ISSN 2601-8683 (Print)
ISSN 2601-8675 (Online)
European Journal of
Formal Sciences and Engineering
July -December 2020
Volume 3, Issue 2

How Artificial Intelligence Can Augment the Collection of Scientific Literature

Mokeddem Allal

University Algiers 3, Department of management and Technology, Algeria

Abstract

This article describes the contribution of artificial intelligence (AI) to the literature collection process, which has become more efficient and more homogeneous. In this context, the researcher will receive his literature not only according to his field. Moreover, the literature is strongly linked to scientific and academic ambitions. AI through its deep learning techniques offers the possibility of speeding up the process of collecting augmented literature via an approach based on the annotation of scientific names and none-scientific names related to the field. AI provides original or reproduced research avenues with reliable and precise results. In this article, we have highlighted how to develop conceptual framework based on scientific and none-scientific names related to the area of expertise, all ensuring the reproducibility, reliability and accuracy of the study.

Keywords: artificial intelligence (ai), augmented literature, reproducible literature, reliable literature, accuracy literature.

Introduction

In the era of numeric approved by scientific databases, scientific social network (researchgate, linkedin) and professional reports..., the volume of data increases considerably Cassel and al (2016). In a competitive world, having a data structure is not enough to obtain reliable results over time. Now, the scientific decision making is linked with the augmented data structure. A data structure requires more than the computing power with machine. On the other hand, the use of intelligent machines to organize and group the developed data increases the quality of the stored data.

In this context, a study conducted by Kaak (2019) confirms that almost 70% of the companies surveyed say that the quality of their data has an impact on the smooth running of their activities. At this point, the scientists are inspired, extrapolating the data source to the prospects their research is highly recommended.

The issue of this paper is how to build and improve the literature collection process by integrating the AI approach. The interest is to offer purely refined literature resources for scientific and academic purposes.

Conceptual Framework

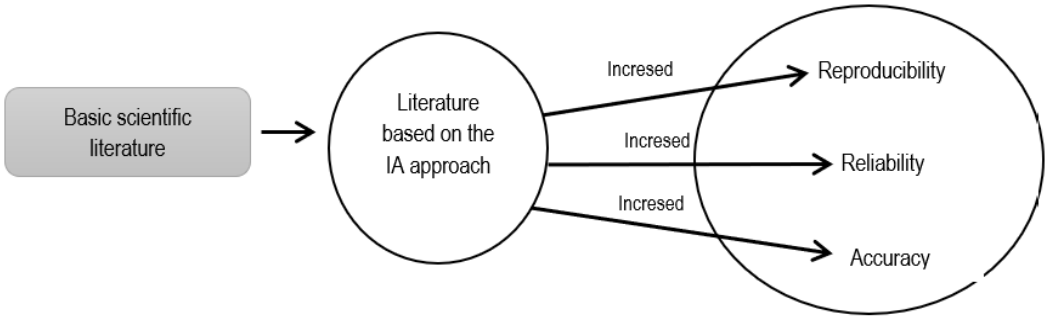


Fig 1. Framework of the study: Author developed

This study was divided into two parts. The first part carries the descriptive meaning via the definition of augmented literature and how is it constructed following the AI based literature approach, in ordered to extract the scientific and none-scientific names related to the field. In the second part, we focus on performance criteria for the development of the literature based on the AI approach.

Approach to Enhance Literature

The scientist's mission is to question each phenomenon observed continuously. In other words, the scientist keeps asking questions and looking for data that not only meets the requirements of the present but also about the future of society. In order to increase and adjust the human decision, the Brain Behind MECBot group defines augmented data as dynamic and agile information resources. At this level, the data must receive ingestion, cleaning, unification, integration, extraction and hydration processing with an almost real-time influx of new data developed.

