Textual Analysis of the Marie Skłodowska-Curie Actions Evaluation Summary Reports. Assessing Strengths and Weaknesses of Funded and Non-Funded Proposals

Ilaria Rodella, Andrea Sciandra**, Arjuna Tuzzi****

Author information

* IInternational Research Office, University of Padova, Italy. Email: ilaria.rodella@unipd.it

** Department of Philosophy, Sociology, Education and Applied Psychology & Department of Statistical Sciences, University of Padova, Italy. Email: andrea.sciandra@unipd.it

*** Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Italy. Email: arjuna.tuzzi@unipd.it

Article first published online

March 2025

HOW TO CITE

Rodella I., Sciandra A., Tuzzi A. (2025) "Textual Analysis of the Marie Skłodowska-Curie Actions Evaluation Summary Reports. Assessing Strengths and Weaknesses of Funded and Non-Funded Proposals" Italian Journal of Sociology of Education, 17(1), 247-266.
DOI: 10.25430/pupj-IJSE-2025-1-12

PADOVA UP

Textual Analysis of the Marie Skłodowska-Curie Actions Evaluation Summary Reports. Assessing Strengths and Weaknesses of Funded and Non-Funded Proposals

Ilaria Rodella, Andrea Sciandra, Arjuna Tuzzi

Abstract. This study analyses Evaluation Summary Reports (ESRs) of Marie Skłodowska-Curie Actions (MSCA) Individual and Postdoctoral Fellowships proposals at the University of Padua (Unipd), spanning Horizon 2020 and Horizon Europe from 2015 to 2022. The aim is to identify recurring strengths and weaknesses in the evaluation process, recognizing the most important and recurrent features of successful proposals. The use of artificial intelligence is also discussed in the paper. Nearly 400 ESRs were analysed by employing keyword extraction and correspondence analysis (CA) to map relationships between words and variables such as project success. While CA did not clearly distinguish between successful and unsuccessful proposals, machine learning was applied. The coordinates from CA were used to predict project outcomes. Comparisons were made with models using only textual features and those employing transformers, specifically, BERT contextualised embeddings. Results showed that using a Large Language Model (LLM) for text representation improved prediction accuracy compared to other methods. However, it highlighted challenges in interpretability and emphasised the need for explicable methods in the absence of words. Overall, the study provides valuable insights for refining support services and training at Unipd, highlighting the effectiveness of LLMs in prediction while acknowledging the interpretive challenges associated with their use.

Keywords: Marie Skłodowska-Curie Actions, evaluation, large language models, text classification

1. Introduction

Since 1996 and with a budget of 6.6 billion under Horizon Europe, "Marie Skłodowska-Curie Actions (MSCA)" gained a reputation among the research community as one of the most popular European Commission Research Funding Programmes (European Commission, 2023). The MSCA Postdoctoral Fellowships aim to support the creative and innovative potential of postdoctoral researchers wishing to acquire new skills through advanced training, international, interdisciplinary and inter-sectoral mobility[1]. The Individual Fellowships and Postdoctoral Fellowships grants (respectively under Horizon 2020 - 2014/2020 - and Horizon Europe - 2021/2027 - funding schemes) support international experienced researchers (post-doctoral researchers). They comprised two main different schemes, of which the Intra-European Fellowships (EF or Standard Fellowship) address mobility within Europe and the Global Fellowships (GF) that provide mobility from and to the European Union and Third Countries (Postdoctoral Fellowships can take place in Europe, i.e. in an EU Member State or a Horizon Europe Associated Country, or in a Third Country not associated to Horizon Europe; European Commission, 2022).

This study aims to elaborate, through the text analysis of the Evaluation Summary Reports (ESRs), a representation of the most relevant and frequent linguistic features in the evaluations of Marie Skłodowska-Curie Actions (MSCAs) Individual and Postdoctoral Fellowship proposals. This work intends to highlight both the strengths and weaknesses of ESRs related to proposals having the University of Padua (Unipd, Italy) as the Host Institution. Overall, this work assesses recurring and significant lexical features and modalities in the evaluation process and, from a practical perspective, it aims to assess the evaluation criteria and sub-criteria based on the Standard Evaluation Form (HE MSCA; European Commission, 2021)[2].

To achieve these goals, we applied some text mining techniques to extract the most relevant keywords. We then carried out an exploratory multivariate statistical analysis (Correspondence Analysis) and text classification models, focusing on the results of the proposals. Regarding the classification, we compared different models and sets of predictors, namely the most frequent nouns, the coordinates of the Correspondence Analysis and the pre-trained vectors of BERT.

[1] https://rea.ec.europa.eu/funding-and-grants/horizon-europe-marie-sklodowska-curie-actions/msca-postdoctoral-fellowships_en

[2] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/ef/ef_he-msca_en.pdf

We believe that our study can contribute to the existing literature regarding the relationship between textual analysis and evaluation in higher education. In particular, it can:

- Provide guidance for those who wish to submit a proposal for a MSCA project.
- Offer novel comparative evidence regarding the efficacy of Large Language Models.
- Promote a discussion on the application of machine learning methods with an explainable artificial intelligence approach.

The paper is structured as follows: section 2 presents the conceptual background of this work, while section 3 outlines the data and methods used. Section 4 displays the results of the various analyses, which are discussed in section 5, which also includes conclusions.

This paper is an extended version of the proceeding presented at JADT 2024 – 17th International Conference on the Statistical Analysis of Textual Data, which had been held in Brussels (Belgium) on June, 2024 (Rodella et al., 2024).

2. Conceptual Background

The literature on peer review grant evaluation analysis covers various aspects and methods. Some studies were survey-based (Gallo et al., 2021), while others analysed literature and research databases (Demicheli & Di Pietrantonj, 2007; Tricco et al., 2017). While some works present critical aspects of grant peer review, such as lack of reliability and time-wasting (Roumbanis, 2021; Dinov, 2019), other recurring themes include gender issues (Marsh et al., 2009; Magua et al., 2017; Tricco et al., 2017; Gallo et al., 2021) and mobility related to the postdoctoral grant system (Reale et al., 2019; Cattaneo et al., 2019).

