UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Using complex networks and natural language processing to characterize and classify scientific items

Jorge Andoni Valverde Tohalino

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Jorge Andoni Valverde Tohalino

Using complex networks and natural language processing to characterize and classify scientific items

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
March 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

T627u

Tohalino, Jorge Andoni Valverde
Using complex networks and natural language processing to characterize and classify scientific items / Jorge Andoni Valverde Tohalino; orientador Diego Raphael Amancio . -- São Carlos, 2023.
147 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Research Grant Classification. 2. Keyword Extraction. 3. Complex Networks. 4. Natural Language Processing. 5. Bibliometric Analysis. I. Amancio , Diego Raphael , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Jorge Andoni Valverde Tohalino

Usando redes complexas e processamento de línguas naturais para caracterizar e classificar itens científicos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
Março de 2023

Este trabalho é dedicado
aos meus queridos pais
pela sua ajuda e amor incondicional

ACKNOWLEDGEMENTS

Agradeço primeiramente a Deus, pelo seu amor infinito, paciência e pelas diferentes oportunidades que ele me deu na vida e para terminar este trabalho de doutorado.

Agradeço a minha família, epecialmente a meus pais Jorge Antonio e Juana Irenia, pelo seu grande apoio e compreensão nestes anos que eu estive longe do Perú. Todo esse trabalho é dedicado a vocês meus queridos pais. Eu amo muito vocês!

Aos meus amigos peruanos, que me apoiaram muito em suas mensagens de incentivo e motivação durante todo esse período que eu estive fora do Peru.

Aos meus colegas e amigos do NILC, pelo apoio, amizade e aprendizado que tive com vocês.

Ao meu orientador Dr. Diego Raphael Amancio, pelo seu apoio, confiança e todos os ensinamentos ao longo deste projeto. Também aprecio sua paciência e sua disposição em me ajudar em qualquer momento.

A CAPES pelo apoio financeiro durante este trabalho de doutorado.

Ao ICMC e à USP por me darem a oportunidade de estudar e crescer como pessoa através deste trabalho de doutorado.

“Posso todas as coisas em Cristo que me fortalece.”
(Filipenses 4:13)

RESUMO

VALVERDE, J. A. **Usando redes complexas e processamento de línguas naturais para caracterizar e classificar itens científicos**. 2023. 147 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Processamento de Linguagem Natural (PLN) surgiu como uma área crítica de estudo para analisar grandes quantidades de dados textuais. No entanto, com o crescimento exponencial de big data, a análise de textos de diferentes tipos e tamanhos tornou-se mais desafiadora. Métodos existentes podem funcionar bem para conjuntos de dados específicos, mas podem não funcionar de maneira ideal para outras aplicações de texto. Por exemplo, analisar textos curtos, como títulos ou resumos de artigos científicos, pode ser desafiador porque esses textos podem conter uma quantidade limitada de informações, tornando difícil extrair insights valiosos usando abordagens de PLN tradicionais. Nesta tese, propomos uma nova metodologia que integra PLN, Redes Complexas (RC) e cienciometria/bibliometria para classificar e extrair tópicos importantes em textos científicos. Combinamos os conceitos de cada área de diversas maneiras para as tarefas de classificação de propostas de projetos de pesquisa e extração de palavras-chave. As abordagens de PLN forneceram diferentes maneiras de obter representações matemáticas de palavras e textos. Por exemplo, as representações vetoriais de palavras foram úteis para encontrar relações semânticas e contextuais para extração de palavras-chave, enquanto a representação vetorial de textos completos foi usada para tarefas de classificação. Também usamos abordagens baseadas em redes complexas para modelar relacionamentos entre textos como redes. Isso nos permite extrair informações relevantes por meio de propriedades estruturais e topológicas de redes. Em seguida, as métricas de centralidade de rede ajudaram a encontrar as palavras mais importantes em resumos e artigos de pesquisa, enquanto os métodos de detecção de comunidades foram eficientes em encontrar grupos de resumos de artigos com conteúdo semelhante. Também usamos conceitos de cienciometria e bibliometria para dois propósitos. Primeiro, extraímos características bibliométricas de pesquisadores brasileiros para a tarefa de classificação de propostas de projetos de pesquisa. Também usamos os padrões de citação de artigos científicos como fonte importante de informação para auxiliar nossa abordagem de extração de palavras-chave. Nossa pesquisa demonstra a importância de usar várias metodologias de diferentes áreas para extrair informações valiosas de textos curtos. A metodologia proposta nesta pesquisa pode ser usada posteriormente para outras aplicações de PLN e mineração de textos, como classificação de textos, agrupamento de textos e sumarização de documentos, especialmente quando os textos-alvo são pequenos e limitados em conteúdo.

Palavras-chave: Classificação de Projetos de Pesquisa, Extração de Palavras-chave, Redes Complexas, Processamento de Linguagem Natural, Análise Bibliométrica, Análise Cienciométrica.

ABSTRACT

VALVERDE, J. A. **Using complex networks and natural language processing to characterize and classify scientific items**. 2023. 147 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Natural Language Processing (NLP) has emerged as a critical area of study to analyze large amounts of textual data. However, with the exponential growth of big data, analyzing texts of different types and sizes has become more challenging. Existing methods may work well for specific datasets but may not perform optimally for other text applications. For example, analyzing short texts such as titles or abstracts of research papers could be challenging because these texts can contain a limited amount of information, making it difficult to extract valuable insights using traditional NLP approaches. In this thesis, we propose a new methodology that integrates NLP, Complex Networks (CN), and scientometrics/bibliometrics to classify and extract important topics in scientific texts. We combined the concepts from each area in various ways for research grant classification and Keyword Extraction (KE) tasks. NLP approaches provided different ways to obtain mathematical representations of words and texts. For example, word vector representations were useful in finding semantic and contextual relationships for keyword extraction, while vector representation of full texts was used for classification tasks. We also used complex network-based approaches to model relationships between texts as networks. This enables us to extract relevant information through structural and topological properties of networks. Then, network centrality metrics helped to find the most important words in abstracts and research papers, while community detection methods were efficient in finding groups of paper abstracts with similar contents. We further employed scientometric and bibliometric concepts for two purposes. First, we extracted bibliometric features from Brazilian researchers for the grant classification task. We also used the citation patterns from research papers as an important source of information to assist our keyword extraction approach. Our research demonstrates the importance of using multiple methodologies from different areas to extract valuable information from short texts. This framework can be further used for other NLP and text mining applications such as text classification, text clustering, and document summarization, particularly when the target texts are small and limited in content.

Keywords: Research Grant Classification, Keyword Extraction, Complex Networks, Natural Language Processing, Bibliometric Analysis, Scientometric Analysis.

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BERT	Bidirectional Encoder Representations from Transformers
CN	Complex Networks
KE	Keyword Extraction
NLP	Natural Language Processing
tf-idf	Term frequency – Inverse document frequency
VSM	Vector Space Model

CONTENTS

CHAPTER 1

INTRODUCTION

In recent years, the amount of information available on the Internet has grown exponentially. This information includes a vast array of databases containing images, videos, music, and textual documents. In this Ph.D. program, our focus was on analyzing text-oriented information. The prevalence of internet access has led to a surge in user-generated content, resulting in a wide variety of texts of varying types and sizes. Examples of longer texts include literary books, research papers, encyclopedic websites, blog posts, theses, and dissertations, among others. In contrast, shorter texts include abstracts of research papers, social media comments (such as those on Twitter, Facebook, or YouTube), product reviews, emails, and more (LI *et al.*, 2019). To handle this massive volume of textual information, the fields of Natural Language Processing (NLP) and text mining have emerged. These areas aim to process, organize, extract important topics, summarize, and classify large amounts of textual data (RIBALDO *et al.*, 2012).

Text classification and keyword extraction are two key tasks in natural language processing (NLP) that have been extensively studied by the research community to help manage large text datasets. While text classification involves extracting features from texts and categorizing them using machine learning algorithms (KOWSARI *et al.*, 2019), keyword extraction focuses on identifying the most relevant words or topics that best describe the content of a document or set of documents (TIMONEN *et al.*, 2012). Although a wide range of methods have been proposed in the literature for addressing these tasks in larger texts, analyzing shorter texts remains a significant challenge. Traditional NLP methods rely on statistical models that require a considerable amount of data to train, and they typically assume that the input text is composed of complete sentences or paragraphs. Short texts, on the other hand, may contain incomplete or fragmented sentences, and their brevity makes it difficult to extract meaningful information using traditional NLP approaches. Additionally, short texts are often sparse in terms of word frequency, and traditional NLP methods may not perform well when there are limited contextual cues available to disambiguate the meaning of the words (CHEN; HOU; GAO, 2020). Therefore, working with short texts for various NLP applications may require additional steps compared to larger texts. For example, several works have considered using external information such as semantics, word relations, or background knowledge to analyze smaller texts (LI *et al.*, 2019; CHEN; HOU; GAO, 2020). In this sense, exploring methodologies for short texts can provide new research opportunities, leading to the development of novel methods and techniques that can be applied to other text mining and NLP tasks. In this Ph.D. program, we focused on text classification and keyword extraction tasks for scientific texts, ranging from short texts such as paper abstracts and research grant summaries to longer texts such as full research papers. We chose scientific texts as they are relatively unexplored in NLP applications, providing ample scope for research. Moreover, we employed scientometric and bibliometric concepts to develop the text classification and keyword extraction tasks. We also used NLP approaches and complex networks to characterize the input texts and extract relevant patterns for our proposed methods.

Over the years, a variety of methods have been explored for representing texts as numerical vectors for use in different NLP applications. The classical Vector Space Model (VSM) was the first approach used for information retrieval tasks (SALTON; WONG; YANG, 1975). These models were based on word frequency and the Term frequency – Inverse document frequency (tf-idf) weighting to assign an importance value to each element of a text vector. However, these methods have several weaknesses, as they ignore the semantics of words and the word order in texts. As a result, related words or synonyms may have completely different vectors, and shuffling the text would result in a meaningless version having the same vector representation as the original text (MIKOLOV *et al.*, 2013b). To overcome the weaknesses of vector space models, methods based on word embeddings and BERT embeddings were proposed (MIKOLOV *et al.*, 2013b; MIKOLOV *et al.*, 2013a; KENTON; TOUTANOVA, 2019). These approaches consider the semantic meaning of words to encode their meanings, such that words with similar meanings have similar vector representations. These methods have been successfully used in various NLP applications, such as information retrieval, text classification, question answering, document summarization, and keyword extraction (STEIN; JAQUES; VALIATI, 2019; ZHOU *et al.*, 2015; MOHD; JAN; SHAH, 2020; WANG; LIU; MCDONALD, 2015). In our research, we use various approaches for different purposes. For instance, we consider the concepts of word embeddings to enrich the relationships between similar words in a word co-occurrence network for keyword extraction. We also use frequency and tf-idf models for text clustering, text classification, and keyword detection tasks. Vector space models can be used as features for a supervised classifier and are also useful for finding important words in keyword extraction applications (LI; FAN; ZHANG, 2007).

In recent years, studies in complex networks have gained more attention as they have been found to be an efficient tool for representing real-world phenomena. Complex networks are graphs with particular statistical and topological properties, as observed in graph models like the Watts-Strogatz Small-World model and the Barabási-Albert Scale-Free model (WATTS; STROGATZ, 1998; ALBERT; BARABÁSI, 2002). Complex network concepts have been widely used in text mining applications to model and process texts. For instance, lexical networks have been used to construct spell checkers, where each node represents a word and edges are based on the orthographic distance between two words (CHOUDHURY *et al.*, 2007). Word co-occurrence networks have been used for text classification and keyword extraction, where centrality metrics can be used as features for machine learning algorithms (QUISPE; TOHALINO; AMANCIO, 2021). Sentence networks have been used for automatic document summarization, where the most central sentences can be included in a final summary (TOHALINO; AMANCIO, 2018). Syntactic networks have been used to investigate language acquisition (ANTIQUEIRA *et al.*, 2007), among other applications. In this research, we examined word co-occurrence networks for keyword extraction and studied the effect of adding virtual edges to the network via word embeddings. Our results showed that the inclusion of these edges improved the performance of the keyword extraction methods. We evaluated several centrality measurements to assign importance values to each word. Furthermore, we used complex network and bibliometric concepts to model research papers as citation networks. We applied community detection methods to find groups of related research papers and extracted keywords according to their importance inside and outside these groups. Our approach allowed us to identify important topics and trends in research fields and discover new potential collaborators. Overall, complex network concepts have proven to be valuable tools for text mining applications, allowing us to explore and discover relationships and patterns in textual data that might be difficult or impossible to detect through traditional methods.

Several NLP applications have emerged to organize large textual databases, while scientometrics and bibliometrics have become relevant for carrying out quantitative and qualitative studies of scientific activity (VINKLER, 2010). By analyzing the scientific literature, these areas can track the evolution or decline of scientific fields and identify the emergence of new areas. They consider quantitative indicators, such as the number of publications, and impact indicators, which are reflected in the number of citations obtained from published articles (MINGERS; LEYDESDORFF, 2015). Bibliometric networks, including citation analysis, co-citation analysis, bibliographic coupling, co-author analysis, and co-word analysis, are also studied to characterize the importance of scientific production of researchers, scientific papers, and scientific journals (ZUPIC; ČATER, 2015). This research applies concepts from scientometric and bibliometric analyses to grant classification and keyword extraction. For the former, we extract bibliometric information from the primary investigators of research grants to generate features for a supervised classifier. In the latter approach, we utilize citation networks to identify groups of related papers and develop a methodology to extract the most prominent keywords based on their relevance to each paper group.

This thesis raises several research questions, which are outlined below:

- How can we effectively handle the sparse and noisy nature of short text data in NLP tasks such as text classification and keyword extraction?

- What challenges arise when working with smaller texts in NLP applications, and how can they be addressed through the development of new methodologies and techniques?

- How can we identify important entities or concepts in short text data without access to the full context or background knowledge?
- How can we incorporate external knowledge or context to improve the performance of NLP models on short text data?
- How can we efficiently cluster and group short text data based on similarities and differences, especially when the number of clusters is not known beforehand?
- What centrality measures in word co-occurrence networks can be used as features for machine learning algorithms in text classification and keyword extraction tasks?
- Can the addition of virtual edges to word co-occurrence networks via word embeddings improve the performance of keyword extraction methods?
- How can scientometric and bibliometric analysis be applied to grant classification and keyword extraction?
- How can machine learning models be trained on scientific papers to predict scientific impact or identify emerging areas of research?

To address these research questions, this monograph is organized as a collection of four research papers, which have been published or submitted. Each chapter of the monograph corresponds to a research paper that focuses on a specific aspect related to the research questions. These papers addressed the grant classification and keyword extraction tasks for scientific items. In each chapter, before presenting the article, we provided the motivation and contributions derived from each paper. The articles are presented chronologically and they are organized according to the main topic they addressed. In relation to the background information, all papers are self-contained, however, we also supplied a background chapter presenting some concepts about natural language processing, complex networks, and scientometrics that are relevant for the carried study. This manuscript is organized in the following Chapters:

- The Chapter 2 presents the background information, where the main concepts related to natural language processing, complex networks, and scientometrics are briefly explained.
- The Chapter 3 and the Chapter 4 included two research papers that addressed the grant classification task. The goal of these papers was to classify research grants according to their productivity or success they achieved over the years. The paper presented in the Chapter 3, extracted text features from the abstracts of the research proposals, while the paper displayed in the Chapter 4 considered bibliometric features extracted from the academic history of the main investigators of each research grant.

- The Chapter 5 and the Chapter 6 comprised two articles that focused on the keyword extraction task. In Chapter 5, we presented a method that modeled the texts as word co-occurrence networks. We then used word embeddings to enrich the relationships between words. This approach used centrality measurements to find the most central (relevant) words. In Chapter 6, we described a method based on citation networks and community detection for extracting keywords from paper abstracts.

- Finally in Chapter 7 we presented the conclusions which included the main contributions of this research, limitations, and possible future works.

CHAPTER 2

BACKGROUND

In this chapter, we briefly explain the most relevant concepts of the three areas that we covered in this research work. In Section 2.1, we detailed the main approaches about Natural Language Processing and text mining. The concepts related to complex networks are presented in Section 2.2. Finally, in Section 2.3, we mentioned some approaches related to scientometric and bibliometric analysis. In some cases, we briefly explain some important concepts covered in this research, however, other approaches were only mentioned in this section because their definitions were already given in the following chapters.

2.1 Natural Language Processing and text mining

Natural Language Processing (NLP) is a subfield of artificial intelligence, and linguistics focused on the study of automatic generation and understanding of natural human languages (GUIDA; MAURI, 1986). The goal of the NLP area is to interpret human language by evaluating text, speech, or grammatical syntax. In this sense, NLP seeks to extract the grammatical structure and syntactic meaning from texts and speech (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). On the other hand, text mining is an area derived from data mining with the aim of extracting information about textual content. Text mining uses various methods based on statistics, patterns or correlations between words to analyze each text. While NLP focuses on the meaning of the content, text mining considers the structure of texts (HOTHO; NÜRNBERGER; PAASS, 2005). Therefore, these two areas have been used together for many text-oriented applications.

A widely researched stage in these areas is the transformation of each input text into a form that can be understood by computers. In this sense, several approaches have been proposed to obtain the mathematical representation of textual documents. Most of these methods extracted representative vectors from each text to be used in different NLP and text mining applications. In this chapter we will briefly detail some methods, from the simplest ones such as vector space models, to more recent methods that used sophisticated methodologies such as BERT embeddings.

2.1.1 Vector Space Models (VSM)

One of the first approaches to model texts as numerical vectors was the vector space model. Sentences, paragraphs or complete texts are represented by vectors of identifiers in a multidimensional linear space of size N, where N is the total number of unique terms that appear in the document set (SALTON; WONG; YANG, 1975). Each element from the vector represents the contribution of each term for the representation of the whole dataset. The contribution or weight of each term could be computed in different ways. In this sense, there were proposed three algorithms derived from the vector space model: i) the Boolean model (LANCASTER; GALLUP, 1973), which is based on the presence or absence of words in the document; ii) the word frequency model, which counts the frequency of words in the whole set of documents; and the tf-idf model (SALTON; WONG; YANG, 1975), which is an improvement of the frequency-based model. Because of their simplicity, both the Boolean and frequency model presented several weaknesses, for this reason, the tf-idf model was proposed as an improvement to these techniques. The tf-idf which stands for Term Frequency - Inverse Document Frequency, is an approach that quantifies how relevant is a word in a document collection. The tf-idf weight increases proportionally to the number of times a word appears in the text, but it is compensated by the word frequency in the whole dataset. In this sense, the tf estimates how often a word occurs in a text, and idf estimates if the word is common or rare in the document collection.

Despite their popularity and good performance in some NLP applications, vector space models have several weaknesses. For example, these models ignore the semantics of words, in this sense, words with similar meanings could have very different representative vectors. Furthermore, vector space models present vectors with high dimensionality, since the size of the vectors depends on the number of words in the vocabulary. In this sense, these models would have a low performance for classification tasks (EMINAGAOGLU, 2022). Word embedding models emerged as an improvement to the weaknesses of methods based on vector space models.

2.1.2 Word embeddings and BERT embeddings

Word embeddings are a set of algorithms to represent words as fixed-size vectors. The main strength of these methods is that words with similar meanings, their vectors could have similar representations as well. In this sense, related words such as synonyms or words that belong to the same context, it is possibly their representations are closer in the vector space. The approaches based on word embeddings are mainly divided into three methods: techniques using neural networks; techniques based on dimensionality reduction using word co-occurrence matrices; and probabilistic methods (ALMEIDA; XEXÉO, 2019). One of the first word embedding approaches to come out is the Word2Vec method (MIKOLOV *et al.*, 2013b; MIKOLOV *et al.*, 2013a). This method uses a three-layer neural network, composed of an input layer, a hidden layer, and an output layer. This technique focuses on learning word vector representations based on predicting a word given its neighboring words. Such words are surrounding or context words. For example, for the sentence "while cats meow, dogs ..." , the goal of the neural network is to predict the word "bark". In this sense, the input layer from the neural network is composed of signals for context words and the output layer corresponds to signals for the predicted target word. The neural network is then trained with the documents from the corpus, and for each word, it is computed neighborhood probability considering the rest of the words in the vocabulary. Once the neural network is trained, the weights of the hidden layer are considered as the vectors of each word. Before the training stage, the size of the vectors representing each word can be defined. The selection of such a size depends on the dataset and the text application to be considered.

Despite their efficiency for various NLP tasks, word embedding models generate a single representation for every word in the vocabulary, regardless of a word's polysemy. For example, for the word "apple", Word2Vec will generate a single vector representation, regardless of whether the word refers to the type of fruit or the Apple technology company. Due to this weakness presented in several traditional word embeddings models, the model based on BERT was proposed. The Bidirectional Encoder Representations from Transformers (BERT) is an improvement to Word2Vec, because it generates different vectors for each word according to the context in which the word appears (KENTON; TOUTANOVA, 2019). BERT generates vector representations considering the context of a word in a bidirectional way, that is, before and after the target word appears. In Chapter 5 we provided a more detailed explanation of this approach.

2.1.3 Main applications

In this section, we briefly describe some of the most commonly investigated tasks in NLP and text mining that are related to our research, as well as their importance for various real-world applications. For example, a widely researched application is text classification, where each text can be grouped according to its category or class (KOWSARI *et al.*, 2019). These texts are grouped by identifying common features between them. Such features are extracted according to several patterns found in each text. In this sense, several supervised classification algorithms use the features extracted from these texts to classify them according to the corresponding class. This task is useful for authorship attribution, where books are grouped according to the author who wrote them (QUISPE; TOHALINO; AMANCIO, 2021). Text classification methods were also appropriate for information retrieval tasks and spam detection in emails (CRAWFORD *et al.*, 2015).

Other important applications for processing large amounts of information are the automatic summarization of documents and the keyword extraction task. The document summarization task consists of automatically finding a general description of the content that is mentioned in one or more text documents. A summary can be made up of text segments extracted directly from the documents (extractive summarization) or by content that is not necessarily explicitly cited in the original texts (abstractive summarization) (ANTIQUEIRA, 2007; TOHALINO; AMANCIO, 2018). On the other hand, keyword extraction aims to find a set of words that best represent the entire content of one or more texts. The most important words or phrases are called keywords or keyphrases (LI; FAN; ZHANG, 2007). Both automatic document summarization and keyword extraction are of paramount importance for several applications because they are capable of extracting relevant information from large text datasets in a concise, organized, and summarized way (WAN; YANG; XIAO, 2007). For example, through the keywords obtained from opinions, comments or suggestions of consumers, it is possible to obtain valuable information for the development of products. Keywords are also quite useful for examining how public opinion changes over time on a given topic. In the case of scientific articles, keywords help to understand which are the most relevant topics that a research paper is considering (BELIGA, 2014).

2.1.4 Final considerations

In this research we focused on two text mining applications based on the text classification and keyword extraction from scientific texts such as research proposals and paper abstracts. For text classification, we generated the numerical representations of abstracts of research grants based on the vector space model. Then, we used the extracted vectors as features for several machine learning algorithms. On the other hand, we used the vector representations based on word embeddings of each word to compute the semantic similarity between all word pairs for the keyword extraction task.

2.2 Complex Networks

A graph or network $G = \{V,E\}$ is composed by a set $V = V(G)$ of elements called nodes or vertices, and another set $E = E(G) \subset VxV$ of elements called links or edges that join the network nodes (LÜ *et al.*, 2013). An adjacency matrix is commonly used to represent the connectivity patterns in the graph. Networks are also represented by a weighted matrix which contains the values or weights assigned to each network edge. Networks can be directed or undirected. Complex networks are graphs with special topological properties. For example, many complex networks exhibit a scale-free degree distribution, where the majority of nodes have very few connections, but a small number of nodes have many connections. Additionally, complex networks often exhibit small-world properties, meaning that the average path length between any two nodes is relatively short, despite the vast size of the network. Another important property of complex networks is community structure, where nodes can be grouped into clusters or communities based on their connections within the network (Costa *et al.*, 2007).

In recent years, we have seen tremendous progress in the study of the structural and dynamic properties of complex networks (CHERIFI, 2014). During this time, hundreds of articles on this topic have been published in scientific research journals covering different disciplines. These disciplines include physics, biology, sociology, neurology, economics, medicine, computer science, among others (Costa *et al.*, 2007). The interest in this area increased due to the fact that any phenomenon that occurs in our real world can be modeled as a network (Costa *et al.*, 2007). Examples of graphs with complex network properties are found in biological networks, communication and computer networks, ecological networks, and even social networks.

The first investigations in complex networks arose with the study of the Erdos-Renyi model (ERDÖS; RÉNYI, 1959). This model consisted on the creation of a random graph. In such a graph, for each edge was established a probability p, where each pair of nodes had the same probability p of being connected. Differently from the traditional research which was focused on random networks, many works began to study and represent networks that model real and complex systems. Scientists concluded that the properties of these new network models are very different from the attributes observed in random graphs. Therefore, these special graphs were called as *complex networks* (WATTS; STROGATZ, 1998; STROGATZ, 2001). These graph types possess particular structural properties which are very different from random networks. Such properties could refer to degree distributions that follow power laws, hierarchical structures, community structures, high local cohesiveness, among others. Therefore, several researches found that complex network concepts are a powerful tool to model any real-world phenomena related to our social interactions, the environment we live or our own biological behavior (Costa *et al.*, 2007).

2.2.1 Structural properties on complex networks

Two concepts commonly studied in the area of complex networks are the centrality measurements and methods for community detection. Centrality metrics refer to the importance or prominence of the nodes within a network (NEWMAN, 2010). The goal of these indices is to quantitatively determine and compare the relative importance of an actor (network node) within the structure defined by the network. In this sense, the use of these metrics allows comparing and ranking each node according to its topological importance. Centrality is not an intrinsic attribute of the nodes of a network, but rather a structural attribute, that is, an assigned value that depends on the actor's relationships with the other actors in the network(BORGATTI, 2005).

The identification of the most central nodes plays a fundamental role in different applications: in social networks, for example, centrality indices allow to analyze the influence of a person; they assist in finding how good a path is in transportation networks; and they also allow to detect important web pages in a network (Costa *et al.*, 2007). There are hundreds of centrality indices that have been proposed to characterize the topological importance of a node. Node degree, strength, clustering coefficient, pageRank, or betweenness are some examples of measurements commonly used for various applications. A detailed description of these measurements is presented in Chapter 5.