An augmented data source serves to strengthen and transform the value of the research model developed. In this context, according to Miller and Brown (2018) faced with the technological means available; the extraction of knowledge from complex data generates information that optimizes concepts, searches for models, follows trends and associations, discovers the inefficiencies and predict outcomes. The following figure illustrates the evolutionary framework of the augmented scientific literature:

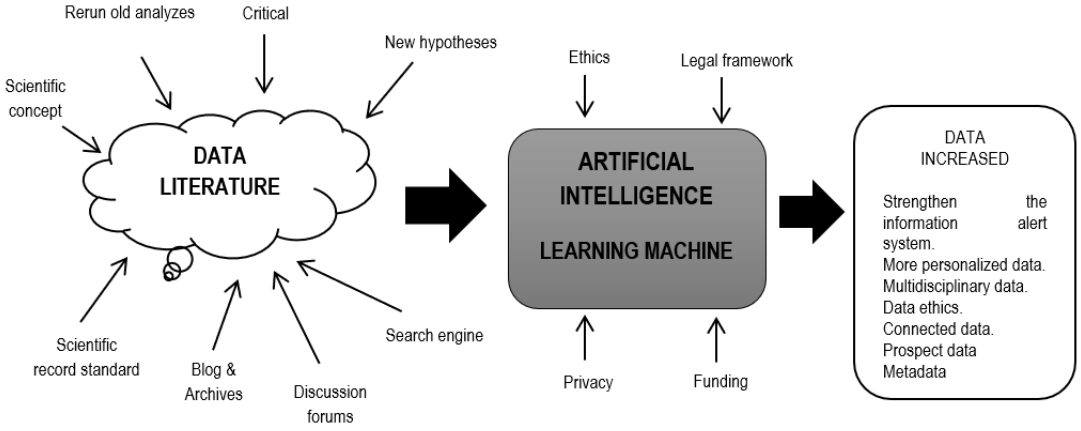


Fig. 2. The structure of the augmented literature

The development of discovery science Leach and (2009), has made scientific activity more valuable. However, with the advent of AI systems that can represent hypotheses and design data collection techniques based on scientific discovery. At this point, AI techniques speed up the process of analyzing gigantic data, which can be disseminated on all laboratory equipment. According to Gil and al (2014), analysis process includes the reduction of dimensionality and the functionality extractor to create high speed classifiers based on machine learning approaches, such as Bayesian networks or support vector machines.

The scientific activity is not limited in time or in space. It is considered as a continuous discovery circuit. At the time of scientific writing, the researcher can collect a set of conceptual information via automatic reading. This option can refine the search process and therefore the writing style. In this regard, renowned research begins with an in-depth analysis of the various sources of information selected.

The process of a literature analysis system begins with the extraction of scientific names. This will be very useful for bringing together all research contexts in the form of a set of resources: sentences, paragraphs, models or mathematical equations. These resources can help to enrich existing content or add new content to existing research projects such as the Encyclopedia project. In this context, the construction of augmented scientific literature is an operation in three stages:

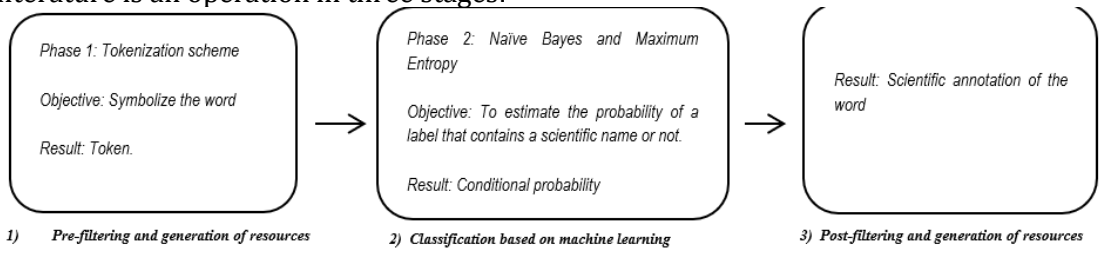


Fig. 2. The design process for augmented literature, Source Akella and al (2012)

The development of an augmented literature begins with a validation of the existing source of information. First, the text is symbolized via a trigram word which groups together three tokens. The tokens will be subjected to a qualitative evaluation to see if the words are well capitalized, abbreviated and check if the trigram has no common English words. Each filtered trigram will then be classified by a machine learning classifier as "scientific name" or " none-scientific name". The validation system takes into account the structural and contextual characteristics of the trigram. If the trigram is rejected by the rule filtering system, the first two tokens of the bigram of the same previous trigram will then be estimated. If it fails, the bigram analyzed can become a first name which will be classified accordingly if it is considered as a higher-order resource.