The MSCA evaluation process consists of the five main steps[3], of which the most important for the purpose of this study are: i) allocation of each proposal to a set of (at least) 3 external reviewers, based on the best possible match between their expertise and the scientific field of the proposal (also checking conflicts of interest, a fair representation of reviewers' nationalities and gender balance); ii) a remote evaluation phase, where reviewers assess the proposals allocated to them, against evaluation criteria, and draft an Individual Evaluation Report (IER) for each proposal, so that each proposal has (a minimum of) 3 IERs; iii) consensus meeting out of which a Consensus Report (CR) is prepared, with the comments and scores commonly agreed by

[3] https://rea.ec.europa.eu/system/files/2022-08/MSCA%20evaluation%20in%20Horizon%20Europe.pdf

all 3 reviewers. The Evaluation Summary Report (ESR) is the final version of the CR sent to applicants and represents the basilar material to compile the text corpus of this study.

In the field of MSCA, detailed analyses of proposal peer-reviews demonstrated a high-level of agreement among reviewers, especially among their IER and the final ESR (Pina et al., 2015; Pina et al., 2021), and an increasing performance and coherence in the evaluation process in relation to the reviewing experience (Seeber et al., 2021). Other studies on MSCAs instead focused on host institutions (Falk & Hagsten, 2020), while in relation to the specific scope of our study, there are examples of guidelines (Baumert et al., 2022) that provide simple rules for success in MSCA calls. Furthermore, a very recent analysis (Buljan et al, 2023; also citing from REA[4]) found that the final status of the proposals (i.e. main-listed or rejected) can be predicted by the linguistic characteristics of the reviewer's comments, especially the tone related to the identified weaknesses, indicating that weaknesses may be crucial in proposal evaluation. In fact, a recent study has indicated that proposal weaknesses have a greater effect on the ranking, compared to proposal strengths (Hren et al., 2022). Therefore, REA suggested to employ a quantitative text analysis, sentiment analysis or analysis of the text tone as useful methods for assess the proposal evaluations and to determine any differences in the evaluation performed by distinct reviewers.

From a methodological perspective, studies that use content analysis approaches via text mining are highly relevant to our work (Hren et al., 2022; van den Besselaar et al., 2018; Ma et al., 2020; Magua et al., 2017; Bornmann et al., 2012; Luo et al., 2021). In particular, several of these researches employ sentiment analysis through ontological dictionaries (lexicons), focusing on the positive and negative aspects found in ESRs regarding the success or failure of project proposals. The aim is to analyse reviewers' reports to distinguish successful from unsuccessful proposals. Hren et al. (2022) use machine learning methods specifically to highlight relevant aspects of the language of reviewers' comments, thus supporting the idea of using artificial intelligence (AI) in this context. The potential of AI to support publishing and peer review was also investigated (Kousha & Thelwall, 2023). In this research, we aim to explore the use of a Large Language Model to enhance the classification of success for MSCA proposals and provide insights for prospective project submitters.

[4] Is numerical scoring important in the evaluation of grant proposals? - European Commission (europa.eu)

3. Materials & Methods

Almost 400 ESRs related to proposals submitted from 2015 to 2022 were collected and analysed. The data collected has been used for filling a dataset, in which we organised the variables used for this analysis:

1. Year of proposal submission (from 2015 to 2022),
2. Score (from 0 to 100),
3. Excellence score (Threshold: 0/5.00, Weight: 50.00%),
4. Excellence strengths (text),
5. Excellence weaknesses (text),
6. Impact score (Threshold: 0/5.00, Weight: 30.00%),
7. Impact strengths (text),
8. Impact weaknesses (text),
9. Implementation score (Threshold: 0/5.00, Weight: 20.00%),
10. Implementation strengths (text),
11. Implementation weaknesses (text),
12. Project funded (Yes/No),
13. Original Panels (e.g., EF-ECO, GF-PHY, MSCA-IF-EF-CAR, MSCA-IF-EF-RI, etc.). Proposals must be submitted to only one of eight 'main evaluation panels': Chemistry (CHE), Social Sciences and Humanities (SOC), Economic Sciences (ECO), Information Science and Engineering (ENG), Environment and Geosciences (ENV), Life Sciences (LIF), Mathematics (MAT), Physics (PHY) (European Commission, 2022). MSCA-IF-EF-CAR: Career Restart (CAR) panel in Horizon 2020 - CAR includes proposals from any of the 8 scientific areas and provides financial support to individual researchers who wish to resume research in Europe after a career research break (e.g. unemployment, periods of employment outside research, parental or long-term sick leave etc.); MSCA-IF-EF-RI: The Reintegration Panel (RI) panel in Horizon 2020 - RI includes proposals from any of the 8 scientific areas and is dedicated to researchers who wish to return and reintegrate in a longer term research position in Europe (European Commission, 2020).
14. Panels modified (here the panels MSCA-IF-EF-CAR, MSCA-IF-EF-RI have been replaced by the corresponding disciplinary panels according to the project topics (e.g., MSCA-IF-EF-RI= EF-SOC),
15. European vs Global Fellowship (here the Original Panels have been replaced from the categories EF or GF; CAR and RI have been replaced by EF),
16. Panel clusters (here the Original Panels have been replaced by the macro-disciplinary panels Physics and Engineering (PE that merged the panels CHE, ENV, ENG, MAT, PHY), Social Sciences and Humanities (ECO, SOC), and Life Sciences (LIF).

The corpus consists of proposal evaluation texts written in English by referees, which corresponded to variables 4-5, 7-8, and 10-11 listed above. These texts exhibit specific characteristics, such as length and structure, due to the rigid form in which the evaluations are entered. Reviewers are obliged to observe the following guidelines: provide substantial, explanatory comments; avoid comments that merely give a description or a summary of the proposal; use dispassionate, analytical and unambiguous language; use grammatically correct, complete, clear sentences with no jargon; provide polite comments; critical comments should be constructive and not offensive; avoid self-declaration of insufficient expertise (personal or panel) or non-confidence in the proposal; avoid any comments about your expertise that may reveal your identity; avoid reference to the applicant's age, nationality, gender, or personal matters; marks should be consistent with the comments. Therefore, evaluators provide succinct explanatory comments substantiating each evaluation criterion score. Comments take the form of a statement and explanation of key strengths and key weaknesses of the proposal, in the light of the evaluation elements. Furthermore, the texts display genre-specific characteristics as they are project evaluations, also written by non-native speakers, with a specific purpose, resulting in the use of a restricted code[5].

