WPOM

Working Papers on Operations Management.
Vol. 15 (172-208)
ISSN: 1989-9068

https://doi.org/10.4995/wpom.22282

Protocol paper: From Chaos to Order. Augmenting Manual Article Screening with Sentence Transformers in Management Systematic Reviews

Juan A. Marin-Garcia [a], Juan Martinez-Tomas[b] Amable Juarez-Tarraga[c] and Cristina Santandreu-Mascarell[d]

[a]ROGLE - Departamento de Organización de Empresas - Universitat Politècnica de València jamarin@omp.upv.es [b]Departamento de Organización de Empresas - Universitat Politècnica de València juamarto@upv.es [c]Departamento de Organización de Empresas - Universitat Politècnica de València amjua@omp.upv.es and [d]Departamento de Organización de Empresas - Universitat Politècnica de València crisanma@omp.upv.es

Recibido: 2024-08-18 Aceptado: 2024-12-02

To cite this article: Marin-Garcia, J.A.; Martinez-Tomas, J.; Juarez-Tarraga, A. & Santandreu-Mascarell, C. (2024). *Protocol paper: From Chaos to Order. Augmenting Manual Article Screening with Sentence Transformers in Management Systematic Reviews. WPOM-Working Papers on Operations Management,* 15, 172-208. doi: https://doi.org/10.4995/wpom.22282

Abstract

A spanish version of the article is provided (see section before Acknowledgements)

As scientific output grows, systematic reviews have become essential yet increasingly challenging. Our approach to this protocol aims to make this process more effective, efficient and accessible to researchers worldwide, including those in developing countries.

We developed a tool to complement human judgment in the screening phase using pre-trained language models and natural language processing techniques. This tool generates text embeddings and calculates semantic similarities, prioritizing potentially relevant articles. The goal is to utilise the similarity ranking instead of reviewing articles randomly or following the relevance sort option of search engines like WOS or Scopus. Coders can start with those closest to the category/categories of interest and progressively move towards the more distant ones. This approach would save time and effort while reducing the fatigue and biases of the coders.

The models we have tested in this research are all-MiniLM-L6-v2, all-distilroberta-v1, all-mpnet-base-v2, paraphrase-multilingual-mpnet-base-v2, distiluse-base-multilingual-cased-v1, all-MiniLM-L12-v2, allenai-specter, allenai/scibert_scivocab_uncased, distilbert-base-nli-mean-tokens, roberta-base-nli-stsb-mean-tokens, distiluse-base-multilingual-cased-v2, paraphrase-multilingual-MiniLM-L12-v2, stsb-roberta-large, bert-base-nli-mean-tokens.

The method was implemented using limited computational resources and open-source software, ensuring accessibility for research teams with restricted economic resources.

(cc) BY

Results indicate a possible reduction in screening time and improved consistency in article selection. The tool demonstrated utility in classifying relevant studies and would facilitate more comprehensive reviews.

By providing a low-cost solution, we aim to level the playing field in global research, enabling researchers from diverse economic backgrounds to participate more fully in producing scientific knowledge.

Keywords: *protocol paper; Systematic Literature Review; Management Research; Sentence Transformers; Natural Language Processing; Article Screening; Open Science; Research Accessibility; Machine Learning; Text Embeddings; Research Methodology*

Introduction

A systematic literature review is a rigorous and structured way to locate, select, analyse, and synthesise existing research results related to a particular research question (Page et al., 2021; Saunders et al., 2016). The review process must be transparent and replicable, this involves completing the following steps (Booth et al., 2019; Marin-Garcia, 2021; Medina-López et al., 2010; Ogbonnaya & Brown, 2023; Page et al., 2021; Saunders et al., 2016; Tranfield et al., 2003):

1. Preparation of a protocol containing at least:
    1. Identification of the need for the review
    2. Proposal of an explicit research question(s) (as far as possible, indicating the population to which the answer is to be generalised, or who needs it or will use it; the intervention, practice, service or action on which you want to decide; the conditions under which the intervention/service/action/practice is carried out (place, who performs it, etc.); the groups against which it is going to be compared, if any; the outcomes that are of interest to the decision or what is to be evaluated; why are you going to need to answer this question)
    3. The specific inclusion and exclusion criteria of the studies to be analysed in the review
2. Develop a search strategy that generates as few false positives and false negatives as possible
3. To assess the adequacy of the studies found (screening).
4. Assess the quality of the studies screened or the quality of their reporting
5. Data extraction and encoding following a protocol
6. Analyze and synthesise findings in a systematic manner (following the protocol)
7. Write the scientific report of the review
8. Guides to putting evidence into practice

(cc) BY

This process aims to minimise bias in the review process and provide a comprehensive view of the current state of knowledge, identifying areas for future research. This approach allows us to reach informed conclusions about what is known and what is still unknown about the research question, thus facilitating the advancement of knowledge in the field of study (Cooper, 1991; Saunders et al., 2016).

For us, the concept of systematic literature review involves the process with which the bank of documents to be processed is built and not how to resolve or analyse the documents. In this sense, a systematic review could be solved qualitatively with a content analysis, which can be solved narratively (Aguinis et al., 2020; Friese et al., 2018; Krippendorff, 2018), or quantitatively, with a bibliometric study or a meta-analysis (Borenstein et al., 2009; Van Rhee et al., 2018). The systematic review of literature has been used in different scientific fields for the advancement of knowledge (e.g. Alfalla-Luque et al., 2013 and 2023; Bayonne et al., 2020; Marín-García et al, 2018; Sánchez-Cazorla et al., 2016).

Although artificial intelligence can support any of the steps discussed above, in this protocol, we will focus exclusively on the screening phase of the documents obtained by applying the automatic search strategy. Our objective is to propose a working methodology with artificial intelligence that allows improving the effectiveness and efficiency of the screening phase of a systematic literature review compared to that done only with humans, especially when the number of documents to be processed is high and without draining the future capacity of the participating researchers (Ma & Su, 2024). To do this, we will develop an executable Python code and check if the affinity ranking obtained through vectorisation represents the association between the texts and the target topic of the systematic review.

The screening phase in a systematic literature review allows us to guarantee the relevance of the studies included in the final analysis. This stage involves the systematic evaluation of the titles, abstracts and, sometimes, full texts of the articles identified in the initial search, applying predefined inclusion and exclusion criteria (Higgins et al., 2021; Marin-Garcia, 2021; Medina-López et al., 2010; Moher et al., 2016). In addition, with the volume of existing scientific production, it is foreseeable that the results of automatic searches will yield large amounts of documents. This can compromise the review project's viability or the results' consistency if the screening phase is not handled efficiently (Borah et al., 2017). Since this process can be highly laborious and error-prone if the task is performed manually with human coders, we find it important to develop and use automated tools to support this phase (O'Mara-Eves et al., 2015).

However, as we will discuss in the conclusions of this work, the possibilities are not closed with the screening phase, and the use of artificial intelligence can be applied to other stages of the systematic literature review process, although its potential has to be tested in future research.

Previous research

Natural Language Processing is a field of artificial intelligence focusing on the interaction between computers and human language. Its primary purpose is to enable machines to understand, interpret and generate human language naturally and effectively (Goldberg, 2022; Rayhan, 2024; Speer et al., 2024). It uses algorithms and statistical models to analyse and process large amounts of linguistic data. In addition, machine learning provides the techniques and tools that allow Natural Language Processing systems to

(cc) BY

improve their performance based on experience or iterations of the task (Do et al., 2024; Patil & Gudivada, 2024).

On the other hand, transformers are a deep learning model designed especially for natural language processing tasks with an architecture based on attention mechanisms (Vaswani et al., 2017), which allows the capture of wide-context relationships in input data. Transformers have proven to be exceptionally effective in various language processing tasks, such as machine translation, text summarisation, and language generation (Binh et al., 2023; Devika et al., 2021; Troxler & Schelldorfer, 2024; Young et al., 2018).

Within Natural Language Processing classification applications/tasks, a quick search in Web Of Science (WOS) (Figure 1) makes us suspect that, by far, "sentiment analysis" is one of the most represented applications in scientific publications. The popularity of "topic modelling" seems to be a long way off. However, there are hardly any publications on "relevance classification".

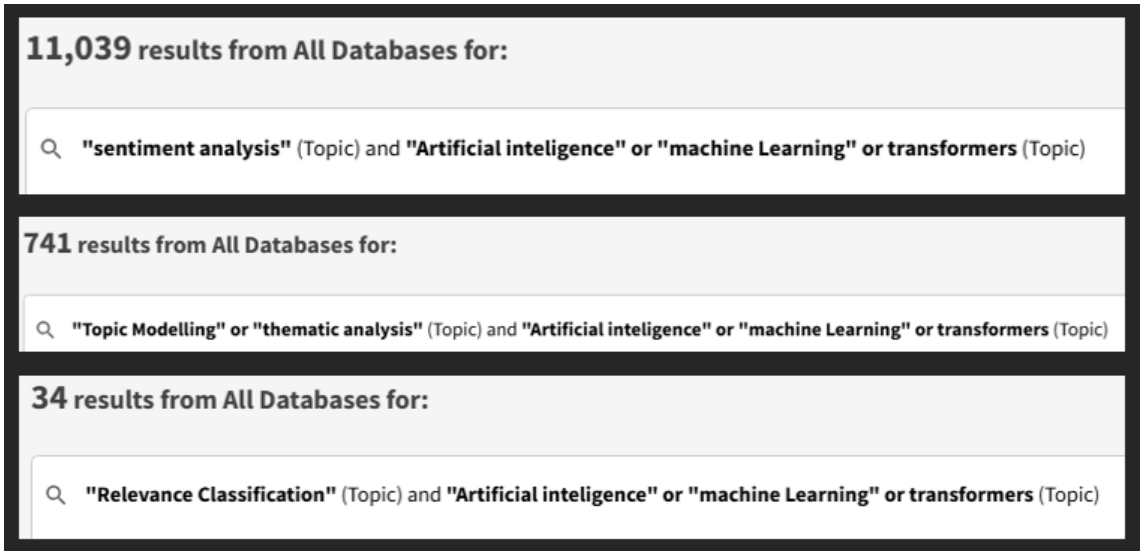


Figure 1. Comparison of the popularity of three NLP applications

"Sentiment Analysis" aims to identify and extract subjective information (opinions, feelings, attitudes or emotions) expressed in a text. The solution sought is a classification of feelings (positive, negative, neutral) or polarity indices between the extremes (positive/negative) (Liu, 2012). The level of analysis focuses on words or phrases.

On the other hand, Topic Analysis aims to identify the main themes or topic categories in a collection of documents, based on the frequency at which words or word patterns appear together. The expected output is a set of topics, each comprising keywords or word distributions (Asmussen & Moller, 2019; Blei et al., 2001; Bruno et al., 2022; Odacioglu et al., 2023). In this case, the analysis level is usually the collection of documents.

(cc) BY

Operationally, we can define the screening phase in a literature review as a situation where, on the one hand, we have a definition or description of the topic that is the object of our interest and, on the other hand, we have objects to classify, defined by title and abstract. Our decision is based on the proximity of the objects to be classified concerning the subject pursued.

In this sense, the screening phase fits better into the "Relevance Classification" category. This category focuses on determining the relevance (semantic similarity) of a document (in this case, a scientific article) concerning a specific query or criterion (in our case, the objective of the systematic review). The metrics of interest in this case are a binary classification (included/excluded) or a relevance (or similarity) score to the topic to be addressed (Manning et al., 2008). This task usually requires comparing the semantic content of the objects to be classified with that of the topic against which the relevance is evaluated. It is not only about the similarity between the words, but it involves considering the deep meaning of the text. The level of analysis is the document or its presentation through the title and abstract (Mikolov, Sutskever, et al., 2013; O'Mara-Eves et al., 2015; Sebastiani, 2002).

These text classification tasks, and especially those of "relevance classification", can use different Natural Language Processing techniques (Hickman et al., 2024; Hickman et al., 2020; Patil & Gudivada, 2024; Rayhan, 2024):

- Bag of Words. They use count-based methods (unigrams, bigrams, trigrams) and generate superficial representations that do not take context into account or capture semantic meaning.
- Recurrent Neural Networks. They are helpful for context-based classification tasks where word order adds to or alters meaning. They have variants such as Long Short-Term Memory Networks (LSTMs), adapted to situations where context must be retained in a long text block.
- Convolutional Neural Networks. They capture local patterns in phrases or n-grams. They are helpful for sentiment analysis and topic modelling.
- Pretrained Language Models. They are language models trained with large text corpora to learn general representations/patterns of language. Examples of these are ELMo, ULMFiT, XLNet, T5 (Text-to-Text Transfer Transformer), and GPT (Generative Pre-trained Transformer). Among these models, some are transformer architecture, others are not. Some are also autoregressive. In this block, we would include the "sentence transformers" that generate vector representations (embeddings) that capture semantic meaning and are context-sensitive. For example, BERT (Bidirectional Encoder Representations from Transformers) or RoBERTa.

Considering the above, three concepts are relevant to our use case. Therefore, we are going to briefly expand their definition because it will help to highlight why they have been part of the characteristics that we will look for in the chosen solution:

- Word Embeddings are dense vector representations of words in a multidimensional space. In that space, semantically similar words are close to each other. These representations capture semantic and syntactic relationships between words, allowing mathematical operations to be carried out on vocabulary that respect meanings (Mikolov, Chen, et al., 2013).

(cc) BY

- Sentence Transformers are language models based on the Transformer architecture, specifically designed to generate embeddings of sentences or entire paragraphs. These models are trained to produce vector representations of text that capture the semantic meaning of entire sentences, allowing comparisons of semantic similarity between texts of varying lengths (Binh et al., 2023; Devika et al., 2021; Galli et al., 2024; Su et al., 2023). This is just what we need for the screening task.
- "Cosine Similarity" is a measure of similarity between two objects, where each of them is represented by putting an n-dimensional vector and calculating the cosine of the angle between them. The smaller the angle, the more similar the vectors are considered. In natural language processing, it is used to compare the semantic similarity between embeddings of words or longer texts (sentences or paragraphs). The cosine similarity produces a value between -1 and 1, where 1 indicates vectors in the same direction (regardless of length), 0 indicates orthogonality, and -1 indicates vectors in opposite directions. This measurement is helpful in our case study because it is independent of the magnitude (length) of the vectors and only considers the orientation (Bugorski, 2014; Hanifi et al., 2022; Xia et al., 2015).

Contribution

In this work, we will check the technical feasibility of running a functional prototype on personal computers or free Cloud Computing platforms for any researcher to use, regardless of the budget available.

Our goal is to offer a step-by-step guide and code (python script) that anyone, even without programming training, can run on a personal computer or tablet with a free Gmail account that gives them access to Google Colab. This material will be available on GitHub (https://github.com/jamg-upv/LLMforSLRscreening). It will be complemented by a collection of videos (https://media.upv.es/#/portal/channel/222a2680-53d2-11ef-a0f7-e57947840c32) that show how to perform each of the stages, from the construction of the dataset to the execution, download of the results and interpretation of them.

In the current context, where the extension of the use of AI is generating an intense debate that questions its benefits and challenges in various sectors, Budhawar et al. (2023) highlight that it is not clear whether these technologies help transform work by replacing tasks that do not add value, or if they generate a displacement, where people will deal only with trivial or practically irrelevant decisions. In this scenario of uncertainty, our research aims to reduce the barriers to entry for the co-piloted use of advanced language models, thus contributing to a more equitable distribution of these tools in the academic research field without renouncing the human decision-makers participation in the process.

