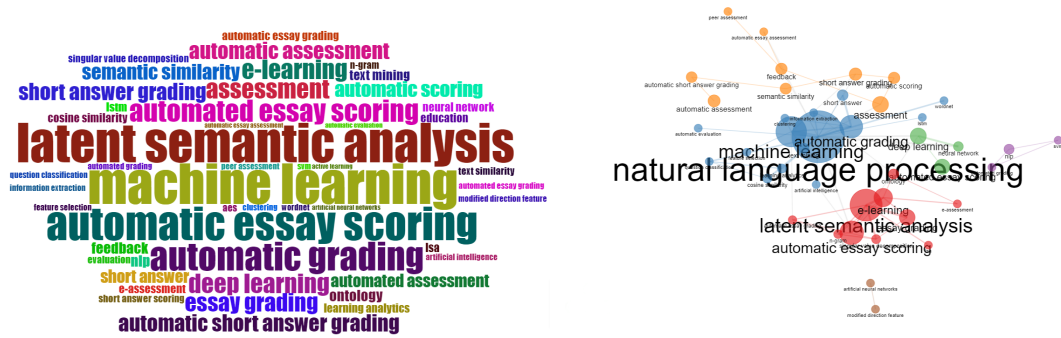


Figure 7: Topics Trend over years.

The keywords in the search query have, as we would expect, great relevance; however, further keywords of interest are still noticeable. E-learning results to be relevant as environment of application on the automated marking; however, many keywords refers to the most used techniques: latent semantic analysis, machine learning and deep learning.

Figure 8b shows with a Fruchterman-Reingold network layout [62] the co-occurrence matrix of authors' keyword. The network locates four different clusters. The orange-colored cluster incorporates the "assessment" keywords; the blue-colored cluster involves "natural language processing" and "machine learning"; lastly, the red-colored cluster incorporates the "latent semantic analysis" and "essays".



(a) Wordcloud of most used studies' keywords. (b) Co-occurrence network of studies' keywords.

Figure 8: Keyword used by the identified papers.

5. Conclusion

Automatic grading of questionnaires has been studied for a long time for both open-ended and closed-ended questions. Still, handling open-ended questionnaires' responses where students' can write essays and short-answers (not available in a list pre-defined of possibilities) represents a challenge for academia. Thus, over the years, there have succeeded a series of ever-more advanced techniques and approaches. In this respect, we have proposed a systematic mapping study on automatic grading techniques for open-ended questionnaires' responses encompassing 488 papers published between 1984 and 2021.

Overall, it emerges that the research area is not *mature* enough. A mature field is one that: (a) is well-documented (i.e., codified) and broadly accessible, (b) is agreed upon by a distinct research community, (c) is differentiated from other research areas, (d) is robust across research paradigms, research methods/approaches, contingent factors, and application contexts, (e) has an impact on the research community, i.e., is cited by other research areas, and (f) is put into practice [63]. Automatic grading of open-ended questionnaires' responses is still an emerging and immature framework. Most scientific production occurs in the last five years, and there is great interest in experimenting with different techniques. The evolution of this field strictly follows the evolution of Natural Language Processing techniques – of which is strictly dependent – and machine learning. The complexity of problems and issues in evaluating different kinds of open text answers, spanning between short-answer on specific domains to long essays, make the research field challenging and requires tailored methodologies that are difficult to generalize and partially disperse the efforts of a growing but still small research community. Although there is broad research on issues such as “explainability” and “interpretability” [64, 65, 66], it is interesting to notice that these keywords do not appear in the research works found, or anyway represent a negligible fraction so that they do not appear in Figure 7 or Figure 8. This holds true for “visualization” [67, 68, 69, 70, 71], a crucial part of the learning analytics endeavor [72]. It is worth noting that approaches to automatically assign a grade to students represent *high risk* applications as recently reported by EU Commission[73] and must guarantee (in different ways) transparency and interpretability explaining the *how* and the *why* of their outcomes. Still, these aspects appear overlooked from our analysis.

Another aspect worth mentioning regards the applications and tools developed with such a set of techniques. We have not found keywords somehow related to techniques put in practice, i.e. “system”, “tool”, “software”, or “framework”. Part of the effort in this research area should be put into developing feasible solution and supports for teachers, and performing user study to evaluate the usability of proposals in real-world with end-users.

In the next future, we aim at deepening this systematic mapping and work on the challenges arising from the survey carried.

Acknowledgments

This research was financially supported by University of Foggia within the project “TILD (Teaching and Learning Development)”.