The second phase of the augmented literature design is estimated by the conditional probability of a label, belonging either to the rubric of scientific names or none-scientific names. This is based on contextual information about the meaning of the word. Once we get these conditional probabilities, we simply choose the label with the highest probability for a given string. This presents the closest name in terms of conceptual meaning related to the context of the study. This classification can be used in many natural language processing tasks such as text segmentation Neche and al (2019), markup of part of speech, language modeling and text classification.

The finished product is symbolized in a book which counts the chains derived from a structure of unigrams, bigrams and trigrams of words in the text. The neighborhood of scientific names is calculated according to the context of the words in the sentence. At this point, each researcher will have access to the documentation not only according to his area of expertise, but beyond his area. Behind each concept or terminology exists a model and an approach is therefore billions of resources that can be consulted and developed using the augmented literature approach.

The performance criteria for augmented literature

The literature in the various research fields evolves both in terms of theoretical and empirical models developed and even in terms of constructed statistical models. In this context, the scientific community is looking for a way to overcome the obstacles of research and to push the researchers to build scientific models and approaches of a higher order of reflection. At this stage, according to Wu and al (2018) the traditional guidelines for literature reviews are based on human screening, which may be subject to bias, lack of transparency, reproducibility and to human error. However, in the face of AI developments, learning machines combined with automatic natural language processing techniques can be used to ensure the reproducibility, reliability and accuracy of constructed literature.

Increase Reproducibility

The scientific community has not even reached a consensus, in which part of the scientific document will be placed the reproduction procedure. In standards, the peer-

reviewed article differs from other resources. First, the concept of the reproducibility of the literature does not mean the replica of the work. On the contrary, the literature will be subjected to a new experiment either according to new specific developed values, a new standard of calculation and constructed measures and either confirmation or none-confirmation of an extended hypothesis as a function of time and space. At this stage, the study conducted by [Cohen and (2016); Goodman and al (2016); Branco and al (2017); Clemens and (2017); Avidan and al (2019)] presents a number of conditions that must be met in order to reproduce a study via a natural language processing. A paper that receives natural language processing requires a set of criteria:

Methodology: the system must be clear and articulate, or there must be enough detail available to reconstruct the system exactly.

Result: there is an increasingly urgent call for validation and verification of the results of published research, both within the academic community and the general public.

Robustness and generalizability: the reproducibility will be obtained by strengthening and generalizing the body of the study, according to the variation of the basic hypotheses or in the experimental procedures.

Inferential reproducibility: the inferential reproducibility refers to the ultimate objective. According to which different scientists analyzing the same set of data on a larger population and who should reach similar conclusions.

Assessment: the reproducibility based on the assessment is used to ensure that all primary studies contain the appropriate information for the analysis and relevant to the field of research.

In a community with different cultural conventions, the need to enrich the literature with cases and experimental situations is very useful in many scientific fields. In this regard, Goodman and al (2016) suggest that the reproducibility of the methods, aims to capture the original meaning of the study. In another way, that is to say the possibility of implementing the experimental and technical procedures as exactly as possible, with the same data and the same tools. This is in order to obtain the same results broken down by demographic factors. In practice, things happen differently. The degree of reproducibility of a methodology is based on more detailed information. This information is not always kept by the investigator, such as on which machine have the samples being tested? ; In what order and on what day of execution?