The relevance of this research lies in addressing two challenges. On the one hand, the technical complexity and programming knowledge required to use transformers can be a barrier, especially for people who research social sciences. On the other hand, discrimination could be generated between research teams with budgets to pay for high-performance computing resources compared to groups or individuals without access to research funds.

(cc) BY

By providing accessible and free tools, our research seeks to level the playing field, allowing a wider group of researchers to benefit from using advanced technologies to make systematic reviews more effective and efficient. This not only democratises access to sentence transformers' technology but can also contribute to a deeper and more diverse understanding of its implications in different fields of study, thus contributing to the academic and practical debate about the future of artificial intelligence and its impact on society.

Methods

The use case we will use is the classification of a set of articles, characterised by their title and abstract obtained from WOS, based on their relevance concerning the theoretical definition of a topic or concept.

The result to be obtained is an Excel file, which includes sheets with (1) the comparison of models used (time spent in process and tokens), (2) the cosine similarity table between objects and classification category(s), (3) the embeddings calculated for each text (either objects or categories) and (4) the ranking of articles closest to the category of interest as an average of the classification of ranks in each of the models used.

The Python code has been generated through several iterations with the help of a chatbot on the poe.com platform using the Claude-3.5-sonnet model. At the time of this work, the programming and scientific reasoning capabilities of Claude-3.5-sonnet (from Anthropic) exceeded those of the best available model of OpenAI (ChatGPT4o). We have ruled out classifying the articles directly with generative artificial intelligence models. We did tests using the OpenAI API to obtain a reliable classification of each article via the prompt. The result was inadequate, even after performing several iterations of prompt engineering and several tests during two months of work. On the other hand, even if the results had been promising, they would always lack transparency and replicability. That is why we decided to approach the project from a different point of view by calculating the embeddings and the cosine distance, which is at least completely replicable.

For this work, we have selected 14 models of sentence transformers:

1. all-MiniLM-L6-v2
2. all-distilroberta-v1
3. all-mpnet-base-v2
4. paraphrase-multilingual-mpnet-base-v2
5. distiluse-base-multilingual-cased-v1
6. all-MiniLM-L12-v2
7. Allenai-Specter
8. Allenai/scibert_scivocab_uncased
9. distilbert-base-nli-mean-tokens
10. roberta-base-nli-stsb-mean-tokens
11. distiluse-base-multilingual-cased-v2
12. paraphrase-multilingual-MiniLM-L12-v2
13. Stsb-Roberta-Large
14. bert-base-nli-mean-tokens.

(cc) BY

At the time of this research, we do not have a clear justification for choosing these models and not others. We have selected those, suggested by the generative artificial intelligence chatbot with Claude-3.5-sonnet (accessed in August 2024), that we have been able to verify have been used in a scientific article (Galli et al., 2024; Dhini et al., 2023; Guesmi et al., 2023; Kim et al., 2024; Kulkarni et al., 2023; Kurek et al., 2024; Muñoz & Iglesias, 2023; Naing et al., 2024; Scarpino et al., 2022; Singh et al., 2021; Troxler & Schelldorfer, 2024; Wu et al., 2022).

We believe no sentence transformer model is a clear winner for any "Relevance Classification" task. The performance/accuracy of each model varies depending on the classification task to be performed and the specific dataset available (the definition of the categories and summaries of the objects to be classified) (Perone et al., 2018; Wang et al., 2018). Therefore, it is necessary to test several models and choose the one that best suits the specific use case. We have not found any comparison in articles published in WOS or Scopus that analyses whether there is a universally superior sentence transformer model for the screening task in a systematic literature review.

Distance metrics directly applicable to embeddings include cosine similarity (defined in the previous section of this article); Euclidean distance, which measures the direct distance between two points in vector space and is sensitive to the magnitude of the vectors; Pearson's correlation coefficient; the Manhattan Distance, which adds up the absolute differences in the coordinates of the vectors; and Soft Cosine Similarity, an extension of cosine similarity (Bugorski, 2014; Cha, 2007; Chandrasekaran & Mago, 2021; Elekes et al., 2017; Levy et al., 2015; Nursalman et al., 2018; Sidorov et al., 2014; Xia et al., 2015).

Other metrics that could be used require simple transformations applicable to embeddings. Among them, we could have the Mahalanubis Distance, based on the correlation between vectors and requires the calculation of the inverse covariance matrix; Jaccard Similarity, which measures the intersection divided by the joining of the word sets of two documents, can be adapted to work with discretised or binary versions of embeddings; Hamming distance, helpful in comparing binary or categorical strings of equal length, can be applied to binary versions of embeddings; the Sørensen-Dice Similarity, which gives more weight to coincidences than Jaccard, can also be adapted for use with discretised embeddings.

Compared to the other alternatives, cosine similarity has the advantage of being scale-invariant since it only considers the direction of the vectors, not their magnitude. This is useful when comparing texts of different lengths. However, unlike the Euclidean Distance or the Manhattan Distance, it does not capture information about the magnitude of differences between vectors. On the other hand, the Mahalanobis distance could be more robust in cases where the dimensions of the embeddings are correlated (Cha, 2007). In this research, we will choose cosine similarity as distance and leave it for future research to compare whether any other distances might be more suitable to represent the similarity between articles and the review target in the screening phase.

The first objective of this protocol is to obtain an executable Python code that offers the tables of the Excel workbook discussed above. A second objective is to check if cosine similarities and affinity ranking extracted through vectorisation are helpful and represent the true association between texts and the review's objective.

(cc) BY

To do this, we will consider the results of each sentence transformer model as if they were an independent rater. We will take as "Gold Standard" the classification made by a human coder (in this case, the consensus of the team of authors of this article who act as human raters).

The first author will carry out a visual inspection of the results of each model (the order in which the works have been placed with the consensus distance, compared to the order previously agreed upon by all the authors), also analysing the Recall (sensitivity) (proportion of true positives ordered above the negatives, with respect to the total of real positives), Precision (proportion of true positives ordered above the negatives, with respect to the total number of classified objects until reaching the position of the last of the true positives) and Accuracy (total of true positives and true negatives with respect to the total number of classified objects) (O'Mara-Eves et al., 2015).

Pilot results

We will use as a pilot test a systematic literature review on the concept of High Involvement Work Programs or some of its components (Santandreu Mascarell et al., 2024). The categories for classification will be a definition of High Involvement Work Programs (Marin-Garcia, 2013), the 12 definitions of each of the human resources programs that have usually been identified as high-involvement practices (Marin-Garcia, 2013; Perello-Marin & Ribes-Giner, 2014) and an aggregate definition of High Involvement Work Programs as the concatenation of the global definition and that of the 13 practices. In addition, we have incorporated the definition of other categories used as distractions as classification topics. In them, we include three human resources practices which are not High Involvement Work Programs (Green HRM, Remote Work, Non-Traditional Work Arrangements); The definition of Operations Management (Marin-Garcia et al., 2021), the definition of Scholarship Of Teaching and Learning and the definition of Higher Education Course Management. These are the topics on which the 15 articles were selected in the pilot deal. Each category for classification has a short description (150-200 words), although in some cases, it may be longer.

As objects to be classified in this pilot test, we have selected 15 articles, and the team of authors of this work has agreed on a classification of them (Table 1). The first three (Hauff et al., 2022; Marin-Garcia & Martinez Tomas, 2016; Song et al., 2021) are articles that deal with High Involvement Work Programs (1hiwp). The fourth (Marin-Garcia & Martínez-Tomás, 2022) deals with remuneration systems. However, it focuses on the salary structure of people hired in Spain and its evolution over time, so it is an HRM article and not a specific one of high involvement (2hrm2). (Ahmad et al., 2022; Houeland & Jordhus-Lier, 2022) are Green HRM articles. (Becker et al., 2022) is from Remote Work. (Burbano & Chiles, 2022) it is a Non-Traditional Work Arrangement, although it can also be considered Remote Work. (Mora-Valentin et al., 2024) It can be considered Course Management, although due to the subject's topics, it could fit with Green HRM or be in the category of Training (2hrmteach). (Rincon & Zorrilla, 2017; Rincón et al., 2023) is Course Management (3teach). (Marin-Garcia, Garcia-Sabater, Garcia-Sabater, et al., 2020) it is Course Management but when dealing with continuous improvement teams/kaizen topics, it can be linked to Operations Management or Empowerment (4omteach). The penultimate rows (Aguilar-Escobar et al., 2024; Marin-Garcia, Garcia-Sabater, Ruiz, et al., 2020; Nguyen et al., 2024) are clearly from Operations Management (4om), although NgocNguyen2024 talks about sustainability and could create confusion with

(cc) BY

Green HRM. The last row (Aznar-Mas et al., 2023) is Course Management (5 teach), but when evaluating the subjects' performance, I could have some confusion with Performance Management.

Table 1. Articles to classify

Component	Title	Classification
Hauff2022	High-performance Work Practices, employee well-being, and supportive leadership: spillover mechanisms and boundary conditions between HRM and leadership behavior	1hiwp
Song2021	High involvement work systems and organizational performance: the role of knowledge combination capability and interaction orientation	1hiwp
Marin 2016	Deconstructing AMO framework: a systematic review	1hiwp
Marin2022	What does the wage structure depend on? Evidence from the national salary survey in Spain	1-2hrm
Ahmad2022	The impact of green HRM on green creativity: mediating role of pro-environmental behaviors and moderating role of ethical leadership style	2hrn
Houeland2022	Not my task': Role perceptions in a green transition among shop stewards in the Norwegian petroleum industry	2hrn
Becker2022	Surviving remotely: How job control and loneliness during a forced shift to remote work impacted employee work behaviors and well-being	2hrn
Burbano2022	Mitigating Gig and Remote Worker Misconduct: Evidence from a Real Effort Experiment	2hrn
Black2024	Integrating the SDG into university teaching: an application in human resources subjects	2hrmteach
rincon2023	The impact of active learning on entrepreneurial capacity	3teach
Marin2020triplediam	Protocol: Triple Diamond method for problem solving and design thinking. Rubric validation	4omteach
Aguilar2024	Factors influencing nurse satisfaction with Automated Medication Dispensing Cabinets	4om
NgocNguyen2024	Can sustainable supply chain strategies of company enhance for mitigation of risk damages and long-term resilience?	4om
Marin2020Omcovid	Operations Management at the service of health care management: Example of a proposal for action research to plan and schedule health resources	4om
Aznar2023	Effectiveness of the use of open-ended questions in student evaluation of teaching in an engineering degree	5teach

The generated code (accessible on GitHub: https://github.com/jamg-upv/LLMforSLRscreening) has worked correctly. The loading of the necessary libraries takes about 2 minutes, and the process of calculating embeddings, with each of the 14 models tested for the 20 categories and the 15 articles, is done in about 8 minutes. The total processing time does not exceed 11 minutes until the desired Excel file is obtained.

Some models run in less than 4 seconds, and others take more than a minute (Figure 2). The smaller models (384 or 512 tokens) are generally the fastest. Among the 768-token models, some resolve quickly, and others do not. The only model tested with 1024 tokens is among those that require the most processing time. We have repeated the vectorisation several times, and the times, although varying, are always in the same range of values, respecting the process time ranking. Models that are faster (or slower) are faster in all repetitions.

(cc) BY

	A	B	C	D
1	model	best_matches	processing_time_seconds	embedding_size
2	all-MiniLM-L6-v2	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPshortDescrip', 'Compensations', 'GrenHRM', 'GrenHRM', 'RemoteWork', 'FairJob', 'CourseManag', 'HIWPLongDescrip', 'HIWPLongDescrip', 'Empowerment', 'Compensations', 'OpMange', 'SotlhigEdu']	4,54	384
3	all-distilroberta-v1	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPLongDescrip', 'HIWPLongDescrip', 'GrenHRM', 'GrenHRM', 'RemoteWork', 'Communication', 'SotlhigEdu', 'Training', 'HIWPLongDescrip', 'HIWPLongDescrip', 'OpMange', 'OpMange', 'PerformEval']	39,31	768
4	all-mpnet-base-v2	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPshortDescrip', 'Compensations', 'GrenHRM', 'GrenHRM', 'WorkLifeBalance', 'PositiveCulture', 'HIWPLongDescrip', 'Training', 'Empowerment', 'HIWPLongDescrip', 'GrenHRM', 'OpMange', 'SotlhigEdu']	54,04	768
5	paraphrase-multilingual-mpnet-base	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPLongDescrip', 'HIWPLongDescrip', 'GrenHRM', 'GrenHRM', 'RemoteWork', 'HIWPLongDescrip', 'CourseManag', 'Empowerment', 'HIWPLongDescrip', 'Empowerment', 'Compensations', 'OpMange', 'PerformEval']	13,29	768
6	distiluse-base-multilingual-cased-v1	['HIWPshortDescrip', 'HIWPshortDescrip', 'GrenHRM', 'Compensations', 'GrenHRM', 'Empowerment', 'Empowerment', 'RemoteWork', 'GrenHRM', 'Training', 'HIWPLongDescrip', 'Compensations', 'OpMange', 'OpMange', 'Communication']	7,73	512
7	all-MiniLM-L12-v2	['HIWPshortDescrip', 'HIWPshortDescrip', 'GrenHRM', 'Compensations', 'GrenHRM', 'GrenHRM', 'RemoteWork', 'FairJob', 'CourseManag', 'Training', 'Career', 'PerformEval', 'PositiveCulture', 'OpMange', 'SotlhigEdu']	3,74	384
8	allenai-specter	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPshortDescrip', 'Career', 'GrenHRM', 'GrenHRM', 'WorkLifeBalance', 'Communication', 'SotlhigEdu', 'SotlhigEdu', 'CourseManag', 'OpMange', 'OpMange', 'OpMange', 'SotlhigEdu']	66,32	768
9	allenai/scibert_scivocab_uncased	['HIWPLongDescrip', 'HIWPshortDescrip', 'HIWPLongDescrip', 'Communication', 'GrenHRM', 'Empowerment', 'HIWPLongDescrip', 'HIWPLongDescrip', 'SotlhigEdu', 'Training', 'Communication', 'Empowerment', 'OpMange', 'Empowerment', 'SotlhigEdu']	73,69	768
10	distilbert-base-nli-mean-tokens	['HIWPshortDescrip', 'HIWPLongDescrip', 'Select&Recruitment', 'HIWPLongDescrip', 'GrenHRM', 'Communication', 'WorkLifeBalance', 'Empowerment', 'SotlhigEdu', 'Communication', 'HIWPshortDescrip', 'Communication', 'Training', 'Training', 'SotlhigEdu']	6,67	768
11	roberta-base-nli-stsb-mean-tokens	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPshortDescrip', 'PerformEval', 'GrenHRM', 'NonTradWork', 'RemoteWork', 'Empowerment', 'SotlhigEdu', 'Training', 'HIWPshortDescrip', 'Communication', 'Communication', 'Communication', 'SotlhigEdu']	14,68	768
12	distiluse-base-multilingual-cased-v2	['HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPshortDescrip', 'HIWPLongDescrip', 'GrenHRM', 'Communication', 'Empowerment', 'FairJob', 'GrenHRM', 'Select&Recruitment', 'Empowerment', 'Compensations', 'Empowerment', 'OpMange', 'Select&Recruitment']	8,85	512
13	paraphrase-multilingual-MiniLM-L12-	['HIWPshortDescrip', 'HIWPshortDescrip', 'GrenHRM', 'Career', 'GrenHRM', 'Empowerment', 'RemoteWork', 'RemoteWork', 'SotlhigEdu', 'HIWPLongDescrip', 'HIWPLongDescrip', 'PositiveCulture', 'Communication', 'OpMange', 'SotlhigEdu']	3,65	384
14	stsb-roberta-large	['HIWPshortDescrip', 'HIWPshortDescrip', 'SotlhigEdu', 'HIWPLongDescrip', 'GrenHRM', 'Communication', 'RemoteWork', 'RemoteWork', 'CourseManag', 'HIWPLongDescrip', 'HIWPLongDescrip', 'Communication', 'OpMange', 'Communication', 'Communication']	47,8	1024
15	bert-base-nli-mean-tokens	['HIWPshortDescrip', 'HIWPshortDescrip', 'SotlhigEdu', 'Empowerment', 'GrenHRM', 'Compensations', 'NonTradWork', 'Compensations', 'SotlhigEdu', 'Training', 'HIWPshortDescrip', 'Training', 'Compensations', 'SotlhigEdu', 'SotlhigEdu']	13,46	768