A common feature found in complex networks is the presence of communities. A network is said to have a community structure if the nodes of the network can be easily grouped into potentially overlapping groups of nodes (RADICCHI *et al.*, 2004). These groups of nodes are more densely connected internally than the rest of the network. This heterogeneity of connections suggests that the network has certain natural divisions (GIRVAN; NEWMAN, 2002). The identification of communities is quite useful for several applications because it is very likely that the nodes that belong to the same community have properties and dynamics in common. The features of the community structure of networks also help to understand their dynamic evolution and organization (Costa *et al.*, 2007). For example, in social networks each community could represent locations, common interests, common occupations, etc. Metabolic networks have communities based on functional groups. In citation networks, communities are formed by research topics (GIRVAN; NEWMAN, 2002).

The detection of the optimal community structure of a network is a fairly complex task. The most used methods for finding the ideal division in communities are based on modularity maximization. Modularity, proposed by Girvan and Newman (2002), is a measurement that evaluates the quality of a particular division of the network in communities. Then, all possible divisions of the network are computed until maximum modularity is reached. The modularity metric is useful to analyze the number of edges within communities in relation to the number of edges present between communities. Several methods for finding communities were proposed in the literature. In this work, we evaluated the following methods: Multilevel, Label Propagation, Infomap, Fast Greedy, and Leiden method. In Chapter 6, we briefly explained these community extraction techniques.

2.2.2 Complex networks applications

The concepts of complex networks have been widely used in several applications because they are capable of modeling any phenomenon that occurs in the real world. Complex networks can be divided into four categories: social networks, information networks, technological networks, and biological networks (NEWMAN, 2010). In social networks, each node is represented by a person and the edges could be friendly relationships between people. For information networks, the citation networks were widely studied, where nodes represent the authors of research papers, which are linked through the references that are given to the papers of other authors. Examples of technological networks are commonly observed in electrical and Internet networks. Biological networks represent information patterns between different biological elements. For example, neurons can be modeled as nodes, and their relationships with other neurons are determined by chemical reactions between cells (NEWMAN, 2010; Costa *et al.*, 2007).

Natural Language Processing has also benefited from the use of complex networks since graphs are a powerful tool for modeling texts. Researchers developed a myriad of methods for representing texts as networks for different NLP applications. For instance, lexical networks are commonly used for the construction of spell checkers, where each edge is set from the orthographic distance between two words (CHOUDHURY *et al.*, 2007). Syntactic networks are used for the study of language acquisition (ANTIQUEIRA *et al.*, 2007). According to Costa *et al.* (2011), several NLP applications employed linguistic networks, where nodes could be syllables, words, sentences or paragraphs. The interconnections between nodes from these networks could be established in several ways. For example, word co-occurrence networks were used for the evaluation of the quality of machine translators, keyword extraction, and text classification tasks (AMANCIO *et al.*, 2008). In Antiqueira (2007), Tohalino and Amancio (2018), the authors modeled each text as sentence networks for extracting the most important text segments for automatic document summarization. According to Costa *et al.* (2011), texts could also be modeled as language networks, which are divided into semantic and surface networks. The semantic networks are constructed from dictionaries of lexicons and they usually contain information about semantic relationships between words, such as synonyms or antonyms. On the other hand, surface networks are based on the internal structure of words such as their morphological properties or their position in sentences or syntactic structures Costa *et al.* (2011).

2.2.3 Final considerations

In this research we modeled the document texts as networks for the keyword extraction task. We first represented each text as word co-occurrence network, and then we applied centrality measurements to get the most relevant words in each text. We also modeled the texts from the dataset as a citation network. In this approach, we applied community detection methods to get the best representative words from each cluster. In the Chapters 5 and 6 we presented two approaches the addressed the keyword extraction task using complex network concepts.

2.3 Scientometric and bibliometric analysis

Scientific research is the set of systematic and empirical processes dedicated to the study of a phenomenon, it is dynamic, changing and evolutionary (MINGERS; LEYDESDORFF, 2015). It produces knowledge and theories and it proposes to solve practical problems. The term big science has prospered since the significant growth of available scientific research covering different areas (MINGERS; LEYDESDORFF, 2015). The areas of scientometrics and bibliometrics emerged to study, measure and analyze scientific production. The study of these areas allowed to investigate various phenomena that occur in the scientific literature, such as the rise and decline of research areas, as well as the analysis of a discipline over the years.

Both scientometric and bibliometric analysis have been extensively investigated to study the progress of scientific research, however, there is some confusion in the literature about which methods are covered in each area. Scientometrics is a discipline that uses mathematical methods to quantify scientific research to reveal the process of scientific development, and can provide a scientific basis for decision-making and management. It commonly uses citation analysis and other quantitative methods to evaluate scientific research activities and thus guide the policy of science (NALIMOV; MULCHENKO, 1971). Bibliometrics is related to the application of mathematics and statistical methods to books or scientific journals, whose analyzes are linked to the management of libraries and databases (PRITCHARD *et al.*, 1969). Bibliometrics and scientometrics differ in subject background but are the same in theories, methods, technologies, and applications (SILUO; QINGLI, 2017). Below we briefly explain the main metrics related to scientometric and bibliometric analysis.

2.3.1 Metrics related to scientometric analysis

The main metrics commonly used in the scientometric analysis are the indicators of productivity, indicators of citation impact of papers or researchers, and the indicators of journal quality. The indicators of productivity are mainly based on: i) the number of papers produced by an author or research unit, ii) the number of papers journals produced on a particular subject, and iii) the number of keywords that texts generate (MINGERS; LEYDESDORFF, 2015).

In the case of indicators of citation impact, these metrics focus on analyzing the number of citations that papers or researchers received over a period of time. As follows we detail some indicators of impact (MINGERS, 2008):

- *Citation patterns:* they are related to the analysis of the number of citations per year received by a paper over time. These patterns generally show a birth-death process. In this process, initially, there are few citations; then the number increases to a maximum value; finally, they die away as the content becomes obsolete. There are many variants to this pattern, for example, the term "shooting stars" indicates papers that are highly cited but die quickly, and “sleeping beauties” which are ahead of their time (RAAN, 2004).

- *h-index:* the total number of citations is strongly affected by the number of papers, however, does not provide any information on this. The h-index, proposed by Hirsch (2005) combined in a simple way the impact (number of citations) and productivity (number of papers). The h-index is defined as: "a scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p \breve{} h)$ papers have $<= h$ citations each. So h represents the top h papers, all of which have at least h citations. Therefore, the h-index combines both number of citations and number of papers. However, the h-index ignores all the other papers below h, and it also ignores the actual number of citations received above h.

As follows we briefly describe some indicators of journal quality(MINGERS; LEYDESDORFF, 2015):

- *JIF (Impact factor)*: it represents the mean citations per paper for a journal over a two year period. For example, the impact factor of the year 2014 is the number of citations in 2014 to papers published in a journal in 2012 and 2013, divided by the number of such papers (GARFIELD *et al.*, 1994).
- *SJR (SCimagoJournal Rank)*: it is a metric that weights based on the prestige of a journal. It equally distributes the prestige of a journal among the total number of citations of the journal and normalizes the differences in the behavior of the citation of the different thematic fields (FALAGAS *et al.*, 2008).
- *SNIP (SourceNormalizedImpact per Paper)*: it measures the impact of a citation according to the characteristics of the subject on which it is investigated. It levels the differences in citation between the different thematic fields, and also levels the differences in their coverage by providing a standardized metric that allows the comparison of journals of different categories (LEYDESDORFF; OPTHOF, 2010).
- *I3*: it combines relative citation impact with productivity in terms of the numbers of papers but is normalised through the use of percentiles (LEYDESDORFF, 2012).

2.3.2 Network-based methods used in bibliometric analysis

As follows we mention the main approaches that used network analysis methods commonly used in bibliometric studies (HOOD; WILSON, 2001; PRICE, 1965; REUTERS, 2008; ZUPIC; ČATER, 2015).

- *Citation analysis:* A citation network is a directed graph in which each node represents a scientific article and each edge represents a citation from the current paper to another paper (PRICE, 1965). The citations are used as a measure of influence, in this sense, the most cited authors, papers, or journals could be the most influential scientific agents. Through the analysis of the citation networks we can understand the following issues: Which authors most influenced research in a journal?, Which journals and disciplines have the most impact on a research survey?, Who are the experts in a research area?, What articles should we read from a certain area?
- *Co-citation analysis*: Two papers (authors or journals) are co-cited if they are both cited by the same, third, paper (author or journal) (SMALL, 1973). According to this analysis, the more two works are cited together, the more their content is related. Through the analysis of the co-citation networks we can understand the following issues: Who are the central and peripheral organizations in a research field?, What is the group of authors cited systematically by a specific group of works?, What works are referenced together?
- *Bibliographic coupling*: When two works refer to common work(s), the relation between two referring documents is called bibliographic coupling (KESSLER, 1963). The nodes

of the network could be scientific publications, authors, or journals. According to this measurements, the more two works cite similar works, the more their content is related. Through the analysis of the bibliographic coupling we can understand the following issues: Who are the central and peripheral organizations in an emerging research group?, Which authors produced summaries that approached a body of literature?, How does the structure of a research aspect reflect the diversity of theoretical approaches?

- *Co-author analysis*: For the co-authorship networks, the nodes are scientists (or institutions), and two scientists are connected if they have coauthored a paper (NEWMAN, 2004). This analysis can be used to evaluate collaboration at institutional and national level. Through the analysis of the co-authorship networks we can understand the following issues: What researchers work together?, How is the social structure of a research area?, Which institutions or countries collaborate in a specific research field?

- *Co-word analysis*: It is a technique that uses pattern of co-occurrence of words (generally keywords) from a corpus of scientific papers. It establishes relation between ideas and concepts within the subject area, presented in the corpus. Occurring of two keywords within the same paper indicates a relationship between the topics to which they refer (KOSTOFF, 1993). Through the co-word analysis we can understand the following issues: What keywords are being the most used in each given period of time?, What words are used together?, How the research interest changed?

2.3.3 Final considerations

We used the concepts studied in scientometrics and bibliometrics for the two tasks proposed in this research. First, we extracted bibliometric features according to the academic activity of each researcher for the research grant classification task. We also modeled research papers as a citation network for the keyword extraction task of paper abstracts.

CHAPTER 3

ANALYZING THE RELATIONSHIP BETWEEN TEXT FEATURES AND GRANTS PRODUCTIVITY

Title	Analyzing the relationship between text features and grants productivity
Authors	Jorge Valverde Tohalino, Laura Cruz Quispe, and Diego Amancio
Year	2021
Journal	Scientometrics
Link	<https://link.springer.com/article/10.1007/s11192-021-03926-x>
Situation	Published

3.1 Motivation

Research on the factors that contribute to the success of scientific items, such as research papers, theses, and grants, has been extensively conducted. These factors include the number of citations, author visibility, and textual content analysis. However, most of these studies have been limited to large datasets focused on foreign researchers, particularly those from the United States and China. Therefore, for this study, we sought to evaluate the productivity of research grants funded by Brazilian government agencies, as there were no datasets available for assessing the academic research carried out by Brazilian investigators. Our approach analyzed only the abstract of each research grant, with the main objective of assessing whether linguistic patterns found in a short text are capable of capturing predictive features of the success of a research project.

This research was also motivated by the need to improve the effectiveness of funding decisions in research by identifying productive research grants. Many research proposals are not funded due to limitations in resources, and this may affect the success and diffusion of important ideas. Therefore, there is a need to identify promising research proposals that are more likely to yield productive results. We aimed to investigate whether text features extracted from project titles and abstracts can be used as predictors of productivity and assist in identifying relevant research ideas. This research is important for research funding bodies, scientific entities, and government agencies who want to make more informed funding decisions and avoid wasting resources on research proposals that are less likely to yield productive results.

3.2 Contributions

This paper investigates whether text features extracted from the titles and abstracts of research grant proposals can be used to identify productive grants in the fields of Medicine, Dentistry, and Veterinary Medicine. We used complexity and topical features to identify predictors of productivity and we found that there is a statistically significant relationship between text features and grant productivity, although the dependence is weak. The abstract text length and metrics derived from lexical diversity were among the most discriminative features. We found that text features should be used in combination with other features to assist in the identification of relevant research ideas. This study used a dataset of research grants funded by São Paulo Research Foundation (FAPESP-Brazil) and limited the sense of productivity by considering that productive grants are those yielding at least one publication. We concluded that future research should consider other productivity criteria, including the total number of publications, the reputation of the respective journals and conferences, and other measurements derived from citation and usage counts. This paper also suggests that there is a large space for improvement in performance by incorporating additional features, such as those based on recent authors' performance, text network-based attributes, and features related to researchers and their respective institutes. This paper concludes that the results provide a basis for developing automatized tools to assist funding bodies, scientific entities, and government agencies in identifying promising research ideas.

Scientometrics
https://doi.org/10.1007/s11192-021-03926-x

Analyzing the relationship between text features and grants productivity

Jorge A. V. Tohalino[1] · Laura V. C. Quispe[1] · Diego R. Amancio[1]

Received: 17 September 2020 / Accepted: 24 February 2021

Abstract
Predicting the output of research grants is of considerable relevance to research funding bodies, scientific entities and government agencies. In this study, we investigate whether text features extracted from projects title and abstracts are able to identify productive grants. 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 productivity. The results indicate that there is a statistically significant relationship between text features and grants productivity, however such a dependence is weak. A feature relevance analysis revealed that the abstract text length and metrics derived from lexical diversity are among the most discriminative features. We also found that the prediction accuracy has a dependence on the considered project language and that topical features are more discriminative than text complexity measurements. Our findings suggest that text features should be used in combination with other features to assist the identification of relevant research ideas.

Keywords Language analysis · Productivity · Grants productivity · Text analysis

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 (Fortunato et al. 2018). 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 (Ding 2011). 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 (Thelwall and Nevill 2018; Zeng et al. 2017). Many studies have also sought linguistic patterns in the scientific literature (McKeown et al. 2016). Similar

✉ Diego R. Amancio
diego@icmc.usp.br

[1] Institute of Mathematics and Computer Science, Department of Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil

Published online: 06 March 2021

studies have used paper metadata to analyze and understand the behavior of authors, including their collaboration/citation patterns and contributorship patterns (Corrêa Jr et al. 2017). Another important area in science of science concerns the studies devoted to make predictions in many scenarios (Acuna et al. 2012). 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 projects.

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 before and after it is funded (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 and style of texts have been assessed using machine learning methods (Antiqueira et al. 2007; Silva and Amancio 2012). A pattern recognition approach applied in the context of grants assessment could also shed light on the understanding of which factors are associated with strong grants. This could be particularly useful for early career scholars, as many of them have received little or no feedback. In the current study, we touch these points by probing whether information retrieved from projects can be used to predict grants productivity.

While many factors may affect the perceived quality of projects (Markowitz 2019; Boyack et al. 2018), in this study we focus on the analysis of textual features of funded projects. Our main objective is to analyze if we can automatically (i.e. using machine learning methods) predict grants productivity via linguistic patterns. We focused on two types of textual attributes, namely:

1. *Topical features*: our hypothesis here consists in probing whether projects on specific subtopics are more likely to yield at least one publication. While certainly there are differences in publication/citation patterns across different fields (Piro et al. 2013), our insight is that a similar pattern may also appear in subfields (?).
2. *Complexity features*: complexity features are affected by many psycholinguistic factors (Graesser et al. 2004). Examples of complexity features includes lexical diversity and function words counts (Markowitz 2019; Graesser et al. 2004). Several works have shown that complexity measurements could influence the perceived quality of scientific works (Markowitz 2019; Letchford et al. 2015, 2016; Wager et al. 2016). Our hypothesis regarding complexity is that similar complexity patterns could also appear in research projects. In other words, if the results obtained from grants are disseminated using similar linguistic patterns, one could expect that the patterns could also influence the perception of the results/conclusions being analyzed, and thus influence grants productivity. For example, if grants abstracts are written in a clear, simple and more assertive

way, one could expect that the same properties could appear in manuscripts submitted for publication and those patterns could affect acceptance decisions.

Our analysis was conducted in a subset of research grants funded by the São Paulo Research Foundation (FAPESP-Brazil). We selected research grants in three areas comprising the largest number of projects funded by FAPESP. We considered projects in the following areas: Medicine, Dentistry and Veterinary Medicine. Here we classified if a grant yielded or not at least one publication. Therefore, in our classification system, two classes were considered, according to the number of observed publications: (i) zero publications; and (ii) one or more publications. While the term productivity might be allusive to distinguishing a few from many publications, hereafter productivity is used in the context of discriminating class (i) from (ii).

Several interesting results could be found in our analysis. By considering only a balanced version of the datasets, we found that there is a relationship between text features and grants productivity in all considered areas. However, we only found a weak correlation. When comparing complexity and topical features, the latter turned out to be more effective to predict grants productivity. We also found that the ability to predict productivity depends on the considered language (English or Portuguese). The accuracy was higher when analyzing texts written in researchers' native language (Portuguese). A feature importance analysis revealed that the measurements capturing lexical diversity of abstracts are relevant features for identifying productive grants in all three considered datasets. Our analysis also revealed that the best classifiers for the adopted features were those based on Decision Trees. All in all, the adopted framework provides evidence that text features might be relevant in the identification of productive grants. 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 "Feature relevance"section, we present related works on features used to predict the output of scientific papers and projects. In "Dataset"section, we describe the methodology used in the machine learning framework. The obtained results are discussed in "Results and Discussion"section. Perspectives for works extending our approach are presented in "Conclusion"section.

Related works

Several studies have investigated the factors leading to the success of scientific items (Markowitz 2019; Wang et al. 2008; Boyack et al. 2018; Xie et al. 2015). In the case of scientific papers, many factors have found to play a role in defining their visibility. Eom and Fortunato (2011) 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 (Eom and Fortunato 2011). 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 (Didegah and Thelwall 2013; Onodera and Yoshikane 2015; Silva et al. 2016).

Text factors have also been found to affect the visibility of papers (Amancio et al. 2012b; Letchford et al. 2015; Paiva et al. 2012; McKeown et al. 2016). Amancio et al.

(2012b) 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 (Letchford et al. 2015; Paiva et al. 2012). 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 (Paiva et al. 2012).

The factors affecting the success of research proposals have also been analyzed in the last few years (Boyack et al. 2018; Hörlesberger et al. 2013; Cabezas-Clavijo et al. 2013; Fang et al. 2016; Li and Agha 2015). Boyack et al. (2018) found that researchers productivity can not be used to predict proposals 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.

Markowitz (2019) studied if word patterns extracted from National Science Foundation (NSF) proposals are able to predict the received amount of funding. As descriptors of text complexity the author used word counts, words per sentence, the percentage of common words and the complexity of thinking as measured via function words. Several interesting results were found showing a relationship between text variables and the amount of money received. Larger grant abstracts with fewer common words were among the main patterns correlating with funding success. Markowitz (2019) advocated that the observed patterns contradict NSF guidelines, since more complex textual structures are more correlated to funding success.

Another feature that could be used to predict research proposal success are those related to peer review scores. In (Cabezas-Clavijo et al. 2013), 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. Fang et al. (2016) 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 (Germain 2015). Leading to a different conclusion, Li and Agha (2015) argue 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 projects success, here we focus on text features to predict productivity.

Material and methods

The dataset used in the current paper is described in "Dataset"section. The framework proposed to classify grants comprises the following three main steps:

1. *Feature extraction*: this phase is responsible for extracting topical and complexity features from textual fragments of research projects. This is detailed in "Feature Extraction"section. While we test the influence of topical features, our main focus here is to analyze the influence of text complexity on the predictability of grants productivity.
2. *Pattern recognition*: the features extracted are used as input for traditional machine learning methods. An overview of methods is provided in "Machine Learning Methods"section. A more detailed reference on machine learning and pattern recognition methods can be found in (Duda et al. 2012).
3. *Feature relevance analysis*: this phase is responsible for identifying the most relevant (i.e. discriminative) features. A brief description of the adopted method in provided in "Textual complexity measurements relevance"section.

Dataset

The main objective of this work is to analyze whether textual features can be used to predict the productivity of research grants. 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[1] (FAPESP). 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 projects: their title and abstract. The data were retrieved from the *Biblioteca Virtual* website[2]. The association between papers and projects are automatically extracted by the *Web of Science*[3] dataset, and this information is also made available by the *Biblioteca Virtual* website. We could not use information from other datasets since the link between paper and projects is not available. Once a project funded by FAPESP leads to a research paper, it is required that the authors acknowledge São Paulo Research Foundation. Specifically, the funding agency provides a format that every funded project must comply with. A grant must be mentioned using the explicit form “aaaa/nnnnn-d”, where “aaaa” represents the year and “nnnnn-d” is the grant number[4].

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, which are grants under the responsibility of a Principal Investigator associated with higher education (or research institutions) in the State of São Paulo. Regular grants are usually limited to a duration of 2 years[5]. Financial resources associated to each grant are usually limited to a maximum of 200, 000 *Brazilian Reais*. 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 grants of this specific type in the dataset. Considering this type of research project, we

[1] fapesp.br/en.

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

[3] webofknowledge.com.

[4] This information is available from this link https://fapesp.br/11789/referencia-ao-apoio-da-fapesp-em-todas-as-formas-de-divulgacao (in Portuguese).

[5] Details regarding regular grants are available at fapesp.br/apr (in Portuguese).

Table 1 Fraction of grants with a respective number of papers. In column #P, n+ corresponds to n or more papers being published. Note that, e.g. in the Medicine VET dataset, only a small fraction of grants published three or more papers (5.5%)

#P	Research areas		
	MED	DENT	VET
2+	17.8%	25.6%	12.7%
3+	9.3%	12.9%	5.5%
4+	4.9%	7.3%	3.2%
5+	3.1%	4.3%	1.5%
6+	2.5%	3.3%	0.9%
7+	1.6%	2.4%	0.6%
8+	1.1%	1.3%	0.4%

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 grants productivity. This could be the number of published papers, the number of citations, and several other metrics commonly used in quantifying success in academia (Wang et al. 2013). Because most of these distributions are skewed, we decided to simplify the criteria to consider a research project as productive. To avoid an extreme unbalancing in the number of positive and negative examples (Li et al. 2010), we consider a project as productive if it yielded at least one publication. While this criterion still generates unbalanced datasets, a considerable number of both positive and negative 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 positive examples in each area was: 41.27%, 48.48% and 31.96% for Medicine, Dentistry and Veterinary Medicine, respectively. In order to probe the consistency of the dataset, we conducted a *manual* analysis. More specifically, we randomly selected 200 research projects and *manually* analyzed how many publications were missing in the FAPESP dataset. For each selected research project, we analyzed all publications of the respective grantees up to 3 years after the project was finished. This time lag is compatible with the similar findings observed in NIH projects ?. This study revealed that less than 4% of the projects were missed in the FAPESP dataset. More details regarding Brazilian research agencies are discussed in McManus and Neves (2020).

Additional basic statistics regarding grants productivity is available in Table 1. 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 (Duda et al. 2012). 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.

Feature extraction

Several studies have used linguistic features to analyze scientific items (Larrimore et al. 2011; Markowitz et al. 2014) We used two distinct approaches to analyze the abstract of the grants. Topical and complexity features. While topic features are intended to analyze if specific topics are associated with productive grants, complexity measurements analyzes whether language simplicity is associated with a lower or higher degree of productivity.

In this paper, 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 grants output 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:

$$\text{tf-idf}(w,d) = \frac{f(w,d)}{n_d} \cdot \frac{\log N}{\log (N_w)}, \tag{1}$$

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, that can be measure in different ways. A statistical approach based on structural features of text modeled as networks was described by Amancio et al. (2012a). While this approach was able to assess the readability of texts, the co-occurrence can only be effectively applied in larger texts (Amancio 2015b). The measurements used in the current are a subset of metrics adapted from the English version of Coh-Metrix (Graesser et al. 2004). We used Coh-Metrix because it encompasses many different degrees of text complexity, including words, sentence, textbase and situation model features (Graesser et al. 2004). Some examples of textual complexity features used here are:

1. *Basic counts*: total number of sentences, words, adjectives, adverbs and verbs. Larger pieces of texts, such as word counts or more words per sentence, are associated with higher degree of complexity in texts (Markowitz 2019; Kincaid et al. 1975).
2. *Logic operators*: this feature quantify the number of logical operators. Note that logical operators such as “if” and “or” could denote texts with a certain degree of uncertainty (Larrimore et al. 2011). In the context of grants, a higher degree of uncertainty could mean taking more risks, and this could affect the productivity related to the grant.
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). According to Pennebaker et al. (2014), particular function words reflect complex and analytic thinking. A style avoiding complex thinking could be associated with more clarity when expressing ideas. If such a simplicity is also reflected when

writing papers, clarity could increase the likelihood of a paper being published. A link between function words and concreteness is discussed by Larrimore et al. (2011). In other words, counting function words is a different approach to capture the number of concrete words in texts.

4. *Preposition diversity*: this corresponds to the same counting in *function word diversity*, but applied to prepositions only. The number of prepositions could be linked to with high academic performance in college and thus potentially could be useful to identify academic performance in research (Pennebaker et al. 2014).
5. *Punctuation diversity*: this corresponds to the same counting in *function word diversity*, but applied to punctuation marks only. Texts with long sentences and few punctuation marks may suggest that they are harder to understand (Graesser et al. 2004). Sentences with a few pauses may require a higher cognitive effort to be processed.
6. *Noun SD*: this corresponds to the standard deviation in the number of nouns per sentence. Nouns can refer to concepts, and a text involving many different concepts could indicate a higher degree of complexity (Graesser et al. 2004).
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 β corresponds to a high lexical diversity. Thus higher values of β indicate a richer language (Brunet 1978).
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. The interpretation of this measure in terms of complexity is similar to the one provided in *Noun SD*. The difference is that here we are also considering modifiers along with nouns.
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’. As discussed by Larrimore et al. (2011), the use of concrete words is related to a better contextualization of concepts and is linked to a reduction of uncertainty. Thus, more confident language could be linked to stronger results, which could make it easier for authors to publish papers. While issues with some semantic psycholinguistic variables have been reported Pollock (2018), we decided to use this measurement because it has been useful in other contexts (Diller et al. 2014). A less context-dependent approach to operationalizing concreteness was used by Markowitz (2019). The approach employed by Markowitz (2019) relies on function words, rather than a selection of concrete words. The method is consistent with function word features used in this work (Graesser et al. 2004).
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 (Nadeau and Sekine 2007). In the scientific context, named entities could be linked e.g. to different methodologies or datasets. Texts with more different methodologies and/or named concepts involved could indicate a more detailed research, which in turn could facilitate publications from the respective grant.