In addition, the reproducibility of methodology requires to understand, how many analyzes were carried out and how the particular analyzes reported in a published article were chosen. All these parameters will be configured in the machine learning system. This is in order to take advantage from the level of detail required in the measurement process during the study.

According to Branco and al (2017), reproducing the results means obtaining the same results using the same procedures and the same experimental conditions. At this level, the reproduction of results will vary depending on the research area. For example, in the IT field; the results are determined by the initial conditions. On the contrary, in the social sciences, the models are subject to important stochastic components. This means the accumulation of evidence and data that can generate different results.

Indeed, two conditions must be fulfilled for an effective reproduction of the results Branco and al (2017). First, reproduce the phenomenon studied outside their original context. This situation could generate a false diagnosis, wrong procedures, a measurement error, a biased conception or a fraud, which constitutes the way towards a refusal to achieve the same produced results. The second efficient condition is that the reproduction of a study is not limited to the replication of the presence or absence of statistical significance, but beyond the evaluation of cumulative evidence and the evaluation of its sensitivity with significant biases.

The third important factor is presented by robustness and generalizability of the study. According to Clemens and (2017), a robust and generalized study signified the transportability of the empirical model of the study to other experimental contexts. In this context, through machine learning techniques, the investigator will be able to generate different tests and therefore have the possibility of extrapolating the study to an extended scale. The robustness of a study take different forms, as an example : a conceptual replication or pseudoreplication based on an open question, recoding or re-periodization of the same set of data by modifying the set of co-variables, the method of calculation of the standard error and updating of the data source with a set of original observations.

The robust and original scientific production should not only be classified in the category of new and distinctive studies. Therefore, the inferential studies are also recognized as studies which could also be classified among the most important studies. An inferential study is based on the exploitation of the content. This involves drawing qualitative conclusions from an independent replication of a study or thorough re-analysis of the original study. An inferential study differs in its objective from reproducible studies based on results and methods. As a result, the researchers can extract the same conclusions in all pre-existing studies, or draw different conclusions from the same specified original data structure. The figure below summarizes the different options using inferential studies in the reproductive process.

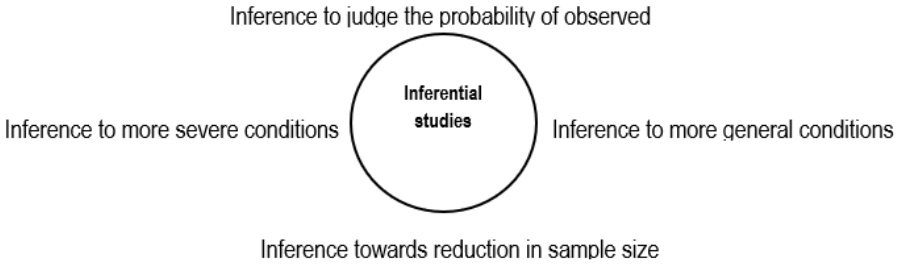


Fig. 3. The different cases based on reproducible inferential studies, Source Avidan and al (2019)

The last part of a reproducibility process is embodied in the systemic assessment of paper quality. In this step, we discover the scientific trends that have occurred in recent times. According to McKinnel and al (2019), in order to introduce scientific paper into the evaluation process, a set of conditions must be fulfilled:

Construction of the algorithm: a paper must integrate a section namely the implication of AI concept and machine learning in its conceptual structure. This part is described by a source code based on the taxonomy of scientific research carried out during the literature collection process.

Context: the contextual data must be provided in paper. In order to properly analyze the literature related to the field of study.

Composition of the algorithm: this criterion is the most essential for the analysis. It involves the analysis of the independent and dependent variables and if they are clearly established and indicated in the document itself.

Data: the data set used for the evaluation of the model must be explicitly described. If the private data set is used, it is necessary to explain the composition of this data and how it has been normalized for use in an AI scenario.

Performance: the performance of each model, algorithms or applications must be measured and presented with precision in the document.