Exploratory analyses were carried out, in particular by means of different keyword extraction methods, in order to gain possible insights, especially for multiword expressions. In particular, we focused on the most frequent terms and employed the Rapid Automatic Keyword Extraction (RAKE) algorithm for terms and multiword expressions. Additionally, we tried to identify which words were most related to the positive or negative outcome of the proposals through log-odds ratios.

A correspondence analysis (CA) was also used for content mapping to represent the system of relationships between words and supplementary variables (such as project success or failure). CA is a classical method of textual data analysis from an exploratory perspective. It enables us to create a content map that depicts the relationships between texts and words, contributing to the emergence of possible patterns. CA transforms word frequencies into coordinates on a multidimensional Cartesian axis system. CA presents texts/words in a low-dimensional space by transforming the chi-squared distance into a specific Euclidean distance, then mapping them into Cartesian planes. The technique relies on singular value decomposition (eigenvalues/eigenvectors) and allows to represent the similarity between

[5] https://newlearningonline.com/new-learning/chapter-5/supproting-materials/basil-bernstein-on-restricted-and-elaborated-codes

lexical profiles. In particular, if two words are close on the graph, it indicates they have similar profiles in terms of text usage.

Within a machine learning framework, we decided to use the coordinates of the CA to predict the outcome of the projects. Furthermore, the CA results were compared with those derived from models utilising only textual features and with those achieved through a representation of texts obtained via transformers, specifically by exploiting the pre-trained vectors of a Large Language Model (LLM). Here, BERT contextualised embeddings were used to obtain a text mapping applied as a set of predictors. Contextualised word embeddings consider the context in which a word is used. BERT (Bidirectional Encoder Representations from Transformers) is based on transformer architecture (Devlin et al., 2019) and produces word representations that are dynamically informed by the surrounding words. This enables BERT to learn a more profound and comprehensive representation of the input text by using both left and right contexts. In this procedure, we concatenated the 12 BERT layers that represent the same token and used the mean to aggregate the embeddings from different tokens to represent a text. We chose contextualised embeddings instead of the fine-tuning procedure of pre-trained vectors to compare the results with other feature extraction techniques.

To conduct this comparison, we selected some of the most widely used machine learning models, namely:

- Random Forest (Breiman, 2001),
- Logistic Regression,
- Support Vector Machine (Scholkopf et al., 1997),
- Extreme Gradient Boosting (Chen et al., 2019).

We selected these models because they enable us to examine the impact of various approaches (GLM, bagging, boosting, etc.).

Logistic regression (Logistic), which is part of the GLM models, aims to establish a causal relationship between the predictors and the response variable *Y*. *Y* can only take on the values 0/1, where 0 represents 'unfunded' and 1 represents 'funded'.

On the other hand, Random Forests (RF) is a member of the bagging techniques family, which is used to decrease the variance of an estimated prediction function. RF (Breiman, 2001) is a variation of the bagging method that constructs a large set of uncorrelated trees and then calculates the average.

Support Vector Machine (SVM) is a classification model that maps observations as points in space to divide the categories by the largest possible space. The SVM algorithm finds the optimal separation hyperplane, using linear or non-linear mapping, defined by the observations that lie within an optimised margin determined by a cost hyperparameter. These observations are called support vectors. The cost hyperparameter penalises large residuals.

XGBoost, or eXtreme Gradient Boosting, is a machine learning algorithm based on the gradient boosting algorithm developed by Chen et al. (2019). At a fundamental level, the algorithm employs a sequential approach to improve the subsequent model based on gradient descent. Among the features that make the performance of this technique high, we stress: regularisation to avoid overfitting (through L1 Lasso Regularization and L2 Ridge Regularization), and tree pruning (as XGBoost builds the tree until the designated parameter, 'max depth', starts pruning backward).

The importance of each feature in the classification models was initially measured by a ROC curve analysis performed on each predictor. As we are dealing with a two-class problem, we apply a series of cut-offs to the features in order to predict the class. In particular, for Linear Models, we employ the absolute value of the t-statistic for each model parameter. Concerning the Random Forest model, we measure the prediction accuracy for each tree on the out-of-bag portion of the data. After permuting each predictor variable, the same process is repeated. The difference between the two accuracies is subsequently averaged over all trees and normalised by the standard error. For SVM and XGBoost classification models, the default process is to calculate the area under the ROC curve.

We additionally utilised the shapr R package to implement the Kernel SHAP (SHapley Additive exPlanations) procedure for estimating Shapley values (Lundberg and Lee, 2017), which proved to be valuable in explaining complex machine learning models. Although Kernel SHAP operates under the assumption of independent features, Aas, Jullum, and Løland (2021) extend this method to dependent features. SHAP analysis is a model-agnostic explanation method that takes into account each feature importance and the interactions with other features. This goal is achieved by calculating the importance of a feature and comparing a model's predictions with and without the feature.

4. Results

The initial step in conducting exploratory text analysis of the ESRs is to create a corpus from the textual fields that cover strengths and weaknesses related to Excellence, Impact, and Implementation criteria. The choice to merge all available text for each proposal was driven by the fact that some ESRs have empty fields (e.g., in 'weaknesses') or only the statement 'none'. Approximately 10% of the text fields were empty, while 4% included only the word 'none' or a similar short text. Additionally, referees may not always differentiate between the strengths and weaknesses of excellence, impact, and implementation in the same manner.

