Analyzing the relationship between text features and research proposal success

Jorge A. V. Tohalino, Laura V. C. Quispe, and Diego R. Amancio

Institute of Mathematics and Computer Science, Department of Computer Science,
University of São Paulo, São Carlos, SP, Brazil

(Dated: May 19, 2020)

Abstract

Predicting the success of research proposals is of considerable relevance to research funding bodies, scientific entities and government agencies. In this study, we investigate whether text features extracted from proposals title and abstracts are able to identify successful funded research proposals. Our analysis was conducted in three distinct areas, namely Medicine, Dentistry and Veterinary Medicine. Topical and complexity text features were used to identify predictors of success. The results indicate that both topical and complexity measurements are relevant for identifying successful proposals in the considered dataset. A feature relevance analysis revealed that abstract text length and metrics derived from lexical diversity are among the most discriminative features. We also found that the prediction accuracy has no significant dependence on the considered proposal language. Our findings suggest that textual information, in combination with other features, are potentially useful to assist the identification of relevant research ideas.

*diego@icmc.usp.br
I. INTRODUCTION

Science of science has emerged, in the last few years, as the research area devoted to study the mechanisms underlying research and its related aspects [25]. This area has investigated a large number of important questions, including the evolution of science, and more specifically patterns of collaboration, citation and contribution among scientific entities [21]. Many studies have shed light on several important issues related to many processes involved in the creation and dissemination of scientific manuscripts. For example, studies on the behavior of paper citation networks not only have characterized these evolving networks, but also have developed models to predict their behavior [53, 57]. Many studies have also sought linguistic patterns in the scientific literature [39]. Similar studies have used paper metadata to analyze and understand the behavior of authors, including their collaboration/citation patterns and contributorship patterns [17]. Another important area in science of science concerns the studies devoted to make predictions in many scenarios [1]. Those studies are important because they favor more informed decisions, thus improving the design of research policies. While the focus of most investigations in science of science, especially those in the predictive area, use data from papers, in this paper we probe whether it is possible to make predictions regarding research output using data extracted from research proposals.

Writing research proposals represents an important part of scientists’ work. While proposals themselves are usually not intended to be published, they are equally relevant because they may ultimately decide whether novel ideas are going to be further developed and possibly disseminated. Deciding thus which proposals are going to be funded is of paramount importance for the advancement of science. Those decisions should be as fair as possible and, in many desired situations, they should be devoid of any personal bias other than the expected quality criteria. In this sense, it becomes interesting if an automatic approach could assist (but not replace) the traditional evaluation of research proposals (at least in some criteria). While being less prone to personal bias, another advantage associated with automatic approaches is their ability to make decisions in a much short period of time, when compared to the traditional human classification. Similar approaches have already been employed with success in other areas. For example, the quality of essays and translations have been assessed using machine learning methods [5]. A pattern recognition approach applied in the context of proposals assessment could also shed light on the understanding of which
factors are associated with strong proposals. This could be particularly useful for early career scholars, as many of them have received little or no feedback regarding previous proposals submissions. In the current study, we touch these points by probing whether information retrieved from research proposals can be used to predict their success.

While many factors may affect the perceived quality of a research proposals [12], in this study we focus on the analysis of textual features. More specifically, we focused on two types of textual attributes. We first analyzed the influence of topical features. We also used complexity measurements – such as lexical diversity and words concreteness – to characterize the research proposals. While the latter is intended to capture linguistic patterns that are topic independent, the former is used to investigate whether proposals on specific topics are more likely to be successful. Our analysis was conducted in a subset of research proposals funded by the São Paulo Research Foundation (FAPESP-Brazil). We selected research proposals in three research areas comprising the largest number of proposals funded by FAPESP. We considered proposals in the following areas: Medicine, Dentistry and Veterinary Medicine. A proposal was considered successful if it yielded at least one publication.

Several interesting results could be found in our analysis. By considering only a balanced version of the datasets, we could find a maximum accuracy rate of roughly 83% in predicting proposals success in the Dentistry dataset. A slight lower accuracy was found for the other areas. These results suggest that both topical and complexity measurements plays a relevant role in identifying successful proposals in the considered dataset. We also found that the results for complexity measurement are not dependent on the considered language (English or Portuguese). A feature importance analysis revealed that the measurements capturing lexical diversity of abstracts are relevant features for identifying successful proposals in all three considered datasets. Our analysis also revealed that the best classifiers for the adopted features were those based on Decision Trees and Support Vector Machines. All in all, the adopted framework provides evidence that text features seem to be relevant in the identification of successful funded projects. Therefore, we believe that text features could be combined with other features in future works to improve the discriminative rate of the classification systems.

This manuscript is organized as follows. In Section IV C, we present related works on features used to predict the success of scientific papers and research proposals. In Section III A
we describe the methodology used in the machine learning framework. The obtained results are discussed in Section IV. Perspectives for works extending our approach are presented in Section V.

II. RELATED WORKS

Several studies have investigated the factors leading to the success of scientific items [12, 55, 56]. In the case of scientific papers, many factors have found to play a role in defining their visibility. In [23], the authors show that the number of citations received in recent years can be an indication of future success. The authors proposed a linear preferential attachment with time dependent initial attractiveness that can recover not only the distribution of citations, but also the citation burstiness effect [23]. Similar models have extended this idea to characterize and predict researchers’ impact. Other factors affecting the popularity of papers include the visibility of authors, journals, universities and the interdisciplinarity of fields and subfields [20, 44, 49].

Text factors have also been found to affect the visibility of papers [6, 34, 39, 45]. In [6], the authors proposed a model to describe the evolution of papers citation networks. In addition to the age and visibility factor, they found that the similarity with other papers also represents a factor that cannot be disregarded. The impact of text features has also been discussed in some works [34, 45]. Recent results have pointed out that journals publishing papers with short titles tend to be more visible, as measured by the average citation counts. This is consistent with the idea that the use of a less complex linguistic style in papers leads to a better paper understanding. The influence of other textual factors on citations including question marks and titles describing results has also been reported [45].