The full list of the considered features and a detailed description of each feature can be found in (Scarton and Aluısio 2010).

Machine learning methods

The textual features extracted from the abstract of the research projects are used in the classification process (Duda et al. 2012). For each instance, we consider two possible classes: (i) zero publications; and (ii) one or more publications. 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 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 (Duda et al. 2012). The evaluation was based on the F1-Score measurement, a traditional measure in the area of information retrieval (Manning et al. 2008). 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 (Amancio et al. 2014). 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 methods is the kernel used to create the discriminative hyperplane. In this paper, we used the optimization strategy described in (Amancio et al. 2014).
3. *Naive Bayes*: this method relies on the Bayesian optimal decision rule to perform a classification. Let $m = \{f_1, f_2, \dots\}$ be the set of features used to characterize research grants (i.e., the features described in Section 3.2). The class $\mathbf{c}$ assigned to a grant satisfies the following condition:

$$P(\mathbf{c}|m) \geq P(c_k|m), \tag{2}$$

for every class $c_k \neq \mathbf{c}$, 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 $\mathbf{c}$:

$$\mathbf{c} = \arg\max_{c_k} \frac{P(m|c_k)}{P(m)} P(c_k). \tag{3}$$

$P(m)$ is the same for every class c_k, therefore the above equation can be simplified to:

$$\begin{aligned} \mathbf{c} &= \arg\max_{c_k} P(m|c_k)P(c_k) \\ &= \arg\max_{c_k} \left[\log P(m|c_k) + \log P(c_k) \right]. \end{aligned} \tag{4}$$

Assuming attribute independence, the class assigned to a new instance from the test dataset is computed as:

$$\mathbf{c} = \arg\max_{c_k} \left[\sum_{f_i} \log P(f_i|c_k) + \log P(c_k) \right]. \tag{5}$$

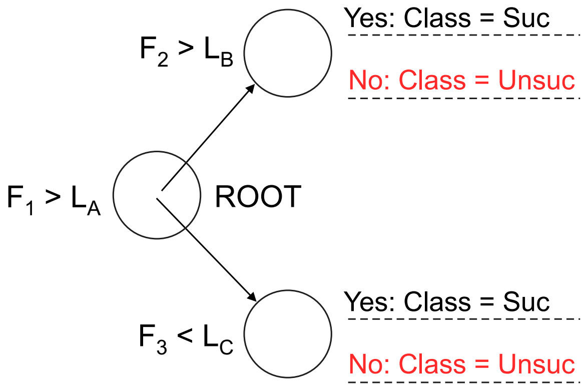


Fig. 1 Example of decision tree used for classification. The classification process for a new instance starts at the root node. Consider a new instance that should be classified. This new instance is described by the vector of features $(f_1 = x > L_A, f_2 = y < L_B, f_3 = z)$. The first test $(f_1 > L_A)$ 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 yielding zero publications. 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 productive grant

For the particular case of balanced datasets, $P(c_k)$ is uniform. Therefore,

$$\mathbf{c} = \arg\max P(m|c_k). \tag{6}$$

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 (Cormen et al. 2009). 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 Fig. 1. The classification process starts at the root node (see Fig. 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:

$$\mathcal{K}(D_{\mathrm{TR}}, f_i) = \mathcal{H}(D_{\mathrm{TR}}) - \mathcal{H}(D_{\mathrm{TR}}|f_i). \tag{7}$$

   where $\mathcal{H}(D_{\mathrm{TR}})$ is the entropy of the training dataset and $\mathcal{H}(D_{\mathrm{TR}}|f_i)$ is the entropy of the training dataset considering the separation of classes obtained with the i-th feature (Duda et al. 2012; Garreta and Moncecchi 2013).

   In addition to traditional decision trees, we also used random forests (Breiman 2001). The latter has the advantage of avoiding the tendency of decision trees to overfit the training set (Breiman 2001). 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 (LeCun et al. 2015). 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 transfers functions (Hassoun et al. 1995).

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 (Hassoun et al. 1995). 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 (Hassoun et al. 1995). 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) (Hassoun et al. 1995), a simple yet effective approach in many scenarios (Amancio et al. 2014).

Textual complexity measurements relevance

In order to evaluate the relevance of features when identifying productive grants, 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 (Nembrini et al. 2018). 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:

$$\mathcal{G} = \sum_{i \in \mathcal{C}} p_i (1 - p_i), \tag{8}$$

where $\mathcal{C}$ is the set of classes. In our study, $\mathcal{C} = \{\text{productive}, \text{zero publications}\}$. 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 ($\Delta\mathcal{G}$) for the considered dataset. The decrease in impurity for each tree node is computed as

$$\Delta\mathcal{G} = \mathcal{G}_{\mathrm{B}} - \beta_L \mathcal{G}_{\mathrm{L}} - \beta_R \mathcal{G}_{\mathrm{R}}, \tag{9}$$

where $\mathcal{G}_{\mathrm{B}}$ is the Gini impurity before the dataset is split in the respective node and $\mathcal{G}_{\mathrm{L}}$ and $\mathcal{G}_{\mathrm{R}}$ are the Gini impurity obtained in the left and right child nodes, respectively. β_L and β_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 Fig. 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. $\mathcal{G}_B = 50\%$). 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 eq. 8:

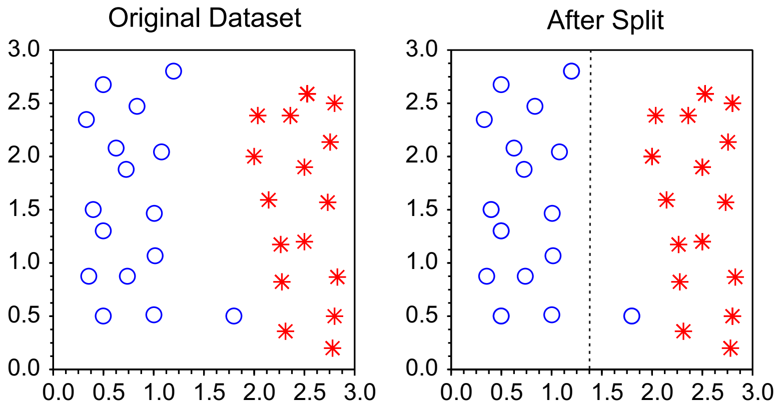


Fig. 2 Computing the decrease in Gini impurity for a small dataset with two classes. For each class, there are 16 instances. In the original dataset, the probability of misclassification is high for a randomly drawn instance is high, i.e. $\mathcal{G} = 0.50$. After the original dataset is split in two subsets, the discrimination of classes becomes almost perfect. This leads to a high variation in the Gini impurity, i.e. $\Delta\mathcal{G} = 0.4412$

$$\mathcal{G}_{\mathrm{R}} = \frac{1}{17}\left(1 - \frac{1}{17}\right) + \frac{16}{17}\left(1 - \frac{16}{17}\right) = 0.1107. \tag{10}$$

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

$$\Delta\mathcal{G} = \mathcal{G}_B - \frac{15}{32}\mathcal{G}_L - \frac{17}{32}\mathcal{G}_R = 0.4412. \tag{11}$$

In other words, the split for the considered feature yield a reduction of $\Delta\mathcal{G} = 0.4412$ in the impurity of the original dataset.

Results and discussion

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

Performance analysis

In this section, we start the discussion of results by considering the F1-Score obtained with complexity measurements extracted from title and abstracts (in Portuguese). The obtained results are shown in Table 2. We show, for each considered dataset (Medicine, Dentistry and Veterinary Medicine) the performance obtained from the machine learning methods considered in this study. We computed the statistical significance of the obtained results. If n_g is the number of corrected classified grants, the p-value corresponds to the probability of correctly classifying at least n_g instances just by chance. To compute the p-value we took into account that classes are imbalanced. This means that each single instance, in a random classification, is correctly classified with probability p_i, where p_i is the fraction of

Table 2 F1-Score rate obtained when classifying grants productivity using Coh-metrix features (Graesser et al. 2004) for Portuguese. Three different datasets were considered: Medicine (MED), Dentistry (DENT) and Veterinary Medicine (VET). The best results for each dataset are highlighted. We also show for each F1-Score the corresponding significance of the classification. The best results were obtained with decision trees

Method	Medicine	Dentistry	Vet. Med.
	F1-Score	F1-Score	F1-Score
DTrees	**0.5673**	**0.5725**	**0.6008**
SVM	0.5417	0.5437	0.5657
kNN	0.4997	0.5384	0.5336
Bayes	0.4182	0.4800	0.5115
MLP	0.5177	0.5142	0.5278

the dominant class in the considered dataset. The best results for each dataset were found to be statistically significant. They are highlighted in Table 2. The best result in predicting the productivity of a research grant (according to the adopted productivity criteria) was found in the area of Veterinary Medicine. In this case a F1-Score = 0.6008 has been found. Significant results were also found for Medicine and Dentistry, however with a lower discriminability result. In the best scenario, the F1-Score obtained for these areas were 0.5673 and 0.5725, respectively. These results suggest that the complexity of texts plays a statistically significant role in predicting research grants productivity. While the best results are significant, they reveal only a weak discriminative power.

When considering different strategies for supervised classification and the set of Coh-metrix features, in all three considered datasets, the best results were obtained with Decision Trees. The second best classifier was SVM in all considered datasets. The worst classification systems, were found to be kNN and Naive Bayes. Interestingly, the performance of the Naive Bayes was much lower than the F1-Score obtained with decision trees, and such a difference turned out to be more prominent in the Medicine area. This result suggests that the choice of classification systems plays an important role in the performance of the considered classification task, even when the same family of features are considered.

While the results shown in Table 2 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 grants productivity. For this analysis, we considered the tf-idf representation of texts. Here, the classification considered different parts of the research project. 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 $X = 1100$ and $X = 7196$ words, respectively.

In Table 3, we show the results obtained for different classifiers. The best results found for each dataset are significant. The best F1-Score were 0.5766, 0.6246 and 0.6395 respectively for the MED, DENT and VET datasets. The discriminability observed in the Veterinary Medicine area once again was found to be slightly higher than the discriminability found in other areas. These results also suggest that the frequency-based features also play a role in predicting productivity of the considered datasets. Topical features were found to

Table 3 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. The best significant results for each dataset are highlighted

Features	DTrees F1-Score	SVM F1-Score	kNN F1-Score	Bayes F1-Score	MLP F1-Score
Research projects on *medicine*					
Title	0.5189	0.4972	0.4773	0.5228	0.5115
Subject	0.5376	0.4349	0.5163	0.5372	0.5320
Tit. + Sub.	0.5404	0.4979	0.4978	0.5254	0.5158
Abstract$^{(1)}$	**0.5660**	**0.5695**	0.5293	**0.5588**	0.5470
Abstract$^{(2)}$	**0.5766**	**0.5649**	0.5314	**0.5696**	0.5505
Research projects on *dentistry*					
Title	0.5541	0.5055	0.4755	0.5606	0.5636
Subject	0.5664	0.4680	0.5466	0.5636	0.5646
Tit. + Sub.	**0.5916**	0.5133	0.5284	**0.5772**	**0**.5771
Abstract$^{(1)}$	**0.5979**	**0.6033**	0.5502	0.5697	**0.5913**
Abstract$^{(2)}$	**0.6246**	**0.5770**	0.5548	**0.6026**	**0.5869**
Research projects on *veterinary medicine*					
Title	0.5485	0.5115	0.5127	0.5543	0.5235
Subject	0.5566	0.4885	0.508	0.5517	0.5496
Tit. + Sub.	0.5721	0.5503	0.5128	0.5579	0.5407
Abstract$^{(1)}$	0.5890	0.5903	0.5371	0.5799	0.5636
Abstract$^{(2)}$	**0.6395**	0.5886	0.5076	0.5958	**0.6010**

be more discriminative than complexity measurements, when considering the best results. The highest improvement in performance was found in the DENT dataset: the F1-Score improved from 0.5725 (with complexity measurements) to 0.6246 (with topic features). Interestingly, the best performance was obtained with Decision Trees: in all three datasets, the highest accuracy was obtained with this classifier.

Table 3 also reveals that particular fragments of research projects 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. In general, when more features are considered (i.e. Abstract$^{(2)}$), a higher discriminability rate is obtained. A lower performance is observed when considering both the title and the grant subject. However, the performance of Abstract$^{(1)}$ and Tit. + Sub. is similar for Dentistry and Veterinary Medicine. It is also worth noting that an extremely low discriminative power obtained with kNN in all three datasets. Regardless of the chosen set of features, the discriminability is always too low. This is consistent with the results observed for complexity measurements.

Language dependence

As mentioned in "Dataset"section, the abstract of each research project is available in two languages: Portuguese and English. The results reported in "Performance analysis"section 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 4. The best results for each language and research area are highlighted. When comparing the

Table 4 F1-Score obtained when discriminating research projects productivity. We used complexity features to characterize the texts. The results reveal that a difference in performance is observed when comparing Portuguese and English abstracts. The best significant results for each dataset and language are highlighted

Method	Portuguese F1-Score	English F1-Score
Projects on *medicine*		
DTrees	**0.5673**	0.5322
SVM	0.5417	0.4855
kNN	0.4997	0.5069
Bayes	0.4182	0.4876
MLP	0.5177	0.5180
Projects on *dentistry*		
DTrees	**0.5725**	0.5104
SVM	0.5437	0.4858
kNN	0.5384	0.5087
Bayes	0.4800	0.5048
MLP	0.5142	0.4953
Projects on *vet. med.*		
DTrees	**0.6008**	0.5168
SVM	0.5657	0.4843
kNN	0.5336	0.5038
Bayes	0.5115	0.4948
MLP	0.5278	0.5308

Table 5 F1-Score obtained when discriminating research project productivity. We used tf-idf features to characterize the project abstracts. The best results for each dataset and language are highlighted

Method	Portuguese F1-Score	English F1-Score
Projects on *medicine*		
DTrees	**0.5766**	0.5488
SVM	**0.5695**	0.5555
kNN	0.5314	0.5406
Bayes	**0.5696**	0.5354
MLP	0.5505	0.5155
Projects on *dentistry*		
DTrees	**0.6246**	0.5164
SVM	**0.6033**	0.4934
kNN	0.5548	0.4979
Bayes	**0.6026**	0.5543
MLP	**0.5869**	0.5413
Projects on *vet. med.*		
DTrees	**0.6395**	0.5231
SVM	0.5903	0.4932
kNN	0.5371	0.5299
Bayes	0.5958	0.5509
MLP	**0.6010**	0.5653

best results for Portuguese and English, we found no a difference in performance, especially in both Dentistry and Veterinary Medicine areas. In all three datasets, the best discriminability was found for the Portuguese language. In the Veterinary Medicine area, the best discriminability rate found for Portuguese is roughly 13% higher than the best score obtained using abstracts in English.

In Table 5, 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 4: in all three datasets, we found a 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.

We note that abstracts in English are less discriminative, and this might be related to the fact that almost all research projects are written by Portuguese native speakers. Because English can be viewed as a second language in the analyzed project, a lower discriminability might be a consequence of a lower linguistic variety, both at the complexity and topical levels. In other words, textual properties observed in projects might have a higher variability when the text is written in the researcher's native language. As a consequence, this effect can cause significant differences in the considered classification task. This result suggests that predicting productivity with text features should also take into account whether the language being analyzed is a first or second language.

Feature relevance

The results in the previous section showed that there is a dependence between text features extracted from research projects and the output of the respective grants. The productivity of specific grants 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 (McKeown et al. 2016; Silva et al. 2016). 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 (Leydesdorff et al. 2019; Leydesdorff and Rafols 2011). The importance of text complexity (i.e. topic independent) features is not as clear. In order to better understand in future works if particular features plays a more relevant role in predicting the productivity of grants, in this section we provide an analysis of the main complexity features responsible for identifying productive grants.

For the analysis of features relevance, we used the strategy described in "Performance analysis"section, 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 $\Delta\mathcal{G}$, 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.

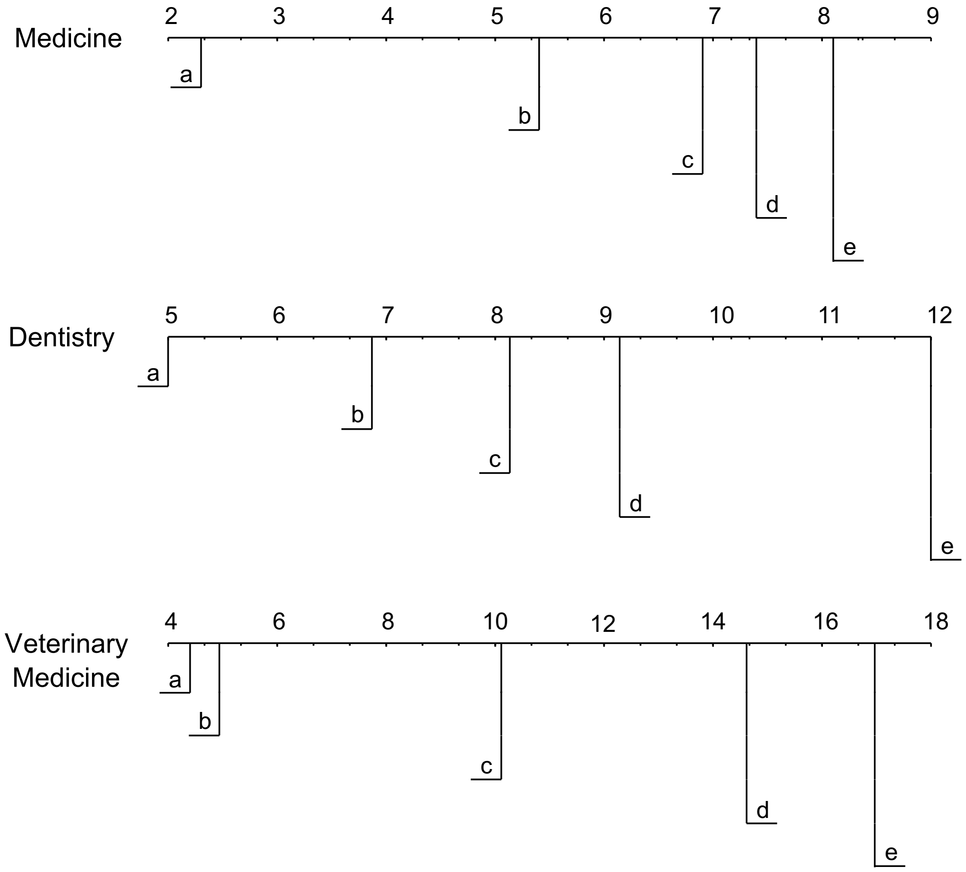


Fig. 3 Feature ranking diagram for the classification of research projects according to their productivity. For each dataset, we show the average ranking obtained by each of the considered Coh-metrix features. In Medicine, the best features were: (a) function word diversity; (b) nouns SD; (c) total number 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

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 (Demšar 2006) (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 grants 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 abstract length is important, but also the diversity of lexical items. This finding is compatible with studies correlating lexical diversity and writing quality (Antiqueira et al. 2007). The relevance of preposition diversity reveals that not only the diversity of semantic concepts might be relevant to discriminate productive grants, but also stopwords (prepositions), i.e. words conveying no semantical meaning. This result reinforces the importance of style when performing a text analysis in grants (Markowitz et al. 2014; Markowitz 2019). Because the presence of function words could indicate a lack of concrete words (see e.g. Larrimore et al. (2011)), we can infer that concrete words are also a relevant feature for the task. Therefore, it seems that concrete words are not only important to detect the amount of funding (Markowitz 2019), but also if the grant will yield a paper. In addition, the 'concreteness' of words – as quantified by Coh-Metrix – also seems to play a role in the identification of productive DENT grants. This means that concreteness might be an important feature less dependent of the considered research area.

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 grants. 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 productive grants.

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, 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 (Kassirer and Campion 1994). As it happens in other bibliometrics contexts, it is still interesting to provide automatized tools that can assist humans in particular issues (Silva et al. 2016; Amancio 2015a; Daud et al. 2015). In this context, in this paper, we analyzed whether textual features extracted from research projects can be used to identify productive grants. Given the nature of our dataset, we considered a machine learning setting where productive research grants 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 grants 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 there is indeed a relationship between text features and grants productivity. However, the use of text features alone showed only a *weak discriminability*. Interestingly, the subject being approached (i.e. topical features) seems to be more relevant than the style (i.e. complexity) of the text provided in the title and abstract of the respective project. We also found that the obtained results do not depend on the considered language, since similar differences in performance were found for projects written in 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 not strongly correlated* with productive research grants, according to the adopted criteria for productivity. Therefore, there is a large space for improvement in performance, since other features can be used to characterize projects. Examples of possible features are the ones based on recent authors' performance. In this paper, we limited the sense of productivity by considering that productive grants are those yielding at least one publication. In future works, it is interesting to analyze other productivity criteria, including e.g. the total number of publications, the reputation of the respective journals and conferences and other measurements derived from citation and usage counts (Ruan et al. 2020; Hou and Yang 2020). We also intend to incorporate additional features in order to improve our predictions, including text network-based attributes (Stella and Zaytseva 2020; Stella et al. 2019; Amancio et al. 2015; Stella 2019) and other features related to researchers and their respective institutes (Correa Jr et al. 2018; Arruda et al. 2016).

Acknowledgements D.R.A. acknowledges CNPq-Brazil (Grant No. 304026/2018-2) for financial support. This study was financed in part by the CoordenaçÃo de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Predicting scientific success. *Nature, 489*(7415), 201–202.

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

Amancio, D. R. (2015b). Probing the topological properties of complex networks modeling short written texts. *PLoS ONE, 10*(2), e0118394.

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

Amancio, D. R., Oliveira, O. N., Jr., & Costa, L. F. (2012b). 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.

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & Costa, L. F. (2014). A systematic comparison of supervised classifiers. *PLoS ONE, 9*(4), e94137.

Amancio, D. R., Silva, F. N., & Costa, L. F. (2015). Concentric network symmetry grasps authors' styles in word adjacency networks. *EPL (Europhysics Letters), 110*(6), 68001.

Antiqueira, L., Nunes, Md. G. V., Oliveira, O., Jr., & Costa, Ld. F. (2007). Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications, 373,* 811–820.

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

Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics, 114*(2), 449–461.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brunet, E. (1978). Le vocabulaire de Jean Giraudoux. Structure et évolution.

Cabezas-Clavijo, A., Robinson-Garcia, N., Escabias, M., & Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? *PLoS ONE, 8*(6), e68258.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. Cambridge: MIT press.

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

Correa, E. A., Jr., Lopes, A. A., & Amancio, D. R. (2018). Word sense disambiguation: A complex network approach. *Information Sciences, 442,* 103–113.

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

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7,* 1–30.

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

Diller, J. W., Salters-Pedneault, K., & Gallagher, A. R. (2014). Effective dissemination requires effective talk: A comparison of behavior-analytic journals. *Behavior Analysis in Practice, 7*(2), 103–106.

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

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. Hoboken: John Wiley & Sons.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6*(9), e24926.

Fang, F. C., Bowen, A., & Casadevall, A. (2016). Nih peer review percentile scores are poorly predictive of grant productivity. *Elife, 5*(e13), 323.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., et al. (2018). Science of science. *Science, 359*(6379), eaao0185.

Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn: Machine learning in python*. Birmingham: Packt Publishing Ltd.

Germain, R. N. (2015). Healing the nih-funded biomedical research enterprise. *Cell, 161*(7), 1485–1491.

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

Hassoun, M. H., et al. (1995). *Fundamentals of artificial neural networks*. Cambridge: MIT press.

Hörlesberger, M., Roche, I., Besagni, D., Scherngell, T., François, C., Cuxac, P., et al. (2013). A concept for inferring 'frontier research' in grant proposals. *Scientometrics, 97*(2), 129–148.

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

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

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech rep: Naval Technical Training Command Millington TN Research Branch.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research, 39*(1), 19–37.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

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

Letchford, A., Preis, T., & Moat, H. S. (2016). The advantage of simple paper abstracts. *Journal of Informetrics, 10*(1), 1–8.

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

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

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

Li, D. C., Liu, C. W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine, 40*(5), 509–518.

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

Markowitz, D,, Powell, J., & Hancock, J.T. (2014). The writing style of predatory publishers. In: ASEE Annual Conference and Exposition, Indianapolis, IN.

Markowitz, D. M. (2019). What words are worth: National science foundation grant abstracts indicate award funding. *Journal of Language and Social Psychology, 38*(3), 264–282.

McKeown, K., Daume, H., III., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., , et al. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology, 67*(11), 2684–2696.

McManus, C., & Neves, A. A. B. (2020). Funding research in brazil. *Scientometrics, 126*(1), 801–823.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the gini importance? *Bioinformatics, 34*(21), 3711–3718.

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

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

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS one, 9*(12), e115844.

Piro, F. N., Aksnes, D. W., & Rørstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology, 64*(2), 307–320.

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods, 50*(3), 1198–1216.

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

Scarton, C., & Aluısio, S.M. (2010). 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, sn, vol 10.

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

Silva, T. C., & Amancio, D. R. (2012). Word sense disambiguation via high order of learning in complex networks. *EPL (Europhysics Letters), 98*(5), 58001.

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

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

Stella, M., De Nigris, S., Aloric, A., & Siew, C. S. (2019). Forma mentis networks quantify crucial differences in stem perception between students and experts. *PLoS ONE, 14*(10), e0222870.

Thelwall, M., & Nevill, T. (2008). Could scientists use altmetric. com scores to predict longer term citation counts? *Journal of Informetrics, 12*(1), 237–248.

Wager, E., Altman, D. G., Simera, I., & Toma, T. P. (2016). Do declarative titles affect readers' perceptions of research findings? a randomized trial. *Research Integrity and Peer Review, 1*(1), 1–5.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.

Wang, M., Yu, G., & Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications, 387*(18), 4692–4698.

Xie, Z., Ouyang, Z., Zhang, P., Yi, D., & Kong, D. (2015). Modeling the citation network by network cosmology. *PLoS ONE, 10*(3), e0120687.

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

CHAPTER 4

ON PREDICTING RESEARCH GRANTS PRODUCTIVITY VIA MACHINE LEARNING

Title	On predicting research grants productivity via machine learning
Authors	Jorge Valverde Tohalino, and Diego Amancio
Year	2022
Journal	Journal of Informetrics
Link	<https://www.sciencedirect.com/science/article/pii/S1751157722000128>
Situation	Published

4.1 Motivation

In our previous work, we evaluated the potential of textual features from project titles and abstracts as productivity indicators for Brazilian research grants. Our findings indicated the need to supplement these features with other types of data. However, one significant challenge we encountered was the limited availability of valuable information for each research grant. Only the project title, a brief summary, publications, and some data on the researchers involved are publicly accessible. Therefore, in this study, we aimed to generate bibliometric features from the available data on the primary researchers associated with each grant. To accomplish this goal, we added new information to the dataset we used for the paper presented in the Chapter 3.