References

- [1] C. Glas, Item response theory in educational assessment and evaluation, *Mesure et évaluation en éducation* 31 (2008) 19–34. doi:<https://doi.org/10.7202/1025005ar>.
- [2] G. Alexandron, S. Lee, Z. Chen, D. E. Pritchard, Detecting cheaters in moocs using item response theory and learning analytics, in: UMAP, 2016.
- [3] J. P. Meyer, S. Zhu, Fair and Equitable Measurement of Student Learning in MOOCs: An Introduction to Item Response Theory, Scale Linking, and Score Equating., *Research & Practice in Assessment* 8 (2013) 26–39.
- [4] M. Uto, D.-T. Nguyen, M. Ueno, Group Optimization to Maximize Peer Assessment Accuracy Using Item Response Theory and Integer Programming, *IEEE Transactions on Learning Technologies* 13 (2020) 91–106. doi:10.1109/TLT.2019.2896966.
- [5] M. Uto, M. Ueno, Empirical comparison of item response theory models with rater’s parameters, *Heliyon* 4 (2018) e00622. doi:<https://doi.org/10.1016/j.heliyon.2018.e00622>.
- [6] N. González-Castro, P. J. Muñoz-Merino, C. Alario-Hoyos, C. Delgado Kloos, Adaptive learning module for a conversational agent to support MOOC learners, *Australasian Journal of Educational Technology* 37 (2021) 24–44. doi:10.14742/ajet.6646.
- [7] R. Ferguson, Learning analytics: drivers, developments and challenges, *International Journal of Technology Enhanced Learning* 4 (2012) 304–317.
- [8] W. Greller, H. Drachler, Translating learning into numbers: A generic framework for learning analytics, *Journal of Educational Technology & Society* 15 (2012) 42–57.
- [9] J. Wong, M. Baars, D. Davis, T. Van Der Zee, G.-J. Houben, F. Paas, Supporting self-regulated learning in online learning environments and moocs: A systematic review, *International Journal of Human–Computer Interaction* 35 (2019) 356–373.
- [10] K. Mangaroska, M. Giannakos, Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning, *IEEE Transactions on Learning Technologies* 12 (2018) 516–534.
- [11] F. Sciarrone, M. Temperini, Learning analytics models: A brief review, in: 2019 23rd International Conference Information Visualisation (IV), IEEE, 2019, pp. 287–291.
- [12] P. Ardimento, M. L. Bernardi, M. Cimitile, G. De Ruvo, Mining developer’s behavior from web-based ide logs, in: 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), IEEE, 2019, pp. 277–282.
- [13] F. R. Castelli, M. A. Sarvary, Why students do not turn on their video cameras during online classes and an equitable and inclusive plan to encourage them to do so, *Ecology and Evolution* 11 (2021) 3565–3576.
- [14] F. Almeida, J. Monteiro, The challenges of assessing and evaluating the students at distance, *arXiv preprint arXiv:2102.04235* (2021).
- [15] B. De Carolis, F. D’Errico, N. Macchiarulo, M. Paciello, G. Palestra, Recognizing cognitive emotions in e-learning environment, in: International Workshop on Higher Education Learning Methodologies and Technologies Online, Springer, 2020, pp. 17–27.
- [16] H. Harada, M. Nakayama, Estimation of reading ability of program codes using features of eye movements, in: ACM Symposium on Eye Tracking Research and Applications, 2021, pp. 1–5.
- [17] Z. Liu, C. Yang, S. Rüdian, S. Liu, L. Zhao, T. Wang, Temporal emotion-aspect modeling