The factors affecting the success of research proposals have also been analyzed in the last few years [12, 14, 24, 30, 37]. In [12] the authors found that researchers productivity can not be used to predict research proposal success. Likewise, institutional research strengths are not strong indicators of success. The success of research proposal was found to be more correlated with the topic similarity between the proposal references and the respective applicant publications.

Another feature that could be used to predict research proposal success are those related to peer review scores. In [14], the correlation between peers’ scores and visibility indexes
was analyzed for Spanish researchers in 23 fields. The study found that correlations are strongly dependent on the field being analyzed. Moreover, this study revealed that the main indicators that are associated to the acceptance of research proposals are the total number of publications and the number of papers published in prestigious journals. In [24], the authors studied the correlation between future research productivity and peers’ scores of grants funded by the U.S. National Institutes of Health (NIH). They found that assigned scores are poor discriminators of success. As a consequence, they argue that this finding might increase the lack of discontentment with the peer review evaluation [27]. Leading to a different conclusion, the study carried out in [37] argues that good peer review rating are correlated with better research outcomes, even when some specific controls are considered in the analysis, including authors and institutions visibility. This conclusion was reached in a dataset comprising 130,000 research projects funded by NIH.

While many studies have focused on a variety of features to predict research proposals success, here we focus on text features, and more specifically on the complexity/style related features of language.

III. MATERIAL AND METHODS

The dataset used in the current paper is described in Section [III A]. The framework proposed to classify research proposals comprises the following three main steps:

1. **Feature extraction**: this phase is responsible for extracting topical and complexity features from textual fragments of research proposals. This is detailed in Section [III B]. While we test the influence of topical features, our main focus here is to analyze the influence of text complexity on the predictability of proposals success.

2. **Pattern recognition**: the features extracted are used as input for traditional machine learning methods. An overview of methods is provided in Section [III C]. A more detailed reference on machine learning and pattern recognition methods can be found in [22].

3. **Feature relevance analysis**: this phase is responsible for identifying the most relevant (i.e. discriminative) features. A brief description of the adopted method is provided in Section [III D].
A. Dataset

The main objective of this work is to analyze whether textual features can be used to predict the success of research projects. The adopted dataset consists in a subset of research projects carried out by researchers in Brazil (São Paulo State) and funded by the São Paulo Research Foundation (FAPESP) [42]. While it would be of interest to analyze the full content of research projects, this information is not public available. For this reason, most of the text analysis was based in two parts of the research proposals: their title and abstract. The data were retrieved from the Biblioteca Virtual website [43]. The research projects are written originally in Portuguese. This is the reason why we focus our analysis on Portuguese textual data. However, because several abstracts are also available in English, we also provide an analysis of the dependence of the results on the considered language.

We focused our analysis on regular grants. We decided to analyze this type of grants for two main reasons: regular grants have a duration of at least 18 months (most of them lasts for 24 months). Therefore, some publications can be expected after this period. The other reason for choosing regular grants is the fact there are several regular proposals in the dataset. Considering this type of research project, we could retrieve textual information from more than 31,000 instances. We considered projects funded between 1989 and 2015. More recent projects were disregarded because papers resulting from the projects may take several months to be published.

There are several useful bibliometric metrics to gauge research proposals success. This could be the number of published papers, the number of citations, and several other metrics commonly used in quantifying success in academia [54]. Because most of these distributions are skewed, we decided to simplify the criteria to consider a research project as successful. To avoid an extreme unbalancing in the number of positive and negative examples [38], we consider a project as successful if it yielded at least one publication. While this criterion still generates unbalanced datasets, a considerable number of both positive (successful) and negatives (unsuccessful) examples can be recovered.

In order to avoid bias when comparing different research areas, we compared only projects belonging to the same area. In particular, we considered the following three areas comprising most of the research projects funded by FAPESP: Medicine (MED), Dentistry (DENT) and Veterinary Medicine (VET). According to the adopted criterion, the percentage of pos-
itive examples in each area was: 41.27%, 48.48% and 31.96% for Medicine, Dentistry and Veterinary Medicine, respectively. Note that, in all cases, the number of positive examples is lower than the number of negative examples. In order to balance the data, the following standard procedure was applied [22]. Before training the models, we randomly draw from the set of negative examples $X$ instances, where $X$ is the number of positive instances in the dataset. This procedure was repeated 10 times for each area. The reported results therefore represents an average over these 10 generated balanced datasets.

**B. Feature Extraction**

For each research project, we extracted textual features from both Portuguese and English text versions of research project abstract and titles. We are particularly interested in analyzing if there is an association between text structure (or complexity) and the observed research output. For comparison purposes, we also studied how predictable are proposal outputs when texts are characterized with topical features.

The first feature used is the frequency of specific words. For each text, this generates a sparse vector whose $i$-th element stores the frequency of $i$-th word of the vocabulary. We also used a normalized version of this strategy, the so-called term frequency–inverse document frequency (tf-idf) approach. According to this strategy, the relevance of a word $w$ in each document depends not only on the frequency of $w$ in the document, but also on how many documents of the dataset. More specifically, the tf-idf representation of a word $w$ in a document (i.e research project) $d$ is given by:

$$
\text{tf-idf}(w, d) = \frac{f(w, d)}{n_d} \cdot \frac{\log N}{\log (N_w)},
$$

where $f(w, d)$ is the frequency of $w$ in $d$, $n_d$ is the number of words in $d$, $N$ is the number of documents in the dataset and $N_w$ is the number of documents in which $w$ occurs at least once.

A different approach to characterize texts is via complexity analysis [3]. The measurements used in the current are a subset of metrics adapted from the English version of Coh-Metrix [28]. Some examples of textual complexity features used here are:

1. **Basic counts**: total number of sentences, words, adjectives, adverbs and verbs.
2. **Logic operators**: this feature quantify the number of logical operators, such as “if”, “and”, “or” and negations.

3. **Function word diversity**: this corresponds to the total of function word types (i.e. function word vocabulary size) normalized by the total number of different words (vocabulary size).

4. **Preposition diversity**: this corresponds to the same counting in function word diversity, but applied to prepositions only.

5. **Punctuation diversity**: this corresponds to the same counting in function word diversity, but applied to punctuation marks only.