The primary objective of this study was to determine whether bibliometric features extracted from each researcher can serve as distinguishing factors of productivity. We extracted these features by examining the academic history of each researcher using the available data. We believed that bibliometric features, which can quantify the characteristics of research projects, the experience of researchers, and the importance of host institutions, could be used to predict the productivity of research grants via machine learning. We aimed to investigate whether bibliometric features could be used to predict grant productivity and to identify which features are most relevant in predicting grant success. Overall, our motivation was to develop a more efficient and effective method for evaluating research grant proposals.

4.2 Contributions

This paper aimed to investigate whether bibliometric features could predict the success of research grants of Brazilian researchers in Medicine, Dentistry, and Veterinary Medicine. We extracted features related to the researchers' academic history, including research topics, affiliations, number of publications, and visibility. We then used machine learning to predict grant productivity. We found that research subject and publication history play a role in predicting productivity, and institution-based features were relevant when combined with other features. However, while the best results outperformed the text-based attributes examinated in the previous research, the evaluated features were not highly discriminative. Therefore, we concluded that predicting grant success, at least with the considered set of bibliometric features, is not a trivial task, and one cannot rely solely on machine learning to make predictions with high accuracy. We suggested that our findings may spark further discussions about the evaluation of research proposals and that further research could investigate whether wider contexts or other productivity and impact criteria could lead to improved results. We also stress that our conclusions are based on a dataset of research projects funded by São Paulo Research Foundation and that it remains to be seen whether the considered features display similar discriminative performance in other datasets.

Journal of Informetrics 16 (2022) 101260

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

On predicting research grants productivity via machine learning

Jorge A.V. Tohalino, Diego R. Amancio*

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

ABSTRACT

Understanding the reasons associated with successful proposals are of paramount importance to improve evaluation processes. In this context, we analyzed whether bibliometric features are able to predict the success of research grants. We extracted features aiming at characterizing the academic history of Brazilian researchers, including research topics, affiliations, number of publications and visibility. The extracted features were then used to predict grants productivity via machine learning in three major research areas, namely Medicine, Dentistry and Veterinary Medicine. We found that research subject and publication history play a role in predicting productivity. In addition, institution-based features turned out to be relevant when combined with other features. While the best results outperformed text-based attributes, the evaluated features were not highly discriminative. Our findings indicate that predicting grants success, at least with the considered set of bibliometric features, is not a trivial task.

1. Introduction

In recent years, *Science of Science* emerged as an important application of big data analysis (Fortunato et al., 2018). Owing to the availability of large data sets derived from the scientific literature, several studies have been conducted to shed light on how science is organized and evolves as a complex system (Zeng et al., 2017). Examples of approached topics include science evolution (Silva et al., 2016), collaboration/citation patterns and measures to evaluate science (Bar-Ilan, 2008). More recently, studies in *Science of Science* have also focused on predictive tasks, which has become very important in different scenarios (Acuna et al., 2012). For example, automatic approaches have been used to predict when a new topic will emerge (Salatino et al., 2017). In a similar fashion, neural networks representations have been able to predict outcomes of scientific research (Bagrow et al., 2018). Mobility trajectories of researchers have also been studied using computational methods (He et al., 2019). Equally important are those studies predicting scientific success, including the prediction of papers and scholars' impact (Wang et al., 2019).

More recently, several studies have focused on analyzing the factors underlying grants success (Boyack et al., 2018; Letchford et al., 2016; Paiva et al., 2012; Tohalino et al., 2021). Understanding the factors that may lead to successful grants are ultimately important to determine which proposals are the most promising and relevant to be funded. While machine-based techniques are not meant to replace an expert, thorough analysis of proposals, they may assist the analysis of a large number of documents and other metadata extracted from research proposals. Potential advantages associated with the use of machine learning methods to assist the analysis of proposals include an analysis less prone to personal bias, and a much faster review compared to traditional human evaluations. In addition, automatic analyses could also be used to understand the factors correlated to successful grants.

Text- and reference-based features have been used to predict the success of research proposals (Boyack et al., 2018; Letchford et al., 2016; Paiva et al., 2012; Tohalino et al., 2021). Boyack et al. (2018) found that proposals success depends on the topic being approached. More specifically, Boyack et al. (2018) found that subjects that have already been studied by the researcher are more likely to yield successful grants. The topic similarity, in this case, was computed by comparing proposal references and the respective applicant publications. A text analysis was conducted by Markowitz (2019). The authors studied if text complexity measurements extracted from NSF projects correlate with the amount of funding received by the researchers. They found that larger abstracts comprising a low number of common words are among the main patterns associated with larger funding values. In a similar study,

* Corresponding author.
E-mail address: diego@icmc.usp.br (D.R. Amancio).

https://doi.org/10.1016/j.joi.2022.101260
Received 20 June 2021; Received in revised form 30 December 2021; Accepted 27 January 2022

Tohalino et al. (2021) found that topical and complexity textual features play a role in grants predicting grants productivity, but the prediction values were not very high.

Different from other approaches, here we use machine-learning methods applied to features extracted from researchers, institutions and publications to analyze whether those features can be used to predict the productivity of grants. We used several features such as total number of publications, citations, relevance of PI's institution and diversity of the approached subjects. Because we are interested in identifying grants that produced any piece of valuable scientific work, we considered as criteria for productivity the publication of at least one scientific paper (Tohalino et al., 2021). While considering a larger threshold might also be interested, in terms of accuracy, we could not identify an improvement of performance when dealing with larger threshold values. Using a dataset of research grants from the *São Paulo Research Foundation* (FAPESP-Brazil), our analysis was conducted in three distinct research areas, namely Medicine, Dentistry and Veterinary Medicine.

Several interesting results have been found from our analysis. The analysis based on classifiers with a single feature showed that there is a relationship between the studied features and future productivity of grants. In this single-feature analysis, features based on research subjects and on publication/citation counts were the most effective to predict grants productivity. The analysis combining different features in the same classifier also showed an improvement in performance. The highest accuracy rates were found for the Veterinary Medicine area. In this case, we could discriminate productive grants with an accuracy higher than 67%. While the results are significant, the typical prediction accuracy was not very high. They were typically higher, though, than approaches based on textual features alone (Tohalino et al., 2021). Our analysis also revealed that both *Support Vector Machines* and *Multilayer Perceptron* were the classifiers yielding the highest accuracy rates. Despite being a challenging task, we believe that the studied features could be combined with additional information to allow a better understanding of the factors correlated with grants success.

This manuscript is organized as follows. In Section 2, we describe the methodology, including the description of features and the machine learning framework. We discuss the obtained results in Section 3. Finally, in Section 4, we present the conclusions and perspectives for future works.

2. Material and methods

In order to classify research grants according to their productivity, the following steps were taken:

1. *Dataset collection*: the dataset we used comprises research projects supported by *São Paulo Research Foundation* (FAPESP-Brazil). The dataset is available from the *Biblioteca Virtual* website (see Section 2.1). In addition to the information regarding research projects (number of papers derived from the grant, title, abstract etc), the dataset also provides information to characterize PIs (e.g., their publication history) and institutions (e.g. universities and research institutes).
2. *Feature extraction*: this step is responsible for extracting features from researchers that are used to predict grants productivity. Our hypothesis is that the success of a grant could be dependent on PIs features, such as previous success in other grants and publication/citation history. Several features were extracted to characterize authors. Examples of extracted features are: number of funded projects, number of publications and citations yielded by the researcher's grants, affiliation and diversity of subfields studied by the researcher. Section 2.2 describes the features we used to perform the classification.
3. *Classification*: the aim of this phase is automatically identify productive research proposals according to the established criteria for productivity. We considered a binary classification task. The features extracted from the previous step were used as input for traditional machine learning algorithms. We also performed several tests in order to find the best combination of features. In Section 2.3, we describe the classification step. This phase includes the training and evaluation phase.

In Fig. 1, the architecture for research grant classification is shown. First, we collected relevant information from the FAPESP Dataset (FAPESP Virtual Library). This includes information from PIs that are related to their previous research experience and other features linked to their professional activity. All information from researchers are collected in the researcher dataset. Examples of features extracted are the number of publications obtained in previous grants (*pubFeat*), number of citations received by these publications (*pubCitFeat*) and other features that are detailed later on. These features are used to train supervised classifiers in a binary classification task to predict whether a grant will be productive. We use the number of publications resulting from the grant as the criterion to measure productivity.

2.1. Dataset collection

The considered dataset comprises a subset of grants offered by *São Paulo Research Foundation* (FAPESP) (Tohalino et al., 2021). FAPESP is an important public research foundation in Brazil and is fully funded bythe State of São Paulo.[1] The metadata regarding PIs and research grants were retrieved from *Biblioteca Virtual* website.[2] We focused our analysis on *regular grants*, which are grants with an average duration of 18–24 months. All FAPESP regular grants are conducted under the supervision of a principal investigator, who must be associated with a university (or research institution) from São Paulo. We decided to study this type of grant because the *Biblioteca Virtual* has a large number of regular grants (roughly 31,000 instances). We selected research grants starting before 2016 with duration between 23 and 24 months. Because recent grants were disregarded, all considered projects had at least 3 years to yield at least one publication after the grant is finished.

[1] https://fapesp.br/en/about.

[2] https://bv.fapesp.br/pt/.

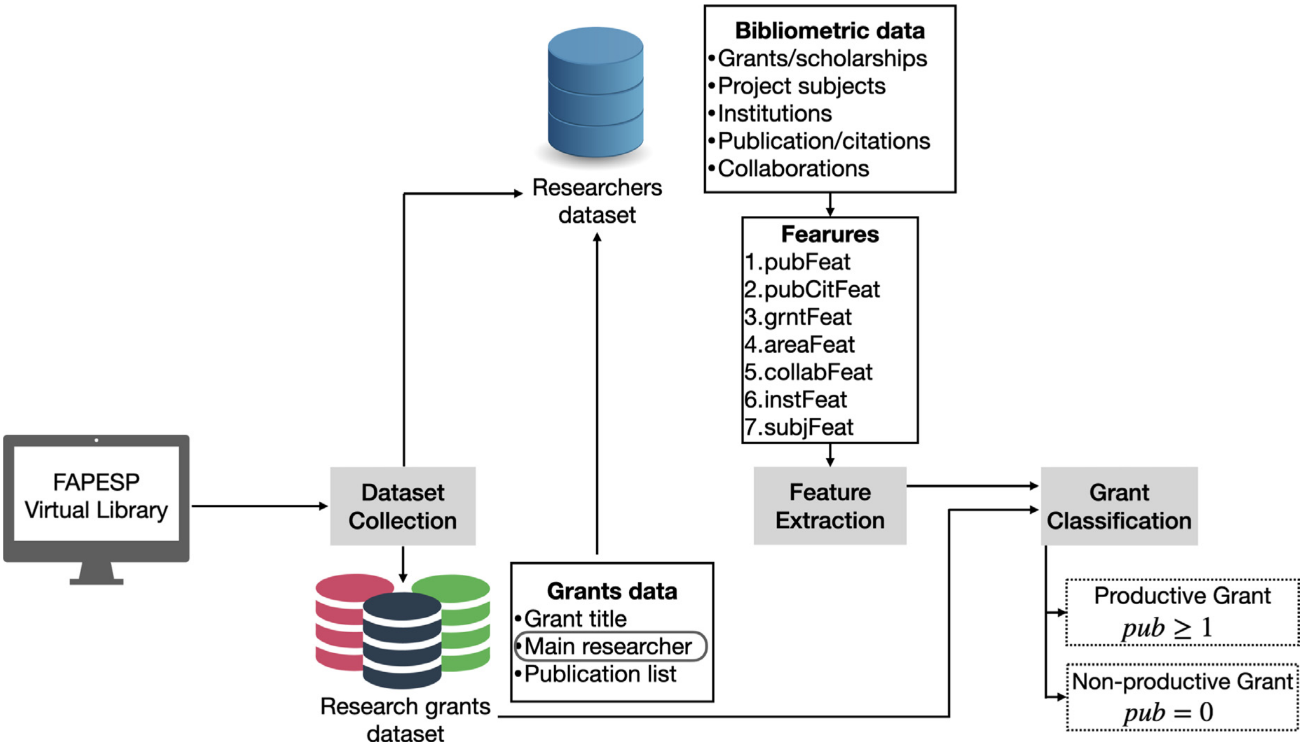


Fig. 1. Architecture of the proposed methodology for research grant classification.

Each research grant has a list of associated publications. This information is automatically extracted from the Web of Science (Clarivate Analytics) dataset by the *Biblioteca Virtual* website. This automatic extraction is possible because any paper published in the context of a FAPESP research grant must acknowledge FAPESP in a specific format. Regular grants particularly are acknowledged using the format “yyyy/nnnnn-d”, where “yyyy” represents the year when the research project was submitted and “nnnnn-d” is the grant number.

Grants funded by FAPESP cover a wide variety of research areas, including, e.g. Health, Biological and Earth Sciences. Because distinct areas have different publication patterns (Rafols and Meyer, 2010), we decided to compare grants *within* the same field. This analysis was not carried out in several areas due to the low number of relevant projects to our study. The limitation of these areas was the very low number of grants with at least one publication. In this sense, our analysis was conducted in the three largest fields: Medicine (MED), Dentistry (DENT) and Veterinary Medicine (VET).

2.2. Feature extraction

For each research grant, we extracted several features related to the respective PI. The features are meant to characterize researchers’ academic trajectory just before the grant started. The features used to characterize PIs were grouped into the following groups:

1. *Publication-based features* (*pubFeat*): here we use features that are related to PIs publications. Our hypothesis here is that a good previous performance related to publications could be a good indicator of future performance (Lu et al., 2019). The following publication-based features were used for this analysis:
   (a) total number of publications;
   (b) number of grants yielding at least one publication;
   (c) maximum number of publications resulting from a single grant received by the PI;
   (d) average productivity (in number of publications); and
   (e) number of grants at least one publication divided by the total number of grants received by the PI.
2. *Features based on both publications and citations* (*pubCitFeat*): While scientific publications denote researchers effort to provide new pieces of knowledge, citations can be considered as a metric of relevance and visibility (Kong et al., 2020; Siudem et al., 2020). Our hypothesis is that citations can be used as a proxy to PIs scientific influence (Ioannidis et al., 2020). Thus, we investigate whether influential researchers are more likely to conduct productive research grants. The following PIs measurements were considered for this set of features:
   (a) total number of articles published;
   (b) total number of citations accrued by the researcher;
   (c) average number of publications per year; and
   (d) average number of citations per year.

3. *Features based on the number of grants and scholarships received by the PI* (*grntFeat*): our hypothesis is that more experienced researchers are more likely to have a productive grant (Larrimore et al., 2011; Markowitz, 2019). The degree of researcher's experience was measured in terms of the total number of grants and scholarships received by the researcher. In addition to regular grants, we also considered as features the number of undergraduate, master's, doctoral and post-doctoral degree scholarships supervised by the researcher.
4. *Features based on the diversity of research areas* (*areaFeat*): our hypothesis here is that PIs might have experience on diverse research areas, and this could be an indication of future productivity. Some studies have shown that interdisciplinary journals and papers are more visible in the sense that they tend to attract more citations than more specialized research (Rinia et al., 2002). In a similar fashion, our hypothesis is that interdisciplinary research could be more visible and this could facilitate the publication of papers since more journals and scholars could be interested in the interdisciplinary results being disseminated. The features used to quantify the degree of interdisciplinary encompasses three different granularity levels. We considered the number of areas in each of the first three levels. Examples of top-levels areas include Exact and Health Sciences. Examples of second-level hierarchy areas for Health Sciences include e.g. Nursing, Pharmacy, Medicine and Dentistry. Finally, examples of third level areas for Medicine include Medical Clinic, Maternal Health, Surgery and Psychiatry.
5. *Collaboration-based features* (*collabFeat*): in this set of features we evaluate whether the number of different collaborators in the past might be correlated with grants success. A large number of collaborators could be a proxy to quantify researchers' experience (or even seniority) and thus collaboration-based features could indicate if an author is able to gather researchers with different backgrounds to conduct scientific research. Because more collaborations could be also correlated with more distinct contributions (Corrêa Jr et al., 2017), we could also expect that joint effort could be correlated with higher quality research (Franceschet and Costantini, 2010), which in turn could positively contribute to the success of a research grant. The following features were used to quantify the diversity of PIs collaborations:
   (a) total number of local collaborators in research grants. Local researchers are all researchers affiliated to Brazilian research (or higher-education) institutions;
   (b) total number of abroad collaborators;
   (c) total number of grants received by the PI with one or more associated researchers;
   (d) total number of distinct co-authors in scientific publications; and
   (e) average number of co-authors per article.
6. *Institution-based features* (*instFeat*): institution-based features are used to probe whether PIs affiliation plays a role in predicting the success of research grants. The hypothesis is that grants conducted at larger (or more visible) institutions are more likely to yield a productive grant. More prestigious institutions could favour productivity given that more prestigious institutions could themselves host more productive researchers (Bauder, 2020). In addition, more prestigious universities could also have more access to the material and resources to conduct high-quality research. The visibility and importance of institutes were measured in terms of the following features:
   (a) total number of projects hosted by the PI's institution;
   (b) total number of publications associated with the PI's institution; and
   (c) total number of productive grants hosted by the PI's institution. We used here the criteria discussed in Section 2.1 to classify a grant as productive.

   The set of features mentioned in (a)–(c) are henceforth referred to as *instFeat*$_A$. We also considered an additional data representation, where we do not consider the features extracted from the institution, but the host institution becomes a feature. More specifically, a vector is used to represent if the PI belongs to a specific institution. The ith element of the vector takes the value 1 if the PI is affiliated to the ith institution. Otherwise, the value stored is zero. This representation is henceforth referred to as *instFeat*$_B$. We also used a different representation referred to as *instFeat*$_C$. This representation uses a vector that is similar to the previous version, but instead of assigning the value 1, we assigned the value of the success rate of the researcher's university or institution.
7. *Success of research subjects* (*subjFeat*): each research project in the dataset comprises keywords (or keyphrases) that help to describe the main topics approached by the research. Our hypothesis here is that topics may play an important role in predicting the productivity of grants, since particular research lines might have higher levels of productivity (Tohalino et al., 2021). In order to analyze whether productivity has a dependency on the research subject, we considered two measurements taking into account the success history of different subjects. The *global* importance considers the success of a subject in the whole dataset. Differently, the *local* importance considers the success of subjects in grants conducted by the PI being analyzed. The success rate of a subject X is computed as the number of productive grants approaching X divided by the total number of grants associated with X. The criteria used to characterize productivity is detailed in Section 2.1. Three different sets of features were considered to represent the success of research subjects:
   (a) $subjFeat_A$: each researcher was characterized using a vector summarizing the local and global success of the approached subjects. Because many subjects might be related to a PI in previous projects, we summarized the success rate of the approached subjects observed for keywords. In particular, we considered the average, the standard deviation and the maximum success rate observed for the keywords. Therefore, for each PI, six features were considered: both local and global strategies were used to compute the success rate, and three summarization strategies were applied.
   (b) $subjFeat_B$: we first obtained the k-most frequent subjects of the researcher. Then we calculated the global success rate and local success rate vectors for these subjects. The generated vectors were considered feature vectors. We evaluated with k ranging between 10 and 50 subjects.

(c) $subjFeat_C$: This version is similar to $subjFeat_A$, but instead of considering the success rate, we considered the frequency count of the researcher's subjects.

2.3. Classification

The main purpose of this study is to probe whether bibliometric information of researchers (e.g. their publication history and participation in previous research grants) are relevant factors to predict productivity of their research grants. In order to address the problem of class unbalancing in the classification scheme (Li et al., 2010), we considered a grant as productive if it yielded at least one publication. Thus, for each considered research field, the total number of positive (i.e. grants with at least one publication) and negative instances are more regularly distributed. This is compatible with previous related research (Tohalino et al., 2021). If a higher threshold was considered to label an instance as positive, only a small percentage of grants would be considered as positive and this effect would lead to a high level of class unbalancing (see Fig. B.3 in Appendix B). The fraction of positive instances found for MED, DENT and VET are 41.6%, 49.5% and 32.9%, respectively. According to a previous study, we found that the dataset is consistent with regard to the fraction of positive instances (Tohalino et al., 2021). The consistency of the dataset considered papers acknowledging grant support in an interval of 3 years after the grant is finished. This is consistent with similar analyses conducted with NIH grants (Fang et al., 2016).

To analyze the relationship between the extracted features and grants productivity, we used the following machine learning algorithms: k-Neatest Neighbors (kNN), Support Vector Machines (SVM), Naive Bayes (NB), Neural Networks (MLP) and Decision Trees (DTrees). The algorithms hyperparameters were optimized using the strategy described in Amancio et al. (2014), Rodriguez et al. (2019). We used these algorithms because they use different strategies to identify data patterns and therefore can complement each other (Amancio et al., 2014). In addition, optimized results can be obtained even if a very large dataset is not available for training (Amancio et al., 2014). Such a variety of classifiers has also been used in related works (Tohalino et al., 2021). A description of the used algorithms and the strategy used to balance the classes are described in Appendix A.

The evaluation of the classification was based on the 10-fold cross-validation method to split the dataset into training and test datasets (Duda et al., 2012). We also tested the significance of the classification results by using a permutation test (Ojala and Garriga, 2010). In this test, a null distribution is generated by computing the accuracy of the classification in versions of the dataset comprising the same set of features, but with random permutations in instances label (Ojala and Garriga, 2010). In addition, in order to compare the performance of classifiers, we used ranking diagrams to evaluate whether differences in performance are statistically significant. When comparing classifiers performance over multiple datasets, the average ranks are considered different if the average rank difference is higher than the value computed via the Nemenyi post-hoc test (Demšar, 2006). This value is the critical distance (CD). Graphically, the critical distance is represented as a line above the x-axis. An example of ranking diagram is illustrated in Fig. B.4 (Appendix B). In this study, rank diagrams were generated for the Nemenyi test with $p - values = 5 \times 10^{-2}$.

3. Results and discussion

In this section, we discuss the results we obtained when evaluating whether the considered features can be used to predict the productivity of grants. We divided our analysis into the following sections: In Section 3.1, we report the performance of different features when they are individually evaluated. In Section 3.2, we discuss the results we obtained when researcher features are combined. The relevance of features for the classification task is also analyzed. In Section 3.3, we report the results obtained when combining several classifiers via voting method.

3.1. Performance analysis of single features

In this section, we discuss the results when each family of features is individually analyzed. The obtained results for each of the considered research areas are shown in Fig. 2. For the Medicine area, we observed that the best result was found with subject-based features ($subjFeat_B$), meaning that the average success of the topic being approached plays a role in predicting productivity. In the best scenario, the accuracy rate reached roughly 62% ($p - value < 1.0 \times 10^{-3}$). While this is not a very high accuracy rate, this result turned out to be significant. We also found that, for this research area, publication and citation history also play a role in predicting productivity. Both previous productivity (*pubFeat*) and impact (*pubCitFeat*) have a significant role in predicting grants productivity (both p-values were lower than 1.0×10^{-3}). All other features were found to be not statistically discriminative ($p - value \geq 5.0 \times 10^{-2}$). Particularly, we found a very low discriminative performance for a particular institution feature (*instFeat*$_A$) that considers the total number of projects hosted by the institution. Because this feature might be related to the size of the institution, this result suggests that being in a larger university is not necessarily linked to a higher productivity. Surprisingly we found that the diversity of areas can not be used as a source of productivity. While interdisciplinary researchers usually attain better performance in funding (Sun et al., 2021), we did not observe any significant relationship between grants interdisciplinarity and productivity.

When analyzing the results obtained for the Dentistry area, we also found similar results. The productivity rate of the approached subject (*subjFeat*$_C$) was found to be the most relevant feature to identify productive Dentistry grants. In this case, the highest accuracy rate reached 63.0%. We also found that publication and citation-based features also generated statistically significant results. All other features were not able to predict productivity.

The highest accuracy rates were found for the Veterinary area. An accuracy of roughly 66% was obtained with the MLP and SVM methods. Once again we observe that the previous success of a topic is correlated with future success. In a similar fashion, publication

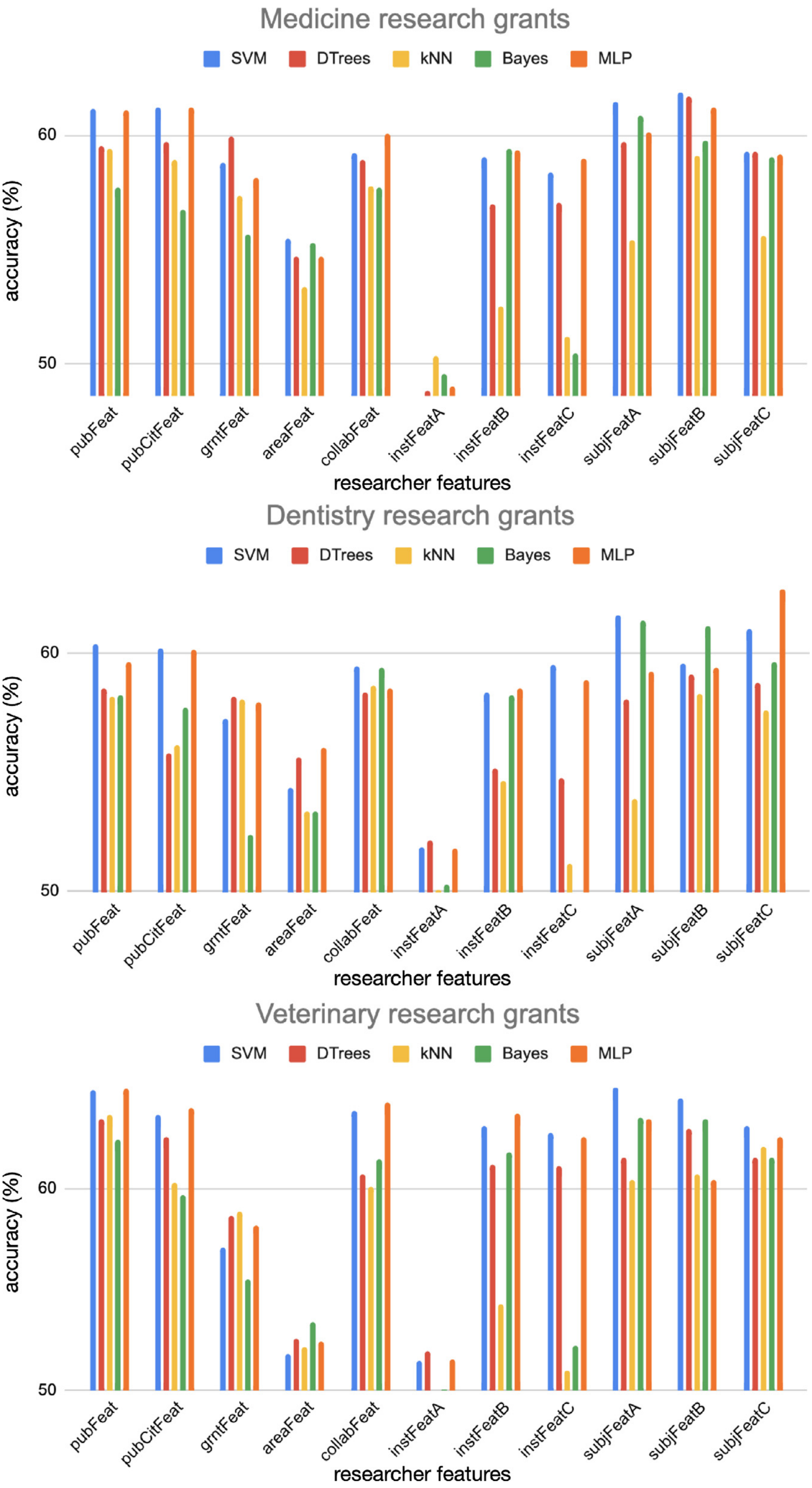


Fig. 2. Results based on accuracy rate obtained from the evaluation of each researcher feature. We considered projects from the following areas: Medicine, Dentistry and Veterinary Medicine. From each researcher we considered seven features and their variations.