After parsing the text, tokens were extracted with minimal intervention. This involved the removal of punctuation and the conversion of tokens to lowercase. The Document-term matrix (DTM) comprises 392 documents with 8,139 words (word types), resulting in a total of 245,970 tokens. The Type/Token ratio (TTR) is 0.033, and the hapax percentage is 42%. Based on the number of tokens, it appears that the language used is limited and redundant. Subsequently, the corpus was lemmatized to reduce the size of the DTM and focus the analyses on relevant parts of speech (POS). The treebank for English GUM (the Georgetown University Multilayer corpus; Zeldes, 2017) was used for this purpose. GUM is suitable for the collected texts as it was also trained on scientific texts. After lemmatization, only nouns were selected, resulting in a new DTM that contains 3,409 nouns with a total of 73,197 tokens, which were used for the subsequent analysis.

To identify the most important keywords, the following graphs display:

- The most frequent nouns (Fig. 1);
- The most important nominal compounds selected using the RAKE algorithm for nouns only (Fig. 2).

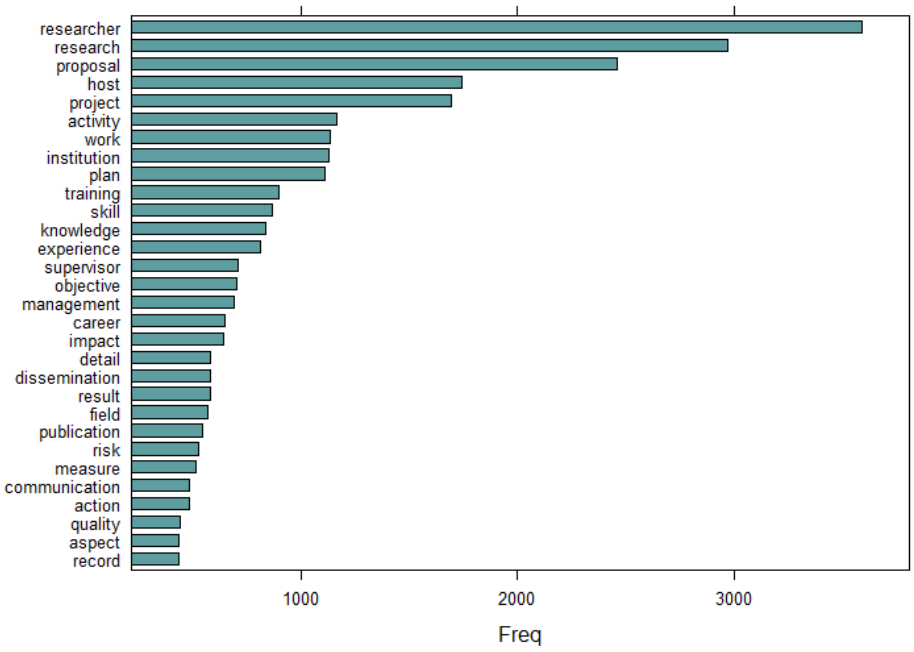

Fig.1 Most frequent nouns obtained after the lemmatization process (3,409 nouns found in total)

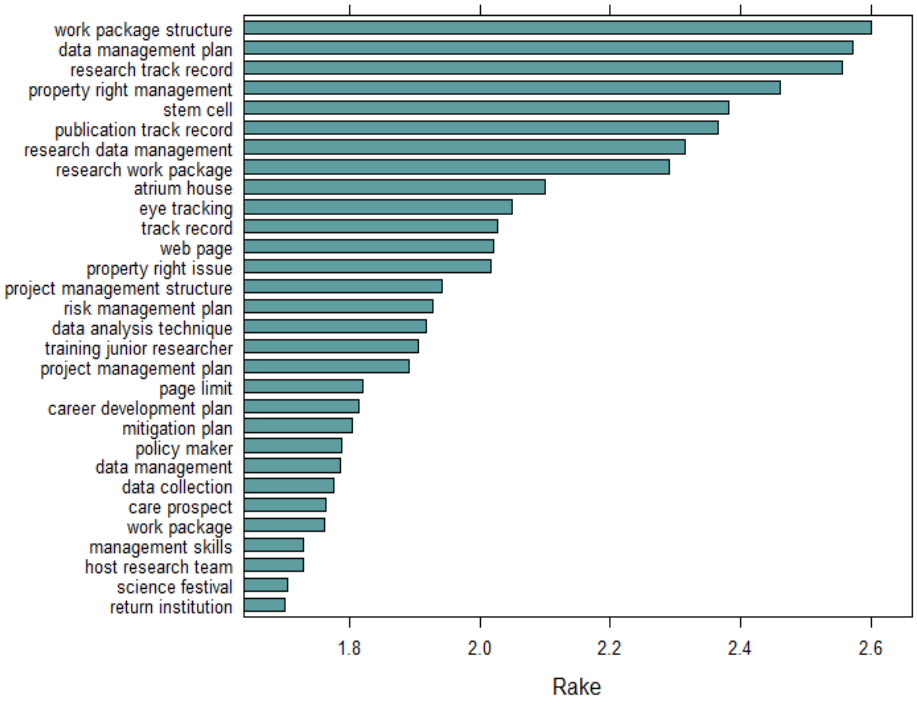

Fig. 2 Most important nominal compounds selected using the RAKE algorithm

Text mining was also used to extract keywords associated with the proposals' outcomes. Here, the outcome of the projects was treated similarly to the result of sentiment analysis by identifying the words most associated with the positive/negative outcome in terms of odds ratios. In particular, the analysis of odds ratios displays the lemmas that are most likely to be associated with the funded/not-funded types. The log odds ratio is calculated using the formula:

$$log\ odds\ ratio = ln\left(\frac{\left[\frac{n+1}{total+1}\right]_{funded}}{\left[\frac{n+1}{total+1}\right]_{not\_funded}}\right) \quad (1)$$

where n represents the number of times a particular noun is used in each subset, and *total* indicates the total number of nouns in each subset.