6. **Noun SD**: this corresponds to the standard deviation in the number of nouns per sentence.

7. **Brunet index**: this index quantifies the lexical diversity in the text. It is computed as $\beta = v^\alpha$, where $\alpha = n^{-0.165}$, $v$ is the vocabulary size and $n$ is the total number of words in the text. Typically, $10 \leq \beta \leq 20$. A high value of $\beta$ corresponds to a high lexical diversity.

8. **Mean noun phrase**: this corresponds to the average number of noun phrases in sentences. A noun phrase usually includes a noun and its modifiers.

9. **Concreteness SD**: this index quantifies the number of concrete words in the text. A concrete word is defined as a word representing concepts and events that can be measured and observed. Examples of concrete words are ‘car’ and ‘beans’. Conversely, examples of abstract words include ‘faith’ and ‘chaos’ [7].

10. **NE ratio text**: this index corresponds to the proportion of named entities in the text. A named entity is any real-world entity, such as persons, locations, organizations, products etc [40].

The full list of the considered features and a detailed description of each feature can be found in [48].
C. Machine Learning Methods

The textual features extracted from the abstract of the research projects are used in the classification process [22]. For each example (research proposal), we consider two possible classes (successful or unsuccessful). In a typical classification task, the dataset is divided into two parts: the training and test datasets. The training dataset is used to create the model (e.g., a Decision Tree), while the test dataset is used to evaluate the performance of the model. Here we used a standard procedure to split the original dataset into training and test datasets, the so-called 10-fold cross validation scheme [22]. To perform the classification, the following algorithms were used:

1. **k-nearest neighbors (kNN):** In order to classify an unknown (unlabeled) instance, the algorithm first selects the $k$ nearest instances in the training dataset. The class associated to the unknown instance corresponds to the majority class observed in the selected $k$-set. The parameter $k$ is a parameter to be optimized [4]. In the results section, we report the best results obtained for different values of $k$.

2. **Support Vector Machines (SVM):** In this method, instances from different classes are divided by different spaces. These spaces are generated during the training phase. The main objective of this class of methods is to find a separation hyperplane between two or more classes. One of the main parameters of this method is the kernel used to create the discriminative hyperplane. In this paper, we used the optimization strategy described in [4, 46].

3. **Naive Bayes:** This method relies on the Bayesian optimal decision rule to perform a classification. Let $m = \{f_1, f_2, \ldots\}$ be the set of features used to characterize a research proposal (i.e., the features described in Section III B). The class $c$ (successful or unsuccessful) assigned to a research proposal satisfies the following condition:

$$P(c|m) \geq P(c_k|m),$$

for every class $c_k \neq c$, where $P(c_k|m)$ is the probability of the $k$-th class to have a set of features $m$. Because $P(c_k|m)$ is not available in most cases, the Bayes’ theorem can
be used to find $c$:
\[
c = \arg \max_{c_k} P(m|c_k) \frac{P(c_k)}{P(m)} P(c_k).
\] (3)

$P(m)$ is the same for every class $c_k$, therefore the above equation can be simplified to:
\[
c = \arg \max_{c_k} P(m|c_k)P(c_k) = \arg \max_{c_k} \left[ \log P(m|c_k) + \log P(c_k) \right].
\] (4)

Assuming attribute independence, the class assigned to a new instance from the test dataset is computed as:
\[
c = \arg \max_{c_k} \left[ \sum_{f_i} \log P(f_i|c_k) + \log P(c_k) \right].
\] (5)

For the particular case of balanced datasets, $P(c_k)$ is uniform. Therefore,
\[
c = \arg \max_{c_k} P(m|c_k).
\] (6)

4. **Decision Trees**: the method based on Decision Trees uses a data structure composed of nodes and edges to represent the recognized patterns. In particular, a tree is a particular type of connected graph with the restriction that there is no cycle in such structure [15]. Nodes represent attributes and edges correspond to the decision taken in different tests performed on the respective node. An example of decision tree is provided in Figure 1. The classification process starts at the root node (see Figure 1) and continues until a leaf node (i.e. a node with no children) is reached. The class assigned to the instance in the test set corresponds to the one stored in the respective leaf node. While this process is used to classify a new instance, a decision tree should be created during the training phase. This requires the definition of a measurement to identify the most discriminative attribute at each phase (i.e. node) of the classification process. A well-known measure used to identify the relevance of features is the Kullback–Leibler divergence. In the training dataset $D_{TR}$, the relevance of each feature $f_i$ is computed as:
\[
\mathcal{K}(D_{TR}, f_i) = \mathcal{H}(D_{TR}) - \mathcal{H}(D_{TR}|f_i).
\] (7)
FIG. 1. Example of decision tree used for classification. The classification process for a new instance starts at the root node. Consider a new instance that should be classified. This new instance is described by the vector of features \((f_1 = x > L_A, f_2 = y < L_B, f_3 = z)\). The first test \((f_1 > L_A)\) leads the decision to the upper child node. Because the result of the current test fails (i.e. \(f_2 > L_B\)), this new instance is classified as an unsuccessful research proposal. In a similar fashion, an instance described by \((f_1 = q < L_A, f_2 = u, f_3 = v < L_C)\) would be classified as a successful research proposal.

where \(H(D_{TR})\) is the entropy of the training dataset and \(H(D_{TR}|f_i)\) is the entropy of the training dataset considering the separation of classes obtained with the \(i\)-th feature \([22, 26]\).

In addition to traditional decision trees, we also used random forests \([13]\). The latter has the advantage of avoiding the tendency of decision trees to overfit the training set \([13]\). All results obtained with decision trees and random forests are reported as DTrees in the Results sections.