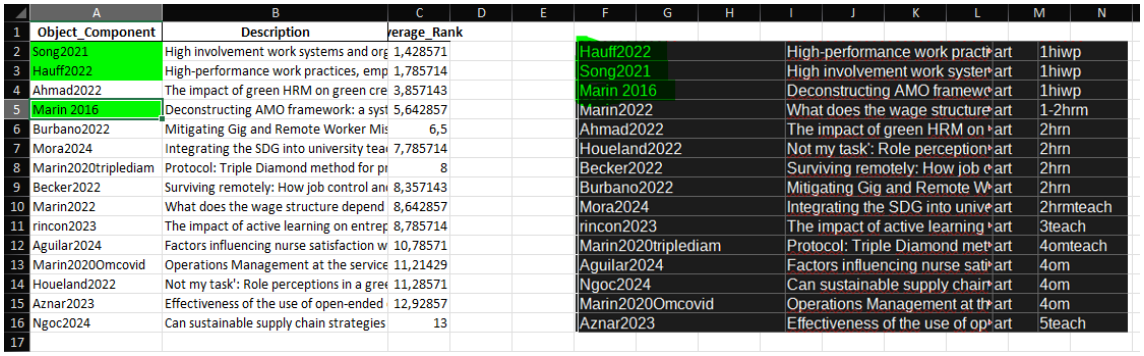
Figure 2. Comparing processing time and size of different models

We have tested the processing time with different text corpora. On the one hand, by embedding only the categories (about 14000 characters) or the categories along with the items to be sorted (about 39500 characters). In fast models, almost tripling the number of characters to be processed multiplies the processing time by 1.5. In slower models, it triples the processing time. These data must be interpreted with caution because they result from a single repetition with a single corpus of text, and we do not know if they are replicable if they are repeated more times or if the corpus to be analysed is changed. However, they can serve as a first estimate of the processing time depending on the corpus size to be analysed.

Figure 3, on the left, shows the result of averaging the positions of each model of the rankings from highest to lowest cosine similarity concerning the "HIWPshortDescript" category (giving equal weight to all models). On the right side (in black) is the classification made by human experts, where the true positives (articles that deal with High Involvement Work Programs) have been highlighted in green. Next, those dealing with HR issues have been placed, and there are articles on teaching or operations management. Our human classification is sorted by blocks of categories. Those tagged as 1hiwp are our target items. Those

(cc) BY

labelled 2hrm are articles that do not fit our review but should be more akin than the others. The tags 3teach, 4om or 5teach, form a block of articles even more distant from the objective of our review. Within each block, there is no order; any of the items in the block may or may not be better positioned in the ranking, and that is not considered a classification error. What we will consider as a false positive is the insertion of an article from one of the non-objective blocks between the articles highlighted in green. In the same way, we will consider a false negative if any of the articles marked as green (target block) are buried in the classification or are surpassed by an article that does not belong to the 2hrm category (which is a category-related in content to our target category).



	Object_Component	Description	erage_Rank
2	Song2021	High involvement work systems and org	1,428571
3	Hauff2022	High-performance work practices, emp	1,785714
4	Ahmad2022	The impact of green HRM on green cre	3,857143
5	Marin 2016	Deconstructing AMO framework: a syst	5,642857
6	Burbano2022	Mitigating Gig and Remote Worker Mis	6,5
7	Mora2024	Integrating the SDG into university tea	7,785714
8	Marin2020tripledіam	Protocol: Triple Diamond method for pı	8
9	Becker2022	Surviving remotely: How job control an	8,357143
10	Marin2022	What does the wage structure depend	8,642857
11	rincon2023	The impact of active learning on entrep	8,785714
12	Aguilar2024	Factors influencing nurse satisfaction w	10,78571
13	Marin2020Omcovid	Operations Management at the service	11,21429
14	Houeland2022	Not my task': Role perceptions in a gre	11,28571
15	Aznar2023	Effectiveness of the use of open-ended	12,92857
16	Ngoc2024	Can sustainable supply chain strategies	13
17			

Hauff2022	High-performance work practi	art	1hiwp
Song2021	High involvement work syster	art	1hiwp
Marin 2016	Deconstructing AMO framewo	art	1hiwp
Marin2022	What does the wage structure	art	1-2hrm
Ahmad2022	The impact of green HRM on	art	2hrn
Houeland2022	Not my task': Role perception	art	2hrn
Becker2022	Surviving remotely: How job c	art	2hrn
Burbano2022	Mitigating Gig and Remote W	art	2hrn
Mora2024	Integrating the SDG into unive	art	2hrmteach
rincon2023	The impact of active learning	art	3teach
Marin2020tripledіam	Protocol: Triple Diamond met	art	4omteach
Aguilar2024	Factors influencing nurse sati	art	4om
Ngoc2024	Can sustainable supply chain	art	4om
Marin2020Omcovid	Operations Management at th	art	4om
Aznar2023	Effectiveness of the use of op	art	5teach

Figure 3. Global ranking taking into account results of the 14 sentence transformer models

In Table 2, we display the ranking performance metrics for the rankings aggregate and for the ranking of each of the models separately when the target classification category is "HIWPshortDescript". The solution representing all 14 models has acceptable ranking metrics, with high Acuracy and Recall, although the Accuracy is not perfect. However, we can see that the performance of the models is quite different. Some models classify perfectly (in total agreement with human classification), and others have very low yields (especially in precision and accuracy). Therefore, the aggregate ranking is penalised by incorporating results from models that do not perform very well in the classification task. This result is important because it opens a line for future research to confirm whether the same models always work well or whether it is necessary to detect which one works best for each situation (for each corpus to be analysed). In the event that the best models change with the corpus, we should analyse whether it is worthwhile to detect the best model or whether it would be less expensive to accept the solution of the aggregation of the 14 models, which is imperfect but perhaps good enough from a practical point of view.

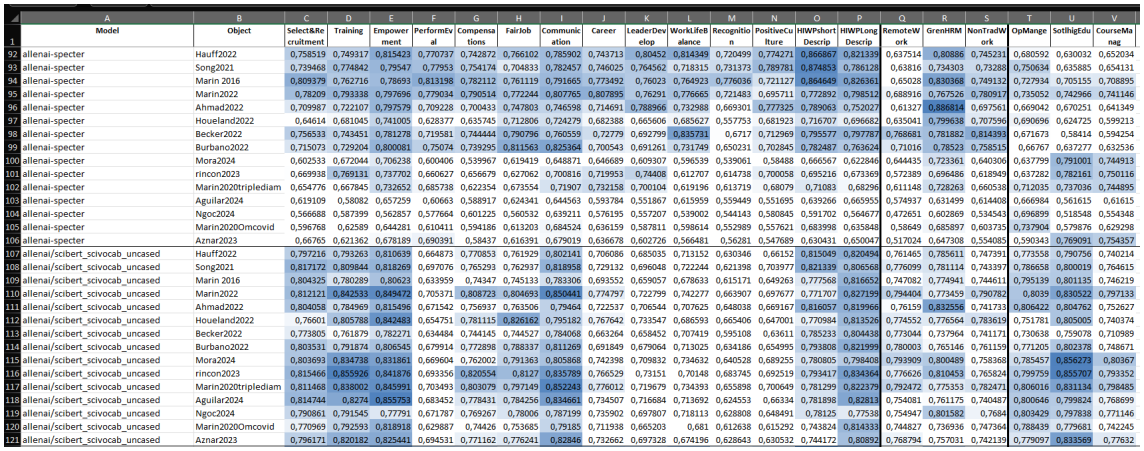
(cc) BY

Table 2. classification performance metrics (total positives=3; Total objects=15)

Model	Recall	Precision	Accuracy	True Positive	True Negatives	False Positive	False Negative	Position Of Last Positive
Average14Models	100,0%	75,0%	93,3%	3	11	1	0	4
all-distilroberta-v1	100,0%	100,0%	100,0%	3	12	0	0	3
Allenai/scibert_scivocab_uncased	66,7%	18,2%	86,7%	2	11	1	1	11
Allenai-Specter	100,0%	100,0%	100,0%	3	12	0	0	3
all-MiniLM-L12-v2	100,0%	100,0%	100,0%	3	12	0	0	3
all-MiniLM-L6-v2	100,0%	100,0%	100,0%	3	12	0	0	3
all-mpnet-base-v2	100,0%	100,0%	100,0%	3	12	0	0	3
bert-base-nli-mean-tokens	66,7%	20,0%	33,3%	2	3	9	1	10
distilbert-base-nli-mean-tokens	66,7%	18,2%	33,3%	2	3	9	1	11
distiluse-base-multilingual-cased-v1	66,7%	28,6%	26,7%	2	2	10	1	7
distiluse-base-multilingual-cased-v2	100,0%	75,0%	40,0%	3	3	9	0	4
paraphrase-multilingual-MiniLM-L12-v2	100,0%	100,0%	93,3%	3	11	1	0	3
paraphrase-multilingual-mpnet-base-v2	100,0%	75,0%	53,3%	3	5	7	0	4
roberta-base-nli-stsb-mean-tokens	66,7%	33,3%	20,0%	2	1	11	1	6
Stsb-Roberta-Large	66,7%	20,0%	20,0%	2	1	11	1	10

By reviewing the results of the cosine similarities of the different models, we can analyse in detail the performance of the models against different classification categories with the text corpus of this pilot (Figure 4). As we have already mentioned, some models rank better than others. Taking allenai-specter and allenai/scibert_scivocab_uncase as examples, we see that the former ranks better than the latter. allenai-Specter identifies the three true positives as High Involvement Work Programs articles (deeper blue in the category column) and ranks Ahmad2022 and Houeland2022 as GreenHRM articles (their proximity to that category is superior to the others). Becker2022 clearly associates it with Non-Traditional Work Arrangements and Work-Life Balance. The articles in Mora2024, Rincon2023, Marin2020tripleDiam and Aznar2023, identify them as teaching articles. Aguilar2024, Ngoc2024 and Marin2020omcovid, identifies them as Operations Management (in the row the highest value is in that category). However, allenai/scibert_scivocab_uncased is not able to discriminate the articles well, and practically everything is similar, with some exceptions, within each category. The columns have a very similar colour in all 15 cells, which indicates that the articles are very similar for this model, and you cannot discriminate which one is more similar. When the reality is that there are big differences between items from different blocks. We recall that these results cannot be considered conclusive until replicated with another corpus.