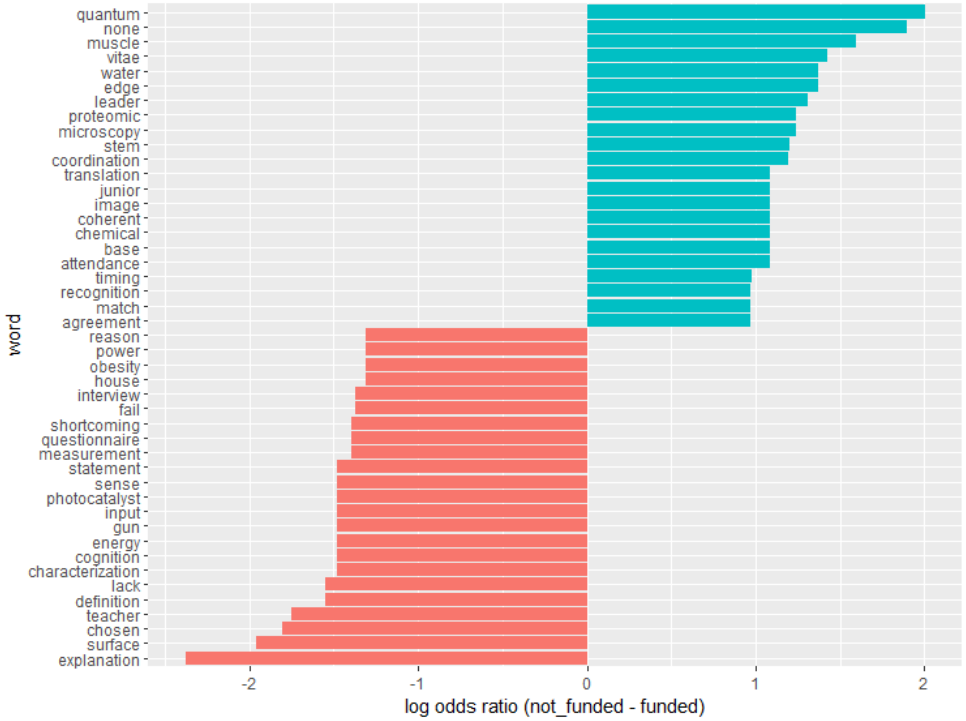

Fig. 3 Log-odds-ratios of nouns associated with funded/not-funded proposals

This analysis only considers nouns or nominal compounds. No nominal compound appears among the terms most likely to be associated with funded or unfunded projects (Fig. 3). The term 'none' frequently appears in funded projects' weaknesses, as the only description. Potentially interesting terms for funded projects include 'leader', 'coordination', 'coherent', 'timing', and 'agreement'. Potentially interesting terms for unfunded projects include 'reason', 'fail', 'shortcoming', 'measurement', 'lack', 'characterisation', and 'explanation'. This analysis displays results different from the previous analyses, with the presence of topic-related words associated with funded or non-funded proposals. For instance, proposals with a focus on energy topics have been funded less than proposals on water issues.

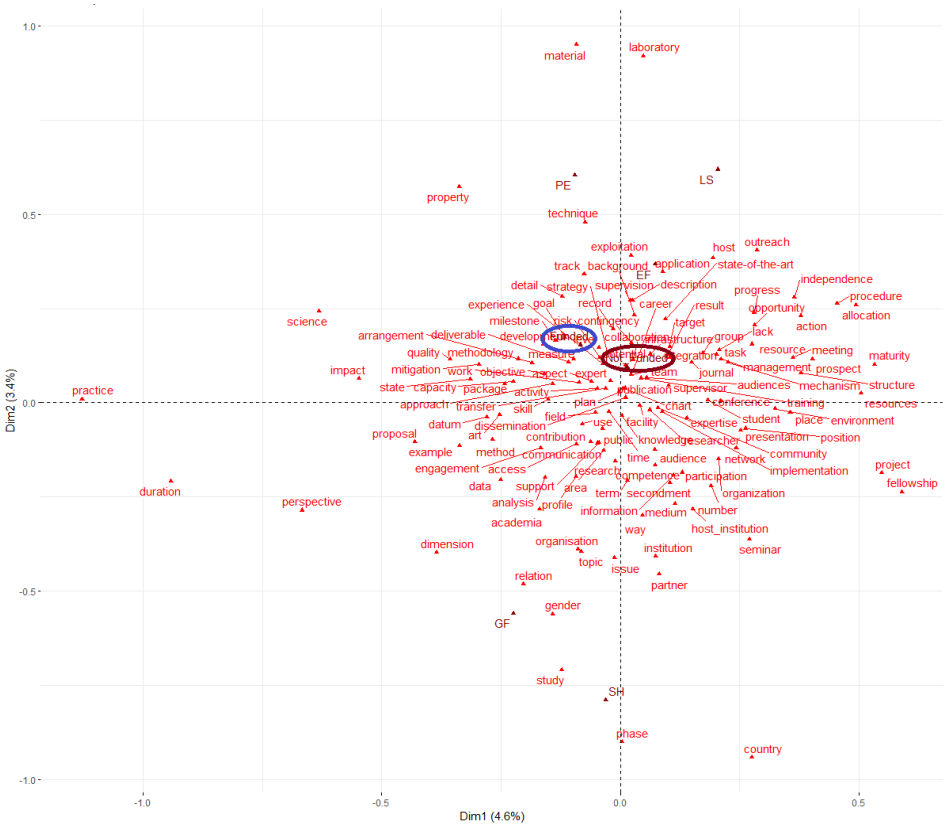

Fig. 4 Correspondence Analysis – first two dimensions – nouns and n-grams

Following the keyword-based analysis, we applied correspondence analysis (CA) to the ESRs texts. Therefore, our aim is to create a lemma mapping through CA and to project the supplementary variables of interest (Panels: LS, PE, and SH; EF or GF fellowships). We have specifically chosen high-frequency nouns (>=100) to conduct this analysis and provide clear results. On the Cartesian plane comprised of the first two dimensions (Fig. 4), we observe that the supplementary variable linked to funded projects is relatively close to that of non-funded projects (encircled in blue and red, respectively). Hence, this analysis does not unambiguously distinguish different terms associated with the success/failure of the projects. Nevertheless, the nouns that are most closely related to funded projects are: 'goal', 'milestone', 'measure', 'deliverable', 'development', 'experience', 'strategy', 'detail', 'expert', 'risk', 'methodology', and 'objective'. The nouns closest to non-funded projects are: 'potential', 'career', 'collaboration', 'contingency', 'dissemination', 'infrastructure', 'audience', 'training', and 'team'. To sum up, the nouns closest to funded projects seem to be more relevant to the project ('measure', 'deliverable', 'objective', 'methodology'); conversely, the nouns closest to

non-funded projects seem to be more relevant to the researcher ('training', 'career', 'team'). At the panel level, PE is closer to funded projects than LS and SH. The CA coordinates were then extracted for further analysis.