5. Artificial Neural Networks (ANN): artificial neural networks are not a recent approach in the machine learning area, but have been widely used in recent years owing to the recent advancements in the deep learning area \([33]\). The most basic unit in a neural network is the perceptron. According to this model, the activation of a neuron depends on both input signals and transfer functions \([29]\). The activation can be considered as the perceptron output. Let \(a_i\) be the \(i\)-th input and \(w_i\) the weight associated to \(a_i\). The output depends on the linear combination of input as weights, according to the value \(s = \sum_i w_i a_i + b\), where \(s\) is the input used as reference to the transfer learning function.
and \( b \) is a constant value. The transfer learning function may assume many different forms \[29\]. If one chooses the Heaviside function, for example, the neuron if activated if \( s \) is above an established threshold. The adequate choice of weights allows the neural network to effectively process the input in order to yield the expected output (class). Several algorithms have been designed to establish optimized synaptic weights \[29\]. One simple approach is to initially assign random weights and then update the values according to the observed error, i.e. the difference between the generated and expected outputs. Here we considered as neural network approach the multi-layer perceptron (MLP) \[29\], a simple yet effective approach in many scenarios \[1\].

D. Textual complexity measurements relevance

In order to evaluate the relevance of features when identifying successful proposals, we used a feature relevance method that is based on decision trees. The relevance method uses the Gini impurity measurement to decide how discriminative is a partition of the dataset \[41\]. The Gini impurity is defined as the probability of incorrectly classifying an instance if it were randomly classified according to the class distribution observed in the dataset. It is computed as:

\[
G = \sum_{i \in C} p_i (1 - p_i),
\]

where \( C \) is the set of classes. In our study, \( C = \{ \text{successful, unsuccessful} \} \). \( p_i \) is the probability of choosing an instance from the \( i \)-th class in the considered subset.

As depicted in Figure 1, each tree node is associated with a feature. A feature is relevant in a node if it yields a decrease in the Gini impurity (\( \Delta G \)) for the considered dataset. The decrease in impurity for each tree node is computed as

\[
\Delta G = G_B - \beta_L G_L - \beta_R G_R,
\]

where \( G_B \) is the Gini impurity before the dataset is split in the respective node and \( G_L \) and \( G_R \) are the Gini impurity obtained in the left and right child nodes, respectively. \( \beta_L \) and \( \beta_R \) are normalization factors to account for the number of instances falling in the left and right child nodes. This means that a higher weight is associated to the split region comprising more examples. Finally, the relevance of a given feature \( m_i \) is computed as the average
decrease in impurity observed in all nodes in which \( m_i \) is used.

To illustrate the process of computing the Gini impurity for a given split of the dataset, we provide an example in Figure 2. The original dataset with two classes and two features is shown in the left panel. Because there are 16 positive and 16 negative examples, the probability of misclassifying a randomly selected instance is 50% (i.e. \( G_B = 50\% \)). After the dataset is split (see right panel), two subsets are created. In the left subset, the impurity is zero, because all instances belong to the same class. In the right subset, the impurity is computed according to equation 8:

\[
G_R = \frac{1}{17} \left(1 - \frac{1}{17}\right) + \frac{16}{17} \left(1 - \frac{16}{17}\right) = 0.1107.
\]

The proportion of data in the left and right subsets are respectively \( 15/32 \) and \( 17/32 \). Thus, the decrease in impurity, \( \Delta G \), as defined in equation 9, is given by:

\[
\Delta G = G_B - \frac{15}{32} G_L - \frac{17}{32} G_R = 0.4412.
\]

In other words, the split for the considered feature yield a reduction of \( \Delta G = 0.4412 \) in the impurity of the original dataset.

IV. RESULTS AND DISCUSSION

In this section, we discuss the obtained results. Our analysis is divided into three sections. In Section IV A, the performance for different features and machine learning methods is reported. In Section IV B, we discuss whether the discriminability varies significantly when considering different languages (Portuguese and English). Finally, in Section IV C, we perform an analysis of features relevance.

A. Performance analysis

In this section, we start the discussion of results by considering the accuracy rates obtained with complexity measurements extracted from title and abstracts (in Portuguese). The obtained results are shown in Table [I]. We show, for each considered dataset (Medicine,
FIG. 2. Computing the decrease in Gini impurity for a small dataset with two classes. For each class, there are 16 instances. In the original dataset, the probability of misclassification is high for a randomly drawn instance is high, i.e. $G = 0.50$. After the original dataset is split in two subsets, the discrimination of classes becomes almost perfect. This leads to a high variation in the Gini impurity, i.e. $\Delta G = 0.4412$.

Dentistry and Veterinary Medicine) the accuracy rate obtained from the machine learning methods considered in this study. The best results for each dataset were found to be statistically significant. They are highlighted in Table I. The success of a research proposal (according to the adopted success criteria) could be predicted with an accuracy of roughly 80%, for proposals in the areas of Medicine and Veterinary Medicine. An even better prediction rate was found for proposals in the Dentistry area ($\simeq 83\%$). These results suggest that the complexity of texts could be a component that can be used to predict the scientific output of research proposals.

TABLE I. Accuracy rate obtained when classifying research projects as successful or unsuccessful using Coh-metrix features [28] for Portuguese. Three different datasets were considered: Medicine (MED), Dentistry (DENT) and Veterinary Medicine (VET). The best results for each dataset are highlighted. In most cases, the best accuracy rate is obtained with decision trees.

| Method | Medicine Accuracy (%) | Dentistry Accuracy (%) | Vet. Medicine Accuracy (%) |
|--------|-----------------------|------------------------|----------------------------|
| DTrees | 79.66 ± 0.85          | 83.33 ± 0.68           | 79.61 ± 1.17               |
| SVM    | 79.12 ± 1.26          | 78.97 ± 1.06           | 80.32 ± 1.51               |
| kNN    | 66.08 ± 1.37          | 66.91 ± 1.89           | 63.72 ± 2.62               |
| Bayes  | 51.16 ± 0.93          | 51.93 ± 0.98           | 52.06 ± 1.38               |
| MLP    | 76.00 ± 1.34          | 79.92 ± 1.07           | 76.01 ± 1.75               |
When considering the different strategies for supervised classification and the set of Cohmetrix features, apart from the Veterinary Medicine dataset, the best results were obtained with Decision Trees. For the VET dataset, the highest accuracy was obtained with the SVM classifier. However, the difference between Decision Trees and SVM was found to be not significant. The worst classification system, in this case, were found to be kNN and Naive Bayes. Interestingly, the performance of the Naive Bayes was not even significantly different from the performance expected for a random classifier.