Table 1
Results based on accuracy rate obtained by performing feature combinations. We show the results from the best combination of features for each research area (Medicine, Dentistry and Veterinary Medicine). The best results for each research area are highlighted. The SVM classifier always achieved the highest scores for all cases.

Method	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
DTrees	61.96 ± 1.12	62.08 ± 0.90	65.46 ± 1.59
SVM	**62.82 ± 0.67**	**62.50 ± 1.22**	**66.57 ± 1.57**
kNN	59.29 ± 0.88	61.58 ± 0.91	64.27 ± 1.49
Bayes	60.34 ± 0.41	61.00 ± 0.37	64.06 ± 1.61
MLP	60.99 ± 0.61	60.28 ± 0.61	64.22 ± 0.82

Table 2
Results obtained from the evaluation of feature selection methods. We tested with the most 10, 20, 50 and 100 relevant features. The best results for each research area are highlighted. The accuracy rate was obtained with the SVM method, since this method provided the best results.

k relevant features	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
Top-10 features	59.94 ± 1.09	60.35 ± 1.18	67.35 ± 1.28
Top-20 features	60.24 ± 1.00	60.68 ± 0.52	**67.37 ± 1.14**
Top-50 features	61.99 ± 0.55	61.74 ± 1.15	66.98 ± 1.55
Top-100 features	**62.58 ± 1.11**	**62.77 ± 0.46**	65.77 ± 1.14

and citation-based features also displayed significant accuracy rates. Differently from Medicine and Dentistry, a statistically significant relationship between collaboration features and productivity has also been found. This means that the number of previous PI's collaborators could be an indicator of future grant productivity. A correlation was also found for two institution-based features ($instFeat_B$ and $instFeat_C$). While the total number of projects hosted by the PI's institution is not correlated with productivity, the fraction of productive grants have a higher correlation rate. Similarly, the total number of publications associated with the PI's institution also seems to be correlated with productivity.

Concerning the methods, the performance varied according to the considered feature and dataset. When considering only the best results across all variations of features and methods, we observed that the highest accuracy rate was found with SVM (Medicine and Veterinary Medicine) and MLP (Dentistry). In fact, the ranking diagram of performance (see Fig. B.4 in Appendix B) confirmed that SVM and MLP displayed equivalent performance when considering single features. In general, the worst accuracy rates were obtained with kNN and Naive Bayes. The best results are highlighted in Appendix B.

3.2. Performance analysis of feature combinations

While in the previous section we analyzed the discriminability of each feature when they are used individually, here we analyze whether combinations of features can lead to optimized results. In addition to considering all features, we also considered other feature selection algorithms to find an optimized combination of attributes (Kou et al., 2020). The first approach used to combine features considered random subgroups of features. We considered feature sets of different sizes and we used them as input for the classification systems. In Table 1, we show the best accuracy rate found in each dataset, with different classifiers. We found that all best results were found to be statistically significant, and the best result being found for Veterinary Medicine (66.5%). As observed in the single feature analysis, the results are not highly discriminative. In addition, the results show that the combination of features does not improve the accuracy rates by a very large value when one compares with the results found with single features. Considering the best classifiers, the improvement in performance – when it occurs – is lower than 1%. Regarding the methods, we note that SVM outperformed all other classifiers in all datasets. This is evident in the ranking diagram displayed in Fig. B.5 (Appendix B). However, similar results were obtained with other classifiers, especially with Decision Trees.

In our analysis we investigated whether the threshold for considering a grant as productive can affect the result of the classification. Differently from the results displayed in Table 1, we show in Appendix B (seeTables B.6–B.8) the accuracy rates found when considering that a grant is productive when it yields more than two, three or four papers. The results showed that there is no large improvement in performance if such a threshold is modified. For this reason all subsequent analyses in this paper will classify a grant as productive if it yields at least one paper. Also, as observed in Fig. B.3 (Appendix B), it is not suitable to choose high threshold values, because the datasets would be very unbalanced.

In addition to the approach based on a random selection of features, we also used an approach based on the Gini coefficient (Nembrini et al., 2018). This approach is widely used to find relevant features in methods based on decision trees (Pedregosa et al., 2011) and has also been used in the scientometrics context (Tohalino et al., 2021). Using the Gini index, each feature was given a relevance value and then we selected the top k-features to analyze the gain in discriminability, with k

Table 3
Accuracy rates obtained from the evaluation of voting algorithms. For each research area we show the results of the following methods: All (when the results of all proposed classifiers are combined into the voting system) and Best (when only SVM and MLP are considered). The best results for each research area are highlighted.

Method	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
All	59.67 ± 0.62	57.88 ± 0.36	64.61 ± 1.36
Best	**62.07 ± 0.89**	**62.50 ± 0.74**	**65.00 ± 0.80**

ranging between 10 and 100. Then, we used the selected features along with the SVM algorithm in order to perform the classification process. We considered the SVM classifier because it displayed the highest accuracy rates with the selected features.

The results obtained with the Gini feature selection are displayed in Table 2. Overall the results show that there is no significant improvement in performance when one compares with the results obtained with the results displayed in Table 1. However, a small improvement can be observed mainly for the Veterinary Medicine Area. While this strategy was not useful to improve significantly the performance, this allowed an improved representation since similar results were obtained with a much smaller set of features (10 features in the case of Veterinary Medicine). We also note that a large set of features – even when selected via Gini method – does not necessarily improve the discriminability rate.

In addition to providing a compact representation, the feature selection algorithm allowed us to investigate which features are the most important for the classification task. This analysis is different from the analysis performed in the previous section because different discriminability performances can be observed when features are *combined* (Amancio et al., 2011). According to the Gini coefficient, the most relevant features for each field are:

- *Medicine*: institution-based and subject-based features displayed the highest Gini values. The success rate of the projects hosted by the PI's institution was a relevant feature in the family of features related to the institution ($instFeat_C$). In a similar fashion, the success history associated with the approached subject was an important feature. This result suggests that, when used in combination with other features, the success history of both institution and approached subject are relevant to predict the output of research projects.
- *Dentistry*: $instFeat_C$ and $subjFeat_A$ were found to be the most relevant features. This result is similar to the one found for Medicine. The other variations related to subject-based features were also relevant: both history of global and local success of approached subjects were important for the task.
- *Veterinary Medicine*: here the most relevant feature was the history of the PI's publications (*pubFeat*). The other important variables were institution based features and all features based on the relevant of subject features ($subjFeat_A$, $subjFeat_B$ and $subjFeat_C$).

The vast majority of relevant features are based on subject features (*subjFeat*). Consequently, these results confirm the good performance obtained from the subject-based features when they are only considered for the classification systems (according to the results shown in Section 3.1). We also observed that some variations of the features based on the institution of the researcher (*instFeat*) have a degree of importance to characterize the performance of a researcher. However, it is important to recall that these features performed poorly when they were considered individually, while their performance improved when they were evaluated together with other features in a combined approach. These results indicate that institution-based could be an important factor, but individually they did not display a discriminative power. It is also interesting to note that the publication can also play a role in predicting success. This measurement turned out to be particularly important for the Veterinary Medicine area, even when used as a single feature (see Fig. 2).

3.3. Performance analysis of ensembles: voting algorithm

In the previous section, we analyzed if the combination of different features is able to improve the discriminability rates. Here we combine different classifiers to analyze if evidence from multiple methods can lead to optimized results. For this, we used a voting algorithm (Kiziloz, 2021). Two strategies were considered: (i) the use of all considered classifiers and (ii) the use of the best classifiers (see Material and Methods). According to the results presented in the previous sections, the classification systems that achieved the highest accuracy rates were the SVM and MLP algorithms, while Naive Bayes and kNN obtained the worst performance. We considered all the combinations of features described in the previous sections.

We show in Table 3 the results obtained from these evaluations. The obtained results revealed that the combination of classifiers did not improve the results obtained in previous sections. In particular, using all classifiers is not useful given the low performance achieved especially by Naive Bayes and kNN. The performance of SVM + MLP combination was also not useful to improve the performance of the classification.

In sum, in the considered dataset, we found that combining different evidence from different features is more important than combining different classifiers. While some classifiers perform better than others, this result indicates that additional features could be used to improve the predictive power of the classifiers. The study conducted here showed that the features we used are more relevant than features based on topical or complexity textual features (Tohalino et al., 2021). However, additional text information could also be obtained from project abstracts and complement the characterization of research projects in order to improve the predictability of

Table 4
Summary of the best results for each proposed approach. We highlighted the highest scores for each research area. The highest accuracy was found for the Veterinary Medicine field. Four different approaches were considered: (i) classification based on a single feature; (b) random combination of features; (iii) classification based on feature selection; and (iv) combination of classifiers via voting strategy.

Method	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
Single feature	62.07 ± 0.70	**63.00 ± 0.55**	65.69 ± 1.24
Features combination	**62.82 ± 0.67**	62.50 ± 1.22	66.57 ± 1.57
Feature relevance	62.58 ± 1.11	62.77 ± 0.46	**67.37 ± 1.14**
Voting system	62.07 ± 0.89	62.50 ± 0.74	65.00 ± 0.80

grants productivity. Additional text representations, including those based on network science (Tohalino and Amancio, 2018) could also be used to characterize research grant texts.

4. Conclusion

In this paper, we evaluated whether it is possible to predict the productivity of research projects by considering many different features to describe scientific entities. Our analysis was conducted in three large subareas and considered grants awarded by *São Paulo Research Foundation*, one of the largest research agencies in Brazil. We considered several features that could quantify the characteristics of research projects, PI's experience and the importance of host institutions. The relationship between the features and productivity was analyzed in the context of a traditional classification task. We extracted several features based on the academic activity of Brazilian researchers. In addition to being useful to quantifyproductivity and characterize emerging disciplines, thematic associations and collaboration networks (Ellegaard and Wallin, 2015), the selected features are intended to grasp the experience of the researcher in conducting research. The selected features have been used by *São Paulo Research Foundation* to evaluate research projects via peer review.

Our analysis considered four different approaches to combine features and classifiers. First, we analyzed classifiers created with only a single feature. We then combined features via feature selection and relevance analysis. We also combined classifiers in a voting algorithm. Overall we found that the best results in all four considered approaches are statistically significant, meaning that some of the features play a role in predicting the output of research projects. The main results are summarized in Table 4. All best results were found to be significant, though none of them reached a 70% accuracy rate. The best results in different areas were obtained with distinct strategies. A single feature (the approached subject) was able to provide the highest accuracy rate for the Dentistry area. The combination of features was able to provide the highest accuracy for Medicine. A feature selection algorithm finally provided the best results for Veterinary Medicine. Our analysis also revealed that the voting system combining evidence from multiple classifiers did not provide any improvement in classification performance.

While we found a dependency between the features and the output of projects, in all considered areas, the accuracy rates were not very high. This reinforces the fact that predicting the output of grants is not a trivial task and one can not rely only on machine learning to make predictions with high accuracy, at least with the considered features. The results found in this study were slightly better than the ones found using only textual features (Tohalino et al., 2021).

In terms of evaluations conducted in the context of *Science of Science*, our results may spark further discussions. We found that no single feature is a good predictor of productivity, therefore our results suggest that any of the considered features should be used as a *single factor* to decide whether a research grant should be funded, even if evaluations are performed by experts. Our results also suggest that even a combination of the considered features is not an excellent predictor of productivity. Therefore, a careful analysis of the content or research proposals is essential to assess the quality of the proposed ideas. Similar conclusions have been disseminated with regard to using a single index to predict scientific output at the paper level (Schreiber, 2013). While the considered features alone can not be used in an independent way to predict success, they could be used to assess whether a researcher has experience on a specific field. We also stress that our conclusions are based on a dataset of research projects funded by FAPESP-Brazil. It remains to be probed, therefore, whether the considered features display similar discriminative performance in other datasets.

In future works, it would be interesting to analyze whether the use of wider contexts could lead to improved results. In text analysis, the access to the full content of research projects could provide more information than the title and abstract. A possible analysis could be the extraction of textual patterns via network analysis (Amancio et al., 2012b; Marinho et al., 2016; Teixeira et al., 2021). Unfortunately, full textual information is not currently available in our dataset.

Further extensions of this work could also be investigated. This includes other productivity and impact criteria, such as the total number of citations received by a grant, the reputation of journals and conferences associated with grant publications and other quality and impact criteria. We could also use collaborative network-based approaches (Amancio et al., 2012a; 2015; Corrêa Jr et al., 2017) to analyze whether scientific collaborations and team formation strategies may play a role in grants productivity. In a more general way, it would be interesting to conduct further research to investigate whether some of the considered features could be an indicator of scientific breakthrough. This investigation is analogous to similar works conducted at the paper level (Min et al., 2021; Ponomarev et al., 2014). Predicting scientific breakthroughs is more complex because it would require a much larger and heterogeneous dataset. In a similar fashion, some criterion would be required to objectively identify and characterize scientific breakthroughs.

CRediT authorship contribution statement

Jorge A.V. Tohalino: Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing. **Diego R. Amancio:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Appendix A. List of supervised classifiers

In this section we describe the main classifiers used to predict the productivity of research grants. In addition to the traditional machine learning algorithms, we also used a technique to combine all considered pattern recognition methods.

1. *k-Neatest Neighbors* (kNN): With the aim of classifying an unknown element from the dataset, the kNN method first selects the k-nearest elements from the training dataset. Then, the category assigned to the unknown element corresponds to the majority class which is detected in the selected k-set. k is an important parameter of the algorithm and is chosen via optimization methods (Amancio et al., 2014).
2. *Support Vector Machines* (SVM): Given the training examples, this method constructs a hyperplane with the aim of finding a separation between the classes of the dataset. This method has several parameters, including parameters that sets the kernel function used to create a hyperplane (Amancio et al., 2014).
3. *Naive Bayes* (NB): This classifier is a supervised learning algorithm the uses Bayes' theorem with a strong assumption that features are independent (McCallum et al., 1998). In this sense, the following equation is used to predict the class $\hat{y}$:

   $$\hat{y} = \underset{y}{\text{argmax}}\ P(y) \prod_{i=1}^{n} P(x_i|y) \tag{A.1}$$

   where x_i is a feature. The training dataset can be used to compute the probabilities $P(y)$ and $P(x_i|y)$. Additional parameters related to this method and the optimization process are described elsewhere (Amancio et al., 2014).
4. *Multilayer Perceptron* (MLP): This method is based on a neural network model considering one or more hidden layers which has a training process that usually involves the Back-Propagation algorithm (Haykin, 2008). Two main hyper-parameters exist in this model: (i) the number of layers; and (ii) the number of neurons for each layer. These parameters can also be chosen via optimization.
5. *Decision trees*: decision tree methods create models that are able to predict the value of a target variable by learning decision rules. The rules are inferred from several input variables, i.e. the features. A typical decision tree comprises nodes and edges, where nodes represent features and edges represent the decision chosen for each attribute. Each internal with children is labeled with some input feature, while leaf nodes are labeled with a class. The classification process starts at the root node and ends when a leaf node is reached. As the decision walks through nodes, a rule is applied and the decision on that rule guides the choice of the children to be chosen as next step. Once the decision reaches a leaf node, the predicted label corresponds to the respective label stored in that node (Breiman, 2001)
   The choice of features that are evaluated in each node depends on the feature relevance metric that decides which feature best discriminates the dataset. One important relevance metric is the Gini impurity (Nembrini et al., 2018), a metric that has already been used to evaluate the relevance of features in the context of productivity prediction (Tohalino et al., 2021). The Gini impurity measures how often a randomly selected instance from the dataset would be mislabeled if it was randomly classified according to the distribution of the categories in the considered subset. A feature is considered significant in a given node if the test performed using that feature results in a decrease in the Gini impurity. The relevance of a feature is obtained by averaging the decrease in impurity computed in all tree nodes using the considered feature.
6. *Ensemble learning*: ensemble methods combine the predictions of several machine learning algorithms in order to obtain a better predictive performance over a single method (Dietterich, 2000). Most ensemble methods build several estimators independently, and then they average the predictions of each method. Usually the voting method is a simple, yet effective approach designed to combine the predictions from several supervised classifiers. In this approach, the input classifiers are trained and tested independently. Then the observed predictions from all classifiers are combined by using a majority vote to predict the class labels. Therefore, the class receiving the highest number of votes is chosen as the final predicted class (Ruta and Gabrys, 2005). When a draw occurs, we used the membership strength (Kumbure et al., 2020) provided by each classifier to make a decision.

Appendix B. List of additional results

Table B1
Accuracy rate obtained when considering the classification with single features. We considered projects from the following areas: Medicine, Dentistry and Veterinary Medicine. The best results for each classifier are highlighted.

Features	Research Projects on *Medicine*				
	DTrees	SVM	kNN	Bayes	MLP
	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)
pubFeat	59.48 ± 0.64	61.24 ± 0.40	**59.36 ± 0.81**	57.51 ± 0.43	61.21 ± 0.53
pubCitFeat	59.67 ± 0.85	61.36 ± 0.36	58.80 ± 0.35	56.56 ± 0.42	**61.37 ± 0.39**
grntFeat	59.91 ± 0.75	58.68 ± 0.63	57.14 ± 0.92	55.43 ± 0.53	57.98 ± 1.31
areaFeat	54.45 ± 0.86	55.25 ± 0.93	53.13 ± 1.11	55.05 ± 0.39	54.46 ± 0.95
collabFeat	58.84 ± 0.93	59.14 ± 1.06	57.63 ± 0.73	57.51 ± 0.42	60.06 ± 0.99
$\mathit{instFeat}_A$	48.94 ± 1.25	48.73 ± 1.48	50.28 ± 0.05	49.59 ± 0.47	49.09 ± 1.16
$\mathit{instFeat}_B$	56.80 ± 0.62	58.94 ± 0.67	52.33 ± 2.24	59.37 ± 0.81	59.30 ± 0.74
$\mathit{instFeat}_C$	56.85 ± 1.13	58.25 ± 0.46	51.08 ± 0.53	50.40 ± 0.43	58.90 ± 0.49
$\mathit{subjFeat}_A$	59.70 ± 0.53	61.61 ± 0.34	55.17 ± 2.12	**60.95 ± 0.44**	60.16 ± 0.94
$\mathit{subjFeat}_B$	**61.89 ± 0.93**	**62.07 ± 0.70**	59.05 ± 0.83	59.77 ± 0.36	61.32 ± 1.02
$\mathit{subjFeat}_C$	59.24 ± 0.92	59.24 ± 0.76	55.35 ± 1.98	58.93 ± 0.60	59.06 ± 0.65
Features	Research Projects on *Dentistry*				
	DTrees	SVM	kNN	Bayes	MLP
pubFeat	58.40 ± 0.96	60.41 ± 0.26	58.05 ± 1.03	58.09 ± 0.40	59.56 ± 0.68
pubCitFeat	55.57 ± 0.86	60.24 ± 0.11	55.90 ± 0.39	57.52 ± 0.17	60.12 ± 0.56
grntFeat	58.02 ± 0.59	57.03 ± 0.54	57.92 ± 0.50	52.18 ± 0.55	57.78 ± 0.45
areaFeat	55.35 ± 0.65	54.11 ± 1.01	53.12 ± 0.68	53.11 ± 0.14	55.81 ± 0.83
collabFeat	58.20 ± 0.87	59.38 ± 0.45	**58.50 ± 0.98**	59.32 ± 0.33	58.39 ± 0.61
$\mathit{instFeat}_A$	51.97 ± 0.40	51.67 ± 0.42	50.04 ± 0.14	50.21 ± 0.15	51.61 ± 0.41
$\mathit{instFeat}_B$	54.92 ± 0.57	58.20 ± 0.56	54.37 ± 1.43	58.09 ± 0.39	58.36 ± 0.23
$\mathit{instFeat}_C$	54.48 ± 0.87	59.46 ± 0.56	51.05 ± 0.30	49.93 ± 0.31	58.74 ± 0.67
$\mathit{subjFeat}_A$	57.90 ± 1.10	**61.74 ± 0.48**	53.63 ± 1.01	**61.53 ± 0.17**	59.12 ± 0.43
$\mathit{subjFeat}_B$	**59.01 ± 0.61**	59.50 ± 0.65	58.13 ± 0.83	61.22 ± 0.23	59.34 ± 1.04
$\mathit{subjFeat}_C$	58.61 ± 1.16	61.13 ± 0.74	57.41 ± 0.78	59.59 ± 0.33	**63.00 ± 0.55**
Features	Research Projects on *Veterinary Medicine*				
	DTrees	SVM	kNN	Bayes	MLP
pubFeat	**63.82 ± 1.45**	65.58 ± 0.76	**64.10 ± 1.27**	62.68 ± 0.62	**65.67 ± 0.72**
pubCitFeat	62.82 ± 0.95	64.08 ± 0.88	60.33 ± 0.95	59.66 ± 0.71	64.49 ± 1.14
grntFeat	58.50 ± 1.81	56.90 ± 1.58	58.75 ± 2.40	55.30 ± 0.69	58.00 ± 1.46
areaFeat	52.37 ± 1.95	51.66 ± 2.27	52.01 ± 1.66	53.19 ± 1.04	52.27 ± 2.46
collabFeat	60.75 ± 1.53	64.31 ± 2.00	60.11 ± 0.96	61.61 ± 1.34	64.85 ± 1.51
$\mathit{instFeat}_A$	51.80 ± 1.10	51.34 ± 0.96	49.98 ± 0.39	50.04 ± 0.45	51.42 ± 1.17
$\mathit{instFeat}_B$	61.26 ± 1.07	63.45 ± 1.26	54.02 ± 3.25	61.95 ± 0.93	64.14 ± 1.04
$\mathit{instFeat}_C$	61.18 ± 1.16	63.07 ± 1.43	50.89 ± 0.45	52.07 ± 2.94	62.87 ± 0.81
$\mathit{subjFeat}_A$	61.63 ± 1.44	**65.69 ± 1.24**	60.44 ± 2.39	**63.96 ± 1.09**	63.88 ± 1.88
$\mathit{subjFeat}_B$	63.33 ± 1.63	65.10 ± 1.48	60.78 ± 2.38	63.82 ± 1.42	60.45 ± 1.24
$\mathit{subjFeat}_C$	61.69 ± 0.85	63.49 ± 1.48	62.30 ± 0.93	61.66 ± 1.07	62.83 ± 1.96

Table B2
Results based on accuracy rate obtained by performing feature combinations. We considered as criterium for success research grants yielding at least two publications.

Method	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
DTrees	58.57 ± 2.37	58.71 ± 2.11	62.58 ± 3.45
SVM	59.62 ± 2.68	60.04 ± 2.80	64.33 ± 2.84
kNN	56.07 ± 3.46	57.40 ± 2.11	60.58 ± 2.66
Bayes	58.57 ± 2.14	59.24 ± 1.15	62.42 ± 2.50
MLP	58.46 ± 2.53	57.79 ± 2.63	58.88 ± 3.82

Table B3
Results based on accuracy rate obtained by performing feature combinations. We considered as criterium for success research grants yielding at least three publications.

Method	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
DTrees	59.24 ± 4.51	59.82 ± 3.75	61.25 ± 8.29
SVM	59.10 ± 4.12	60.35 ± 2.13	63.75 ± 6.55
kNN	55.14 ± 3.38	56.10 ± 3.44	60.67 ± 8.63
Bayes	55.66 ± 3.35	57.02 ± 3.00	60.38 ± 5.23
MLP	56.38 ± 4.76	59.08 ± 3.05	58.56 ± 7.50

Table B4
Results based on accuracy rate obtained by performing feature combinations. We considered as criterium for success research grants yielding at least four publications.