In a research context based on the conditions presented above. The ultimate objective is to design a reproducibility approach based on AI. This is to identify the vulnerabilities of pre-existing studies and to what extent will be assessed in real time. Such an evaluation approach will speed up the process of developing theorems relating to the scalability of challenges and with an intention incited to a scientific revolution in each field of research.

Increase Reliability

The scientific landscape is constantly changing. This declaration transforms the research process into a scientific adventure. In this context, it's estimated that almost one million

articles are published each year or 30 articles each 30 seconds. In this situation, the researchers try to ensure reliability and assemble the right combination with regard to the literature introduced. A reliable study identifies the areas of research that could be considered as a reference support for the development of other studies.

According to Extance (2018) Iris.AI is one of a multitude of new AI-based research tools offering targeted navigation of the knowledge landscape. This platform helps researchers to validate existing scientific hypotheses to uncover hidden links between the results which may even suggest new hypotheses to guide experiments.

The researcher must know what the score should be assigned to a study in order to be classified in the categories of reliable studies. In this context, and in his daily column, the professor Seabright (2018) from the Institute for Advanced Studies in Toulouse indicates when a scientific study will have access to resources and how to introduce them into the list of augmented literature. This option strongly depends on a set of factors: (Mohajan (2017); Bobadilla and al (2018); Seabright (2018); Najafabadi and Mahrin (2016))

Sample size: the sample size should be larger than that of the original study sample.

Study measure: the measure must be carefully selected according to the consistency of the internal measure and see the quality of the inter-evaluators according to their judgments in relation to the phenomenon observed.

The quality metric: a reliable study is one that has a high citation index compared to the others.

The control of random errors: no study is perfect. The researcher must control the error structure composed of: material error (an impure sample, poor technical competence, etc.), observation error (instrument not included, perception bias of the observer, sampling error, etc.), conceptual error (calculation error, inappropriate statistical model, poorly specified assumptions, etc.) and discursive error (incomplete reports, erroneous credibility judgments, etc.).

The stability of the results: a reliable study that which treats the data in a uniform way in order to exploit all the stored data, and therefore to avoid falling into the information saturation area.

Demographic factors: a reliable study is strongly broken down by demographic factors such as age, sex and gender. This is to optimize the maximum of resources provided during the experimentation.

Collaborative filtering: a reliable study should not only be calculated on the basis of the problem quality addressed, either original or reproductive. The effective scientific production should rather affect the field of research by creating a new collaborative group. This is to create a common scientific trend.

Increase Accuracy

According to Glasziou and al (2014) a very precise study, serves to control the quality of the methodology, the applicability of the results and to regulate the degree of falling biased before and during the research process. At this point, according to [McGrath and al (2019); Hinojo-Lucena and al (2019)] the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) was developed to improve the quality of research reports and to regulate the meta-analysis process. The PRISMA statement affects many areas of research. The main objective was to help improve the completeness of the reports, compare the validity of the results and organize systematically the improvement notification of the drafting and publication process.

In this context, raising awareness of PRISMA DATA aims to know at what level scientific production progresses over time? And is there a productive relationship between the number of authors and articles? As a result, everyone is expected to be connected and involved in the publishing and writing process, starting with the study authors and going from peer reviewers to journal editors. At this stage, the production of a clear and homogeneous bibliographic study must comply with the editorial guidelines, ie in terms of approval and adherence.