While the results shown in Table I considers as features only complexity factors of language, it would be interesting to analyze if improved results can be obtained when topical features are used to predict proposals success. For this analysis, we considered the tf-idf representation of texts. For this specific analysis, the classification considered different parts of the proposal. In the experiments, we considered the title, the subject, a combination of title and subject, and two variations of the tf-idf representation of the abstracts. The subject is provided by researchers and corresponds to a few words representing the corresponding research area. For the tf-idf representation of abstracts, the adopted approach selected the $X$ most frequent words as features. In the approaches referred to as Abstract$^{(1)}$ and Abstract$^{(2)}$, we used $X = 1, 100$ and $X = 7, 196$ words, respectively.

In Table II we show the results obtained for different classifiers. The best results for each dataset are highlighted. Once again, the best results are significant: 80.3%, 84.0% and 79.9% respectively for the MED, DENT and VET datasets. The accuracy observed in the Dentistry area once again was found to be slightly higher than the accuracy found in the other areas. These results also suggest that the frequency-based features also play an important role in predicting success of the considered datasets. Interestingly, excellent performance was obtained with Decision Trees: in all three datasets, the highest accuracy was obtained with this classifier. The multilayer perceptron also yielded results similar to those obtained with decision trees. Differently from the results obtained with text complexity features, the SVM classifier achieved a low classification performance.

Table II also reveals that particular fragments of the research proposal are more discriminative than others. Considering the best classifier in all three datasets (i.e. the decision tree method), we found that the best results occur when the abstract is taken into account. There is no significant difference, however, between the strategies Abstract$^{(1)}$ and Abstract$^{(2)}$. An excellent performance is also observed when considering both the title and the research pro-
TABLE II. Results based on the frequency (tf-idf) considering different fragments of research projects: the title, the subject, a combination of title and subject and the abstract. For the latter strategy, we selected the X most frequent words as features. In the approaches referred to as Abstract(1) and Abstract(2), we used X = 1,100 and X = 7,196 words, respectively.

| Features     | Research Projects on Medicine |          |          |          |          |
|--------------|-------------------------------|----------|----------|----------|----------|
|              | DTrees | SVM   | kNN     | Bayes    | MLP      |
| Title        | 76.85 ± 0.83 | 55.64 ± 1.04 | 66.75 ± 1.53 | 61.32 ± 0.99 | 77.49 ± 0.67 |
| Subject      | 75.65 ± 1.46 | 53.23 ± 1.35 | 66.64 ± 1.63 | 61.33 ± 1.38 | 75.24 ± 1.56 |
| Tit. + Sub.  | 79.06 ± 0.99 | 56.81 ± 2.18 | 67.50 ± 1.07 | 65.19 ± 1.22 | 77.13 ± 1.02 |
| Abstract(1)  | 79.59 ± 1.15 | 66.37 ± 1.30 | 66.43 ± 3.21 | 66.42 ± 1.37 | 78.90 ± 1.37 |
| Abstract(2)  | 80.31 ± 0.55 | 59.90 ± 1.03 | 64.19 ± 1.47 | 76.78 ± 0.93 | 79.20 ± 1.40 |

| Features     | Research Projects on Dentistry |          |          |          |          |
|--------------|-------------------------------|----------|----------|----------|----------|
|              | DTrees | SVM   | kNN     | Bayes    | MLP      |
| Title        | 81.19 ± 1.84 | 58.90 ± 1.63 | 68.35 ± 1.52 | 61.32 ± 0.99 | 81.88 ± 1.33 |
| Subject      | 80.00 ± 1.74 | 57.58 ± 2.58 | 69.00 ± 1.73 | 61.33 ± 1.38 | 79.78 ± 1.65 |
| Tit. + Sub.  | 81.85 ± 1.07 | 60.74 ± 2.09 | 71.31 ± 1.40 | 65.10 ± 1.22 | 81.82 ± 0.87 |
| Abstract(1)  | 83.36 ± 0.81 | 70.95 ± 1.08 | 69.03 ± 1.03 | 66.42 ± 1.37 | 82.65 ± 0.65 |
| Abstract(2)  | 83.97 ± 1.45 | 62.91 ± 1.89 | 67.73 ± 2.09 | 76.78 ± 0.93 | 82.79 ± 1.26 |

| Features     | Research Projects on Veterinary Medicine |          |          |          |          |
|--------------|------------------------------------------|----------|----------|----------|----------|
|              | DTrees | SVM   | kNN     | Bayes    | MLP      |
| Title        | 76.10 ± 0.99 | 57.39 ± 2.18 | 64.45 ± 2.22 | 60.26 ± 2.37 | 74.97 ± 1.56 |
| Subject      | 72.19 ± 2.07 | 58.55 ± 1.89 | 65.26 ± 3.38 | 60.14 ± 1.82 | 73.46 ± 0.70 |
| Tit. + Sub.  | 77.42 ± 1.56 | 61.88 ± 1.78 | 66.73 ± 2.31 | 63.15 ± 1.75 | 75.68 ± 1.26 |
| Abstract(1)  | 79.93 ± 1.53 | 69.48 ± 1.75 | 65.39 ± 2.57 | 64.70 ± 1.77 | 78.38 ± 1.24 |
| Abstract(2)  | 79.56 ± 1.29 | 64.74 ± 1.83 | 65.84 ± 2.70 | 74.76 ± 1.64 | 79.81 ± 1.11 |

Proposal subject. It is also worth noting that a significant improvement in performance occurs for the Naive Bayes when additional features are considered in the strategy based on the abstract. In all three datasets, we found an improvement of roughly 10%.

B. Language dependence

As mentioned in Section [IllA], the abstract of each research proposal is available in two languages: Portuguese and English. The results reported in Section [IVA] were obtained for textual data in Portuguese. Here we analyze whether there is a significant difference in performance when considering abstracts in English.