(cc) BY



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Model	Object	Select&Recruitment	Training	Empowerment	PerformEval	Compensations	FairJob	Communication	Career	LeaderDevelop	WorkLifeBalance	Recognition	PositiveCulture	HIWPshortDescrip	HIWPLongDescrip	RemoteWork	GrenHRM	NonTradWork	OpMange	SotlhigEdu	CourseManag
92	allenai-specter	Hauff2022	0,758519	0,749317	0,815423	0,770737	0,742872	0,766102	0,785902	0,743713	0,80452	0,814349	0,720499	0,774271	0,868867	0,821339	0,637514	0,80886	0,745231	0,680592	0,630032	0,652034
93	allenai-specter	Song2021	0,739468	0,774842	0,79547	0,77953	0,754174	0,704833	0,782457	0,746025	0,764562	0,718315	0,731373	0,789781	0,874853	0,786128	0,63816	0,734303	0,73288	0,750634	0,635885	0,654131
94	allenai-specter	Marin 2016	0,809379	0,762716	0,78693	0,813198	0,782112	0,761119	0,791665	0,773492	0,76023	0,764923	0,776036	0,721127	0,864649	0,826361	0,65028	0,830368	0,749132	0,727934	0,705155	0,708895
95	allenai-specter	Marin2022	0,78209	0,793338	0,797696	0,779034	0,790514	0,772244	0,807765	0,807895	0,76291	0,776665	0,721483	0,695711	0,772892	0,798512	0,688916	0,767526	0,780917	0,735052	0,742966	0,741146
96	allenai-specter	Ahmad2022	0,709987	0,722107	0,797579	0,709228	0,700433	0,747803	0,746598	0,714691	0,788966	0,732988	0,669301	0,777325	0,789063	0,752027	0,61327	0,886814	0,697561	0,669042	0,670251	0,641349
97	allenai-specter	Houeland2022	0,64614	0,681045	0,741005	0,628377	0,635745	0,712806	0,724279	0,682388	0,665606	0,685627	0,557753	0,681923	0,716707	0,696682	0,635041	0,799638	0,707596	0,690696	0,624725	0,599213
98	allenai-specter	Becker2022	0,756533	0,743451	0,781278	0,719581	0,744444	0,790796	0,760559	0,72779	0,692799	0,835731	0,6717	0,712969	0,795577	0,797787	0,768681	0,781882	0,814393	0,671673	0,58414	0,594254
99	allenai-specter	Burbano2022	0,715073	0,729204	0,800081	0,75074	0,739295	0,811563	0,825364	0,700543	0,691261	0,731749	0,650231	0,702845	0,782487	0,763624	0,71016	0,78523	0,758515	0,66767	0,637277	0,632536
100	allenai-specter	Mora2024	0,602533	0,672044	0,706238	0,600406	0,539967	0,619419	0,648871	0,646689	0,609307	0,596539	0,539061	0,58488	0,666567	0,622846	0,644435	0,723361	0,640306	0,637799	0,791001	0,744913
101	allenai-specter	rincon2023	0,669938	0,769131	0,737702	0,660627	0,656679	0,627062	0,700816	0,719953	0,74408	0,612707	0,614738	0,700058	0,695216	0,673369	0,572389	0,696486	0,618949	0,637282	0,782161	0,750116
102	allenai-specter	Marin2020triplediam	0,654776	0,667845	0,732652	0,685738	0,622354	0,673554	0,71907	0,732158	0,700104	0,619196	0,613719	0,68079	0,71083	0,68296	0,611148	0,728263	0,660538	0,712035	0,737036	0,744895
103	allenai-specter	Aguilar2024	0,619109	0,58082	0,657259	0,60663	0,588917	0,624341	0,644563	0,593784	0,551867	0,615959	0,559449	0,551695	0,639266	0,665955	0,574937	0,631499	0,614408	0,666984	0,561615	0,61615
104	allenai-specter	Ngoc2024	0,566688	0,587399	0,562857	0,577664	0,601225	0,560532	0,639211	0,576195	0,557207	0,539002	0,544143	0,580845	0,591702	0,564677	0,472651	0,602869	0,534543	0,696899	0,518548	0,554348
105	allenai-specter	Marin2020Omcovid	0,596768	0,62589	0,644281	0,610411	0,594186	0,613203	0,684524	0,636159	0,587811	0,598614	0,552989	0,557621	0,683998	0,635848	0,58649	0,685897	0,603735	0,737904	0,579876	0,629298
106	allenai-specter	Aznar2023	0,66765	0,621362	0,678189	0,690391	0,58437	0,616391	0,679019	0,636678	0,602726	0,566481	0,56281	0,547689	0,630431	0,650047	0,517024	0,647308	0,554085	0,590343	0,769051	0,754357
107	allenai/scibert_scivocab_uncased	Hauff2022	0,797216	0,793263	0,810639	0,664873	0,770853	0,761929	0,802141	0,706086	0,685035	0,713152	0,630346	0,66152	0,815049	0,820494	0,761465	0,785611	0,747391	0,773558	0,790756	0,740214
108	allenai/scibert_scivocab_uncased	Song2021	0,817172	0,809844	0,818269	0,697076	0,765293	0,762937	0,818958	0,729132	0,696048	0,722244	0,621398	0,703977	0,821339	0,806568	0,776099	0,781114	0,743397	0,786658	0,800019	0,764615
109	allenai/scibert_scivocab_uncased	Marin 2016	0,804325	0,780289	0,80623	0,633959	0,74347	0,745133	0,783306	0,693552	0,659057	0,678633	0,615171	0,649263	0,777568	0,816652	0,747082	0,774941	0,744611	0,795139	0,801135	0,746219
110	allenai/scibert_scivocab_uncased	Marin2022	0,812121	0,842533	0,849473	0,705371	0,808723	0,804693	0,850441	0,774757	0,722799	0,742277	0,663907	0,697677	0,771707	0,827199	0,794404	0,773459	0,750782	0,8039	0,830522	0,797133
111	allenai/scibert_scivocab_uncased	Ahmad2022	0,804058	0,784969	0,815496	0,671542	0,756937	0,763506	0,79464	0,722537	0,706544	0,707625	0,648038	0,669167	0,816057	0,819966	0,76159	0,832556	0,741733	0,806422	0,804762	0,752627
112	allenai/scibert_scivocab_uncased	Houeland2022	0,76601	0,805788	0,842483	0,654751	0,781115	0,826162	0,795182	0,767642	0,733547	0,686593	0,665406	0,647001	0,770984	0,813526	0,774552	0,776564	0,783619	0,751781	0,805005	0,740374
113	allenai/scibert_scivocab_uncased	Becker2022	0,773805	0,761879	0,782271	0,634484	0,744145	0,744527	0,784068	0,663264	0,658452	0,707419	0,595108	0,63611	0,785233	0,804438	0,773044	0,737964	0,741171	0,730638	0,759078	0,710989
114	allenai/scibert_scivocab_uncased	Burbano2022	0,803531	0,791874	0,806545	0,679914	0,772898	0,788337	0,811269	0,691849	0,679064	0,713025	0,634186	0,654995	0,793808	0,821999	0,780003	0,765146	0,761159	0,771205	0,802378	0,748671
115	allenai/scibert_scivocab_uncased	Mora2024	0,803693	0,834738	0,831861	0,669604	0,762002	0,791363	0,805868	0,742398	0,709832	0,734632	0,640528	0,689255	0,780805	0,798408	0,793909	0,800489	0,758368	0,785457	0,856273	0,80367
116	allenai/scibert_scivocab_uncased	rincon2023	0,815466	0,855926	0,841876	0,693356	0,820554	0,8127	0,835789	0,766529	0,73151	0,70148	0,683745	0,692519	0,793417	0,834364	0,776626	0,810453	0,765824	0,799759	0,855707	0,793352
117	allenai/scibert_scivocab_uncased	Marin2020triplediam	0,811468	0,838002	0,845991	0,703493	0,803079	0,797149	0,852243	0,776012	0,719679	0,734393	0,655898	0,700649	0,781299	0,822379	0,792472	0,775353	0,782471	0,806016	0,831134	0,798485
118	allenai/scibert_scivocab_uncased	Aguilar2024	0,814744	0,8274	0,855751	0,683452	0,778431	0,784256	0,834661	0,734507	0,716684	0,713692	0,624553	0,66334	0,781898	0,82813	0,754081	0,761175	0,740487	0,800646	0,799824	0,768699
119	allenai/scibert_scivocab_uncased	Ngoc2024	0,790861	0,791545	0,77791	0,671787	0,769267	0,78006	0,787199	0,735902	0,697807	0,718113	0,628808	0,648491	0,78125	0,77538	0,754947	0,801582	0,7684	0,803429	0,797838	0,771146
120	allenai/scibert_scivocab_uncased	Marin2020Omcovid	0,770969	0,792593	0,818918	0,629887	0,74426	0,753685	0,79185	0,711938	0,665203	0,681	0,612638	0,615292	0,743824	0,814333	0,744827	0,736936	0,747364	0,788439	0,779681	0,742245
121	allenai/scibert_scivocab_uncased	Aznar2023	0,796171	0,820182	0,825441	0,694531	0,771162	0,776241	0,82846	0,732662	0,697328	0,674196	0,628643	0,630532	0,744172	0,80892	0,768794	0,757031	0,742139	0,779097	0,833569	0,77632

Figure 4. Example of cosine similarity values for two of the models

The difference in the success of the models compared to the human evaluator can be confirmed in Figure 5, which shows the ranking by highest to lowest cosine similarity between the embeddings of each item and that of the "HIWPshortDescript" category. The upper part model correctly classifies the three true positives and identifies the rest as more distant. As an additional detail, it is confirmed that articles related to human resources are considered closer to high-involvement work programs than those related to operations management or teaching. This is a good result. However, the model at the bottom intersperses false positives with true positives (Ahmad2022 is considered more similar to High Involvement Work Programs than Hauff2022) and makes a false negative for Marin2016, which is buried in the similarity ranking. In addition, it was found that some teaching or operations management articles are considered more relevant than the true positives or HRM articles. This would be an example of an undesirable outcome.

(cc) BY

	A	B	D
1	Model	Object_Component	Similarity
62	allenai-specter	Song2021	0,874853
63	allenai-specter	Hauff2022	0,866867
64	allenai-specter	Marin 2016	0,864649
65	allenai-specter	Becker2022	0,795577
66	allenai-specter	Ahmad2022	0,789063
67	allenai-specter	Burbano2022	0,782487
68	allenai-specter	Marin2022	0,772892
69	allenai-specter	Houeland2022	0,716707
70	allenai-specter	Marin2020triplediam	0,71083
71	allenai-specter	rincon2023	0,695216
72	allenai-specter	Marin2020Omcovid	0,683998
73	allenai-specter	Mora2024	0,666567
74	allenai-specter	Aguilar2024	0,639266
75	allenai-specter	Aznar2023	0,630431
76	allenai-specter	Ngoc2024	0,591702
77	allenai/scibert_scivocab_uncased	Song2021	0,821339
78	allenai/scibert_scivocab_uncased	Ahmad2022	0,816057
79	allenai/scibert_scivocab_uncased	Hauff2022	0,815049
80	allenai/scibert_scivocab_uncased	Burbano2022	0,793808
81	allenai/scibert_scivocab_uncased	rincon2023	0,793417
82	allenai/scibert_scivocab_uncased	Becker2022	0,785233
83	allenai/scibert_scivocab_uncased	Aguilar2024	0,781898
84	allenai/scibert_scivocab_uncased	Marin2020triplediam	0,781299
85	allenai/scibert_scivocab_uncased	Ngoc2024	0,78125
86	allenai/scibert_scivocab_uncased	Mora2024	0,780805
87	allenai/scibert_scivocab_uncased	Marin 2016	0,777568
88	allenai/scibert_scivocab_uncased	Marin2022	0,771707
89	allenai/scibert_scivocab_uncased	Houeland2022	0,770984
90	allenai/scibert_scivocab_uncased	Aznar2023	0,744172
91	allenai/scibert_scivocab_uncased	Marin2020Omcovid	0,743824

Figure 5. Example of similarity ranking for two of the models

Conclusion, limitations and further research

The methodology we propose is designed to increase the efficiency and effectiveness of the screening process without replacing human judgment. Throughout the process, we keep researchers at the centre of decision-making and use artificial intelligence only as a support tool, fostering a complementary relationship between technology and coders. In this way, the possible risks of replacement or slavery to AI of the participating research staff are reduced (Ma & Su, 2024).

The goal is to use the similarity ranking generated from the embeddings provided by the sentence transformers to optimise the systematic literature review process. Instead of randomly reviewing articles, coders can start with those closest to the categories of interest and progressively work up to the furthest ones, saving time and effort and reducing coder fatigue and bias.

The pilot test results have allowed us to verify the feasibility of the small-scale proposal. We have also found that some models classify in much the same way as those who have acted as expert coders. We can conclude that the results of the pilot generate an adequate ranking when the models that rank well are chosen.

(cc) BY

The application of this protocol to new datasets (Santandreu Mascarell et al., 2024) and (Marin-Garcia & Martinez Tomas, 2016) from research where we have a large volume of articles coded by human experts will allow us to check the scalability of the procedure when we go from 15 articles to hundreds or thousands. It will also allow us to calculate classification performance metrics compared to humans.

Another aspect to be addressed in the future is whether the approach proposed in this work can be generalised to any type of systematic review or whether the field of research (e.g. Operations Management or Human Resources) can condition its use. The example we have worked with is Human Resources, and we will have to test whether it is generalisable to other examples in that field or even in other scientific fields.

In addition, this protocol opens new lines of research that try to overcome some of the limitations or questions that have arisen while writing this protocol.

Related to sentence transformer models for screening tasks, several questions arise for future research:

1. Model selection: how to choose the suitable models? Are there other relevant models besides the 14 included in this protocol?
2. Comparative model performance: Can we identify models that consistently outperform others in screening classification tasks?
3. Performance generalisation: Is the superiority of specific models maintained by changing the subject of study, the accuracy of category definitions, or by using a new corpus of abstracts for analysis?
4. Improved definitions: Could the results be improved if, instead of using category definitions from previously published research, we refined them with the help of generative AI?
5. Compare the result with models such as Llama3.1 (through, for example, the HuggingFace platform) that could use its wide context window to classify with the full-text content and not just with title and summary
6. To analyse the differences in results between using generic pre-trained models (as in this article) compared to models with a specific fine-tuning for the classification task (scientific field and specific topic of the systematic review)
7. Limitations of sentence transformers to capture meanings, especially in complex texts with context-dependent meanings, which is a common situation in scientific texts in the area of business management

Regarding proximity measurements, future research could compare the results obtained using Euclidean distance or other distance measures between embeddings rather than cosine similarity. This analysis could be cross-referenced with the previous model analysis because perhaps the results of some models are more sensitive to the change in proximity measurement than others.

(cc) BY

One could also compare the effect of using compact definitions (such as our HIWPshortDescript) versus using a concatenation of the compact definitions of each of the components of a multidimensional concept (such as our HIWPLongDescriptshort).

Based on the data from this protocol, another possible research to be developed in the future would be the statistical analysis of consistency between "coders". To do this, the variance between raters (between) and intra-raters (within) could be taken into account (Marin-Garcia et al., 2008) or other measures of ranking performance (O'Mara-Eves et al., 2015). It could also be analysed what is the appropriate number of models to run and integrate to achieve a more consistent solution. One of the questions to be resolved would be whether the ranking offered by a model or the addition of several models is better (even if they are not the models that rank best, their joint ranking still exceeds the ranking of the model with the best individual performance) and, in the latter case, how many models to add to achieve better consistency in classifications of texts other than those of training.

We have verified that processing 39000 characters (titles plus summary of 15 articles) with the seven models that give the best results takes about 3 minutes of script execution in Google Colab with the more limited computing environment of the free version. If these processing time values were linearly extrapolated, processing 1000 articles would take about 20 hours of computation, which would force you to work in blocks or expand the capacity of the Google Colab computing environment with one of its subscription options, or run the script on a local machine with python installed. However, it remains a task for future research to perform a load test and check if the processing times (and RAM resources) grow exponentially with the size of the corpus or not.

To identify the practical usefulness of our proposal, we should compare it with the usual manual procedures. As the current Gold Standard, we can probably consider the screening of results starting from the list ordered by relevance as proposed by WOS or SCOPUS in their search engines. In future research, we could compare the ranking proposed by search platforms with our ranking and human coding. The other alternative is to sort randomly (or alphabetically by the first author) to avoid uncontrolled bias in the screening process. However, this strategy encourages classification errors due to saturation or fatigue in the records at the end of the list. Manually screening 15 articles, selecting them by title and abstract, and checking if they fit the systematic review topic would cost about 15 minutes for each coder. The appropriate thing would be to use at least two encoders and have a third person available to resolve discrepancies (Alfalla-Luque et al., 2023; Losilla et al., 2018). After the independent coding of the two coders, a meeting between them would be necessary to agree on the results. Reviewing results from 15 articles can take about 3 minutes (depending on the degree of agreement between coders). The meeting with the third person coder to resolve disagreements can consume around 1 minute for every 15 articles processed (most do not generate disagreements). The manual processing of 15 articles, accumulating the time of all the people involved, would consume 30 minutes during independent coding, 6 minutes in the consensus meeting and 3 minutes in the disagreement resolution meeting. In total, there were about 40 minutes of research staff. It remains to be seen whether this figure is significantly reduced when the process is co-piloted with our procedure. That is, if having the list automatically sorted by similarity of embeddings, reduces the human classification effort and/or if it improves some of the classification efficiency metrics.

(cc) BY

Moreover, this protocol could be adapted to assess the usefulness of artificial intelligence to achieve effectiveness and efficiency in other steps of the systematic review. In particular, to the phase of extracting information from the selected documents after the screening phase. In this sense, the grounded constructivist theory approach proposed by various authors (Gioia, 2021; Odacioglu et al., 2023) aligns with the capabilities of Sentence Transformers models and the application of embeddings, as we have done in this work.

From the previous literature review conducted in this article, one of the most extensively researched Natural Language Processing classification techniques is Sentiment Analysis. Topic analysis also involves a significant amount of research, although not as much as sentiment analysis. However, the Relevance Classification seems to have the least amount of scientific literature, which could indicate a possible research niche, which our research aims to help solve as much as possible. However, this lower presence could also mean less practical utility than the other techniques. The interest that this article may arouse and its traction in citations will help us elucidate whether we are in one case or the other.

Spanish version of the article

En la versión castellana del articulo se ha optado por no traducir del inglés los términos acuñados que se usan habitualmente en ingles y no suelen traducirse al castellano en la comunicación habitual. Consideramos que esto evitará la ambigüedad de a qué termino nos estamos refiriendo cuando usamos una versión castellana que no estamos seguros de que sea una terminología consolidada (o al menos no tan consolidada como la inglesa).