Method	Medicine Accuracy (%)	Dentistry Accuracy (%)	Vet. Medicine Accuracy (%)
DTrees	60.68 ± 6.38	60.16 ± 4.69	58.39 ± 6.73
SVM	60.21 ± 5.68	61.48 ± 3.84	59.11 ± 10.4
kNN	55.48 ± 5.47	56.09 ± 5.70	58.04 ± 7.11
Bayes	56.71 ± 5.25	58.28 ± 4.82	56.96 ± 10.43
MLP	58.70 ± 5.24	57.11 ± 4.65	54.64 ± 7.20

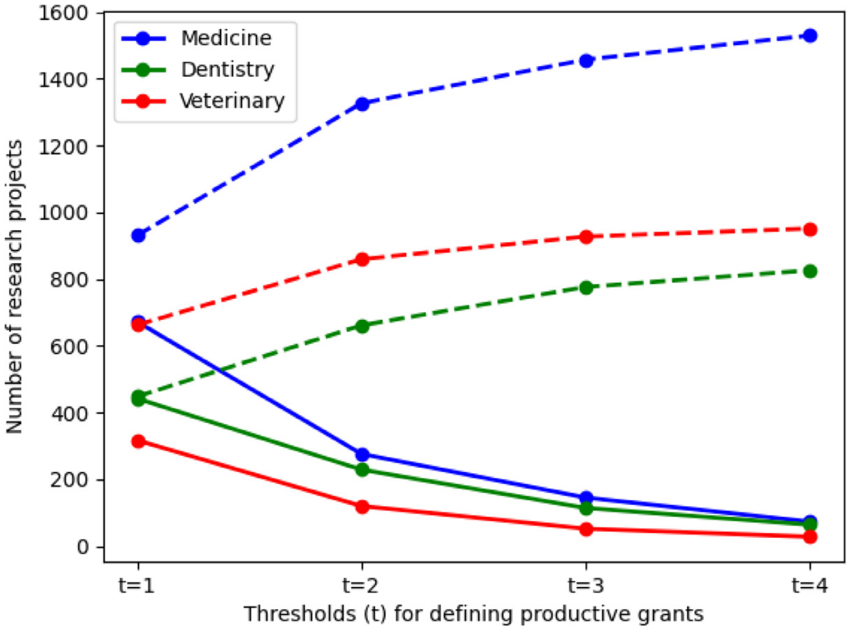


Fig. B1. Comparison of the number of productive and non-productive grants considering different thresholds for the three datasets (Medicine, Dentistry and Veterinary). Continuous lines represent productive projects according to a threshold (t) while dashed lines stand for grants considered as non-productive. t corresponds to at least t published papers.

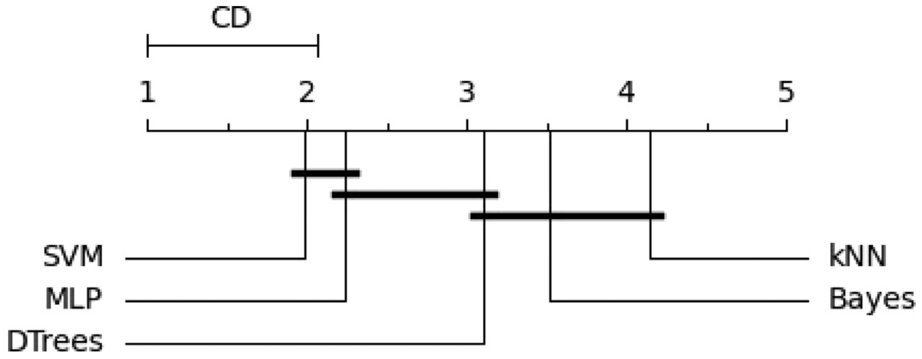


Fig. B2. Ranking diagram based on the classification using only a single feature. When analyzing classifiers trained with only a single feature, SVM and MLP displayed equivalent performance.

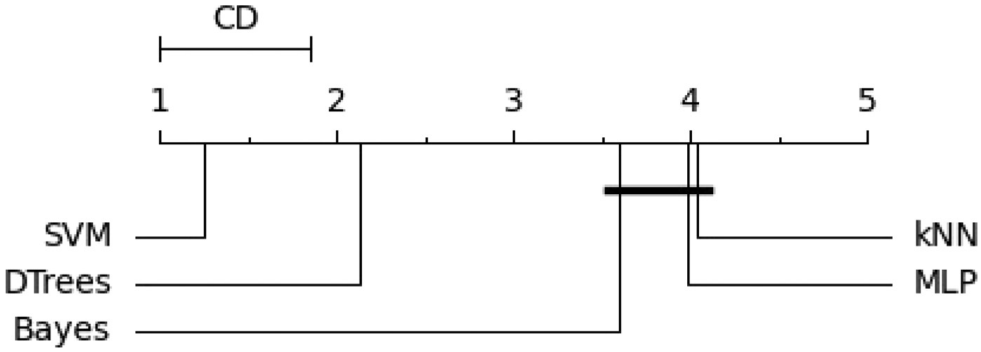


Fig. B3. Ranking diagram created considering the classification with multiple features. While SVM outperformed all other classifiers, Bayes, MLP and kNN displayed equivalent performance.

References

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Predicting scientific success. *Nature, 489*(7415), 201–202.

Amancio, D. R., Altmann, E. G., Oliveira Jr, O. N., & da Fontoura Costa, L. (2011). Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics, 13*(12), 123024.

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & da Fontoura Costa, L. (2014). A systematic comparison of supervised classifiers. *PloS One, 9*(4), e94137.

Amancio, D. R., Oliveira Jr, O. N., & Costa, L. d. F. (2012a). On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks. *EPL (Europhysics Letters), 99*(4), 48002.

Amancio, D. R., Oliveira Jr, O. N., & Costa, L. d. F. (2012b). Unveiling the relationship between complex networks metrics and word senses. *EPL (Europhysics Letters), 98*(1), 18002.

Amancio, D. R., Oliveira Jr, O. N., & Costa, L. d. F. (2015). Topological-collaborative approach for disambiguating authors names in collaborative networks. *Scientometrics, 102*(1), 465–485.

Bagrow, J. P., Berenberg, D., & Bongard, J. (2018). Neural language representations predict outcomes of scientific research. arXiv preprint arXiv:1805.06879.

Bar-Ilan, J. (2008). The h-index of h-index and of other informetric topics. *Scientometrics, 75*(3), 591–605.

Bauder, H. (2020). International mobility and social capital in the academic field. *Minerva*, 1–21.

Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics, 114*(2), 449–461.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

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

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the international workshop on multiple classifier systems* (pp. 1–15). Springer.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics, 105*(3), 1809–1831.

Fang, F. C., Bowen, A., & Casadevall, A. (2016). Nih peer review percentile scores are poorly predictive of grant productivity. *Elife, 5*, e13323.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science, 359*(6379). 10.1126/science.aao0185. https://science.sciencemag.org/content/359/6379/eaao0185

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics, 4*(4), 540–553.

Haykin, S. O. (2008). Neural networks and learning machines. *3rd Edition* (3rd). Prentice Hall. http://gen.lib.rus.ec/book/index.php?md5=0239F16656E6E5E7DB7AAA160CF9F854.

He, Z., Zhen, N., & Wu, C. (2019). Measuring and exploring the geographic mobility of american professors from graduating institutions: Differences across disciplines, academic ranks, and genders. *Journal of Informetrics, 13*(3), 771–784.

Ioannidis, J. P., Boyack, K. W., & Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLoS Biology, 18*(10), e3000918.

Kiziloz, H. E. (2021). Classifier ensemble methods in feature selection. *Neurocomputing, 419*, 97–107.

Kong, X., Zhang, J., Zhang, D., Bu, Y., Ding, Y., & Xia, F. (2020). The gene of scientific success. *ACM Transactions on Knowledge Discovery from Data (TKDD), 14*(4), 1–19.

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing, 86*, 105836.

Kumbure, M. M., Luukka, P., & Collan, M. (2020). A new fuzzy k-nearest neighbor classifier based on the bonferroni mean. *Pattern Recognition Letters, 140*, 172–178.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research, 39*(1), 19–37.

Letchford, A., Preis, T., & Moat, H. S. (2016). The advantage of simple paper abstracts. *Journal of Informetrics, 10*(1), 1–8.

Li, D.-C., Liu, C.-W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine, 40*(5), 509–518.

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., Sugimoto, C. R., Paul, L., & Zhang, C. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829.

Marinho, V. Q., Hirst, G., & Amancio, D. R. (2016). Authorship attribution via network motifs identification. In *Proceedings of the 5th Brazilian conference on intelligent systems (BRACIS)* (pp. 355–360). IEEE.

Markowitz, D. M. (2019). What words are worth: National science foundation grant abstracts indicate award funding. *Journal of Language and Social Psychology, 38*(3), 264–282.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive Bayes text classification. In *Proceedings of the AAAI-98 workshop on learning for text categorization: vol. 752* (pp. 41–48). Citeseer.

Min, C., Bu, Y., & Sun, J. (2021). Predicting scientific breakthroughs based on knowledge structure variations. *Technological Forecasting and Social Change, 164*, 120502.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics, 34*(21), 3711–3718.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research, 11*(6).

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change, 81*, 49–55.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Rinia, E., van Leeuwen, T., & van Raan, A. (2002). Impact measures of interdisciplinary research in physics. *Scientometrics, 53*(2), 241–248.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS One, 14*(1), e0210236.

Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. *Information Fusion, 6*(1), 63–81.

Salatino, A. A., Osborne, F., & Motta, E. (2017). How are topics born? Understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science, 3*, e119.

Schreiber, M. (2013). How relevant is the predictive power of the h-index? A case study of the time-dependent hirsch index. *Journal of Informetrics, 7*(2), 325–329.

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

Siudem, G., Żogała-Siudem, B., Cena, A., & Gagolewski, M. (2020). Three dimensions of scientific impact. *Proceedings of the National Academy of Sciences, 117*(25), 13896–13900.

Sun, Y., Livan, G., Ma, A., & Latora, V. (2021). Interdisciplinary researchers attain better performance in funding. arXiv preprint arXiv:2104.13091.

Teixeira, A. S., Talaga, S., Swanson, T. J., & Stella, M. (2021). Revealing semantic and emotional structure of suicide notes with cognitive network science. *Scientific Reports, 11*(1), 1–15.

Tohalino, J. A. V., Quispe, L. V. C., & Amancio, D. R. (2021). Analyzing the relationship between text features and research proposal productivity. *Scientometrics, 126*, 4255–4275. 10.1007/s11192-021-03926-x.

Tohalino, J. V., & Amancio, D. R. (2018). Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications, 503*, 526–539.

Wang, Y., Jones, B. F., & Wang, D. (2019). Early-career setback and future career impact. *Nature Communications, 10*(1), 1–10.

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

CHAPTER

5

USING VIRTUAL EDGES TO EXTRACT KEYWORDS FROM TEXTS MODELED AS COMPLEX NETWORKS

Title	Using virtual edges to extract keywords from texts modeled as complex networks
Authors	Jorge Valverde Tohalino, Thiago C. Silva, and Diego Amancio
Year	2022
Journal	Submitted to Scientometrics
Link	<https://arxiv.org/abs/2205.02172>
Situation	Submitted on November, 2022

5.1 Motivation

This research is motivated by the need to automatically detect keywords in texts, which is a task of paramount importance for many text mining applications. Keywords are important because they provide a concise representation of the main topics or concepts addressed in a text. However, manual identification of keywords can be a time-consuming and subjective task, especially for large and complex texts.

Modeling texts as word co-occurrence networks has been a common approach for keyword extraction. However, little attention has been paid to using word embeddings to enrich the graph structure. To address this gap, we introduced virtual edges based on the semantic similarity between word vectors. Word embeddings offer valuable contextual and semantic information that can enhance the network's discriminative power. Our study investigates the effectiveness of word embeddings, including BERT embeddings, for representing similarity relationships between words. The paper is motivated by the potential benefits of integrating embeddings into co-occurrence networks, as well as the limitations of existing approaches that rely solely on graph-based techniques or embeddings. The paper aims to contribute to the development of more effective and accurate methods for keyword detection, which could have implications for various text mining applications, such as information retrieval, document clustering, and topic modeling.

5.2 Contributions

This paper addresses the question of whether integrating embeddings to enrich the structure of co-occurrence networks can enhance the quality of extracted keywords from text. The methodology involved representing texts as co-occurrence networks and using two embedding approaches (Word2vec and BERT) to enrich the network structure. The results showed that incorporating a limited proportion of virtual (embedding) edges can effectively enhance the discriminative capacity of co-occurrence networks, with the best performance achieved by using the optimal window length in the co-occurrence network. The degree, PageRank, and accessibility metrics were found to exhibit superior performance in the proposed model, with unweighted versions of the traditional measurements providing better performance than their weighted counterparts in almost all cases. Future improvements to the model could include combining co-occurrence frequency and semantic similarity for edge weighting and handling synonyms before the creation of networks. The proposed approach is limited to finding unigram keywords and could be improved to consider keywords comprising two or more words. The study concludes that using virtual edges can improve the informativeness of co-occurrence networks for keyword detection and could be useful in other network classification scenarios, such as name disambiguation.

Using virtual edges to extract keywords from texts modeled as complex networks

Jorge A. V. Tohalino
Institute of Mathematics and Computer Science,
University of São Paulo, São Carlos, SP, Brazil

Thiago C. Silva
Universidade Católica de Brasília, Brasília, DF, Brazil

Diego R. Amancio*
Institute of Mathematics and Computer Science, Department of Computer Science,
University of São Paulo, São Carlos, SP, Brazil

(Dated: May 5, 2022)

Abstract

The keyword extraction task is an important NLP task in many text mining applications. Graph-based methods have been commonly used to automatically find the key concepts in texts, however, relevant information provided by embeddings has not been widely used to enrich the graph structure. Here we modeled texts co-occurrence networks, where nodes are words and edges are established either by contextual or semantical similarity. We compared two embedding approaches (Word2vec and BERT) to check whether edges created via word embeddings can improve the quality of the keyword extraction method. We found that, in fact, the use of virtual edges can improve the discriminability of co-occurrence networks. The best performance was obtained when we considered low percentages of addition of virtual (embedding) edges. A comparative analysis of structural and dynamical network metrics revealed the degree, PageRank, and accessibility are the metrics displaying the best performance in the model enriched with virtual edges.

* diego@icmc.usp.br

I. INTRODUCTION

In recent years, there has been a large increase in textual information available on the Internet. Examples include newspapers, social network comments, books, encyclopedias and scientific articles. In order to make sense and summarize such a large volume of data, several NLP applications have been proposed. One particular task is the keyword extraction task, which consists of selecting a set of words (or topics) that best represent the content of a document [45]. Finding keywords in multiple documents is important because manually finding the most central words can be an expensive and time-consuming task for human annotators. Since keyword extraction provides a compact representation of the document, many applications can benefit from this task: automatic indexing, automatic document summarization, automatic document classification, document clustering, automatic filtering, among other applications [5, 7, 18].

Different approaches have been considered for the keyword extraction task [7]. The simplest models are the statistical models that study the statistical information regarding the spatial use of words in each text as well as their frequency of use [20]. These methods include for example the well-known Term Frequency (TF) or Term Frequency-Inverse Document Frequency (TF-IDF). Approaches based on linguistic and syntactic analysis have also been used to address this task [44]. Additionally, several features extracted from the previous approaches can be used in machine learning algorithms. The main goal of these methods is to detect keywords via binary classification [22].

Graph-based approaches have also been used to detect keywords [42, 44]. The objective of these methods is to represent each document as a network of words and then apply a set of centrality measurements to assign a relevance value for each network node. In this way, the most central nodes represent the automatic keywords found for each document. Most of these approaches have used word co-occurrence networks, where an edge exists between two words if they are adjacent. However, different strategies to connect words have not been extensively studied, with most of the works considering larger window contexts in the co-occurrence model [30, 44].

In this paper, we propose a graph-based method for keyword extraction, where texts are represented as co-occurrence networks and edges are established in a twofold manner. In addition to word adjacency models, we consider further contexts to connect words. In order to better represent the relationship between words, we also link words that do not necessarily co-occur in the text, but are semantically similar. Our motivation is to enrich the representation by including the so-called virtual edges. Consequently, hidden similarities are explicitly represented in the model. Our hypothesis is that the included virtual edges can be used to improve the traditional co-occurrence network representation based on word adjacency relationships alone. In the proposed model, the virtual edges were constructed from the word vectors generated by the Word2Vec and BERT embedding models [16, 31]. After the networks are constructed, we computed the centrality values for each node (word) of the network. We used several structural and dynamical network measurements to identify the key concepts in texts. We also probed the effect of using the weighted versions of these measurements. The efficiency of our methods was evaluated in different datasets comprising documents of various sizes.

We have found several interesting results from this analysis. First, we observed that including virtual edges can improve the performance in retrieving keywords. The fraction of included virtual edges required to yield optimized results turned out to be relatively low. A negative performance effect was observed, however, when too many virtual edges were included. Concerning the embedding method, both considered strategies – Word2vec and BERT – yielded similar performance. The network metrics with the best performance were the degree, PageRank, and Accessibility. Surprisingly, when the weighted versions of the traditional metrics had a poor performance. Our results reinforce the potential of enriching networks in multiple text network applications [12, 39, 40].

This manuscript is organized as follows: Section II presents a summary of the related works for keyword extraction. The description of the datasets, as well as the proposed methodology, are described in Section III. In this section, we describe the network creation stage and the process of extracting keywords using network centrality metrics. The results are discussed in Section IV. describes the obtained results and the analysis of each network measurement. Finally, in Section V we present the conclusions and perspectives for future works.

II. RELATED WORKS

Studies addressing the keyword extraction problem can be grouped into three main approaches: statistical and network-based methods [19, 29]. The objective of statistical methods is to rank words using their statistical distribution along the text [11, 20]. A very simple approach is described is the one relying on word frequency [27]. According to this approach, words are sorted according to their frequency, and the most and less frequent words are disregarded. Such words are disregarded because they are common words (such as prepositions) or rare (not relevant) words. Note that frequency-based approaches do not consider the structure of the text, since a shuffled, meaningless version of the same text would provide the same set of keywords. To overcome the weaknesses of frequency-based methods, word clustering and word entropy methods were then proposed [20, 32]. Some modifications of these techniques include term-frequency inverse-document frequency approaches [33].

In [32], the authors found that relevant words, which are more closely related to the main topics of the text, are generally concentrated in certain regions of the text. Keywords are usually unevenly distributed along the text and tend to form clusters. Conversely, common words are more regularly distributed along texts. A combination of both spatial clustering and frequency was proposed in [10]. The authors used the Shannon's entropy metric to define a method based on the information content of the sequence of occurrences of each word in the text. They used text partitions to calculate the entropy of all words. Because relevant words are unevenly distributed, the heterogeneity of word distribution in different partitions can be captured via entropy. In comparison to clustering-based methods, an improved performance was reported with entropy-based techniques. The main advantage of the statistical techniques is that they do not require any knowledge of the language and thus can be used any analyze even unknown documents [15].

Graph-based approaches include the representation of the relationship of words as networks [17, 25, 46]. In [17], the authors modeled documents as graphs of semantic relationships between the words. The weight linking two words modeled the semantic relatedness computed as measured via Wikipedia. The strategy considered that the words related to central topics tend to be grouped into densely connected network communities, while common words are organized in weakly connected communities. This method was found to be particularly effective in removing noisy information. A similar study represented texts as word co-occurrence networks considering weighted and directed networks [25]. They used several centrality network measurements to find the relevant words. The authors concluded that network measurements can be successfully used for keyword extraction without the need for large external corpora. They also found that simpler centrality metrics like node degree or strength outperformed more complex and computationally expensive metrics in the proposed methodology. Related strategies have been used to find key concepts for the purpose of text summarization [42].

Finally, word co-occurrence network considering larger co-occurrence contexts was proposed in [44]. Different from other works, an edge is created if words co-occur within a window comprising three words. Using a combination of feature selection and clustering methods to find the best set of keywords, the authors obtained optimized results in comparison to other works that only employed traditional co-occurrence networks. The authors also found that several network metrics are strongly correlated, yielding thus equivalent performance.

Differently from the previous works, here we propose a graph-based method for keyword extraction using *enriched* word co-occurrence networks. In addition to the edges established via word adjacency, we considered edges created via embedding similarity. Several centrality measurements were then used to find the most relevant words. As we shall show, our approach outperforms both the traditional word adjacency model and its modified version considering larger contexts.

III. MATERIAL AND METHODS

The framework proposed to detect keywords via word embeddings and graph modeling comprises the following four main steps: i) text pre-processing; ii) word vectorization; iii) network creation; and iv) word ranking and keyword extraction.

1. *Pre-processing*: This phase comprises the required steps to conveniently pre-process the datasets. This step encompasses sentence segmentation, stopword removal and text stemming. In Section III A, we provide a brief description of the pre-processing steps we applied.

2. *Word vectorization*: we considered the embeddings models to represent the words.

The embeddings are important for identifying similar words that do not co-occur in the text. Section III B provides a detailed explanation of the word embedding methods used in this work.

3. *Network creation*: We modeled the documents as word co-occurrence networks, where nodes represent words and edges are established based on the neighbors of each word. We also considered "virtual" edges, which were generated based on the similarity between two words. The similarity is computed based on the word vectorization. This is an essential step for capturing long-range, semantical relationships. In Section III C, we explain the adopted methodology for network creation.

4. *Keyword extraction*: We used several network centrality measurements to rank the words for each document. Such measurements are used to give an importance value or relevance weight to each node from the network. The top N ranked words were considered keywords. Section III D describes the keyword extraction step.

The workflow we considered for keyword extraction is shown in Figure 1.

A. Pre-processing steps

We applied some pre-processing steps before texts are represented as networks. We first performed sentence segmentation. We defined a sentence as any text portion which is separated by a period, exclamation or question mark. This step is needed because BERT embedding model requires the input documents to be separated into sentences. Next, we removed stopwords and punctuation marks. We finally applied text stemming to the remaining words so that words are converted into their singular, infinitive form. This is important to map related words into the same node. We did not consider text lemmatization because reference keywords from datasets were in their stemmed form.

B. Word vectorization using word embeddings

Word embeddings models are a set of methods to represent words as dense vector representations. The idea behind these models is that words having similar meaning should have similar vector representations [31]. Word embeddings have been successfully used for several

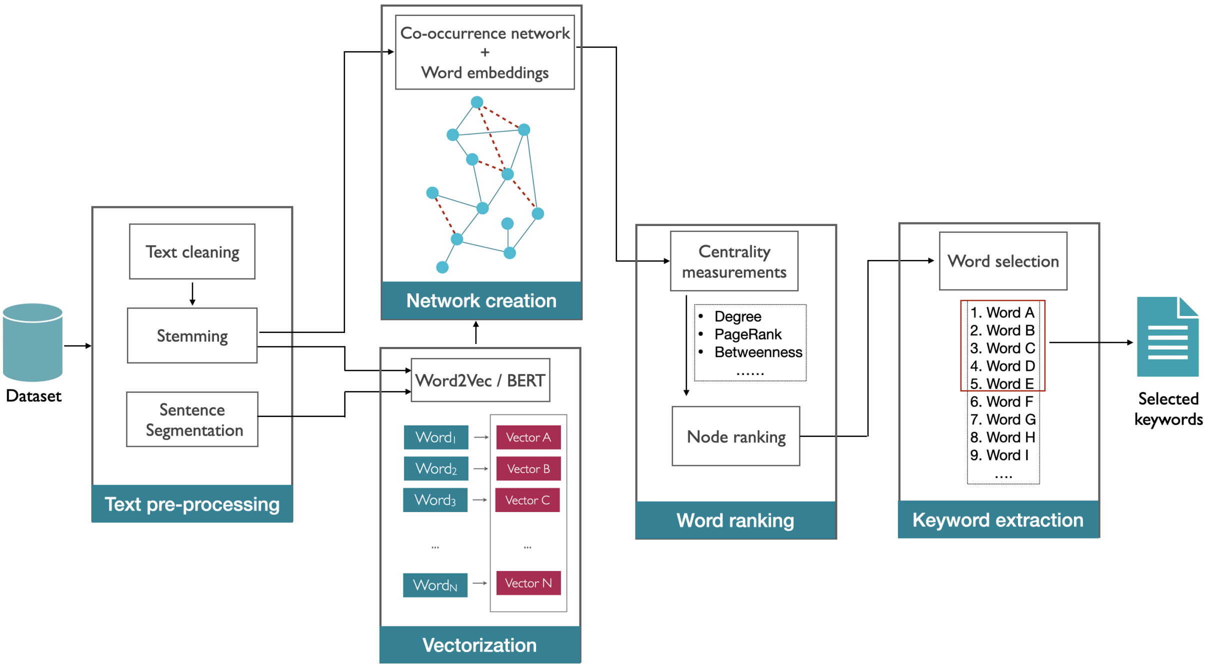


FIG. 1: Architecture of our system for graph-based keyword extraction. The first step consists of pre-processing the input texts. Next, we obtain the vector representation of the words from the pre-processed datasets. Then co-occurrence networks are constructed considering two edges types (co-occurrence and embedding (virtual) edges). Several centrality measurements are calculated to rank words. Finally, the top-ranked words are considered as keywords for each document.

text applications such as information retrieval, question answering, document summarization and text classification [8, 23, 37]. Here we use embeddings to establish links between similar nodes that do not co-occur in the text. The adopted method for including embedding edges as well as the construction process of the networks is described in Section III C.

There is a myriad of approaches to representing words as vectors. Methods to create word vectors include approaches based on neural networks, dimension reduction and probabilistic theory [2]. Here we employed the following methods:

- *Word2Vec*: This method is one of the first models to represent words as vectors [31]. Given a corpus, Word2Vec analyzes the words of each sentence and tries to predict neighbor words. For example, in the sentence "The early bird catches the X", Word2Vec can predict that the next word X is "worm", based on the previous context. This model uses a neural network with a single hidden layer. The neural network is trained with the documents of the corpus, then, for a given word α, it is calculated the probability that each word of the vocabulary is a neighbor of α. Once the network is trained, the model uses the weights of the hidden layer. as word vectors. Before the training stage, we defined different dimensions (d) for the word vectors. We generated vectors with $100 \leq d \leq 1,000$. Despite its simplicity and efficiency for various applications, Word2Vec has a significant weakness: it generates a unique vector for each word, regardless of word meaning and context, and this can generate noisy vectors, especially when representing ambiguous words. For example, the word "apple" will have the same associated vector regardless of whether it refers to the apple fruit or the Apple technology company. The BERT model addresses this problem as it generates different vectors for each word by taking into account the context in which the word appears.

- *Bidirectional Encoder Representations from Transformer (BERT)*: This model creates representations using the context appearing before and after the target word. Then, once previously trained, it can be fine-tuned for several specific tasks [16]. BERT uses a multilayer model of transformers (self-attention modules), and these structures allow learning attention weights of each word appearing before and after the target word. The model is pre-trained in two unsupervised tasks. In the first task, the model hides a percentage of the input tokens (words), and then it learns how to predict them. In the second task, the model selects two sentences, and then it predicts whether they are consecutive or not. Once the model has been pre-trained, it can be adjusted in a different task via a fine-tuning of parameters. We used this model to get the word vector representations of the keyword extraction datasets. We used the pre-trained model of BERT, which was previously trained over millions of texts. For each sentence, we then obtained the representative vectors of the words composing that sentence. In this sense, each occurrence of the same word is represented by a different vector. The context of each occurrence is used to generate the vectors.