The results obtained when considering complexity measurements are shown in Table [III]. The best results for each language and research area are highlighted. When comparing the best results for Portuguese and English, we found no significant difference in performance.
for both MED and VET datasets. A slightly difference in performance was found for the DENT dataset. In this case, the best discriminability rate found for Portuguese is roughly 4% higher than the best accuracy rate obtained using abstracts in English. We also note that good results for English are obtained with SVM, Decision Trees and MLP.

TABLE III. Accuracy rate obtained when discriminating research proposals as successful or unsuccessful. We used complexity features to characterize the texts. The results reveal that only a small difference in performance is observed when comparing Portuguese and English abstracts. The best results for each dataset and language are highlighted.

| Method     | Projects on Medicine | Portuguese | English |
|------------|----------------------|------------|---------|
| DTrees     | 79.66 ± 0.85         | 80.37 ± 1.15 |
| SVM        | 79.12 ± 1.26         | 79.48 ± 1.56 |
| kNN        | 66.08 ± 1.37         | 66.79 ± 3.02 |
| Bayes      | 51.16 ± 0.93         | 53.95 ± 1.55 |
| MLP        | 76.00 ± 1.34         | 78.35 ± 1.83 |

| Method     | Projects on Dentistry | Portuguese | English |
|------------|-----------------------|------------|---------|
| DTrees     | 83.33 ± 0.68          | 77.26 ± 1.90 |
| SVM        | 78.97 ± 1.06          | 79.65 ± 1.80 |
| kNN        | 66.91 ± 1.89          | 65.91 ± 2.41 |
| Bayes      | 51.93 ± 0.98          | 54.49 ± 1.30 |
| MLP        | 79.92 ± 1.07          | 77.81 ± 1.79 |

| Method     | Projects on Vet. Medicine | Portuguese | English |
|------------|---------------------------|------------|---------|
| DTrees     | 79.61 ± 1.17              | 78.64 ± 2.94 |
| SVM        | 80.32 ± 1.51              | 79.48 ± 1.72 |
| kNN        | 63.72 ± 2.62              | 66.68 ± 1.28 |
| Bayes      | 52.06 ± 1.38              | 53.18 ± 3.33 |
| MLP        | 76.01 ± 1.75              | 77.55 ± 2.0  |

In Table IV, we show the results obtained for the English datasets when considering tf-idf features. The best results for each dataset and language are also highlighted. The analysis of the best results reveals a dependence with language that is similar to the one observed in Table III. Apart from the DENT dataset, there no significant difference in performance when comparing the best results obtained for abstracts written in Portuguese and English. The best results for English occur with Decision Trees (MED) and SVM (DENT and VET). The multilayer perceptron classifier also displayed a good performance in abstracts written in English.

The accuracy rates shown in Tables III and IV suggest that the discriminability of research
TABLE IV. Accuracy rate obtained when discriminating research proposals as successful or unsuccessful. We used tf-idf features to characterize the proposal abstracts. The best results for each dataset and language are highlighted.

| Method  | Projects on Medicine |          |          |
|---------|----------------------|----------|----------|
|         | Portuguese           | English  |
| DTrees  | 80.31 ± 0.55         | 80.71 ± 1.42 |
| SVM     | 66.37 ± 1.30         | 57.00 ± 0.79 |
| kNN     | 67.50 ± 1.07         | 76.75 ± 1.72 |
| Bayes   | 76.78 ± 0.93         | 68.57 ± 1.25 |
| MLP     | 79.20 ± 1.40         | 79.72 ± 1.53 |

| Method  | Projects on Dentistry |          |
|---------|-----------------------|----------|
|         | Portuguese            | English  |
| DTrees  | 83.97 ± 1.45          | 78.70 ± 2.32 |
| SVM     | 70.95 ± 1.08          | 48.84 ± 0.05 |
| kNN     | 71.31 ± 1.40          | 67.47 ± 3.71 |
| Bayes   | 76.78 ± 0.93          | 77.26 ± 2.86 |
| MLP     | 82.79 ± 1.26          | 79.53 ± 1.75 |

| Method  | Projects on Vet. Medicine |          |
|---------|---------------------------|----------|
|         | Portuguese                | English  |
| DTrees  | 79.93 ± 1.53              | 77.58 ± 1.93 |
| SVM     | 69.48 ± 1.75              | 52.69 ± 0.99 |
| kNN     | 66.73 ± 2.31              | 67.42 ± 2.10 |
| Bayes   | 74.76 ± 1.64              | 76.98 ± 2.55 |
| MLP     | 79.81 ± 1.11              | 80.62 ± 1.73 |

proposals is weakly dependent upon the language, at least for the considered research areas and languages. This suggests that a machine learning approach applied to other languages is a potential tool to identify proposals more likely to yield at least one publication. Such an approach could consider both complexity and topical features, as no significant different have been found when comparing both types of features.

C. Feature relevance

The results in the previous section showed that there is a dependence between text features and the output of research proposals. The success of specific proposals according to tf-idf features might be a consequence of the fact that some subjects and topics are more visible than others, for several reasons [39, 49]. A similar behavior has been reported at the journal level, since interdisciplinary papers tend to accrue more citations than papers that are specific to a single discipline [35, 36]. The importance of text complexity (i.e.
topic independent) features is not as clear. In order to better understand in future works the reasons why text complexity plays an important role in identifying successful research proposals, in this section we provide an analysis of the main complexity features responsible for the discriminability of research proposals.

For the analysis of features relevance, we used the strategy described in Section IV A, which is based on the Decision Tree algorithm. We used this strategy because Decision Trees displayed excellent results in the previous performance analysis. For each dataset, we ranked in decreasing order the complexity features according to the value of $\Delta G$, which corresponds to the average decrease in impurity for tree nodes involving that feature. Because of the cross-validation and balancing procedures, the ranking obtained by each feature varies in each considered subset of the dataset. In Figure 3 we show the ranking diagram depicting the average rank of the best ranked features for each research area.

An analysis of Figure 3 revealed that the best ranked features (in decreasing order) for each of the considered datasets were:

1. **Medicine**: (a) function word diversity, (b) standard deviation of noun occurrences, (c) total number of words, (d) preposition diversity and (e) Brunet index.

2. **Dentistry**: (a) mean noun phrase, (b) total number of words, (c) preposition diversity, (d) punctuation diversity and (e) standard deviation of words concreteness.