Introducción

Una revisión sistemática de literatura es un modo riguroso y estructurado para localizar, seleccionar, analizar y sintetizar los resultados de la investigación existente relacionada con una pregunta de investigación concreta (Page et al., 2021; Saunders et al., 2016). El proceso de revisión debe ser transparente y replicable, esto implica completar los siguientes pasos (Booth et al., 2019; Marin-Garcia, 2021; Medina-López et al., 2010; Ogbonnaya & Brown, 2023; Page et al., 2021; Saunders et al., 2016; Tranfield et al., 2003):

1. Elaboración de un protocolo que contenga, al menos:
    a. Identificación de la necesidad de la revisión
    b. Propuesta de una/s pregunta/s de investigación explícita/s (en la medida de lo posible indicando la población a la que se quiere generalizar la respuesta, o que la necesita o va a usarla; la intervención, práctica, servicio o acción sobre la que se quiere decidir; las condiciones en las que se realiza la intervención/servicio/acción/práctica (lugar, quien la realiza, etc.); los grupos contra los que se va a comparar, en caso de haberlos; los resultados (outcomes) que interesan para la decisión o lo que se va a evaluar; para qué vas a ser necesario responder a esta pregunta)
    c. Los criterios específicos de inclusión y exclusión de los estudios a analizar en la revisión

(cc) BY

2. Desarrollar una estrategia de búsqueda que genere los menos falsos positivos y falsos negativos posibles
3. Evaluar la adecuación de los estudios encontrados (cribado/screening).
4. Evaluar la calidad de los estudios cribados, o la calidad de su reporte
5. Extracción y codificación de datos siguiendo un protocolo
6. Analizar y sintetizar los hallazgos de manera sistemática (siguiendo un protocolo)
7. Escribir el reporte científico de la revisión
8. Guías para llevar la evidencia a la práctica

El objetivo de este proceso es minimizar sesgos en el proceso de revisión y proporcionar una visión integral del estado actual del conocimiento, identificando áreas de investigación futura. Este enfoque permite llegar a conclusiones fundamentadas sobre lo que se conoce y lo que aún se desconoce en relación con la pregunta de investigación, facilitando así el avance del conocimiento en el campo de estudio (Cooper, 1991; Saunders et al., 2016).

Para nosotros, el concepto de revisión sistemática de literatura tiene que ver con el proceso con el que se construye el banco de documentos a procesar y no con la forma de resolver o analizar los documentos. En este sentido, una revisión sistemática podría resolverse cualitativamente con un análisis de contenido (content analysis), que puede resolverse de forma narrativa (Aguinis et al., 2020; Friese et al., 2018; Krippendorff, 2018), o cuantitativamente, con un estudio bibliométrico o un meta-análisis (Van Rhee et al., 2018) (Borenstein et al., 2009). La revisión sistemática de literatura ha sido empleada en los diferentes campos científicos para el avance del conocimiento (e.g. Alfalla-Luque et al., 2013 and 2023; Bayonne et al., 2020; Marin-Garcia et al, 2018; Sánchez-Cazorla et al., 2016).

Aunque la inteligencia artificial puede apoyar cualquiera de los pasos comentados anteriormente. En este protocolo nos vamos a centrar, exclusivamente, en la fase de cribado (screening) de los documentos que se obtienen como resultado de aplicar la estrategia de búsqueda automática. Nuestro objetivo es proponer una metodología de trabajo, con inteligencia artificial, que permita mejorar la eficacia y eficiencia de la fase de screening de una revisión sistemática de literatura respecto a la que se hace solamente con humanos, especialmente cuando el número de documentos a procesar es elevado y sin drenar la capacidad futura de los investigadores participantes (Ma & Su, 2024). Para ello desarrollaremos un código python ejecutable y comprobaremos si el ranking de afinidad obtenido a través de vectorización representa la asociación entre los textos y el tema objetivo de la revisión sistemática.

La fase de screening en una revisión sistemática de literatura nos permite garantizar la pertinencia de los estudios incluidos en el análisis final. Esta etapa implica la evaluación sistemática de los títulos, resúmenes y, en ocasiones, textos completos de los artículos identificados en la búsqueda inicial, aplicando criterios de inclusión y exclusión predefinidos (Higgins et al., 2021; Marin-Garcia, 2021; Medina-López et al., 2010; Moher et al., 2016). Además, con el volumen de producción científica existente es previsible que los resultados de las búsquedas automáticas arrojen grandes cantidades de documentos. Esto puede comprometer la viabilidad del proyecto de revisión, o la consistencia de los resultados, si no se maneja

(cc) BY

eficientemente la fase de screening (Borah et al., 2017). Dado que este proceso puede ser extremadamente laborioso y propenso a errores si se realiza la tarea manualmente con codificadores humanos, nos parece importante desarrollar y utilizar herramientas automatizadas para apoyar esta fase (O'Mara-Eves et al., 2015).

No obstante, tal como comentaremos en las conclusiones de este trabajo, las posibilidades no se cierran con la fase de screening y el uso de la inteligencia artificial puede aplicarse a otras etapas del proceso de revisión sistemática de literatura, aunque su potencial tiene que contrastarse en investigación futura.

Investigación previa

El Natural Language Processing es un campo de la inteligencia artificial que se enfoca en la interacción entre los ordenadores y el lenguaje humano. Su objetivo principal es permitir que las máquinas entiendan, interpreten y generen lenguaje humano de manera natural y efectiva (Goldberg, 2022; Rayhan, 2024; Speer et al., 2024). Utiliza algoritmos y modelos estadísticos para analizar y procesar grandes cantidades de datos lingüísticos. Además, el aprendizaje automático proporciona las técnicas y herramientas que permiten a los sistemas de Natural Language Processing mejorar su rendimiento a partir de la experiencia o iteraciones de la tarea (Do et al., 2024; Patil & Gudivada, 2024).

Los Transformers, por su parte, son un modelo de aprendizaje profundo diseñado especialmente para tareas de procesamiento del lenguaje natural con una arquitectura basada en mecanismos de atención (Vaswani et al., 2017), que le permite capturar relaciones de contexto amplio en los datos de entrada. Los Transformers han demostrado ser excepcionalmente efectivos en diversas tareas de procesamiento de lenguaje, como la traducción automática, el resumen de textos y la generación de lenguaje (Binh et al., 2023; Devika et al., 2021; Troxler & Schelldorfer, 2024; Young et al., 2018).

Dentro de las aplicaciones/tareas de clasificación de Natural Language Processing, una búsqueda rápida en Web Of Science (WOS) (Figure 1) nos hace sospechar que, de lejos, "sentiment analysis" es una de las aplicaciones más representada en las publicaciones científicas. A mucha distancia parece estar la popularidad de "topic modelling". Sin embargo, apenas hay publicaciones sobre "relevance classification".

El objetivo del "Sentiment Analysis" es identificar y extraer información subjetiva (opiniones, sentimientos, actitudes o emociones) expresadas en un texto. La salida buscada es una clasificación de sentimientos (positivo, negativo, neutral) o unos índices de polaridad ente los extremos (positivo/negativo) (Liu, 2012). El nivel de análisis se centra en palabras o frases.

Por otra parte, el objetivo del "Topic Analysis" es identificar los temas principales o categorías temáticas presentes en una colección de documentos, basándose en la frecuencia en la que las palabras o patrones de palabras aparecen juntos. Siendo la salida esperada un conjunto de temas, cada uno integrado por palabras clave o distribuciones de palabras (Asmussen & Moller, 2019; Blei et al., 2001; Bruno et al., 2022; Odacioglu et al., 2023). El nivel de análisis en este caso, suele ser la colección de documentos.

Operativamente, podemos definir la fase de screening en una revisión de literatura como una situación donde, por una parte, tenemos una definición o descripción del tema que es objeto de nuestro interés y, por

(cc) BY

otra parte, tenemos unos objetos a clasificar, definidos por título y resumen. Nuestra decisión se basa en la proximidad de los objetos a clasificar respecto al tema perseguido.

En este sentido, la fase de screening, encaja mejor en la categoría de "Relevance Classification". Esta categoría se centra en determinar la relevancia (similitud semántica) de un documento (en este caso, un artículo científico) con respecto a una consulta o criterio específico (en nuestro caso, el objetivo de la revisión sistemática). Las métricas de interés en este caso es una clasificación binaria (incluido/excluido) o una puntuación de relevancia (o similitud) con el tema a tratar (Manning et al., 2008). Esta tarea suele precisar la comparación del contenido semántico de los objetos a clasificar con el del tema contra el que se evalúa la relevancia. No se trata solo de la similitud entre las palabras, sino que implica considerar el significado profundo del texto. El nivel de análisis es el documento o su presentación a través del título y el resumen (Mikolov, Sutskever, et al., 2013; O'Mara-Eves et al., 2015; Sebastiani, 2002).

Estas tareas de clasificación de textos y en especial las de "relevance classification" pueden utilizar diferentes técnicas de Natural Language Processing (Hickman et al., 2024; Hickman et al., 2020; Patil & Gudivada, 2024; Rayhan, 2024):

- Bag of Words. Utilizan métodos basados en conteo (unigrams, bigrams, trigrams) y generan representaciones superficiales que no tienen en cuenta el contexto ni capturan significado semántico.
- Recurrent Neural Networks (RNNs). Son útiles para tareas de clasificación basadas en el contexto donde el orden de las palabras aporta o altera el significado. Tienen variantes como los Long Short-Term Memory Networks (LSTMs), adaptadas a situaciones donde hay que retener el contexto en un bloque largo de texto.
- Convolutional Neural Networks (CNNs). Capturan patrones locales en frases o n-grams. Son utiles para "sentiment analysis" y "topic modelling".
- Pretrained Language Models (PLMs). Son modelos de lenguaje que han sido entrenados con grandes corpus de texto para aprender representaciones/patrones generales del lenguaje. Por ejemplo, ELMo, ULMFiT, XLNet, T5 (Text-to-Text Transfer Transformer), GPT (Generative Pre-trained Transformer). Entre estos modelos, algunos son de arquitectura transformers, otros no. También algunos son, además, autorregresivos. En este bloque incluiríamos los "sentence transformers" que generan representaciones vectoriales (embeddings) que capturan el significado semántico y son sensibles al contexto. Por ejemplo, BERT (Bidirectional Encoder Representations from Transformers) o RoBERTa.

Teniendo en cuenta todo lo anterior, para nuestro caso de uso son relevantes tres conceptos. Por ello, vamos a ampliar brevemente su definición porque ayudará a poner de manifiesto por qué han formado parte de las características que buscaremos en solución elegida:

- Los "Word Embeddings" son representaciones vectoriales densas de palabras en un espacio multidimensional. En ese espacio, las palabras semánticamente similares se encuentran cercanas entre sí. Estas representaciones capturan relaciones semánticas y sintácticas entre palabras,

(cc) BY

permitiendo realizar operaciones matemáticas sobre el vocabulario que respetan los significados (Mikolov, Chen, et al., 2013).

- Los "Sentence Transformers" son modelos de lenguaje basados en la arquitectura Transformer, diseñados específicamente para generar embeddings de frases o párrafos completos. Estos modelos están entrenados para producir representaciones vectoriales de texto que capturan el significado semántico de oraciones enteras, permitiendo comparaciones de similitud semántica entre textos de longitud variable (Binh et al., 2023; Devika et al., 2021; Galli et al., 2024; Su et al., 2023). Esto es justo lo que necesitamos para la tarea de screening.
- "Cosine Similarity" es una medida de similitud entre dos objetos, donde cada uno de ellos está representado pon un vector n-dimensional y se calcula el coseno del ángulo entre ellos. Cuanto más pequeño sea el ángulo, más similares se consideran los vectores. En el contexto del procesamiento del lenguaje natural, se utiliza para comparar la similitud semántica entre embeddings de palabras, o de textos más largos (frases o párrafos). La similitud del coseno produce un valor entre -1 y 1, donde 1 indica vectores en la misma dirección (independientemente de la longitud), 0 indica ortogonalidad, y -1 indica vectores en direcciones opuestas. Esta medida es útil en nuestro caso de estudio porque es independiente de la magnitud (longitud) de los vectores y solo considera la orientación (Bugorski, 2014; Hanifi et al., 2022; Xia et al., 2015).

Contribución

En este trabajo vamos a comprobar la viabilidad técnica de ejecutar un prototipo funcional en ordenadores personales o en plataformas de Cloud Computing gratuitas, con el objetivo de que pueda ser utilizado por cualquier persona investigadora, independientemente del presupuesto del que disponga.

Nuestro objetivo es ofrecer una guía paso a paso y un código (script de python) que cualquier persona, incluso sin formación en programación, pueda ejecutar en un ordenador personal o una tableta, simplemente con una cuenta gratuita de Gmail, que le dé acceso a Google Colab. Este material estará disponible en GitHub (https://github.com/jamg-upv/LLMforSLRscreening) y se complementará con una colección de videos (https://media.upv.es/#/portal/channel/222a2680-53d2-11ef-a0f7-e57947840c32) que muestran cómo realizar cada una de las etapas, desde la construcción del dataset hasta la ejecución, descarga de los resultados e interpretación de los mismos.

En el contexto actual, donde la extensión del uso de la IA está generando un intenso debate que cuestiona sus beneficios y desafíos en diversos sectores, Budhawar et al. (2023) resaltan que no está claro si estas tecnologías ayudan a transformar el trabajo sustituyendo las tareas que no aportan valor, o si generan un desplazamiento, donde las personas se ocuparán sólo de las decisiones triviales o prácticamente irrelevantes. En este escenario de incertidumbre, nuestra investigación se propone reducir las barreras de entrada para el uso copilotado de modelos de lenguaje avanzados, contribuyendo así a una distribución más equitativa de estas herramientas en el ámbito académico de investigación, sin renunciar a la participación de personas tomando decisiones en el proceso.

La relevancia de esta investigación radica en abordar dos desafíos. Por un lado, la complejidad técnica y los conocimientos de programación requeridos para utilizar transformers, que pueden ser una barrera,

(cc) BY

especialmente para las personas que investigan en el área de ciencias sociales. Por otro lado, la discriminación que podría generarse entre equipos de investigación con presupuestos para costearse recursos de computación de altas prestaciones, frente a grupos o personas sin acceso a fondos de investigación.

Al proporcionar herramientas accesibles y gratuitas, nuestra investigación busca nivelar el campo de juego, permitiendo que un grupo más amplio de investigadores/as pueda beneficiarse del uso de tecnologías avanzadas para hacer revisiones sistemáticas más eficaces y eficientes. Esto no solo democratiza el acceso a la tecnología de sentence transformers, sino que también puede contribuir a una comprensión más profunda y diversa de sus implicaciones en distintos campos de estudio, contribuyendo así al debate académico y práctico acerca del futuro de la inteligencia artificial y su impacto en la sociedad.

Método

El caso de uso que vamos a utilizar es la clasificación de un conjunto de artículos, caracterizados por su título y resumen obtenido de WOS, en base a su relevancia respecto a la definición teórica de un tema o concepto.

El resultado a obtener es un archivo Excel, que incluye hojas con (1) la comparación de modelos utilizados (tiempo empleado en proceso y tokens), (2) la tabla de similitud de coseno entre objetos y categoría/s de clasificación, (3) los embeddings calculados para cada texto (sea objetos o categorías) y (4) el ranking de artículos más cercanos a la categoría de interés como un promedio de la clasificación de rangos en cada uno de los modelos empleados.