In the final step of our analysis, we are comparing the predictive capabilities of three different sets of predictors with respect to the success or failure of project proposals:

- The coordinates of the CA;
- The most frequent nouns;
- The embeddings derived from text mapping via BERT.
- With the aim of a balanced comparison, we decided to limit these sets of predictors to 50 features, thus corresponding to:

- The first 50 coordinates of the CA;
- The relative frequencies of the 50 most frequent nouns;
- As the BERT-base mapping involves 768 vectors, to reduce dimensionality we applied a principal component analysis, extracting the first 50 components.

In order to improve classification through data mash-up, we added other variables to these features such as the year of proposal and the type of fellowship (EF or GF), alongside dummy variables for different panels. Additionally, we included some summary variables of the ESRs' texts, specifically, the count of types, tokens, and sentences. Altogether, every classification model has 57 features.

The training set was created by randomly extracting 80% of the texts, with the remaining texts used for the test set. Each model involves a 5-fold cross-validation repeated 30 times. To evaluate the results, we analysed the accuracy, the number of misclassifications and the F1-score.

The results of the models for the test set are presented in the table below, including RF, Logistic, SVM, and XGBoost. 'NA' means that a model is not able to identify any funded proposal.

Table 1. Results of the classification models for proposal's success.

Model	Accuracy	# errors	F1-score (ref.: YES)
RF CA	0.7625	19	NA
RF nouns	0.7436	20	NA
RF BERT	0.9231	6	0.8333
Logistic CA	0.7436	20	0.2857
Logistic nouns	0.7564	19	0.0952
Logistic BERT	0.9615	3	0.9189
SVM CA	0.7051	23	0.3030
SVM nouns	0.7308	21	0.0870

Model	Accuracy	# errors	F1-score (ref.: YES)
SVM BERT	0.9231	6	0.8235
XGBoost CA	0.7179	22	0.2143
XGBoost nouns	0.7179	22	0.2143
XGBoost BERT	0.8974	8	0.7778

5. Discussion and conclusion

The results indicate that using BERT to represent texts improves predictions compared to the other two predictor sets. Indeed, the BERT models consistently outperformed those using CA and nouns. Overall, the best model is the logistic regression with penalty (by means of features derived from BERT), having an accuracy of approximately 96% (3 errors out of 78 in the test set). Results also indicate that CA features demonstrate comparable predictive capacity to those of nouns. The substantial difference between these two sets of predictors is that those derived from CA show higher levels of F1-score, as they are better able to predict successful projects. However, the immediate interpretive capacity of predictive models seems to be compromised in the absence of words and could argue in favour of explainable methods. Certainly, given the nature of this work, which is intended to deeply understand the evaluation process besides the formal information given by the EC, it is important to observe directly which aspects can influence the evaluation. This is of course possible through words, but also by applying CA. For example, taking a dimension that is an important predictor (Dim. 11) of the best model (logistic regression) for CA coordinates, we can observe which nouns contribute most to the dimension (Fig. 5).

Ideally, LLM can still be used to highlight the parts of the text that are most relevant to project success or failure. However, in this work, we used an additional transformation (PCA) of the contextual embeddings for comparison purposes with other sets of variables. Therefore, we cannot identify the most important words related to proposal funding.

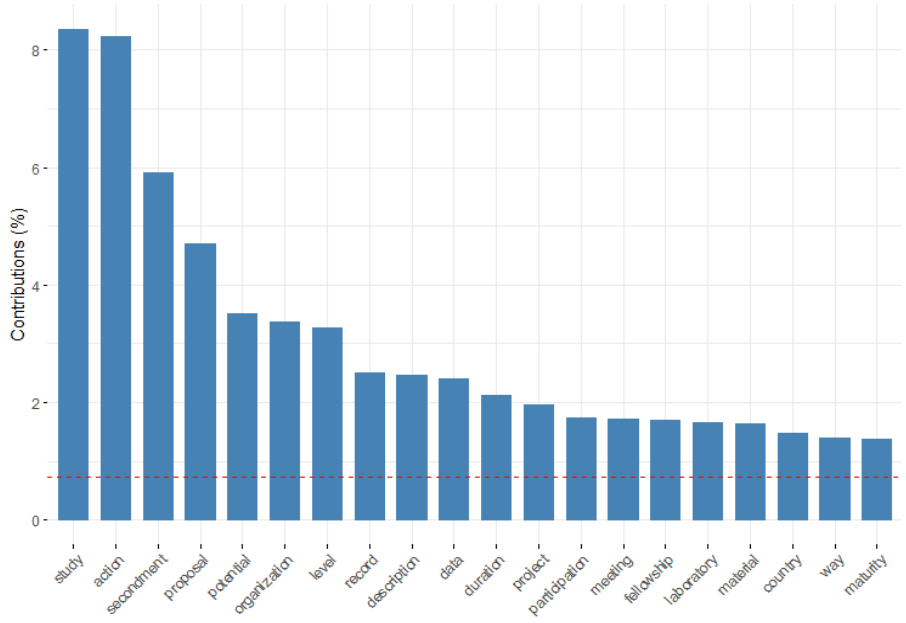

Fig. 5 Nouns that contribute most to determining Dimension 11 of the CA

In contrast, models that exploit textual features may be useful for understanding the variables' importance, primarily through ROC curve analysis of each predictor. Still with a focus on obtaining explainable results of machine learning models, SHAP analysis allowed us to observe the features' importance with respect to the values they show in our dataset. First, we selected the most important textual features of the best prediction model based on nouns (XGBoost) in terms of F1-measure. Next, since SHAP analysis allows local explanations, we wanted to evaluate the impact of the previous best features on some ESRs, in order to highlight why an ESR was classified as funded/unfunded by the model. In this example (Fig. 6), for computational and interpretability reasons we limited this analysis of single ESRs to the best six features highlighted by the global SHAP analysis. For instance, the lack of the term 'quality' results in a negative impact on the feature contribution in Proposal 2, whereas Proposal 4, which has a higher number of types (370), shows a higher contribution for this feature than Proposal 2, which has a lower number of types.