3. **Veterinary Medicine**: (a) named entity ratio text, (b) total number of words, (c) noun ratio, (d) Brunet index and (e) preposition diversity.

While some features, in average, seems to be considerably better than others in the diagram, the Critical Difference [19] (not shown in the diagram) reveals that there is no significant difference among these 5 best ranked features.

Some interesting patterns can be observed from the best ranked features. First, the total number of words seems to be a relevant feature for the classification. However, it is not possible to identify a single pattern (e.g. a correlation) between this feature and the research proposal output, since this feature can be used in different ways in different tree nodes. Other features that were found to be relevant for the classification accuracy are the Brunet index and the preposition diversity. These measurements show that not only the text length is important, but also the diversity of lexical items. This finding is
FIG. 3. Feature ranking diagram for the classification of research projects as successful or unsuccessful. For each dataset, we show the average ranking obtained by each of the considered Coh-metrix features. In Medicine, the best features were: (a) function word diversity; (b) nouns SD; (c) total number of words; (d) preposition diversity; and (e) Brunet index. In Dentistry, the best features were (a) mean noun phrase; (b) total number of words; (c) preposition diversity; (d) punctuation diversity; and (e) concreteness SD. In Veterinary Medicine, the best features were (a) NE ratio text; (b) total number of words; (c) noun ratio; (d) brunet index; and (e) preposition diversity.

compatible with studies correlating lexical diversity and writing quality [9]. The relevance of preposition diversity reveals that not only the diversity of semantic concepts might be relevant to discriminate successful proposals, but also stopwords (prepositions), i.e. words conveying no semantical meaning. This result reinforces the importance of text style when
identifying successful proposals [8, 11]. Surprisingly, the ‘concreteness’ of words also seems to play a role in the identification of successful DENT proposals. Such a relevance, though, is not evident in the other datasets, meaning that some features might be relevant only in particular research areas.

All in all, the results obtained in this section showed that particular word choices and the ability to construct a rich vocabulary might be correlated with the output observed in research proposals. From a linguistic point of view, it should be interesting to investigate in future works if any of the identified relevant features (and respective patterns) can be considered as marks of a high-quality writing. If papers resulting from well-written projects are themselves written in a similar high-quality style, one should expect that they are more likely to be published (provided that all other paper requirements and standards are met). This could explain the fact that the above features are relevant to detect successful proposals.

V. CONCLUSION

The development and advancement of science is fundamental for the evolution of society. A driving force towards the development of science are the preliminary ideas, which often lead to important developments in the near (or distant) future. While many ideas should be developed without restriction, in practice a limitation in resources hinders all research ideas from being developed at their highest potential. In practical terms, this means that many research proposals are not funded, and this may affect the success and diffusion of important ideas. In this context, it is clear that the funding decisions should be as effective as possible in order to avoid the waste of resources that could be otherwise invested in truly strong ideas.

Despite some criticisms, the role of peer review in identifying promising ideas remains undeniable [32]. As it happens in other bibliometrics contexts, it is still interesting to provide automatized tools that can assist humans in particular issues [2, 18, 49]. In this context, in this paper we analyzed whether textual features can be used to discriminate successful from unsuccessful research proposals. Given the nature of our dataset, we considered a machine learning setting where successful research proposals were those yielding at least one publication. As features, we focused on two types of linguistic attributes. First, we used complexity measures that are topic-independent. We also used, for comparison purposes,
a simple frequency-based approach. A dataset of research proposals funded by São Paulo Research Foundation (FAPESP-Brazil) was considered and analyzed in three distinct areas, namely Medicine, Dentistry and Veterinary Medicine.

Our analysis revealed several interesting findings. First, we found a high accuracy when using complexity measurements to characterize research proposals abstracts. We found an accuracy of 83.3% in a binary classification with decision trees in research proposals in the area of Dentistry. Similar results have been found for the other studied areas (Medicine and Veterinary Medicine). This result was found to be as good as the one obtained when classifying texts with tf-idf. Considering complexity measurements, we also found that among the evaluated classifiers, excellent performance was obtained with Decision Tree and SVM for all three considered datasets. We also found that the obtained results are robust to the considered language, since similar accuracy rates were found for proposals written in both English and Portuguese. A feature relevance analysis also revealed that text length and the vocabulary diversity are among the most discriminative features.

The results of this paper suggest that both complexity and topical features are effective in identifying successful research proposals, according to the adopted criteria for research proposal success. As a consequence, we believe that text analysis has a potential to assist the analysis of research proposals. In this paper, we limited the sense of success by considering that successful proposals are those yielding at least one publication. In future works, it is interesting to analyze other success criteria, including e.g. the number of publications, the reputation of respective the journals and conferences and other measurements derived from citation and usage counts [31, 47]. We also intend to incorporate additional features for the prediction, including text network-based attributes [8, 50, 52] and other features related to researchers and their respective institutes [10, 16].

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.
[1] D. E. Acuna, S. Allesina, and K. P. Kording. Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[2] D. R. Amancio. Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*, 105(3):1763–1779, 2015.

[3] D. R. Amancio, S. M. Aluisio, O. N. Oliveira Jr, and L. F. Costa. Complex networks analysis of language complexity. *EPL (Europhysics Letters)*, 100(5):58002, 2012.

[4] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. F. Costa. A systematic comparison of supervised classifiers. *PLoS ONE*, 9(4), 2014.

[5] D. R. Amancio, M. d. G. V. Nunes, O. Oliveira Jr, T. A. S. Pardo, L. Antiqueira, and L. F. Costa. Using metrics from complex networks to evaluate machine translation. *Physica A: Statistical Mechanics and its Applications*, 390(1):131–142, 2011.

[6] D. R. Amancio, O. N. Oliveira Jr, and L. F. Costa. Three-feature model to reproduce the topology of citation networks and the effects from authors’ visibility on their h-index. *Journal of Informetrics*, 6(3):427–434, 2012.

[7] D. R. Amancio, O. N. Oliveira Jr, and L. F. Costa. Using complex networks to quantify consistency in the use of words. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(01):P01004, 2012.