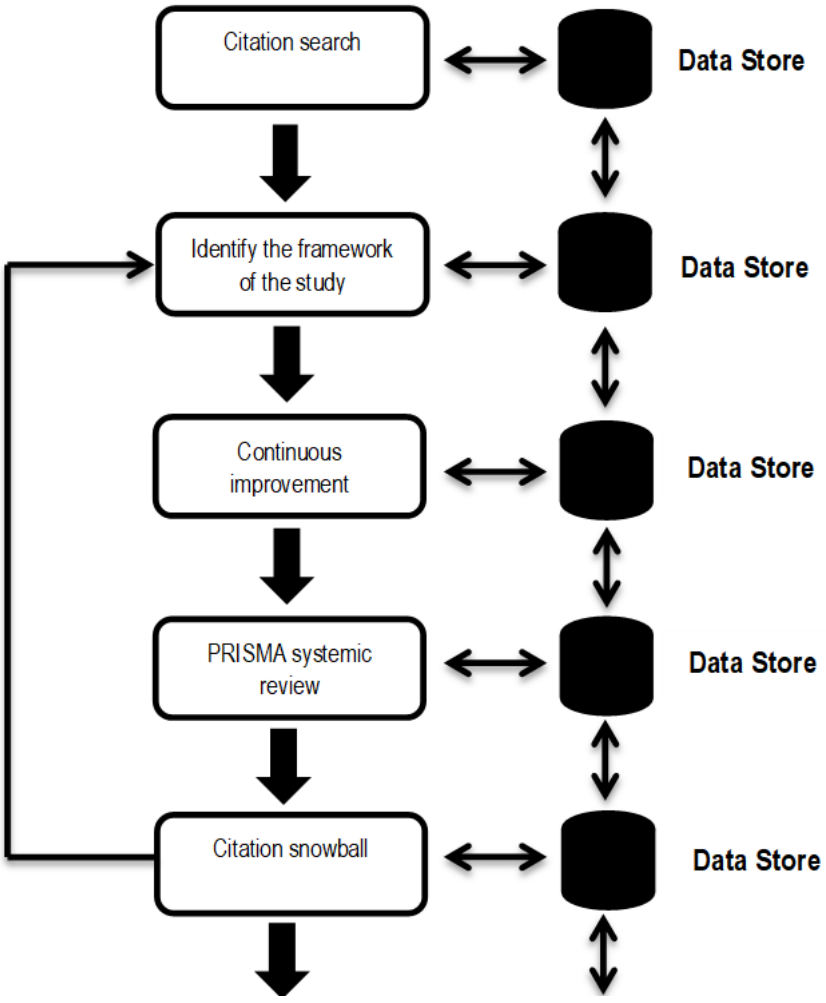
The techniques based on the natural language learning serve the PRISMA research community. At this stage, in order to be sharp in scientific writing, receive in real time the approval recommended by the review, with regard to the writing models which will vary according to the type of research carried out. This approval helps to stimulate the general use of the revision guideline, which could lead to a specialized framework guaranteeing the systemic evaluation of the study and to propose references linked to PRISMA standards.

For scientific purposes, the membership is the responsibility of the author to create well-reported and reproducible research. According to Dewey and al (2019) the studies submitted earlier to the review guidelines will be more adequate to be subjected to a reproductive process. This process is differentiated in a transparent manner according to the number of participants, the factors of comparison of results, the threshold of statistical positivity test adopted and the hierarchical and none-hierarchical intervention methods carried out….etc.

The PRISMA declaration is one of the systematic reviews (RS) in order to manage the evaluation and publication process. In this context, the initiative of International Collaboration for the Automation of Systematic Reviews (ICASR) is an interdisciplinary group with a common interest to maximize the use of technology to facilitate the transfer of scientific results to practice O'Connor and al (2018). The systemic automation is used to reassure continuous improvement of production, compliance with quality standards when writing, flexibility of use and combination of components, exploitation of relational resources and to share the evaluation code with its peers.

According to Beller and al (2018), SR process involves a variety of skills related to the field of information science, librarians, software engineers, statisticians, linguists and artificial intelligence experts. According to Tsafnat and al (2014) several tasks lend themselves to automation are: selecting titles and summaries, the search for full texts of studies, the extraction of data and even the collation of the meta-analysis results, the streamline research through the development of the writing protocol and graph evaluation report. Also the automatic filtering could be useful to quickly determine if a new eligible search has been performed and should trigger an RS update.

The automation of a manuscript is an operation which depends on many tasks. We have to analyze each task separately, from citation of references to reproducing paper. The following diagram summarizes the systemic review for the automating SR process.