El código de python se ha generado a través de diversas iteraciones con ayuda de un chatbot en la plataforma poe.com utilizando el modelo Claude-3.5-sonnet. En el momento de realizar este trabajo las capacidades de programación y de razonamiento científico de Claude-3.5-sonnet (de Anthropic) superaban las del mejor modelo disponible de openAi (ChatGPT4o). Hemos descartado hacer la clasificación de los artículos directamente con modelos de inteligencia artificial generativa. Hicimos pruebas utilizando la API de OpenAI para intentar obtener una clasificación fiable de cada articulo via prompt y el resultado fue totalmente inadecuado incluso realizando varias iteraciones de ingeniería de prompts y varias pruebas durante dos meses de trabajo. Por otra parte, aunque los resultados hubieran sido prometedores, siempre carecerían de transparencia y replicabilidad. Por ello decidimos abordar el proyecto desde un enfoque diferente calculando los embeddings y la distancia de coseno, que al menos es completamente replicable.

Para este trabajo hemos seleccionado 14 modelos de sentence transformers:

1. all-MiniLM-L6-v2
2. all-distilroberta-v1
3. all-mpnet-base-v2
4. paraphrase-multilingual-mpnet-base-v2
5. distiluse-base-multilingual-cased-v1
6. all-MiniLM-L12-v2
7. allenai-specter
8. allenai/scibert_scivocab_uncased

(cc) BY

9. distilbert-base-nli-mean-tokens
10. roberta-base-nli-stsb-mean-tokens
11. distiluse-base-multilingual-cased-v2
12. paraphrase-multilingual-MiniLM-L12-v2
13. stsb-roberta-large
14. bert-base-nli-mean-tokens.

En el momento de realizar esta investigación no disponemos de una justificación clara de porqué elegir estos modelos y no otros. Hemos seleccionado aquellos, sugeridos por el chatbot de inteligencia artificial generativa con Claude-3.5-sonnet (consultado en agosto 2024), que hemos podido comprobar que han sido usados en algún artículo científico (Galli et al., 2024; Dhini et al., 2023; Guesmi et al., 2023; Kim et al., 2024; Kulkarni et al., 2023; Kurek et al., 2024; Muñoz & Iglesias, 2023; Naing et al., 2024; Scarpino et al., 2022; Singh et al., 2021; Troxler & Schelldorfer, 2024; Wu et al., 2022).

Creemos que no hay un modelo de sentence transformer que sea un ganador claro para cualquier tarea de "Relevance Classification". El rendimiento/precisión de cada modelo varía según la tarea de clasificación a realizar y el conjunto de datos específicos disponible (la definición de las categorías y los resúmenes de los objetos a clasificar) (Perone et al., 2018; Wang et al., 2018). Por lo tanto, es necesario probar varios modelos y elegir el que mejor se adapte al caso de uso concreto. De momento, no hemos encontrado ninguna comparativa en artículos publicados en WOS o Scopus que analice si existe un modelo de sentence transformer universalmente superior para la tarea de screening en una revisión sistemática de literatura.

Las métricas de distancia aplicables directamente a embeddings incluyen la similitud de coseno (definida en la sección anterior de este artículo); la distancia euclidiana, que mide la distancia directa entre dos puntos en el espacio vectorial y es sensible a la magnitud de los vectores; el Coeficiente de correlación de Pearson; la Manhattan Distance, que suma las diferencias absolutas de las coordenadas de los vectores; y la Soft Cosine Similarity, una extensión de la similitud del coseno (Bugorski, 2014; Cha, 2007; Chandrasekaran & Mago, 2021; Elekes et al., 2017; Levy et al., 2015; Nursalman et al., 2018; Sidorov et al., 2014; Xia et al., 2015).

Otras métricas que quizás se podrían utilizar requieren transformaciones simples para ser aplicables a los embeddings. Entre ellas podríamos tener, la Distancia de Mahalanobis, basada en la correlación entre vectores y requiere el cálculo de la matriz de covarianza inversa; la Similitud de Jaccard, que mide la intersección dividida por la unión de los conjuntos de palabras de dos documentos, puede adaptarse para trabajar con versiones discretizadas o binarias de embeddings; la Distancia de Hamming, útil para comparar cadenas binarias o categóricas de igual longitud, puede aplicarse a versiones binarias de embeddings; la Similitud de Sørensen-Dice, que da más peso a las coincidencias que Jaccard, también puede adaptarse para su uso con embeddings discretizados.

Comparada con las otras alternativas, la similitud del coseno tiene la ventaja de ser invariante a la escala, pues solo considera la dirección de los vectores, no su magnitud. Esto es útil cuando se comparan textos de diferentes longitudes. Sin embargo, a diferencia de la Distancia Euclidiana o la Manhattan Distance, no captura información sobre la magnitud de las diferencias entre vectores. Por otra parte, la Distancia de Mahalanobis podría ser más robusta en casos donde las dimensiones de los embeddings están

(cc) BY

correlacionadas (Cha, 2007). En esta investigación vamos a elegir la similitud de coseno como distancia y dejamos para investigación futura comparar si alguna de las otras distancias podría ser más adecuada para representar la similitud entre artículos y objetivo de la revisión en la fase de screening.

Un primer objetivo de este protocolo es conseguir un código python ejecutable que ofrezca las tablas del libro Excel comentado anteriormente. Un segundo objetivo es comprobar si las similitudes de coseno y el ranking de afinidad extraído a través de la vectorización, son útiles y representan la verdadera asociación entre textos y el objetivo de la revisión.

Para ello, consideraremos los resultados de cada modelo de sentence transformers como si fuesen un codificador independiente. Tomaremos como "Gold Standard" la clasificación realizada por un codificador humano (en este caso, el consenso del equipo de autores de este artículo que actúan como codificadores humanos).

El primer autor realizará una inspección visual de los resultados de cada modelo (el orden en el que han quedado situados los trabajos con la distancia de coseno, comparado con el orden consensuado previamente por todos los autores), analizando también la Recall (sensitivity) (proporción de verdaderos positivos ordenados por encima de los negativos, respecto al total de positivos reales), Precision (proporción de verdaderos positivos ordenados por encima de los negativos, respecto al total de objetos clasificados hasta llegar a la posición del último de los verdaderos positivos) y Acuracy (total de verdaderos positivos y verdaderos negativos respecto al total de objetos clasificados) (O'Mara-Eves et al., 2015).

Resultados de la prueba piloto

Vamos a utilizar como prueba piloto una revisión sistemática de literatura sobre el concepto de High Involvement Work Programs o alguna de sus componentes (Santandreu Mascarell et al., 2024). Las categorías para clasificación serán una definición de High Involvemente Work Programs (Marin-Garcia, 2013), las 12 definiciones de cada una de las prácticas de recursos humanos que habitualmente han sido identificadas como prácticas de alta implicación (Marin-Garcia, 2013; Perello-Marin & Ribes-Giner, 2014) y una definición agregada de High Involvement Work Programs como la concatenación de la definición global y la de las 13 prácticas. Además, hemos incorporado como temas de clasificación la definición de otras categorías usadas como distractoras. En ellas incluimos tres prácticas de recursos humanos, que no son High Involvement Work Programs (Green HRM, Remote Work, Non Traditional Work Arrangements); la definición de Gestión de Operaciones (Marin-Garcia et al., 2021), la definición de Scholarship Of Teaching and Learning y la defincón de Higher Education Course Management. Que son los temas sobre los que tratan los 15 artículos seleccionados en el piloto. Cada categoría para clasificación tiene una descripción breve (de 150-200 palabras) aunque en algunos casos puede ser un poco más larga.

Como objetos para clasificar en esta prueba piloto hemos seleccionado 15 artículos y, el equipo de autores-as de este trabajo ha consensuado una clasificación de los mismos (Table 1). Los tres primeros (Hauff et al., 2022; Marin-Garcia & Martinez Tomas, 2016; Song et al., 2021) son claramente artículos que tratan de High Involvement Work Programs (1hiwp). El cuarto (Marin-Garcia & Martínez-Tomás, 2022) trata de sistemas de remuneración pero se centra en la estructura salarial de las personas contratadas en España y su evolución a lo largo del tiempo, por lo que es un artículo de HRM y no específico de alta implicación

(cc) BY

(2hrm2). (Ahmad et al., 2022; Houeland & Jordhus-Lier, 2022) son artículos de Green HRM. (Becker et al., 2022) es de Remote Work. (Burbano & Chiles, 2022) es de Non Tradictional Work Arrangements, aunque también puede considerarse de Remote Work. (Mora-Valentin et al., 2024) puede considerarse Course Management aunque por los temas de la asignatura podría encajar con Green HRM o en la categoría de Training (2hrmteach). (Rincon & Zorrilla, 2017; Rincón et al., 2023) es Course Management (3teach). (Marin-Garcia, Garcia-Sabater, Garcia-Sabater, et al., 2020) es Course Management pero al tratar de temas de equipos de mejora continua/kaizen puede ser vinculado a Operations Management o a Empowerment (4omteach). Las penúltimas filas (Aguilar-Escobar et al., 2024; Marin-Garcia, Garcia-Sabater, Ruiz, et al., 2020; Nguyen et al., 2024) son claramente de Operations Management (4om), aunque NgocNguyen2024 habla de sostenibilidad y podría crear confusión con Green HRM. La última fila (Aznar-Mas et al., 2023) es Course Management (5 teach), pero al estar evaluando el rendimiento de las asignaturas podría tener alguna confusión con Performance Management.

El código generado (accesible en GitHub: https://github.com/jamg-upv/LLMforSLRscreening) ha funcionado correctamente. La carga de las librerías necesarias dura unos 2 minutos y el proceso de cálculo de embeddings, con cada uno de los 14 modelos probados, para las 20 categorías y los 15 artículos, se realiza en unos 8 minutos. El tiempo total de procesado no supera los 11 minutos hasta obtener el fichero Excel deseado.

Hay modelos que se ejecutan en menos de 4 segundo y otros que tardan más de un minuto (Figure 2). En general los modelos de menor tamaño (384 ó 512tokens) son los más rápidos. Entre los modelos de 768 tokens, algunos resuelven muy rápido y otros no tanto. El único modelo de los probados con 1024 tokens se sitúa entre los que más tiempo de proceso necesitan. Hemos repetido la vectorización varias veces y los tiempos, aunque variando, se sitúan siempre en el mismo rango de valores, respetando el ranking de tiempo de proceso. Los modelos que son más rápidos (o más lentos), lo son en todas las repeticiones.

Por otra parte, hemos probado el tiempo de procesado con diferentes corpus de texto. Por un lado, haciendo los embeddings sólo de las categorías (unos 14000 caracteres) o de las categorías junto con los artículos a clasificar (unos 39500 caracteres). En los modelos rápidos, casi triplicar el número de caracteres a procesar, multiplica el tiempo de procesado por 1.5. En los modelos más lentos, triplica el tiempo de procesado. Estos datos deben interpretarse con cautela porque son fruto de una única repetición, con un único corpus de texto y no sabemos si son replicables si se repiten más veces, o se cambia el corpus a analizar. No obstante, pueden servir como una primera estimación del tiempo de procesado en función del tamaño del corpus a analizar.

La Figure 3, a la izquierda, muestra el resultado de promediar las posiciones, de cada modelo, de los rankings de mayor a menor similitud de coseno respecto de la categoría “HIWPshortDescrip” (dando el mismo peso a todos los modelos). En la parte de la derecha (en negro) se muestra la clasificación realizada por expertos humanos, donde se ha resaltado en verde los verdaderos positivos (artículos que tratan sobre High Involvement Work Programs). A continuación, se han situado los que tratan temas de RRHH y luego aparecen artículos de docencia o de dirección de operaciones. Nuestra clasificación humana está ordenada por bloques de categorías. Los etiquetados como 1hiwp son nuestros artículos objetivo. Los etiquetados como 2hrm son artículos que no encajan en nuestra revisión, pero que deberían ser más afines que los otros. Las etiquetas 3teach, 4om o 5teach, forman un bloque de artículos aún más distantes respecto al objetivo

(cc) BY

de nuestra revisión. Dentro de cada bloque no hay un orden, cualquiera de los artículos del bloque puede estar mejor posicionado o no en el ranking y eso no se considera error de clasificación. Lo que consideraremos como falso positivo es la intercalación de un artículo de alguno de los bloques no objetivo entre los articulo resaltados en verde. Del mismo modo, consideraremos un falso negativo si alguno de los artículos marcados como verde (bloque objetivo) queda sepultado en la clasificación, o es sobrepasado por algún artículo que no pertenece a la categoría 2hrm (que es una categoría afín en contenidos a nuestra categoría objetivo).

En la Table 2 mostramos las métricas de rendimiento de clasificación para el agregado de rankings para el ranking de cada uno de los modelos por separados cuando la categoría objetivo de clasificación es "HIWPshortDescrip". La solución que representa a los 14 modelos tiene unas métricas de clasificación aceptables, con una Acuracy y Recall elevadas, aunque la Precision no es perfecta. No obstante, podemos observar que el rendimiento de los modelos es bastante diferente. Hay modelos que clasifican a la perfección (acuerdo total con clasificación humana) y otros con rendimientos muy bajos (especialmente en Precision y Acuracy). Por ello, el ranking agregado se ve penalizado por la incorporación de resultados de los modelos que no funcionan muy bien en la tarea de clasificación. Este resultado es importante porque abre una línea de investigación futura para confirmar si son siempre los mismos modelos los que funcionan bien o hay que detectar para cada situación (para cada corpus a analizar) cuál es el que funciona mejor. En el caso de que los mejores modelos cambiaran con el corpus, deberíamos analizar si vale la pena la detección del mejor modelo o si sería menos costoso aceptar la solución del agregado de los 14 modelos, que es imperfecta, pero quizás suficientemente buena desde el punto de vista práctico.

Revisando los resultados de las similitudes de coseno de los diferentes modelos, podemos analizar con detalle, el rendimiento de los modelos frente a diferentes categorías de clasificación con el corpus de texto de este piloto (Figure 4). Como ya hemos comentado, algunos modelos clasifican mejor que otros. Tomando como ejemplo allenai-specter y allenai/scibert_scivocab_uncase, vemos que el primero clasifica mejor que el segundo. allenai-Specter identifica los tres verdaderos positivos como artículos de High Involvement Work Programs (azul más intenso en la columna de la categoría), clasifica bien Ahmad2022 y Houeland2022 como artículos de GreenHRM (su proximidad con esa categoría es superior a las otras). Becker2022 lo asocia claramente con Non Tradicional Work Arrangements y Work Life Balance. Los artículos de Mora2024, Rincon2023, Marin2020triplediam y Aznar2023, los identifica como artículos docentes. Aguilar2024, Ngoc2024 y Marin2020omcovid, los identifica como de Operations Management (en la fila el valor más alto está en esa categoría). Sin embargo, allenai/scibert_scivocab_uncased no es capaz de discriminar bien los artículos y prácticamente todos son similares, con alguna excepción, dentro de cada categoría. Las columnas tienen un color muy parecido en las 15 celdas, lo que indica que los artículos son muy parecidos para este modelo y no puede discriminar cual es más afín. Cuando la realidad es que hay grandes diferencias entre los artículos de bloques diferentes. Recordamos que no se pueden considerar que estos resultados sean concluyentes hasta que hayan sido adecuadamente replicados con otros corpus.