Recently, a large number of word embedding algorithms have been proposed to mathematically represent words and text segments. In this work, we used Word2Vec for being one of the first word embeddings models that were proposed as an improvement to the traditional vector space models. Furthermore, Word2Vec is a simple model whose training stage is fast compared to other techniques. Word2Vec has also been used quite successfully for small and large datasets. BERT is one of the first models to offer significant gain in performance compared to several models based on Word2Vec. Such gain lies in the fact that BERT and related models are capable of producing various vector representations of a word according to its context. In this sense, BERT is able to capture the polysemy of a word, which typically results in more accurate feature representations [16].

C. Network construction

After the pre-processing steps are applied, a graph representation is created. The motivation for representing text as complex networks is the simple, yet competitive and *interpretable* results obtained in related text analysis tasks [1, 6, 13, 14]. The adopted graph represents each word as a node. For the creation of the edges between two words, we defined two edges types: edges based on the neighborhood relationship of two words (co-occurrence edges), and edges based on the semantic similarity of the words (embedding edges). Differently from previous approaches, here we establish long-range edges that can not be obtained from adjacency relationships alone. This approach is a way to link words that are semantically related but do not share the same stem. For co-occurrence edges, the following procedure was applied: we first defined a window value of size w. Edges linking two nodes are established for all words coexisting within the window. To build all edges, the window slides along the document. Figure 2 shows an example of edge formation. Here we considered $w = \{1, 2, 3\}$. Larger values of w were not considered in order to avoid a large complexity in the computation of network measurements. We also did not observe, in preliminary experiments, a significant performance gain when considering larger contexts.

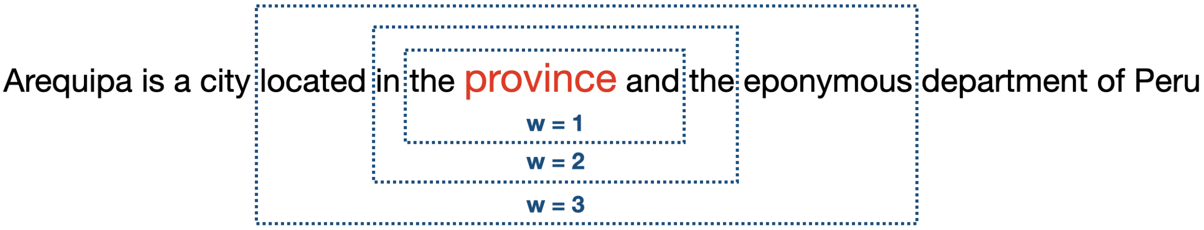

Arequipa is a city located in the province and the eponymous department of Peru
w = 1
w = 2
w = 3

FIG. 2: Example of how to find neighbors of a word for creating co-occurrence edges. In the sentence extracted from Wikipedia, we defined the neighbors of “province” according to a predefined window. If $w = 1$, the immediate left and right side words (“the” and “and”) are considered. The window length $w = 3$ includes the three words at the left and right side of the reference word “province”: “located”, “in”, “the”, “and”, “the”, and “eponymous”.

After the construction of the networks considering the co-occurrence relationships, the next step is the addition of edges established via *embeddings similarity.* This type of edge will also be referred to as *virtual edges.* Let E_t be the number of traditional co-occurrence edges. The number of virtual edges included is $E_v = PE_t$. Here we considered a small percentage P of additional edges, with $0 \leq P \leq 1$. The included *virtual* edges are the E_V most similar ones, according to the cosine similarity index. This strategy of enriching complex networks has been useful to provide more information in related applications [34, 36]. This is particularly useful in short texts [36]. We did not include more edges to avoid the complexity of analyzing denser networks. In addition, we did not find significant improvement in performance when the network is strongly connected.

D. Network characterization

The final step consists in using centrality measurements to rank the words according to the topological significance of the nodes from the network. Therefore, the best-ranked words (nodes) are chosen to be part of the resulting keyword list of the document. Centrality measurements are used to identify the most relevant nodes in a network. They are structural (or dynamical) attributes that indicate how central is a node according to a specific criterion. The identification of central nodes has been successfully used for various text applications. For example, [42] used several traditional network measurements to identify the most important sentences in a sentence network. [44] modeled documents as word co-occurrence networks and used several centrality measurements to rank the words for the keyword extraction task. The network measurements were also used as features for classification and authorship attribution tasks. For example, [34] represented literary books as word co-occurrence networks and the centrality measurements of the most frequent words were considered as feature vectors. Then the selected vectors were used in a machine learning algorithm for authorship identification. Here, we evaluated traditional network measurements and their weighted versions. We also considered the accessibility metric owing to its relative success in text analysis [42]. Apart from the degree, we refer to the weighted version of metric X as $X^{(w)}$.

1. *Degree* (k) and *strength* (s): The node degree of a node is the number of edges that are connected to that node. In the case of weighted networks, the strength represents

the sum of the weights of all the edges that are connected to the reference node.

2. *PageRank* (π): This measurement considers a node i as relevant if it is connected to other relevant nodes. The PageRank can be computed in a recursive way:

$$\pi_i = \gamma \sum_j a_{ij} \frac{\pi_j}{k_j} + \beta, \tag{1}$$

where γ and β are used as damping factors, with $0 \leq \gamma \leq 1$ and $0 \leq \beta \leq 1$ [26]. We also used a variation of this measurement that is based on the eigenvector centrality *Eigenvector centrality* (EV).

3. *Betweenness* (B): This metric is computed as the portion of shortest paths between two nodes that pass through a reference node. The betweenness centrality quantifies the relevance of a node to disseminate information [9]. It can also be used to identify words that are relevant even when they are not frequent [3].

4. *Closeness* (C): This measurement tries to detect the nodes that can efficiently spread information through a network. It is defined as $C_i = N \sum_j 1/d_{ij}$, where d_{ij} is the distance between i and j, and N is the number of nodes in the network. Nodes having high closeness value will have the shortest distances to all other nodes [35]. Distance-based measurements have also been used to analyze texts [3].

5. *Accessibility* ($A^{(h)}$): The accessibility metric quantifies the number of accessible nodes from an initial node using self-avoiding random walks of length h [43]. Nodes having a high accessibility also have effective access to more neighbors. This metric considers both the number of nodes at a given distance and the transition probabilities between the source and neighbor nodes. The accessibility can be evaluated considering different hierarchy levels. The levels can be set by specifying the length h of the random walks [43]. To compute this metric for a reference node i, we first define $p^{(h)}(i,j)$ to denote the likelihood of reaching a node j from an initial node i in a self-avoiding random walk of length h. Then, the accessibility of i is defined as the exponential of the true diversity of $p^{(h)}(i,j)$:

$$A_i^{(h)} = \exp\bigg(- \sum_j p^{(h)}(i,j) \log p^{(h)}(i,j) \bigg). \tag{2}$$

This measurement has been used in several contexts to analyze texts, including in stylometric and semantic tasks [38].

We used each centrality measurement to assign different importance values for each word. Then, the centrality values were used to rank the words. Therefore, the adopted methodology generated various word rankings according to the chosen network metrics. In Section IV, we reported the performance obtained for each network metric. After the word ranking step is performed, we selected the N best-ranked words, where N is the number of reference keywords.

E. Dataset

We used publicly available datasets including the source texts and their gold-standard keywords defined by experts. The following datasets were chosen for their variability in size and sources. The *Hult-2003* contains title, keywords, and abstracts from scientific papers published between 1998 and 2002 [21]. The documents were extracted from the *Inspect Database of Physics and Engineering* papers [21]. This dataset contains 500 abstracts as well as the set of keywords that were assigned by human annotators. The average size of the documents from this dataset is about 123 words. The *Marujo-2012* dataset comprises 450 web news stories on subjects such as business, culture, sport, and technology [28]. The mean document size is 452 words. Finally, we also used the Semeval-2010 [24]. This dataset comprises scientific papers that were extracted from the ACM Digital Library. We considered the full content of 100 papers and their corresponding keywords assigned by both authors and readers [24]. The average document length is $8,168$ words. In Table I we provide a summary indicating the main attributes of each dataset.

TABLE I: Statistical information from datasets for the keyword extraction task. $|D|$ represents the number of documents. We also show the average number of tokens ($\langle W \rangle$), sentences $\langle S \rangle$) and vocabulary size ($\langle U \rangle$). $\langle K \rangle$ is the average number of reference keywords assigned per document.

Dataset	**Description**	$\|D\|$	W_{avg}	U_{avg}	S_{avg}	K_{avg}
Hult-2003	Paper abstracts	500	123.12	73.25	5.14	18.83
Marujo-2012	Web news stories	450	452.36	223.33	20.74	52.79
SemEval-2010	Full papers	100	8168.49	1387.47	393.80	23.34

IV. RESULTS AND DISCUSSION

In this section, we analyze whether our hypothesis that the inclusion of virtual edges can improve the performance of co-occurrence networks in detecting keywords. In Section IV A, an analysis of the effect of parameter variation on the performance is provided. In Sections IV B and IV C, we detail the results obtained with Word2Vec and BERT, respectively. Finally, we show in Section IV D a summary of the obtained results.

A. Parameter analysis

In this section, we investigate whether the proposed extension of traditional word adjacency networks can lead to optimized results. In this section, we analyze if the performance is improved when we vary the model parameters. We are particularly interested in the performance analysis when varying both the window length (w) and the number of virtual edges (P). We focus our analysis on the results obtained for the Word2vec model, since similar results have been found with BERT (see next sections).

In Figure 3 we show, for the Hult-2003 dataset, the performance obtained for different network metrics. We considered distinct model parameters, with the window length being represented by different curves and P represented on the x-axis. The effect of considering edge weights was also considered. Note that the traditional word adjacency (unweighted) model is represented by dotted blue curves. The results observed in this dataset reveal that the best results are achieved with the largest window ($w = 3$) for all of the considered metrics. This means that a wider context does provide a better model for detecting keywords. Concerning the comparison of weighted and unweighted metrics, the best result considering all parameter combinations is always achieved with the unweighted version of the metrics. Most importantly, we also see that, considering the largest context and the unweighted version ($w = 3$), the inclusion of additional edges is also able to improve the performance of the methods. In all considered unweighted measurements, the inclusion of a few virtual edges can lead to optimized results. Interestingly, one should observe though that the inclusion of a large number of edges can cause a loss of relevant information. Whenever $P > 0.60$, the performance tends to decrease. The results observed in the figures also show that the best accuracy rates occurs typically for $0 \leq P \leq 0.20$.

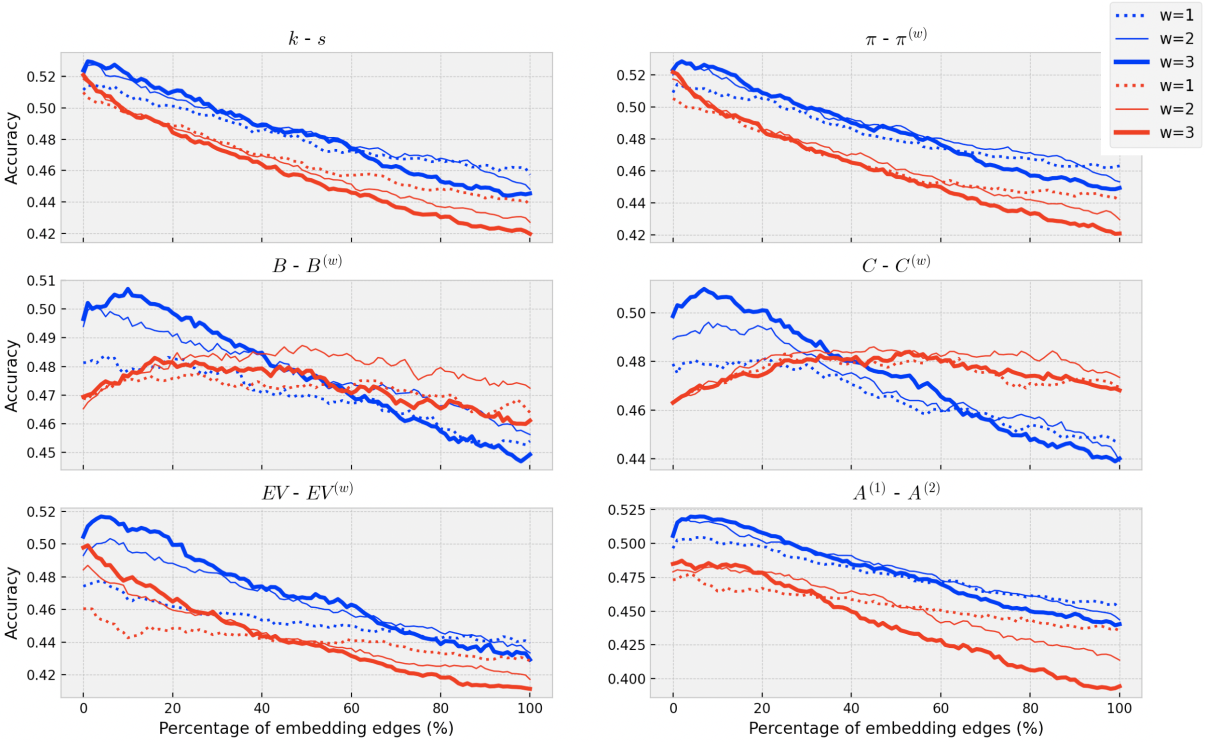


FIG. 3: Comparison of the performance of each centrality measurement based on the Word2Vec model for the *Hult-2003* dataset. For all subplots except the Accessibility metric, the blue lines represent unweighted measurements, while the red lines are weighted measurements. In the case of the Accessibility centrality, the blue lines describe the $A^{(1)}$ metric and the red lines represent the $A^{(2)}$ metric. We also evaluated the window length ($w = \{1, 2, 3\}$) for the network creation step: dotted lines are used when the value $w = 1$ is established, while thicker lines represent values for larger values of w.

When analyzing the Marujo-2012 dataset (see Figure 4), some differences can be observed. While the inclusion of virtual edges can improve the results of some models, one observes that the best results are obtained with the largest window length. Conversely, the role of including additional edges depends on the considered model. For both degree and PageRank, the best results were found with the weighted version and the largest window ($w = 3$) (see the red curve). In this case, the inclusion of additional edges hampers the performance. For all other metrics, the best results were found with the largest window length and the unweighted version when a small percentage of edges is included.

When one observes the results for the SemEval-2010 dataset in Figure 5, all best results were found with the unweighted version of the model considering $w = 3$. However, differently from the other datasets, for almost all metrics the inclusion of virtual edges does not improve the performance of the keyword detection. The performance with PageRank and

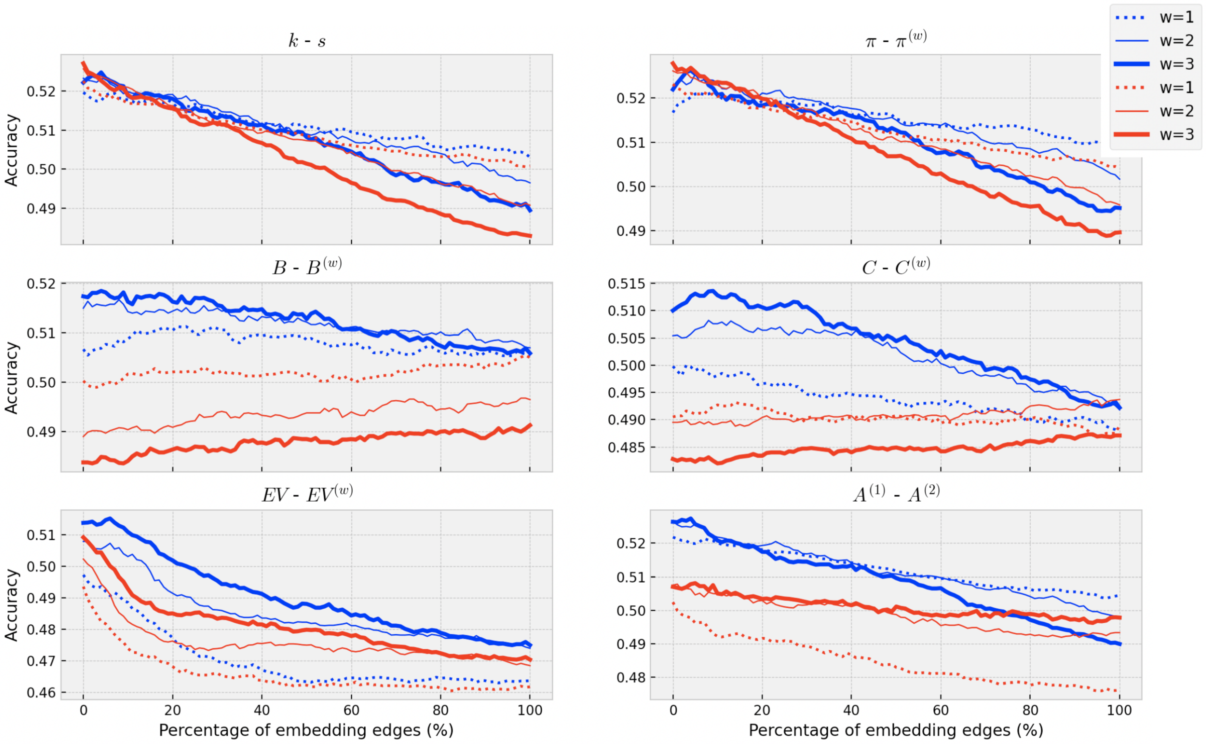


FIG. 4: Comparison of the performance of each centrality measurement based on the Word2Vec model for the *Marujo-2012* dataset. For all subplots except the Accessibility metric, the blue lines represent unweighted measurements, while the red lines stand for measurements based on weights. In the case of the Accessibility centrality, the blue lines describe the $A^{(1)}$ metric and the red lines represent the $A^{(2)}$ metric. We also evaluated the window length ($w = \{1, 2, 3\}$) for the network creation step: dotted lines are used when the value $w = 1$ is established, while thicker lines represent values for w larger than 1.

closeness are not *positively* affected by the inclusion of virtual edges, when analyzing the blue curves with the highest performance. The degree, eigenvector centrality and accessibility are negatively affected if several virtual edges are included. Surprisingly, the informativeness of the model even disappears when more than 50% of virtual edges are included for the eigenvector centrality. The betweenness centrality seems to be the only metric being improved – marginally – when virtual edges are included.

All in all the results show that the parameter behavior seems to depend on the considered dataset. In short texts (Hult-2003), the importance of including virtual edges is clearly observed. This happens because when short texts are modeled as co-occurrence networks, the generated line is almost a graph line. As a consequence, the topological information is not able to detect keywords, since all concepts will have the same topological information. In this case, the use of virtual edges is essential to identify the hidden information in short texts.

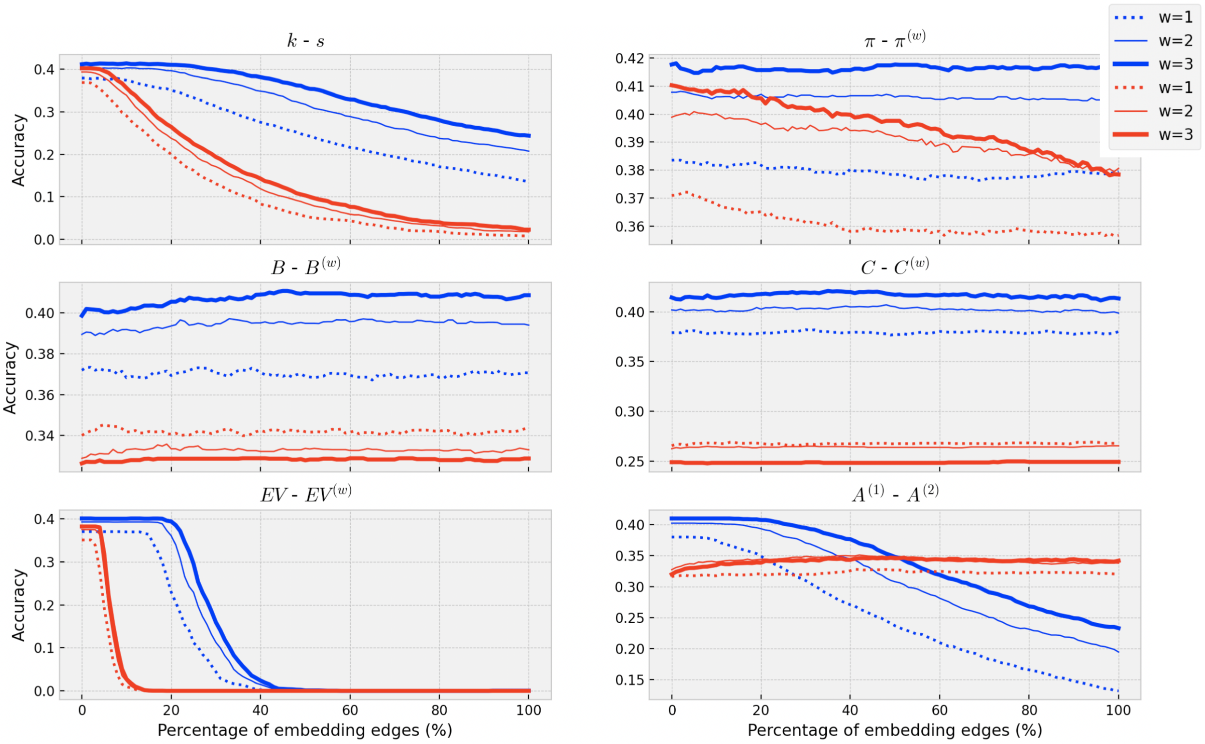


FIG. 5: Comparison of the performance of each centrality measurement based on the Word2Vec model for the *SemEval-2010 dataset*. For all subplots except the Accessibility metric, the blue lines represent unweighted measurements, while the red lines stand for measurements based on weights. In the case of the Accessibility centrality, the blue lines describe the $A^{(1)}$ metric and the red lines represent the $A^{(2)}$ metric. We also evaluated the use of windows ($w = \{1, 2, 3\}$) for the network creation step: dotted lines are used when the value $w = 1$ is established, while thicker lines represent values for w larger than 1.

Therefore, the results suggest that the proposed methodology can be useful to analyze short texts. Despite the above differences, the optimized results are almost always obtained when using a large window length ($w = 3$). The weighted metrics did not provide a significant gain in performance over their unweighted versions.

B. Performance analysis using the Word2Vec model

Table II depicts the results of the evaluation of the Word2Vec model considering 100, 300 and 500 dimensions (result not shown). We did not include the results obtained with larger dimensions because the observed performance decreases compared to smaller dimensions. We also show the performance of each vector size when the window parameter (w) was considered. For each measurement, we also show the percentage of embedding edge insertion that yielded the highest accuracy rates (P) and the highest accuracy observed with the proposed model (Acc.). We defined two additional quantities Γ_1 and Γ_2, which are defined as

$$\Gamma_1 = \frac{\text{Acc} - \text{Acc}^{(\text{tr})}}{\text{Acc}^{(\text{tr})}}, \tag{3}$$

$$\Gamma_2 = \frac{\text{Acc} - \text{Acc}^{(\text{w})}}{\text{Acc}^{(\text{w})}}. \tag{4}$$

$\text{Acc}^{(\text{tr})}$ corresponds to the accuracy obtained with the traditional co-occurrence model [41] (i.e, our model with $P = 0$ and $w = 1$). $\text{Acc}^{(\text{tr})}$ corresponds to accuracy obtained with the model considering only co-occurrence links [44] (i.e., our model with $P = 0$). Thus, Γ_1 and Γ_2 quantify the gain in performance when important features of the model are disregarded.

According to the results shown in Table II, for the Hult-2003 dataset, the vectors having $d = 300$ dimensions yielded the highest accuracy rates in most cases. Considering $d = 300$, the most important results were reached when the percentage of embedding edge insertion was low (less than 10%). However, for dimensions greater than 300, values of P between 0% to 26% yielded high accuracy rates. Conversely, there are some exceptions where percentages of virtual edges larger than 50% yielded the best performance. Regarding the parameter w, in most cases, the best results are reached when the parameter $w = 3$ is considered.

In the case of the Marujo-2012 dataset, Table II shows that, generally, $d = 100$ dimensions are the optimal size for the word vectors. We also observed that the typical optimal percentage of addition of embeddings type edges did not exceed 7%. However, there are some exceptions when high values of P lead to a higher accuracy. However, for these cases ($B^{(w)}$ and $C^{(w)}$ metrics), the addition of edges does not outperform the results obtained with the respective unweighted version of these metrics. Once again the largest context size typically achieved the best performances for the Marujo-2012 dataset. The weighted version of the PageRank ($\pi^{(w)}$) obtained the highest accuracy rate (with $k = 100$, $w = 3$, and 0% of insertion of embedding edges).