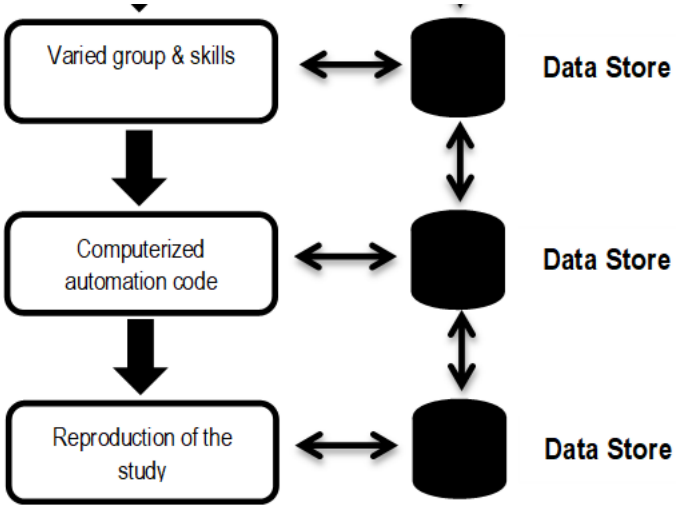


Fig. 4. The reproducible inferential process, Source Beller and al (2018)

The standards of scientific activity begin with an in-depth analysis of the literature related to the field of research. The researcher not only attempts to find a literature, but also to create a conceptual framework of the study. A conceptual framework subjected to an evaluation procedure allows researchers to receive comments that lead to continuous improvement of the literature. At the same time, automating a manuscript can facilitate the systematic evaluation of paper during production, in order to comply with PRISMA standards.

The development of an automation RS system should be flexible and rich in information. The flexibility is calculated based on the number of different components of the study, such as the appearance of new components and new variables, new test constraints and new hypotheses…and so on. All of these constraints transform and make the evaluation operation more efficient and will never be blocked at one level of the process.

A rich and coherent conceptual framework is one that involves a variety of multidisciplinary knowledge. In this context, with technological progress, scientific production continues to progress. Each automation technique should be shared, preferably by making the code available for free. The open source code makes it possible to build on previous work and thus facilitate the mission of reproducibility.

Conclusion and perspectives

In the era of AI, the scientific production finds rigorous and very promising lines of research. The researcher not only needs to access recent documentary databases, but also to find resources in real time and in coherence with the intentions and the research objectives fixed. At this point, the digital science activity has become enriching and more structured with deep learning techniques.

The auteur will be assisted by advanced AI techniques via deep learning approach. The author will be informed which are the most advantageous and ambitious areas of reproducibility and what are the basic scientific concepts which must be articulated via empirical and theoretical tests. At same time, a robust and reliable study is one that receives a high score based on the most important study parameters. The production will be evaluated over time as the study progresses. For this purpose, AI offers a holistic approach to research. The researcher will receive comments that control his production processes from the start of the research activity. Consequently, the researcher will be able to measure the quality of his study before the final submission.

The development of a literature improvement model requires a colossal amount of data (big data). This is in order to detect unusual events in the data and therefore better understand the possibilities for future and original research. An empirical study using artificial intelligence techniques with the Payton software will be conducted with a set of indexed journals in order to bring out a base of enriching literature, all we respect the three performance criteria presented in this paper.

References

[1] Akella, L. M., Norton, C. N., & Miller, H. (2012). NetiNeti: discovery of scientific names from text using machine learning methods. BMC bioinformatics, 13(1), 211.

[2] Avidan, M. S., Ioannidis, J. P., & Mashour, G. A. (2019). Independent discussion sections for improving inferential reproducibility in published research. British journal of anaesthesia, 122(4), 413-420.

[3] Bobadilla, J., Gutiérrez, A., Ortega, F., & Zhu, B. (2018). Reliability quality measures for recommender systems. Information Sciences, 442, 145-157.

[4] Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., ... & Xia, J. (2018). Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic reviews, 7(1), 77.