La diferencia de acierto de los modelos comparado con el evaluador humano se puede confirmar en la Figure 5, donde se muestra la clasificación por mayor a menor similitud de coseno entre los embeddings de cada artículo y el de la categoría "HIWPshortDescrip". El modelo de la pare superior clasifica

(cc) BY

adecuadamente los tres verdaderos positivos e identifica como más distantes el resto. Como detalle adicional, se confirma que los artículos vinculados a Recursos Humanos se consideran más cercanos a High Involvement Work Programs que los de Dirección de Operaciones o de docencia. Esto es un buen resultado. Sin embargo, el modelo de la parte inferior intercala falsos positivos entre verdaderos positivos (Ahmad2022 se considera más similar a High Involvement Work Programs que Hauff2022) y convierte en falso negativo a Marin2016, que queda sepultado en el ranking de similitud. Además, se comprueba que algunos artículos de docencia o Dirección de Operaciones son considerados más relevantes que los verdaderos positivos o que los artículos de HRM. Este sería un ejemplo de resultado poco deseable.

Conclusiones, limitaciones e investigación futura

La metodología que proponemos está diseñada para aumentar la eficiencia y eficacia del proceso de screening sin reemplazar el juicio humano. En todo el proceso, mantenemos al personal investigador en el centro de toma de decisiones y utilizamos la inteligencia artificial sólo como una herramienta de apoyo, fomentando una relación complementaria entre la tecnología y las personas codificadoras. De este modo se reducen los posibles riegos de reemplazo o de esclavitud frente a la IA del personal investigador participantes (Ma & Su, 2024).

El objetivo final es poder utilizar el ranking de similitud, generado a partir de los embeddings proporcionados por los sentence transformers, para optimizar el proceso de revisión sistemática de literatura. En lugar de revisar los artículos de forma aleatoria, las personas codificadoras pueden empezar por los más cercanos a la/s categorías de interés y avanzar progresivamente hacia los más lejanos, lo cual me permitiría ahorrar tiempo y esfuerzo, y reducir la fatiga y los sesgos de los codificadores/as.

Los resultados de la prueba piloto nos han permitido comprobar la viabilidad de la propuesta a pequeña escala. También hemos comprobado que algunos de los modelos clasifican de una manera prácticamente igual que las personas que han actuado como codificadoras expertas. Podemos concluir que los resultados del piloto generan un ranking adecuado cuando se eligen los modelos que clasifican bien.

La aplicación de este protocolo a nuevos conjuntos de datos (Santandreu Mascarell et al., 2024) y (Marin-Garcia & Martinez Tomas, 2016), procedentes de investigaciones donde disponemos de un volumen de artículos elevado, codificados por expertos/as humanos, nos permitirá comprobar la escalabilidad del procedimiento cuando pasamos de 15 artículos a cientos o miles. También nos permitirá calcular las métricas de rendimiento de clasificación comparada con humanos.

Otro aspecto a abordar en el futuro es si el enfoque propuesto en este trabajo se puede generalizar a cualquier tipo revisión sistemática o si el campo de investigación (por ejemplo, Dirección de Operaciones, o Recursos Humanos) puede condicionar su uso. El ejemplo con el que hemos trabajado es de Recursos humanos y habrá que probar si es generalizable a otros ejemplos de ese campo o, incluso, de otros campos científicos.

Además, este protocolo abre nuevas líneas de investigación que intentan superar algunas de las limitaciones o preguntas que nos han surgido durante la escritura de este protocolo.

Relacionado con los modelos de sentence transformers para tareas de screening, nos surgen varias preguntas para investigación futura:

(cc) BY

1. Selección de modelos: ¿cómo elegir los modelos adecuados? ¿Existen otros modelos relevantes además de los 14 incluidos en este protocolo?
2. Rendimiento comparativo de modelos: ¿podemos identificar modelos que consistentemente superen a otros en tareas de clasificación para screening?
3. Generalización del rendimiento: ¿se mantiene la superioridad de ciertos modelos al cambiar el tema de estudio, la precisión de las definiciones de categorías, o al utilizar un nuevo corpus de resúmenes para el análisis?
4. Mejora de definiciones: ¿podrían mejorar los resultados si, en lugar de usar definiciones de categorías provenientes de investigaciones previas publicadas, las refinamos con la ayuda de inteligencia artificial generativa?
5. Comparar el resultado con modelos como Llama3.1 (a través, por ejemplo de la plataforma HuggingFace) que podrían utilizar su amplia ventana de contexto para realizar la clasificación con el contenido de texto completo y no solo con título y resumen
6. Analizar las diferencias de resultados entre usar modelos pre-entrenados genéricos (como en este artículo), respecto a modelos con un fine-tuning específico para la tarea de clasificación (campo científico y temática concreta de la revisión sistemática)
7. Limitaciones de los sentence transformers para capturar significados, especialmente en textos complejos con significados dependientes del contexto, que es una situación habitual en los textos científicos del área de gestion de empresas

Respecto a las medidas de proximidad, la investigación futura podría comparar los resultados que se obtienen al emplear la distancia euclídea, u otras medidas de distancia entre embeddings, en lugar de cosine simmilarity. Este análisis se podría cruzar con el análisis anterior de modelos pues quizás los resultados de algunos modelos son más sensibles al cambio de medida de proximidad que otros.

También se podría comparar el efecto de usar definiciones compactas (como nuestra HIWPshortDescrip) respecto a usar una concatenación de las definiciones compactas de cada una de las componentes de un concepto multidimensional (como nuestra HIWPLongDescripshort).

Otra posible investigación para desarrollar en el futuro, basada en los datos de este protocolo, sería el análisis estadístico de la consistencia entre "codificadores". Para ello se podría tener en cuenta la varianza entre raters (between) e intra rater (within) (Marin-Garcia et al., 2008) u otras medidas de rendimiento de clasificación (O'Mara-Eves et al., 2015). También se podría analizar cuál es el número adecuado de modelos a ejecutar e integrar para lograr una solución más consistente. Una de las cuestiones a resolver sería si es mejor el ranking que ofrece un modelo o el agregado de varios modelos (aunque no sean los modelos que mejor clasifiquen, su ranking conjunto igual supera al ranking del modelo con mejor rendimiento individual) y, en este último caso, cuántos modelos agregar para lograr mejor consistencia en clasificaciones de textos diferentes a los de entrenamiento.

(cc) BY

Hemos comprobado que procesar 39000 caracteres (títulos más resumen de 15 artículo) con los 7 modelos que mejores resultados dan, ocupa unos 3 minutos de ejecución del script en Google Colab con el entorno de computación más limitado de la versión gratuita. Si estos valores de tiempo de procesado fuesen extrapolables linealmente, procesar 1000 artículos llevaría unas 20 horas de computo, lo que obligaría a trabajar por bloques o ampliar la capacidad del entorno de computación de Google Colab con una de sus opciones de suscripción, o pasar a ejecutar el script en una máquina local con python instalado. No obstante, queda como tarea para investigación futura el realizar una prueba de carga y comprobar si los tiempos de procesamiento (y de recurso RAM) crecen exponencialmente con el tamaño del corpus o no.

Para identificar la utilidad práctica de nuestra propuesta, deberíamos compararla con los procedimientos manuales habituales. Probablemente podemos considerar como Gold Standard actual, el hacer el screening de resultados partiendo de la lista ordenada por relevancia tal como la propone WOS o SCOPUS en sus buscadores. En investigación futura podríamos comparar el ranking propuesto por las plataformas de búsqueda con nuestro ranking y la codificación humana. La otra alternativa es ordenar aleatoriamente (o alfabéticamente por primer autor-a) para evitar sesgos no controlados en el proceso de screening. Aunque esta estrategia fomenta los errores de clasificación por saturación o cansancio en los registros del final de la lista. Hacer el screening manual de 15 artículos, seleccionando por título y resumen comprobando si encajan con el tema de una revisión sistemática, costaría alrededor de 15 minutos para cada persona codificadora. Lo adecuado sería usar mínimo 2 codificadoras y tener una tercera persona disponible para resolver discordancias (Alfalla-Luque et al., 2023; Losilla et al., 2018). Tras la codificación independiente de las dos codificadoras sería necesaria una reunión entre ellas para consensuar resultados. Repasar resultados de 15 artículos puede consumir unos 3 minutos (dependiendo del grado de acuerdo entre codificadoras). La reunión con la tercera persona codificadora para resolver desacuerdos puede consumir en torno a 1 minuto por cada 15 artículos procesados (la mayoría no genera desacuerdos). El procesado manual de 15 artículos, acumulando el tiempo de todas las personas implicadas, consumiría 30 minutos durante la codificación independiente, 6 minutos en la reunión de consenso y 3 minutos en la de resolución de desacuerdos. En total unos 40 minutos de personal investigador. Queda pendiente comprobar si esta cifra se reduce significativamente cuando el proceso es copilotado con nuestro procedimiento. Es decir, si el disponer de la lista ordenada automáticamente por similitud de embeddings, reduce el esfuerzo de clasificación humana y/o si mejora la algunas de las métricas de eficiencia de clasificación.

Por otra parte, este protocolo podría adaptarse para evaluar la utilidad de la inteligencia artificial para conseguir eficacia y eficiencia también en otros pasos de la revisión sistemática. En especial, a la fase de extracción de información de los documentos seleccionados tras la fase de screening. En este sentido, el enfoque de grounded theory constructivista propuesto por diversos autores (Gioia, 2021; Odacioglu et al., 2023) está bastante alineado con las capacidades de los modelos de sentence transformers y la aplicación de embeddings como hemos realizado en este trabajo.

De la revisión de literatura previa realizada en este artículo, una de las técnicas de clasificación de Natural Language Processing que ha sido investigada más extensamente es el Sentiment Analysis. El Topic Analysis, también cuenta con una cantidad importante de investigaciones, aunque no tanta como el Sentiment Analysis. Sin embargo, la Relevance Classification parece tener la menor cantidad de literatura científica, lo que podría indicar un posible nicho de investigación, que nuestra investigación pretende

(cc) BY

ayudar a resolver en la medida de lo posible. No obstante, somos conscientes de que esta menor presencia también podría significar una menor utilidad práctica comparada con las otras técnicas. El interés que pueda despertar este artículo y la tracción en citas que genere nos ayudará a dilucidar si estamos en un caso o en otro.

Acknowledgments

No funding has been received for the development of the research.

Conflict of interests

No conflict of interest

Author Contributions

Conceptualization; Data curation (JAMG); Formal analysis (JAMG); Research (JAMG, JMT, CSM, AJT); Methodology (JAMG); Project administration (JAMG); Resources (JAMG, JMT, CSM, AJT); Software (JAMG); Supervision (JAMG); Validation (JMT, CSM, AJT); Visualization (JAMG, JMT, CSM, AJT); Writing - original draft (JAMG); Writing - review & editing (JMT, CSM, AJT).

References

Aguilar-Escobar, V. G., Garrido-Vega, P., Vázquez-Rivas, P. d. V., Monzón-Moreno, A. (2024). Factors influencing nurse satisfaction with Automated Medication Dispensing Cabinets. *WPOM-Working Papers on Operations Management*, *15*, 57-74. https://doi.org/10.4995/wpom.18935

Aguinis, H., Ramani, R. S., Alabduljader, N. (2020). Best-Practice Recommendations for Producers, Evaluators, and Users of Methodological Literature Reviews. *Organizational Research Methods*. https://doi.org/10.1177/1094428120943281

Ahmad, I., Ullah, K., Khan, A. (2022). The impact of green HRM on green creativity: mediating role of pro-environmental behaviors and moderating role of ethical leadership style [Article]. *International Journal of Human Resource Management*, *33*(19), 3789-3821. https://doi.org/10.1080/09585192.2021.1931938

Alfalla-Luque, R., Luján García, D. E., Marin-Garcia, J. A. (2023). Supply chain agility and performance: evidence from a meta-analysis. *International Journal of Operations & Production Management*, *43*(10), 1587-1633. https://doi.org/10.1108/ijopm-05-2022-0316

Alfalla-Luque, R., Medina-Lopez, C., Dey, P.K. (2013). "Supply chain integration framework using literature review". Production Planning and Control, 24 (8–9): 800–817.

Asmussen, C. B., Moller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review [Review]. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0255-7

Aznar-Mas, L. E., Atarés Huerta, L., Marin-Garcia, J. A. (2023). Effectiveness of the use of open-ended questions in student evaluation of teaching in an engineering degree [Teaching evaluation, higher education, student satisfaction, teaching improvement, open-ended questions]. *Journal of Industrial Engineering and Management*, *16*(3), 14. https://doi.org/10.3926/jiem.5620

(cc) BY

Bayonne, E., Marin-Garcia, J.A., Alfalla-Luque, R. (2020). "Partial least squares (PLS) in Operations Management research: Insights from a systematic literature review", Journal of Industrial Engineering and Management, 13(3), pp. 565-597. https://doi.org/10.3926/jiem.3416

Becker, W. J., Belkin, L. Y., Tuskey, S. E., Conroy, S. A. (2022). Surviving remotely: How job control and loneliness during a forced shift to remote work impacted employee work behaviors and well-being [Article]. *Human resource management*, *61*(4), 449-464. https://doi.org/10.1002/hrm.22102

Binh, D., Tung, L., Le-Minh, N. (2023). SubTST: a consolidation of sub-word latent topics and sentence transformer in semantic representation [Article]. *Applied Intelligence*, *53*(11), 13470-13487. https://doi.org/10.1007/s10489-022-04184-x

Blei, D., Ng, A., Jordan, M. (2001). *Latent Dirichlet Allocation* (Vol. 3).

Booth, A., Noyes, J., Flemming, K., Moore, G., Tunçalp, Ö., Shakibazadeh, E. (2019). Formulating questions to explore complex interventions within qualitative evidence synthesis. *BMJ Global Health*, *4*(Suppl 1), e001107. https://doi.org/10.1136/bmjgh-2018-001107

Borah, R., Brown, A. W., Capers, P. L., Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *Bmj Open*, *7*(2), e012545. https://doi.org/10.1136/bmjopen-2016-012545

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons.

Bruno, M., Catanese, E., De Cubellis, M., Fausti, F., Pugliese, F., Scannapieco, M., Valentino, L. (2022). Analyzing textual data through Word Embedding: experiences in Istat.

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., Rogelberg, S., Saunders, M. N. K., Tung, R. L., Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, *33*(3), 606-659. https://doi.org/https://doi.org/10.1111/1748-8583.12524

Bugorski, A. T. (2014). *Cosine Similarity for Article Section Classification: Using Structured Abstracts as a Proxy for an Annotated Corpus.* Electronic Thesis and Dissertation Repository. 2154. https://ir.lib.uwo.ca/etd/2154

Burbano, V. C., Chiles, B. (2022). Mitigating Gig and Remote Worker Misconduct: Evidence from a Real Effort Experiment [Article]. *Organization Science*, *33*(4), 1273-1299. https://doi.org/10.1287/orsc.2021.1488

Cha, S.-H. (2007). Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. *Int. J. Math. Model. Meth. Appl. Sci.*, *1*.

Chandrasekaran, D., Mago, V. (2021). Evolution of Semantic Similarity—A Survey. *ACM Comput. Surv.*, *54*(2), Article 41. https://doi.org/10.1145/3440755

Cooper, H. M. (1991). *Integrating Research. A guide for Literature Reviews* (Vol. 2). Sage Publications, Inc.