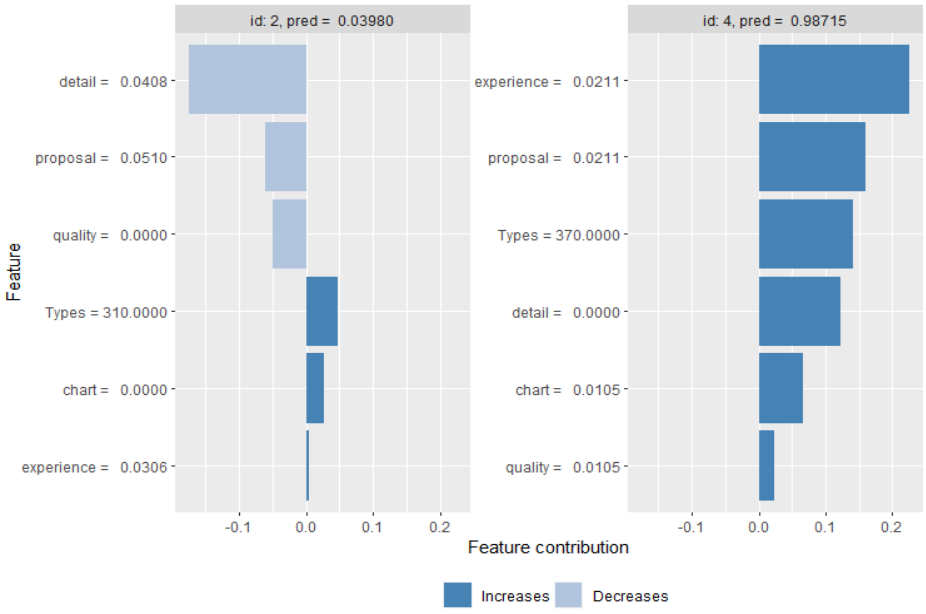

Fig. 6 Shapley values prediction explanation for ESR n. 2 (prediction: not funded) and ESR n. 4 (prediction: funded)

In conclusion, this work highlighted some features that could potentially lead to the success of an MSCA proposal. Classical text mining techniques and machine learning models were used for classification, despite the limited number of ESRs and the limited length of the texts. The results show that different methods can be integrated to obtain an overall assessment of the ESRs, even describing different results. As the main findings of the CA report that nouns closest to funded projects seem to be more relevant to the project ('measure', 'deliverable', 'objective', 'methodology') and nouns closest to non-funded projects seem to be more relevant to the researcher ('training', 'career', 'team'), the Log-odds-ratios allow us to further obtain scientific words related to funded ('quantum', 'water', 'muscle') or non-funded projects ('power', 'surface', 'photocatalysis'). In the same way, XGBoost model displays words (like 'detail', 'quality', 'experience'), not reported from the other methods. In general, this approach not only allows us to identify the most important words in terms of frequency that are associated with the success or failure of proposals, but also provides different insights depending on the type of multivariate analysis used. The most important predictors of the classification models for MSCA proposals do not include subjective evaluations such as 'excellence', 'impact', 'implementation', 'training, 'gender', or 'open-access'. Instead, terms such as 'chart', 'experience', 'detail', 'proposal', and 'quality' emerged. Thus, while the various analyses offer different in-

sights, they also provide different perspectives that enrich the study of ESRs with different dimensions (quantitative and qualitative) within a coherent framework of approaching the research question.

Future developments of this project could exploit a larger database and focus on the use of LLMs, which prove higher predictive performance. In particular, it might be useful to implement the fine-tuning of a pre-trained LLM in relation to the proposal outcome and to highlight the parts of the text that best characterise the success or failure of projects. Moreover, we would like to assess and demonstrate the presence of similarities and differences in the evaluation of the 2 types of fellowships (Standard and Global), their success rate and access to funding, according to the main Horizon Europe evaluation criteria and the submission panels provided in the MSCA framework (see Section 3). Finally, it would be interesting to determine whether there were any differences in the evaluation performed by distinct reviewers for the same proposals.

Acknowledgements

We thank the International Research Office of the University of Padua for the collection of the ESR data and the valuable support, in particular the Head of Office, dr. Francesca Mura and the Head of the Individual Research Unit, dr. Viviana Gialain.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *arXiv Preprint arXiv:1903.10464*.

Baumert, P., Cenni, F., & Antonkine, M. L. (2022). Ten simple rules for a successful EU Marie Skłodowska-Curie Actions Postdoctoral (MSCA) fellowship application. *PLOS Computational Biology*, *18(8)*, e1010371. https://doi.org/10.1371/journal.pcbi.1010371

Bornmann, L., Wolf, M. & Daniel, H. D. (2012). Closed versus open reviewing of journal manuscripts: how far do comments differ in language use?. *Scientometrics*, *91*, 843–856. https://doi.org/10.1007/s11192-011-0569-5

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Buljan, I., Pina, D. G., Mijatović, A., & Marušić, A. (2023). Are numerical scores important for grant proposals' evaluation? A cross sectional study [version 1; peer review: 1 approved]. F1000Research 2023, 12:1216 https://doi.org/10.12688/f1000research.139743.1

Cattaneo, M., Malighetti, P., & Paleari, S. (2019). The Italian brain drain: Cream and milk. *Higher Education*, *77*(4), 603–622. https://doi.org/10.1007/s10734-018-0292-8

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2019). *xgboost: Extreme gradient boosting*. R package version 0.81.0.1, 1–4.