- for discovering what students are concerned about in online course forums, *Interactive Learning Environments* 27 (2019) 598–627.
- [18] J. Psotka, N.-S. Chen, The new potentials for intelligent tutoring with learning analytics approaches, 2019.
- [19] S. Burrows, D. D’Souza, Management of teaching in a complex setting, *Proceedings of the Second Melbourne Computing Education Conventicle* (2005) 1–8.
- [20] D. T. Haley, P. Thomas, A. De Roeck, M. Petre, Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about html, in: *Proceedings of the ninth Australasian conference on Computing education-Volume 66*, Citeseer, 2007, pp. 35–42.
- [21] D. M. Williamson, X. Xi, F. J. Breyer, A framework for evaluation and use of automated scoring, *Educational measurement: issues and practice* 31 (2012) 2–13.
- [22] L. Hirschman, Automated grading of short-answer tests, *IEEE Intelligent Systems, Trends and Controversies section 15* (2000) 22–37.
- [23] P. G. Butcher, S. E. Jordan, A comparison of human and computer marking of short free-text student responses, *Computers & Education* 55 (2010) 489–499.
- [24] M. D. Shermis, J. Burstein, C. Leacock, Applications of computers in assessment and analysis of writing, *Handbook of writing research* (2006) 403–416.
- [25] S. Burrows, I. Gurevych, B. Stein, The Eras and Trends of Automatic Short Answer Grading, *International Journal of Artificial Intelligence in Education* 25 (2015) 60–117. doi:10.1007/s40593-014-0026-8.
- [26] O. Sychev, A. Anikin, A. Prokudin, Automatic grading and hinting in open-ended text questions, *Cognitive Systems Research* 59 (2020) 264–272.
- [27] H. Kazi, P. Haddawy, S. Suebnukarn, Leveraging a domain ontology to increase the quality of feedback in an intelligent tutoring system, in: *International Conference on Intelligent Tutoring Systems*, Springer, 2010, pp. 75–84.
- [28] M. N. Demaidi, M. M. Gaber, N. Filer, Ontopefege: Ontology-based personalized feedback generator, *IEEE Access* 6 (2018) 31644–31664.
- [29] O. Sychev, D. Mamontov, Automatic error detection and hint generation in the teaching of formal languages syntax using correctwriting question type for moodle lms, in: *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*, IEEE, 2018, pp. 1–4.
- [30] A. Mattia, M. Federico, Deep neural machine translation with weakly-recurrent units, in: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d’Alacant, Alacant, Spain, European Association for Machine Translation*, 2018, pp. 119–128.
- [31] D. Schicchi, G. Pilato, G. Lo Bosco, Deep neural attention-based model for the evaluation of italian sentences complexity, 2020, pp. 253–256. doi:10.1109/ICSC.2020.00053.
- [32] D. Schicchi, G. Pilato, A social humanoid robot as a playfellow for vocabulary enhancement, volume 2018-January, 2018, pp. 205–208. doi:10.1109/IRC.2018.00044.
- [33] E. del Gobbo, S. Fontanella, A. Sarra, L. Fontanella, Emerging Topics in Brexit Debate on Twitter Around the Deadlines, *Social Indicators Research* (2020) 1–20. doi:10.1007/s11205-020-02442-4.
- [34] A. Tontodimamma, E. del Gobbo, V. Russo, A. Sarra, L. Fontanella, Facebook Debate

- on Sea Watch 3 Case: Detecting Offensive Language Through Automatic Topic Mining Techniques, in: P. Mariani, M. Zenga (Eds.), *Data Science and Social Research II*, Springer International Publishing, Cham, 2021, pp. 367–378. doi:10.1007/978-3-030-51222-4_29.
- [35] Y. Attali, J. Burstein, Automated essay scoring with e-rater® v. 2, *The Journal of Technology, Learning and Assessment* 4 (2006).
- [36] J. Biolchini, P. G. Mian, A. C. C. Natali, G. H. Travassos, Systematic review in software engineering, System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES 679 (2005) 45.
- [37] M. Petticrew, H. Roberts, *Systematic reviews in the social sciences: A practical guide*, John Wiley & Sons, 2008.
- [38] B. Kitchenham, S. Charters, *Guidelines for performing systematic literature reviews in software engineering* (2007).
- [39] A. Roehrs, C. A. Da Costa, R. da Rosa Righi, K. S. F. De Oliveira, Personal health records: a systematic literature review, *Journal of medical Internet research* 19 (2017) e13.
- [40] M. Aria, C. Cuccurullo, bibliometrix : An R-tool for comprehensive science mapping analysis, *Journal of Informetrics* 11 (2017) 959–975. doi:10.1016/j.joi.2017.08.007.
- [41] S. Graf, B. List, An evaluation of open source e-learning platforms stressing adaptation issues, in: *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)*, IEEE, 2005, pp. 163–165.
- [42] M. Dougiamas, P. Taylor, Moodle: Using learning communities to create an open source course management system, in: *EdMedia+ Innovate Learning, Association for the Advancement of Computing in Education (AACE)*, 2003, pp. 171–178.
- [43] G. R. Alves, M. C. Viegas, M. A. Marques, M. C. Costa-Lobo, A. A. Silva, F. Formanski, J. B. Silva, Student performance analysis under different moodle course designs, in: *2012 15th International Conference on Interactive Collaborative Learning (ICL)*, IEEE, 2012, pp. 1–5.
- [44] M.-D. Dascalu, S. Ruseti, M. Dascalu, D. S. McNamara, M. Carabas, T. Rebedea, S. Trausan-Matu, Before and during covid-19: A cohesion network analysis of students' online participation in moodle courses, *Computers in Human Behavior* 121 (2021) 106780.
- [45] M. Richardson, C. J. Burges, E. Renshaw, MCTest: A challenge dataset for the open-domain machine comprehension of text, in: *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2013, pp. 193–203.
- [46] M. Mohler, R. Mihalcea, Text-to-text semantic similarity for automatic short answer grading, in: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, Association for Computational Linguistics, Morristown, NJ, USA, 2009, pp. 567–575. doi:10.3115/1609067.1609130.
- [47] L. S. Larkey, Automatic essay grading using text categorization techniques, in: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, ACM Press, New York, New York, USA, 1998, pp. 90–95. doi:10.1145/290941.290965.
- [48] T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch, *Handbook of latent semantic analysis*, Psychology Press, 2013.
- [49] P. W. Foltz, D. Laham, T. K. Landauer, *The intelligent essay assessor: Applications to educational technology*, 1999.
- [50] T. Miller, Essay Assessment with Latent Semantic Analysis, *Journal of Educational*