[5] Branco, A., Cohen, K. B., Vossen, P., Ide, N., & Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an LRE special section.

[6] Cassel, L., Dicheva, D., Dichev, C., Goelman, D., & Posner, M. (2016, September). Artificial Intelligence in Data Science. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (pp. 343-346). Springer, Cham.

[7] Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. Journal of Economic Surveys, 31(1), 326-342.

[8] Cohen, K. B., Xia, J., Roeder, C., & Hunter, L. E. (2016, May). Reproducibility in natural language processing: a case study of two R libraries for mining PubMed/MEDLINE. In LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation (Vol. 2016, No. W23, p. 6). NIH Public Access.

[9] Dewey, M., Levine, D., Bossuyt, P. M., & Kressel, H. Y. (2019). Impact and perceived value of journal reporting guidelines among radiology authors and reviewers. European radiology, 29(8), 3986-3995.

[10] Extance, A. (2018). How AI technology can tame the scientific literature. Nature, 561(7722), 273.

[11] Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., ... & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. The Lancet, 383(9913), 267-276.

[12] Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. Science, 346(6206), 171-172.

[13] Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean?. Science translational medicine, 8(341), 341ps12-341ps12.

[14] Hinojo-Lucena, F. J., Aznar-Díaz, I., Cáceres-Reche, M. P., & Romero-Rodríguez, J. M. (2019). Artificial intelligence in higher education: A bibliometric study on its impact in the scientific literature. Education Sciences, 9(1), 51.

[15] Kaak, R. (2019).L'IA au service de la qualité des données, pour des cas d'usage toujours plus innovants, https://www.capgemini.com/fr-fr/2019/11/ia-qualite-donnees/#, last accessed 2020/02/22.

[16] Leach, S. M., Tipney, H., Feng, W., Baumgartner Jr, W. A., Kasliwal, P., Schuyler, R. P., ... & Hunter, L. (2009). Biomedical discovery acceleration, with applications to craniofacial development. *PLoS computational biology*, *5*(3).

[17] Miller, D. D., & Brown, E. W. (2018). Artificial intelligence in medical practice: the question to the answer?. *The American journal of medicine*, *131*(2), 129-133.

[18] Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. Annals of Spiru Haret University. Economic Series, 17(4), 59-82.

[19] McKinnel, D. R., Dargahi, T., Dehghantanha, A., & Choo, K. K. R. (2019). A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment. Computers & Electrical Engineering, 75, 175-188.

[20] McGrath, T. A., Moher, D., & McInnes, M. D. (2019). Steps toward more complete reporting of systematic reviews of diagnostic test accuracy: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA). Systematic reviews, 8(1), 166.

[21] Neche, C., Belaid, A., & Kacem-Echi, A. (2019, September). Arabic Handwritten Documents Segmentation into Text-Lines and Words using Deep Learning. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) (Vol. 6, pp. 19-24). IEEE.

[22] Najafabadi, M. K., & Mahrin, M. N. R. (2016). A systematic literature review on the state of research and practice of collaborative filtering technique and implicit feedback. Artificial intelligence review, 45(2), 167-201.

[23] O'Connor, A. M., Tsafnat, G., Gilbert, S. B., Thayer, K. A., & Wolfe, M. S. (2018). Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic reviews, 7(1), 3.

[24] Seabright, P. (2018). Quand les chercheurs jugent « la fiabilité réelle des études scientifiques », Professeur à l'Institut d'études avancées de Toulouse, https://www.lemonde.fr/idees/article/2018/09/08/quand-les-chercheurs-jugent-la-fiabilite-reelle-des-etudes-scientifiques_5352177_3232.html, last accessed 2020/02/20.

[25] Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. Systematic reviews, 3(1), 74.

[26] Wu, E. Q., Royer, J., Ayyagari, R., Signorovitch, J., & Thokala, P. (2018). PCP 25 Artificial Intelligence Assisted Literature Reviews: Key Considerations For Implementation In Health Care Research. Value in Health, 21, S85.