Demicheli, V., & Di Pietrantonj, C. (2007). Peer review for improving the quality of grant applications. *The Cochrane Database of Systematic Reviews*, *2007*(2), MR000003. https://doi.org/10.1002/14651858.MR000003.pub2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dinov, I. D. (2020). Flipping the grant application review process. *Studies in Higher Education*. https://www.tandfonline.com/doi/full/10.1080/03075079.2019.1628201

European Commission (2020). H2020 Programme Guide for Applicants Marie Skłodowska-Curie Actions Individual Fellowships (IF) Version 1.4 08/04/2020 https://euraxess.ec.europa.eu/sites/default/files/news/h2020-guide-appl-msca-if-2018-20_en_1.pdf

European Commission (2021). Horizon Europe Evaluation Form (HE MSCA), Version 1.0, 18 June 2021. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/ef/ef_he-msca_en.pdf

European Commission (2022). Horizon Europe Work Programme 2023-2024 2. Marie Skłodowska-Curie Actions (European Commission Decision C(2022)7550 of 6 December 2022). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2023-2024/wp-2-msca-actions_horizon-2023-2024_en.pdf

European Commission, Directorate-General for Education, Youth, Sport and Culture, Marie Skłodowska-Curie actions (2023). Developing talents, advancing research. Almost 30 years of European support for researchers' work, Publications Office of the European Union, https://data.europa.eu/doi/10.2766/120204

Falk, M. T., & Hagsten, E. (2021). Potential of European universities as Marie Curie grantee hosts. *High Educ* **81**, 255–272. https://doi.org/10.1007/s10734-020-00540-3

Gallo, S. A., Schmaling, K. B., Thompson, L. A., & Glisson, S. R. (2021). Grant Review Feedback: Appropriateness and Usefulness. *Science and Engineering Ethics*, *27*(2), 18. https://doi.org/10.1007/s11948-021-00295-9

Hren, D., Pina, D. G., Norman, C. R., & Marušić, A. (2022). What makes or breaks competitive research proposals? A mixed-methods analysis of research grant evaluation reports. *Journal of Informetrics*, *16*(2), 101289. https://doi.org/10.1016/j.joi.2022.101289

Kousha, K., & Thelwall, M. (2023). Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*. https://doi.org/10.1002/leap.1570

Luo, J., Feliciani, T., Reinhart, M., Hartstein, J., Das, V., Alabi, O., & Shankar, K. (2021). Analyzing sentiments in peer review reports: Evidence from two science funding agencies. *Quantitative Science Studies*, *2*(4), 1271–1295. https://doi.org/10.1162/qss_a_00156

Lundberg, S. M, and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 4765–74.

Ma, L., Luo, J., Feliciani, T., & Shankar, K. (2020). How to evaluate ex ante impact of funding proposals? An analysis of reviewers' comments on impact statements. *Research Evaluation*, *29*(4), 431–440. https://doi.org/10.1093/reseval/rvaa022

Magua, W., Zhu, X., Bhattacharya, A., Filut, A., Potvien, A., Leatherberry, R., Lee, Y.-G., Jens, M., Malikireddy, D., Carnes, M., & Kaatz, A. (2017). Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers' Critiques. *Journal of Women's Health*, *26*(5), 560–570. https://doi.org/10.1089/jwh.2016.6021

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O'Mara, A. (2009). Gender Effects in the Peer Reviews of Grant Proposals: A Comprehensive Meta-Analysis Comparing Traditional and Multilevel Approaches. *Review of Educational Research*, *79*(3), 1290–1326. https://doi.org/10.3102/0034654309334143

Pina, D. G., Buljan, I., Hren, D., & Marušić, A. (2021). A retrospective analysis of the peer review of more than 75,000 Marie Curie proposals between 2007 and 2018. *eLife*, *10*, e59338. https://doi.org/10.7554/eLife.59338

Pina, D. G., Hren, D., & Marušić, A. (2015). Peer Review Evaluation Process of Marie Curie Actions under EU's Seventh Framework Programme for Research. *PLOS ONE*, *10*(6), e0130753. https://doi.org/10.1371/journal.pone.0130753

Reale, E., Morettini, L., & Zinilli, A. (2019). Moving, remaining, and returning: International mobility of doctorate holders in the social sciences and humanities. *Higher Education*, *78*(1), 17–32. https://doi.org/10.1007/s10734-018-0328-0

Rodella, I., Sciandra, A., & Tuzzi, A. (2024). Analysis of Marie Skłodowska-Curie Actions (MSCA) evaluations and models for predicting the success of proposals. In A. Dister & D. Longrée (eds.), *JADT 2024 Mots comptés, textes déchiffrés*, pp. 783-792.

Roumbanis, L. (2021). The oracles of science: On grant peer review and competitive funding. *Social Science Information*, *60*(3), 356–362. https://doi.org/10.1177/05390184211019241

Scholkopf, B., Sung, K. K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11), 2758-2765.

Seeber, M., Vlegels, J., Reimink, E., Marušić, A., & Pina, D. G. (2021). Does reviewing experience reduce disagreement in proposals evaluation? Insights from Marie Skłodowska-Curie and COST Actions. *Research Evaluation*, *30*(3), 349–360. https://doi.org/10.1093/reseval/rvab011

Tricco, A. C., Thomas, S. M., Antony, J., Rios, P., Robson, R., Pattani, R., Ghassemi, M., Sullivan, S., Selvaratnam, I., Tannenbaum, C., & Straus, S. E. (2017). Strategies to Prevent or Reduce Gender Bias in Peer Review of Research Grants: A Rapid Scoping Review. *PLOS ONE*, *12*(1), e0169718. https://doi.org/10.1371/journal.pone.0169718

van den Besselaar, P., Sandström, U., & Schiffbaenker, H. (2018). Studying grant decision-making: A linguistic analysis of review reports. *Scientometrics*, *117*(1), 313–329. https://doi.org/10.1007/s11192-018-2848-x

Zeldes, Amir (2017) "The GUM Corpus: Creating Multilayer Resources in the Classroom". *Language Resources and Evaluation* 51(3), 581–612.