- Computing Research 29 (2003) 495–512. doi:10.2190/W5AR-DYPW-40KX-FL99.
- [51] D. Pérez, A. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodríguez, B. Magnini, Automatic assessment of students' free-text answers underpinned by the combination of a BLEU-inspired algorithm and latent semantic analysis, in: Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2005 - Recent Advances in Artificial Intelligence, 2005, pp. 358–362.
- [52] T. Kakkonen, N. Myller, E. Sutinen, J. Timonen, Comparison of dimension reduction methods for automated essay grading, *Educational Technology and Society* 11 (2008) 275–288.
- [53] H.-C. Wang, C.-Y. Chang, T.-Y. Li, Assessing creative problem-solving with automated text grading, *Computers & Education* 51 (2008) 1450–1466. doi:10.1016/j.compedu.2008.01.006.
- [54] M. Zahran, A. Magooda, A. Mahgoub, H. Raafat, M. Rashwan, A. Atyia, Word representations in vector space and their applications for Arabic, volume 9041, 2015. doi:10.1007/978-3-319-18111-0_32.
- [55] F. Dong, Y. Zhang, Automatic Features for Essay Scoring – An Empirical Study, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2016, pp. 1072–1077. doi:10.18653/v1/D16-1115.
- [56] F. Dong, Y. Zhang, J. Yang, Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Stroudsburg, PA, USA, 2017, pp. 153–162. doi:10.18653/v1/K17-1017.
- [57] R. Siddiqi, C. J. Harrison, R. Siddiqi, Improving Teaching and Learning through Automated Short-Answer Marking, *IEEE Transactions on Learning Technologies* 3 (2010) 237–249. doi:10.1109/TLT.2010.4.
- [58] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, F. Herrera, An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field, *Journal of Informetrics* 5 (2011) 146–166. doi:10.1016/J.JOI.2010.10.002.
- [59] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [60] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [61] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
- [62] L. Bi, Y. Wang, J.-p. Zhao, H. Qi, Y. Zhang, Social network information visualization based on fruchterman reingold layout algorithm, in: 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), IEEE, 2018, pp. 270–273.
- [63] H. Keathley-Herring, E. Van Aken, F. Gonzalez-Aleu, F. Deschamps, G. Letens, P. C. Orlandini, Assessing the maturity of a research area: bibliometric review and proposed framework, *Scientometrics* 109 (2016) 927–951. doi:10.1007/s11192-016-2096-x.
- [64] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations:

- An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.
- [65] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, R. van Lier, Explainable and interpretable models in computer vision and machine learning, Springer, 2018.
- [66] J. M. Alonso, G. Casalino, Explainable artificial intelligence for human-centric data analysis in virtual learning environments, in: International workshop on higher education learning methodologies and technologies online, Springer, 2019, pp. 125–138.
- [67] D. A. Keim, Information visualization and visual data mining, IEEE transactions on Visualization and Computer Graphics 8 (2002) 1–8.
- [68] D. Leony, A. Pardo, L. de la Fuente Valentín, D. S. de Castro, C. D. Kloos, Glass: a learning analytics visualization tool, in: Proceedings of the 2nd international conference on learning analytics and knowledge, 2012, pp. 162–163.
- [69] S. Charleer, A. V. Moere, J. Klerkx, K. Verbert, T. De Laet, Learning analytics dashboards to support adviser-student dialogue, IEEE Transactions on Learning Technologies 11 (2017) 389–399.
- [70] D. Malandrino, A. Guarino, N. Lettieri, R. Zaccagnino, On the visualization of logic: A diagrammatic language based on spatial, graphical and symbolic notations, in: 2019 23rd International Conference Information Visualisation (IV), IEEE, 2019, pp. 7–12.
- [71] G. Benevento, R. De Prisco, A. Guarino, N. Lettieri, D. Malandrino, R. Zaccagnino, Human-machine teaming in music: anchored narrative-graph visualization and machine learning, in: 2020 24th International Conference Information Visualisation (IV), IEEE, 2020, pp. 559–564.
- [72] O. Viberg, M. Hatakka, O. Bälter, A. Mavroudi, The current landscape of learning analytics in higher education, Computers in Human Behavior 89 (2018) 98–110.
- [73] European Commission, Com(2021) 206 final 2021/0106 (cod) proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative act, 2021. <https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=COM:2021:206:FIN>.