[8] D. R. Amancio, F. N. Silva, and L. F. Costa. Concentric network symmetry grasps authors’ styles in word adjacency networks. *EPL (Europhysics Letters)*, 110(6):68001, 2015.

[9] L. Antiqueira, M. d. G. V. Nunes, O. Oliveira Jr, and L. d. F. Costa. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, 373:811–820, 2007.

[10] H. F. Arruda, L. F. Costa, and D. R. Amancio. Using complex networks for text classification: Discriminating informative and imaginative documents. *EPL (Europhysics Letters)*, 113(2):28007, 2016.

[11] R. Arun, V. Suresh, and C. V. Madhavan. Stopword graphs and authorship attribution in text corpora. In *2009 IEEE International Conference on Semantic Computing*, pages 192–196. IEEE, 2009.
[12] K. W. Boyack, C. Smith, and R. Klavans. Toward predicting research proposal success. *Scientometrics*, 114(2):449–461, 2018.

[13] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[14] A. Cabezas-Clavijo, N. Robinson-Garcia, M. Escabias, and E. Jiménez-Contreras. Reviewers’ ratings and bibliometric indicators: Hand in hand when assessing over research proposals? *PLoS ONE*, 8(6), 2013.

[15] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009.

[16] E. A. Correa Jr, A. A. Lopes, and D. R. Amancio. Word sense disambiguation: A complex network approach. *Information Sciences*, 442:103–113, 2018.

[17] E. A. Corrêa Jr, F. N. Silva, L. F. Costa, and D. R. Amancio. Patterns of authors contribution in scientific manuscripts. *Journal of Informetrics*, 11(2):498–510, 2017.

[18] A. Daud, M. Ahmad, M. Malik, and D. Che. Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, 102(2):1687–1711, 2015.

[19] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[20] F. Didegah and M. Thelwall. Which factors help authors produce the highest impact research? collaboration, journal and document properties. *Journal of informetrics*, 7(4):861–873, 2013.

[21] Y. Ding. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1):187–203, 2011.

[22] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[23] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PLoS ONE*, 6(9), 2011.

[24] F. C. Fang, A. Bowen, and A. Casadevall. Nih peer review percentile scores are poorly predictive of grant productivity. *Elife*, 5:e13323, 2016.

[25] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[26] R. Garreta and G. Moncecchi. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.
[27] R. N. Germain. Healing the nih-funded biomedical research enterprise. *Cell*, 161(7):1485–1491, 2015.

[28] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, 2004.

[29] M. H. Hassoun et al. *Fundamentals of artificial neural networks*. MIT press, 1995.

[30] M. Hörlesberger, I. Roche, D. Besagni, T. Scherngell, C. François, P. Cuxac, E. Schiebel, M. Zitt, and D. Holste. A concept for inferring ‘frontier research’ in grant proposals. *Scientometrics*, 97(2):129–148, 2013.

[31] J. Hou and X. Yang. Social media-based sleeping beauties: Defining, identifying and features. *Journal of Informetrics*, 14(2):101012, 2020.

[32] J. P. Kassirer and E. W. Campion. Peer review: crude and understudied, but indispensable. *Jama*, 272(2):96–97, 1994.

[33] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[34] A. Letchford, H. S. Moat, and T. Preis. The advantage of short paper titles. *Royal Society open science*, 2(8):150266, 2015.

[35] L. Leydesdorff and I. Rafols. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1):87–100, 2011.

[36] L. Leydesdorff, C. S. Wagner, and L. Bornmann. Interdisciplinarity as diversity in citation patterns among journals: Rao-stirling diversity, relative variety, and the gini coefficient. *Journal of Informetrics*, 13(1):255–269, 2019.

[37] D. Li and L. Agha. Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, 348(6233):434–438, 2015.

[38] D.-C. Li, C.-W. Liu, and S. C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, 40(5):509–518, 2010.

[39] K. McKeown, H. Daume III, S. Chaturvedi, J. Paparrizos, K. Thadani, P. Barrio, O. Biran, S. Bothe, M. Collins, K. R. Fleischmann, et al. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696, 2016.

[40] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.
[41] S. Nembrini, I. R. König, and M. N. Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

[42] fapesp.br/en.

[43] bv.fapesp.br/en/6/regular-grants-2-year-grants.

[44] N. Onodera and F. Yoshikane. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4):739–764, 2015.

[45] C. E. Paiva, J. P. S. N. Lima, and B. S. R. Paiva. Articles with short titles describing the results are cited more often. *Clinics*, 67(5):509–513, 2012.

[46] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. F. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PLoS ONE*, 14(1), 2019.

[47] X. Ruan, Y. Zhu, J. Li, and Y. Cheng. Predicting the citation counts of individual papers via a bp neural network. *Journal of Informetrics*, 14(3):101039, 2020.

[48] C. Scarton and S. M. Aluísio. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*, volume 10. sn, 2010.

[49] F. N. Silva, D. R. Amancio, M. Bardosova, L. d. F. Costa, and O. N. Oliveira Jr. Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics*, 10(2):487–502, 2016.

[50] M. Stella. Modelling early word acquisition through multiplex lexical networks and machine learning. *Big Data and Cognitive Computing*, 3(1):10, 2019.

[51] M. Stella, S. De Nigris, A. Aloric, and C. S. Siew. Forma mentis networks quantify crucial differences in stem perception between students and experts. *PLoS ONE*, 14(10), 2019.

[52] M. Stella and A. Zaytseva. Forma mentis networks map how nursing and engineering students enhance their mindsets about innovation and health during professional growth. *PeerJ Computer Science*, 6:e255, 2020.

[53] M. Thelwall and T. Nevill. Could scientists use altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, 12(1):237–248, 2018.

[54] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.
[55] M. Wang, G. Yu, and D. Yu. Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications*, 387(18):4692–4698, 2008.

[56] Z. Xie, Z. Ouyang, P. Zhang, D. Yi, and D. Kong. Modeling the citation network by network cosmology. *PLoS ONE*, 10(3), 2015.

[57] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley. The science of science: From the perspective of complex systems. *Physics Reports*, 714:1–73, 2017.