Devika, R., Vairavasundaram, S., Mahenthar, C. S. J., Varadarajan, V., Kotecha, K. (2021). A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data [Article]. *IEEE Access*, *9*, 165252-165261. https://doi.org/10.1109/access.2021.3133651

Dhini, B. F., Girsang, A. S., Sufandi, U. U., Kurniawati, H. (2023). Automatic essay scoring for discussion forum in online learning based on semantic and keyword similarities. *Asian Association of Open Universities Journal*, *18*(3), 262-278. https://doi.org/https://doi.org/10.1108/AAOUJ-02-2023-0027

(cc) BY

Do, S., Ollion, É., Shen, R. (2024). The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research*, *53*(3), 1167-1200. https://doi.org/10.1177/00491241221134526

Elekes, A., Schaeler, M., Boehm, K. (2017, 19-23 June 2017). On the Various Semantics of Similarity in Word Embedding Models. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL),

Friese, S., Soratto, J., Pires, D. (2018). Carrying out a computer-aided thematic content analysis with ATLAS.ti. *MMG Working Papers Print*(WP 18-02), 1-28. http://www.mmg.mpg.de/publications/working-papers/2018/wp-18-02/?utm_source=CleverReach&utm_medium=email&utm_campaign=16-05-2018+Newsletter+2018-03++-++May&utm_content=Mailing_12459922

Galli, C., Donos, N., Calciolari, E. (2024). Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis [Review]. *Information*, *15*(2). https://doi.org/10.3390/info15020068

Gioia, D. (2021). A Systematic Methodology for Doing Qualitative Research. *The Journal of Applied Behavioral Science*, *57*(1), 20-29. https://doi.org/10.1177/0021886320982715

Goldberg, Y. (2022). *Neural Network Methods for Natural Language Processing.* Springer Cham.

Guesmi, M., Chatti, M. A., Kadhim, L., Shoeb, J., Qurat Ul, A. (2023). Semantic Interest Modeling and Content-Based Scientific Publication Recommendation Using Word Embeddings and Sentence Encoders. *Multimodal Technologies and Interaction*, *7*(9), 91. https://doi.org/https://doi.org/10.3390/mti7090091

Hanifi, M., Chibane, H., Houssin, R., Cavallucci, D. (2022). Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers [Article]. *Engineering Applications of Artificial Intelligence*, *109*. https://doi.org/10.1016/j.engappai.2022.104661

Hauff, S., Felfe, J., Klug, K. (2022). High-performance work practices, employee well-being, and supportive leadership: spillover mechanisms and boundary conditions between HRM and leadership behavior [Article]. *International Journal of Human Resource Management*, *33*(10), 2109-2137. https://doi.org/10.1080/09585192.2020.1841819

Hickman, L., Liff, J., Rottman, C., Calderwood, C. (2024). The Effects of the Training Sample Size, Ground Truth Reliability, and NLP Method on Language-Based Automatic Interview Scores' Psychometric Properties. *Organizational Research Methods*, *0*(0), 10944281241264027. https://doi.org/10.1177/10944281241264027

Hickman, L., Thapa, S., Tay, L., Cao, M., Srinivasan, P. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114-146. https://doi.org/10.1177/1094428120971683

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., Welch, V. (2021). *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021).* Cochrane. Available from www.training.cochrane.org/handbook.

Houeland, C., Jordhus-Lier, D. (2022). 'Not my task': Role perceptions in a green transition among shop stewards in the Norwegian petroleum industry [Article]. *Journal of Industrial Relations*, *64*(4), 522-543. https://doi.org/10.1177/00221856211068500

Kim, K., Kogler, D. F., Maliphol, S. (2024). Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic. *Humanities & Social Sciences Communications*, *11*(1), 603. https://doi.org/https://doi.org/10.1057/s41599-024-03044-y

Krippendorff, K. (2018). *Content analysis : an introduction to its methodology.* SAGE publications Inc.

(cc) BY

Kulkarni, A., Terpenny, J., Prabhu, V. (2023). Leveraging Active Learning for Failure Mode Acquisition. *Sensors*, *23*(5), 2818. https://doi.org/https://doi.org/10.3390/s23052818

Kurek, J., Latkowski, T., Bukowski, M., Świderski, B., Łępicki, M., Baranik, G., Nowak, B., Zakowicz, R., Dobrakowski, Ł. (2024). Zero-Shot Recommendation AI Models for Efficient Job–Candidate Matching in Recruitment Process. *Applied Sciences*, *14*(6), 2601. https://doi.org/https://doi.org/10.3390/app14062601

Levy, O., Goldberg, Y., Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, *3*, 211-225.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining* (Vol. 5). https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Losilla, J.-M., Oliveras, I., Marin-Garcia, J. A., Vives, J. (2018). Three risk of bias tools lead to opposite conclusions in observational research synthesis. *Journal of Clinical Epidemiology*(101), 61-72. https://doi.org/10.1016/j.jclinepi.2018.05.021

Ma, H., Su, M. (2024). Artificial stupidity and coping strategies. *Organizational Dynamics*, 101059. https://doi.org/https://doi.org/10.1016/j.orgdyn.2024.101059

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Marin-Garcia, J. A. (2013). What do we know about the relationship between High Involvement Work Practices and Performance? *WPOM-Working Papers on Operations Management*, *4*(2), 01-15. https://doi.org/http://dx.doi.org/10.4995/wpom.v4i2.1552

Marin-Garcia, J. A. (2021). Publishing in three stages to support evidence based management practice. *WPOM-Working Papers on Operations Management*, *12*(2), 56-95. https://doi.org/10.4995/wpom.11755

Marin-Garcia, J.A. Alfalla-Luque, R., Machuca, J.A.D. (2018). "A Triple-A supply chain measurement model: validation and analysis", International Journal of Physical Distribution & Logistics Management. Vol. 48. No. 10, pp. 976-994. https://doi.org/10.1108/IJPDLM-06-2018-0233

Marin-Garcia, J. A., Garcia-Sabater, J. J., Garcia-Sabater, J. P., Maheut, J. (2020). Protocol: Triple Diamond method for problem solving and design thinking. Rubric validation. *WPOM-Working Papers on Operations Management*, *11*(2), 49-68. https://doi.org/10.4995/wpom.v11i2.14776

Marin-Garcia, J. A., Garcia-Sabater, J. P., Ruiz, A., Maheut, J., Garcia-Sabater, J. J. (2020). Operations Management at the service of health care management: Example of a proposal for action research to plan and schedule health resources in scenarios derived from the COVID-19 outbreak. *Journal of Industrial Engineering and Management*, *13*(2). https://doi.org/10.3926/jiem.3190

Marin-Garcia, J. A., Martínez-Tomás, J. (2022). What does the wage structure depend on? Evidence from the national salary survey in Spain. *WPOM-Working Papers on Operations Management*, *13*(1), 35-63. https://doi.org/10.4995/wpom.16808

Marin-Garcia, J. A., Martinez Tomas, J. (2016). Deconstructing AMO framework: a systematic review. *Intangible Capital*, *12*(4), 1040-1087. https://doi.org/http://dx.doi.org/10.3926/ic.838

Marin-Garcia, J. A., Miralles Insa, C., Marin Garcia, P. (2008). Oral Presentation and Assessment Skills in Engineering Education. *International Journal of Engineering Education*, *24*(5), 926-935. http://www.upv.es/i.grup/repositorio/own/MarinEtAl2008_IEMA_peerassessmentIJEE.pdf

Marin-Garcia, J. A., Vidal-Carreras, I., Maheut, J. (2021). A keyword taxonomy proposal for Operations Management. XII Workshop in Operations Management and Technology (ACEDEDOT OMTech 2021), Granada.

(cc) BY

Medina-López, C., Marin-Garcia, J. A., Alfalla-Luque, R. (2010). Una propuesta metodológica para la realización de búsquedas sistemáticas de bibliografía (A methodological proposal for the systematic literature review). *WPOM-Working Papers on Operations Management*, *1*(2), 13-30. https://doi.org/http://dx.doi.org/10.4995/wpom.v1i2.786

Mikolov, T., Chen, K., Corrado, G. s., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, *2013*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. s., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Estarli, M., Barrera, E. S. A., Martínez-Rodríguez, R., Baladia, E., Agüero, S. D., Camacho, S., Buhring, K., Herrero-López, A., Gil-González, D. M., Altman, D. G., Booth, A., Chan, A. W., Chang, S., Clifford, T., Dickersin, K., Egger, M., Gøtzsche, P. C., Grimshaw, J. M., Groves, T., Helfand, M., Higgins, J., Lasserson, T., Lau, J., Lohr, K., McGowan, J., Mulrow, C., Norton, M., Page, M., Sampson, M., Schünemann, H., Simera, I., Summerskill, W., Tetzlaff, J., Trikalinos, T. A., Tovey, D., Turner, L., Whitlock, E. (2016). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement [Article]. *Revista Espanola de Nutricion Humana y Dietetica*, *20*(2), 148-160. https://doi.org/10.1186/2046-4053-4-1

Mora-Valentin, E.-M., Huertas-Valdivia, I., Garcia-Moreno, M.-B. (2024). Integrating the SDG into university teaching: an application in human resources subjects [Article]. *WPOM-Working Papers on Operations Management*, *15*(1), 1-15. https://doi.org/10.4995/wpom.19824

Muñoz, S., Iglesias, C. Á. (2023). Detection of the Severity Level of Depression Signs in Text Combining a Feature-Based Framework with Distributional Representations. *Applied Sciences*, *13*(21), 11695. https://doi.org/https://doi.org/10.3390/app132111695

Naing, I., Soe Thandar, A., Khaing Hsu, W., Funabiki, N. (2024). A Reference Paper Collection System Using Web Scraping. *Electronics*, *13*(14), 2700. https://doi.org/https://doi.org/10.3390/electronics13142700

Nguyen, T. N. H., Khuu, T. P. D., Nguyen, Q. H., Nguyen, M. C. (2024). Can sustainable supply chain strategies of company enhance for mitigation of risk damages and long-term resilience? An empirical analysis for the context of COVID-19 pandemic. *WPOM-Working Papers on Operations Management*, *15*, 112-131. https://doi.org/10.4995/wpom.21495

Nursalman, M., Kusnendar, J., Fadhila, U. F., Ieee. (2018). Implementation of K-Nearest Neighbor with Cosine Similarity for Classification Abstract International Journal of Computer Science.*International Conference on Information Technology Systems and Innovation (ICITSI)* [2018 international conference on information technology systems and innovation (icitsi)]. 5th International Conference on Information Technology Systems and Innovation (ICITSI), Inst Teknologi Bandung, Sch Elect Engn & Informat, INDONESIA.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, *4*(1), 5. https://doi.org/10.1186/2046-4053-4-5

Odacioglu, E. C., Zhang, L., Allmendinger, R., Shahgholian, A. (2023). Big textual data research for operations management: topic modelling with grounded theory. *International Journal of Operations & Production Management*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/IJOPM-03-2023-0239

Ogbonnaya, C., Brown, A. D. (2023). Editorial: Crafting review and essay articles for Human Relations. *Human relations*, *76*(3), 365-394. https://doi.org/10.1177/00187267221148440

(cc) BY

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Patil, R., Gudivada, A. (2024). *A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)*. https://doi.org/10.20944/preprints202402.0357.v1

Perello-Marin, M. R., Ribes-Giner, G. (2014). Identifying a guiding list of high involvement practices in human resource management. *WPOM-Working Papers on Operations Management*, *5*(1), 31-47. https://doi.org/10.4995/wpom.v5i1.1495

Perone, C., Silveira, R., Paula, T. (2018). *Evaluation of sentence embeddings in downstream and linguistic probing tasks*. https://doi.org/https://arxiv.org/abs/1806.06259

Rayhan, A. (2024). *Advancements in Natural Language Processing: A Comprehensive Review*. https://doi.org/10.13140/RG.2.2.13644.22400

Rincon, V., Zorrilla, P. (2017). Management Protocol: Entrepreneurship in the area of Marketing. Comparing PBL vs active lectures [Article]. *WPOM-Working Papers on Operations Management*, *8*(1), 1-8. https://doi.org/10.4995/wpom.v8i1.6470

Rincón, V., Zorrilla, P., Marin-Garcia, J. A. (2023). The impact of active learning on entrepreneurial capacity [Entrepreneurship, creativity, innovation, competences, marketing]. *Intangible Capital*, *19*(4), 16. https://doi.org/10.3926/ic.2297

Sanchez-Cazorla, A., Alfalla-Luque, R., Irimia-Diéguez, A. (2016): "Risk identification in Megaprojects as a crucial phase of Risk Management: a literature review", Project Management Journal. Vol. 47, No. 6, 75–93.

Santandreu Mascarell, C., Juarez-Tarraga, A., Marin-Garcia Juan, A. (2024). What do we need to research in the future on high involvement work practices? XIII International Workshop on HRM, Sevilla on September 19-20, 2024.

Saunders, M., Lewis, P., Thornhill, A. (2016). *Research methods for business students, 7/e*. Pearson Education.

Scarpino, I., Zucco, C., Vallelunga, R., Luzza, F., Cannataro, M. (2022). Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration. *BioTech*, *11*(3), 41. https://doi.org/https://doi.org/10.3390/biotech11030041

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas*, *18*(3), 491-504. https://doi.org/10.13053/cys-18-3-2043

Singh, I., Scarton, C., Bontcheva, K. (2021). Multistage BiCross encoder for multilingual access to COVID-19 health information. *PLoS ONE*, *16*(9). https://doi.org/https://doi.org/10.1371/journal.pone.0256874

Song, W., Yu, H., Qu, Q. (2021). High involvement work systems and organizational performance: the role of knowledge combination capability and interaction orientation [Article]. *International Journal of Human Resource Management*, *32*(7), 1566-1590. https://doi.org/10.1080/09585192.2018.1539863

(cc) BY

Speer, A. B., Perrotta, J., Kordsmeyer, T. L. (2024). Taking It Easy: Off-the-Shelf Versus Fine-Tuned Supervised Modeling of Performance Appraisal Text. *Organizational Research Methods*, *0*(0), 10944281241271249. https://doi.org/10.1177/10944281241271249

Su, Z., Wang, D., Miao, C., Cui, L. (2023). Multi-Aspect Explainable Inductive Relation Prediction by Sentence Transformer [preprint]. *ArXiv*. https://doi.org/arXiv:2301.01664

Tranfield, D., Denyer, D., Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*, 207--222. https://doi.org/10.1111/1467-8551.00375

Troxler, A., Schelldorfer, J. (2024). Actuarial applications of natural language processing using transformers: Case studies for using text features in an actuarial context. *British Actuarial Journal*, *29*. https://doi.org/https://doi.org/10.1017/S1357321724000023

Van Rhee, H. J., Suurmond, R., Hak, T. (2018). *User manual for Meta-Essentials: Workbooks for meta-analysis (Version 1.4)*. Erasmus Research Institute of Management. www.erim.eur.nl/research-support/meta-essentials

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing [Article]. *Journal of Biomedical Informatics*, *87*, 12-20. https://doi.org/10.1016/j.jbi.2018.09.008

Wu, L., Ali, S., Ali, H., Brock, T., Xu, J., Weida, T. (2022). NeuroCORD: A Language Model to Facilitate COVID-19-Associated Neurological Disorder Studies. *International Journal of Environmental Research and Public Health*, *19*(16), 9974. https://doi.org/https://doi.org/10.3390/ijerph19169974

Xia, P., Zhang, L., Li, F. (2015). Learning similarity with cosine similarity ensemble [Article]. *Information Sciences*, *307*, 39-52. https://doi.org/10.1016/j.ins.2015.02.024

Young, T., Hazarika, D., Poria, S., Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, *13*(3), 55-75. https://doi.org/10.1109/MCI.2018.2840738

(cc) BY