Table II also revealed that $d = 100$ and $d = 300$ dimensions of the word vectors achieved the best accuracy rates for the SemEval-2010 dataset. The edge addition percentages that achieved the best results were higher compared to previous datasets. Such percentages included values ranging between 0 and 45%. However, for the closeness metric ($C^{(w)}$), the optimal value of P was even higher, reaching 64%. Despite this higher level of embedding

TABLE II: Performance based on the Word2Vec model for edge embedding creation. Acc. represents the highest accuracy rate for the considered set of parameters. For this analysis, we experimented with different dimensions (d) of the embedding vectors.

		d = 100					**d = 300**					**d = 500**				
Dataset	**Meas.**	**P**	**w**	Γ_1	Γ_2	**Acc.**	**P**	**w**	Γ_1	Γ_2	**Acc.**	**P**	**w**	Γ_1	Γ_2	**Acc.**
Hult-2003	k	2	3	0.03	0.01	0.5280	1	3	0.04	0.01	0.5296	2	3	0.04	0.02	0.5288
	s	0	3	0.03	–	0.5184	0	3	0.02	–	0.5209	0	3	0.03	–	0.5223
	π	2	3	0.04	0.01	0.5274	2	3	0.04	0.01	0.5283	2	3	0.04	0.01	0.5282
	$\pi^{(w)}$	0	3	0.03	–	0.5178	0	3	0.03	–	0.5216	0	3	0.04	–	0.5237
	B	8	3	0.05	0.01	0.5030	10	3	0.05	0.02	0.5070	7	3	0.04	0.01	0.5036
	$B^{(w)}$	22	2	0.03	0.03	0.4876	49	2	0.04	0.04	0.4873	25	2	0.04	0.04	0.4882
	C	3	3	0.06	0.01	0.5063	7	3	0.06	0.02	0.5098	5	3	0.06	0.02	0.5094
	$C^{(w)}$	26	2	0.04	0.04	0.4879	45	2	0.05	0.05	0.4860	31	2	0.05	0.05	0.4875
	EV	6	3	0.09	0.02	0.5173	4	3	0.09	0.02	0.5169	5	3	0.09	0.02	0.5154
	$EV^{(w)}$	1	3	0.08	–	0.4999	1	3	0.08	–	0.4992	0	3	0.09	–	0.4995
	$A^{(1)}$	5	3	0.05	0.02	0.5199	6	3	0.05	0.03	0.5200	6	3	0.06	0.02	0.5198
	$A^{(2)}$	2	3	0.04	0.01	0.4880	2	3	0.03	–	0.4872	3	3	0.03	0.01	0.4866
Marujo-2012	k	2	3	0.01	0.01	0.5275	4	3	0.01	–	0.5247	2	3	0.01	–	0.5278
	s	0	3	0.01	–	0.5279	0	3	0.01	–	0.5270	0	3	0.01	–	0.5276
	π	4	3	0.02	0.01	0.5258	4	3	0.02	0.01	0.5262	5	3	0.02	0.01	0.5266
	$\pi^{(w)}$	0	3	0.01	–	0.5289	0	3	0.01	–	0.5278	0	3	0.01	–	0.5286
	B	5	3	0.03	–	0.5195	4	3	0.02	–	0.5185	4	3	0.03	–	0.5193
	$B^{(w)}$	43	1	0.01	0.01	0.5029	99	1	0.01	0.01	0.5055	31	1	0.01	0.01	0.5031
	C	7	3	0.03	0.01	0.5143	9	3	0.03	0.01	0.5136	19	3	0.03	0.01	0.5141
	$C^{(w)}$	87	2	–	–	0.4931	100	2	0.01	0.01	0.4937	77	2	–	–	0.4923
	EV	7	3	0.04	–	0.5160	6	3	0.04	–	0.5151	6	3	0.04	–	0.5155
	$EV^{(w)}$	0	3	0.03	–	0.5095	0	3	0.03	–	0.5091	0	3	0.03	–	0.5105
	$A^{(1)}$	1	3	0.01	–	0.5275	4	3	0.01	–	0.5274	2	3	0.01	–	0.5275
	$A^{(2)}$	5	3	0.01	–	0.5094	5	3	0.01	–	0.5081	3	3	0.01	–	0.5097
SemEval-2010	k	0	3	0.09	–	0.4140	10	3	0.09	0.01	0.4140	2	3	0.10	–	0.4144
	s	0	3	0.09	–	0.4039	0	3	0.09	–	0.4024	0	3	0.09	–	0.4035
	π	0	3	0.09	–	0.4177	1	3	0.09	–	0.4181	48	3	0.09	–	0.4185
	$\pi^{(w)}$	3	3	0.10	–	0.4111	0	3	0.11	–	0.4103	1	3	0.10	–	0.4106
	B	74	3	0.10	0.03	0.4100	45	3	0.10	0.03	0.4106	43	3	0.10	0.03	0.4104
	$B^{(w)}$	26	1	0.01	0.01	0.3521	5	1	0.02	0.02	0.3452	23	1	0.01	0.01	0.3458
	C	46	3	0.10	0.01	0.4179	35	3	0.11	0.02	0.4208	35	3	0.11	0.02	0.4200
	$C^{(w)}$	64	1	0.02	0.02	0.2743	93	1	0.01	0.01	0.2690	50	1	0.01	0.01	0.2682
	EV	13	3	0.08	–	0.4011	0	3	0.08	–	0.4007	0	3	0.08	–	0.4007
	$EV^{(w)}$	0	3	0.08	–	0.3807	0	3	0.09	–	0.3820	0	3	0.09	–	0.3835
	$A^{(1)}$	0	3	0.08	–	0.4088	0	3	0.08	–	0.4099	7	3	0.08	–	0.4083
	$A^{(2)}$	29	2	0.09	0.06	0.3450	42	2	0.11	0.07	0.3509	36	2	0.10	0.07	0.3437

enrichment, this metric obtained a poor performance when compared to all other results. Concerning the context parameter for co-occurrence links, both $w = 1$ and $w = 3$ achieved the highest accuracy rates. For the SemEval dataset, the closeness measurement (C) reached the best performance (with $d = 300$, $w = 3$, and $P = 35\%$).

In conclusion, we found that $d \leq 300$ yields competitive performance for the considered datasets. When larger values of d were considered, the accuracy rates did not significantly improve. We also observed that the performance can be improved in several scenarios when larger window length and/or the inclusion of virtual edges are considered.

C. System performance analysis using the BERT model

In this section, we discuss the results we obtained considering the word vectors produced by the BERT model [16]. Because in this model each occurrence of the same word is represented by different vectors, we had to adapt our methodology concerning the insertion of virtual edges. We adopted two approaches to compute the similarity between two words. In the first approach ($BERTSim_1$), the word is represented by averaging the corresponding vector observed in each occurrence. In the second approach ($BERTSim_2$), the similarity $\text{sim}(a, b)$ between nodes a and b is computed as:

$$\text{sim}(a, b) = \frac{1}{f_a f_b} \sum_k \sum_l \cos(v_k^{(a)}, v_l^{(b)}), \tag{5}$$

where $v_k^{(a)}$ is the k-th vector representation of word a, cos is the cosine similarity and f_a is the frequency (i.e. number of occurrences) of a. The results obtained for both approaches are depicted in Table III. The results are shown in terms of w and P.

Concerning the Hult-2003 dataset, the $BERTSim_1$ approach achieved the highest accuracy rates in most cases. However, for various situations, the best results are obtained without using virtual edges or when P is lower than 6%. Only for the weighted closeness metric, the percentage of edge addition was quite high (84%). As for the window length, $w = 1$ and $w = 3$ generally yielded the highest performance. The accessibility metric considering one hierarchy level ($A^{(1)}$) achieved the best performance for the Hult-2003 dataset ($BERTSim_1$ approach and $w = 3$)

The results revealed that the $BERTSim_1$ approach outperformed the $BERTSim_2$ ap-

TABLE III: Performance based on the BERT model for edge embedding creation. Acc. represents the highest accuracy rate for the considered set of parameters. For this analysis, we experimented with different values of window length and fraction of included virtual edges.

		$BERTSim_1$					$BERTSim_2$				
Dataset	**Meas.**	**P**	**w**	Γ_1	Γ_2	**Acc.**	**P**	**w**	Γ_1	Γ_2	**Acc.**
Hult-2003	k	2	3	0.02	–	0.5163	0	3	0.03	–	0.5199
	s	0	2	0.04	–	0.5069	0	3	0.04	–	0.5033
	π	0	3	0.03	–	0.5235	0	3	0.03	–	0.5224
	$\pi^{(w)}$	0	3	0.04	–	0.5080	0	3	0.04	–	0.5065
	B	3	3	0.03	–	0.4973	0	3	0.03	–	0.4957
	$B^{(w)}$	9	2	–	–	0.4727	0	1	–	–	0.4755
	C	6	3	0.04	0.01	0.5016	0	3	0.04	–	0.4984
	$C^{(w)}$	84	1	0.01	0.01	0.4737	3	1	0.01	0.01	0.4722
	EV	5	3	0.08	0.01	0.5097	1	3	0.07	–	0.5057
	$EV^{(w)}$	2	3	0.06	0.01	0.4945	0	3	0.05	–	0.4839
	$A^{(1)}$	0	3	0.03	–	0.5318	0	3	0.03	–	0.5328
	$A^{(2)}$	5	1	0.02	0.01	0.4859	0	3	0.02	–	0.4854
Marujo-2012	k	7	2	0.01	–	0.5256	2	3	0.01	0.01	0.5246
	s	2	2	0.01	–	0.5208	0	3	0.01	–	0.5201
	π	2	2	0.01	–	0.5232	3	2	0.01	–	0.5227
	$\pi^{(w)}$	0	2	0.01	–	0.5183	0	2	0.02	–	0.5179
	B	0	3	0.02	–	0.5174	0	3	0.02	–	0.5173
	$B^{(w)}$	67	2	0.01	0.01	0.5095	67	2	0.01	0.01	0.5081
	C	5	3	0.03	0.01	0.5146	7	3	0.03	–	0.5126
	$C^{(w)}$	86	2	0.02	0.02	0.5066	72	2	0.01	–	0.5027
	EV	11	3	0.04	0.01	0.5177	9	3	0.04	–	0.5147
	$EV^{(w)}$	3	3	0.04	–	0.5096	2	3	0.03	–	0.5061
	$A^{(1)}$	2	2	0.01	–	0.5223	4	3	0.01	–	0.5224
	$A^{(2)}$	8	3	0.01	0.01	0.5043	2	2	0.01	–	0.5032
SemEval-2010	k	1	3	0.09	–	0.4141	8	3	0.09	–	0.4137
	s	0	3	0.07	–	0.4023	2	3	0.07	–	0.4087
	π	0	3	0.09	–	0.4177	0	3	0.09	–	0.4177
	$\pi^{(w)}$	0	3	0.08	–	0.4103	1	3	0.09	–	0.4208
	B	3	3	0.07	–	0.3989	7	3	0.08	0.01	0.4012
	$B^{(w)}$	63	1	0.03	0.03	0.2436	4	1	–	–	0.2759
	C	1	3	0.10	0.01	0.4165	20	3	0.10	0.01	0.4167
	$C^{(w)}$	91	3	0.28	0.04	0.1917	9	3	0.03	–	0.2052
	EV	1	3	0.09	–	0.4022	0	3	0.08	–	0.4007
	$EV^{(w)}$	0	3	0.07	–	0.3861	5	3	0.09	0.01	0.3962
	$A^{(1)}$	2	1	0.01	0.01	0.2830	0	1	–	–	0.2786
	$A^{(2)}$	1	1	–	–	0.0228	5	1	0.06	0.06	0.0237

proach for the Marujo-2012 dataset. We observed that the optimal percentage of edge insertion is typically lower than 15%, but in particular cases, it reached high values between 67% and 86% (for the weighted versions of betweenness and closeness metrics). Regarding the window length, the best results were obtained for $w \geq 2$. The node degree (k) centrality obtained the highest accuracy rate considering the following parameters: $BERTSim_1$ approach, $w = 2$, and a percentage of $P = 7\%$ for the fraction of virtual edges.

Unlike the other datasets, the $BERTSim_2$ approach performed slightly better than the $BERTSim_1$ method for the SemEval-2010 dataset. The optimal value of the edge addition percentage in most cases was less than 20%. Higher values of P, however were used for both weighted versions of Betweenness and Closeness. Once again, the best results were obtained with window length $w = 3$. For the SemEval-2010 dataset, the PageRank metric (π) performed better than the other centrality measurements considering the $BERTSim_2$ approach and $w = 3$.

D. Summary of results and discussion

Table IV displays the highest accuracy rates (Acc.) of Word2Vec and BERT embedding models for each dataset. We also show the values of each parameter that achieved the best performances. We considered the following parameters relevant to our research: optimal dimension d of the vectors produced by Word2Vec, and the best approach (Appr.) employed to calculate the similarity between multiple vectors generated by BERT for the same word. Most importantly, the parameters affecting the network construction (w and P) are also reported.

All in all, our results revealed that both Word2Vec and BERT methods have a similar performance. For the Hult-2003 dataset, the BERT-based methods were slightly better than Word2Vec, while for the Marujo-2012 dataset, the Word2Vec-based methods outperformed the BERT-based methods. Conversely, for the SemEval-2010 dataset, the results of both approaches displayed similar performance. This result allows the use of both techniques based on the desired property of the chosen embedding method. The training of the Word2Vec model is quite fast and has been successfully used for representing documents for different text applications. However, here we needed to detect the optimal size of the vectors. Word2Vec also generates a single vector for a word, regardless of the context of

TABLE IV: Summary of the best results based on accuracy rate (Acc.) for both Word2Vec and BERT models. % represents the optimal fraction of included virtual edges, d is the embedding dimension, and w is the context size adopted to construct co-occurrence networks.

	Word2Vec model				
Dataset	**d**	**w**	**%**	**Meas.**	**Acc.**
Hult-2003	300	3	1	k	0.5296
Marujo-2012	100	3	0	$\pi^{(w)}$	0.5289
SemEval-2010	300	3	35	C	0.4208

	BERT model				
Dataset	**Apr.**	**w**	**%**	**Meas.**	**Acc.**
Hult-2003	Sim_2	3	0	$A^{(1)}$	0.5328
Marujo-2012	Sim_1	2	7	k	0.5256
SemEval-2010	Sim_2	3	1	$\pi^{(w)}$	0.4208

the words. Training documents with BERT is computationally more expensive, especially for larger datasets comprising large documents. However, the main advantage of BERT is that it generates several vectors for a word according to the number of contexts in which the word is used. This fact can lead to enhanced representations and improved performance in different datasets.

Concerning the network creation parameters, we found the best results with Word2vec considered vectors comprising typically less than 500 dimensions. In the BERT approach, the two approaches proposed to handle multiple word vectors for the same concept – namely $BERTSim_1$ and $BERTSim_2$ – had a similar performance. However, the $BERTSim_2$ approach requires a higher computational cost, especially when analyzing large documents. The experiments also showed that in most cases the percentage of addition of virtual edges is typically not very high. The performance of each system considerably decreases when high percentages of addition of virtual edges are considered. In conclusion, we showed – as a proof of principle – that the combination of further window length in the co-occurrence model and virtual edges can improve the quality of the keyword detection [44].

V. CONCLUSION

Identifying keywords is an important task in many text mining applications. In this paper, we addressed this problem by generating different representations for a text using co-occurrence networks. We considered two variations of the word adjacency model: the number of words that can be connected within the same context, and the fraction of virtual edges used to connect similar words. For each generated network, we evaluated several centrality measurements, including a generalization of the node degree centrality considering a network dynamics [43].

Our results revealed that the optimal window length in the co-occurrence network is $w = 3$, while the fraction of embeddings/virtual edges yielding the best results is typically not high. We also observed that the node degree, PageRank, and accessibility metrics reached the highest accuracy rates for the three datasets. The unweighted versions of the traditional measurements turned out to provide better performance than their weighted counterparts in almost all cases.

Our results showed, as a proof of principle, that using virtual edges can improve the informativeness of co-occurrence networks for the keyword detection task. Given that the informativeness of the characterization can be improved in the adopted representation, we believe that the inclusion of virtual edges could be useful in other network classification scenarios, such as in name disambiguation [4]. The proposed methodology could be improved by including other model components. For example, edges weight modeling could be improved if both co-occurrence frequency and semantic similarity are combined, for example, via linear operations.

Another source of improvement to the model could arise if synonyms are handled before the creation of the networks. In this way, words with similar meanings would be represented by a single node, so as to avoid redundancy in the co-occurrence networks. This could be done by taking advantage of the vectors generated by BERT, for example. Finally, our approach is limited to finding unigram keywords (keywords composed of a single word). A more general approach could consider keywords comprising two or more words.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Thiago C. Silva (Grant no. 308171/2019-5, 408546/2018-2) gratefully acknowledges financial support from the CNPq foundation. Diego R. Amancio acknowledges financial support from the (Grant no. 304026/2018-2, 311074/2021-9) and FAPESP (Grant no. 20/06271-0).

[1] C. Akimushkin, D. R. Amancio, and O. N. Oliveira Jr. On the role of words in the network structure of texts: Application to authorship attribution. *Physica A: Statistical Mechanics and its Applications*, 495:49–58, 2018.

[2] F. Almeida and G. Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

[3] D. R. Amancio, E. G. Altmann, O. N. Oliveira, and L. da Fontoura Costa. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, 13(12):123024, 2011.

[4] D. R. Amancio, O. N. Oliveira Jr., and L. F. Costa. Topological-collaborative approach for disambiguating authors' names in collaborative networks. *Scientometrics*, 102(1):465–485, 2015.

[5] J. An and Y.-P. Chen. Keyword extraction for text categorization. In *Proceedings of the 2005 International Conference on Active Media Technology, 2005.(AMT 2005).*, pages 556–561. IEEE, 2005.

[6] H. F. Arruda, F. N. Silva, V. Q. Marinho, D. R. Amancio, and L. F. Costa. Representation of texts as complex networks: a mesoscopic approach. *Journal of Complex Networks*, 6(1):125–144, 2018.

[7] S. K. Bharti and K. S. Babu. Automatic keyword extraction for text summarization: A survey. *arXiv preprint arXiv:1704.03242*, 2017.

[8] S. S. Birunda and R. K. Devi. A review on word embedding techniques for text classification. In *Innovative Data Communication Technologies and Application*, pages 267–281. Springer, 2021.

[9] U. Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[10] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, and J. Oliver. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*, 79(3):035102, 2009.

[11] C. Carretero-Campos, P. Bernaola-Galván, A. Coronado, and P. Carpena. Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A:*

Statistical Mechanics and its Applications, 392(6):1481–1492, 2013.

[12] N. Castro and M. Stella. The multiplex structure of the mental lexicon influences picture naming in people with aphasia. *Journal of Complex Networks*, 7(6):913–931, 2019.

[13] J. Cong and H. Liu. Approaching human language with complex networks. *Physics of life reviews*, 11(4):598–618, 2014.

[14] R. Cremades and M. Stella. Disentangling the climate divide with emotional patterns: a network-based mindset reconstruction approach. *Earth System Dynamics Discussions*, pages 1–34, 2022.

[15] H. F. de Arruda, V. Q. Marinho, L. d. F. Costa, and D. R. Amancio. Paragraph-based representation of texts: A complex networks approach. *Information Processing & Management*, 56(3):479–494, 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, pages 661–670, 2009.

[18] K. M. Hammouda, D. N. Matute, and M. S. Kamel. Corephrase: Keyphrase extraction for document clustering. In *International workshop on machine learning and data mining in pattern recognition*, pages 265–274. Springer, 2005.

[19] K. S. Hasan and V. Ng. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, 2014.

[20] J. P. Herrera and P. A. Pury. Statistical keyword detection in literary corpora. *The European Physical Journal B*, 63(1):135–146, 2008.

[21] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223, 2003.

[22] X. Jiang, Y. Hu, and H. Li. A ranking approach to keyphrase extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 756–757, 2009.

[23] Z. Jiang, M. Srivastava, S. Krishna, D. Akodes, and R. Schwartz. Combining word embeddings and n-grams for unsupervised document summarization. *arXiv preprint arXiv:2004.14119*, 2020.

[24] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, 2010.

[25] S. Lahiri, S. R. Choudhury, and C. Caragea. Keyword and keyphrase extraction using centrality measures on collocation networks. *arXiv preprint arXiv:1401.6571*, 2014.

[26] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton university press, 2011.

[27] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[28] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, and J. Neto. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. pages 399–403, 2012.

[29] Z. A. Merrouni, B. Frikh, and B. Ouhbi. Automatic keyphrase extraction: a survey and trends. *Journal of Intelligent Information Systems*, 54(2):391–424, 2020.

[30] R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[32] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. M. Somoza. Keyword detection in natural languages and dna. *EPL (Europhysics Letters)*, 57(5):759, 2002.

[33] S. Qaiser and R. Ali. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29, 2018.

[34] L. V. Quispe, J. A. Tohalino, and D. R. Amancio. Using virtual edges to improve the discriminability of co-occurrence text networks. *Physica A: Statistical Mechanics and its Applications*, 562, 1 2021.

[35] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[36] L. Santos, E. A. Corrêa Júnior, O. Oliveira Jr, D. R. Amancio, L. Mansur, and S. Aluísio. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1284–1296, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[37] Y. Shen, W. Rong, N. Jiang, B. Peng, J. Tang, and Z. Xiong. Word embedding based correlation model for question/answer matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[38] B. C. Souza, F. N. Silva, H. F. de Arruda, G. D. Silva, L. F. Costa, and D. R. Amancio. Text characterization based on recurrence networks. *arXiv preprint arXiv:2201.06665*, 2022.

[39] M. Stella. Cognitive network science reconstructs how experts, news outlets and social media perceived the covid-19 pandemic. *Systems*, 8(4):38, 2020.

[40] M. Stella, N. M. Beckage, and M. Brede. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, 7(1):1–10, 2017.

[41] E. Sulis, L. Humphreys, F. Vernero, I. A. Amantea, D. Audrito, and L. Di Caro. Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts. *Information Systems*, 106:101821, 2022.

[42] J. V. Tohalino and D. R. Amancio. Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539, 8 2018.

[43] B. Travençolo and L. da F. Costa. Accessibility in complex networks. *Physics Letters A*, 373(1):89 – 95, 2008.

[44] D. A. Vega-Oliveros, P. S. Gomes, E. E. Milios, and L. Berton. A multi-centrality index for graph-based keyword extraction. *Information Processing and Management*, 56(6):102063, 2019.

[45] S. Vijaya Shetty, S. Akshay, S. Reddy, H. Rakesh, M. Mihir, and J. Shetty. Graph-based keyword extraction for twitter data. In *Emerging Research in Computing, Information, Communication and Applications*, pages 863–871. Springer, 2022.

[46] Z. J. Zhan, F. Lin, and X. P. Yang. Keyword extraction of document based on weighted complex network. In *Advanced Materials Research*, volume 403, pages 2146–2151. Trans Tech Publ, 2012.

CHAPTER 6

USING CITATION NETWORKS TO EVALUATE THE IMPACT OF TEXT SIZE ON THE IDENTIFICATION OF RELEVANT CONCEPTS

Title	Using citation networks to evaluate the impact of text size on the identification of relevant concepts
Authors	Jorge Valverde Tohalino, Thiago C. Silva, and Diego Amancio
Year	2023
Journal	Physica A: Statistical Mechanics and its Applications
Link	<https://arxiv.org/abs/2301.06168>
Situation	Submitted on January, 2023

6.1 Motivation

This paper aims to address an important challenge in natural language processing, which is the identification of significant concepts in unstructured data. This task is crucial for many practical applications, such as text classification, information retrieval, and knowledge discovery. In particular, the study investigates the performance of keyword extraction methods (KE), which are used to automatically identify and extract the most important words or phrases from a text. We note that while there are many existing methods for KE, few studies have explored the impact of text size on their performance, which is an important consideration given that many texts are short and do not provide enough context for accurate keyword extraction.

In our previous work (QUISPE; TOHALINO; AMANCIO, 2021), we developed a word co-occurrence network-based methodology for authorship attribution, where we analyzed literary books of varying sizes. We found that the performance of our algorithms decreased as the size of the text segments reduced, while excellent results were achieved with very large text segments. Motivated by this, we proposed a novel approach to keyword extraction based on citation networks and community detection algorithms. We avoided traditional word co-occurrence networks, which are better suited to larger texts. Our method uses the references of each paper's abstract to construct a citation network for the entire dataset. We then use this citation network as a source of external information to assist our keyword extraction method. Our approach is novel and different from previous methods, as we leverage citation networks to obtain more comprehensive information on each paper's content.

6.2 Contributions

This paper explores the impact of text size on the performance of keyword extraction (KE) methods, which is critical for various practical applications. We adopted a network-based approach to evaluate whether keywords extracted from paper abstracts are compatible with keywords extracted from full papers. We employed a community detection method to identify groups of related papers in citation networks and then used these paper clusters to extract keywords from abstracts. The findings suggest that using different sources of information to extract keywords can lead to significant differences in performance. This study also found that citation networks and alternative methods that do not rely on citations demonstrated suboptimal performance. Further research is necessary to investigate whether the observed variations may lead to discrepancies in the analysis of document similarity networks. We suggested several ways to potentially enhance the performance of clustering methods for keyword extraction, such as the incorporation of text embeddings by integrating citation and text-based information when creating paper networks, and implementing synonym handling during the generation of reference keywords. Future studies should explore the impact of text size and source of information on KE methods for different types of data, as well as the impact of reference keyword quality. Additionally, future research could investigate the potential of using machine learning models to improve keyword extraction accuracy.

Using citation networks to evaluate the impact of text size on the identification of relevant concepts

Jorge A. V. Tohalino
Institute of Mathematics and Computer Science,
University of São Paulo, São Carlos, SP, Brazil

Thiago C. Silva
Universidade Católica de Brasília, Brasília, DF, Brazil

Diego R. Amancio[*]
Institute of Mathematics and Computer Science, Department of Computer Science,
University of São Paulo, São Carlos, SP, Brazil

(Dated: January 16, 2023)

Abstract

The identification of the most significant concepts in unstructured data is of critical importance in various practical applications. Despite the large number of methods that have been put forth to extract the main topics of texts, a limited number of studies have probed the impact of the text size on the performance of keyword extraction (KE) methods. In this study, we adopted a network-based approach to evaluate whether keywords extracted from paper abstracts are compatible with keywords extracted from full papers. We employed a community detection method to identify groups of related papers in citation networks. These paper clusters were then employed to extract keywords from abstracts. Our results indicate that while the various community detection methods employed in our KE approach yielded similar levels of accuracy, a correlation analysis revealed that these methods produced distinct keyword lists for each abstract. We also observed that all considered approaches, however, reach low values of accuracy. Surprisingly, text clustering approaches outperformed all citation-based methods. The findings suggest that using different sources of information to extract keywords can lead to significant differences in performance, and this effect can play an important role in applications relying upon the identification of relevant concepts.

[*] diego@icmc.usp.br

I. INTRODUCTION

With the increasing availability of large amounts of textual content on the Internet, the need for efficient analysis of texts has become imperative. Online textual data encompasses a wide range of sizes and types, including books, encyclopedias and newspapers. In the last few decades, user-generated content in the form of short texts has also grown exponentially. Examples of such content include social media messages, product descriptions, online reviews, as well research papers [21]. In order to summarize this large amount of information, the task of keyword extraction (KE) has emerged as a crucial natural language processing (NLP) application. The goal of KE is to identify the most informative and relevant words or topics within a given document [39]. Keywords serve as a useful tool for users, allowing them to quickly understand the overall content of the texts. Moreover, keyword extraction plays an important role in various NLP applications, including text categorization, document summarization, document tagging, recommendation systems, speech recognition, and many more [21, 22].

The KE task has been the subject of numerous studies. These investigations can be broadly classified into three categories: statistical methods, linguistic/syntactic approaches, and graph-based methods. Different paradigms have also been used in a combined approach for supervised classification, where the extracted features are employed in a machine learning algorithm for a binary classification task [18]. While these methods have demonstrated effectiveness in processing large texts, they present significant challenges when applied to short texts with high sparsity [9].

The identification of keywords within short texts, specifically in scientific manuscripts, poses a significant challenge, particularly when utilizing open scholarly datasets that only provide the title and abstract as sources of textual information [17]. The challenge of extracting keywords from short texts, particularly in the case of scientific papers, has motivated the development of some approaches. One proposed method for addressing the sparsity of abstracts is to group abstracts using clustering techniques [37]. This can be accomplished by utilizing citations as a proxy for determining the similarity between papers, thereby circumventing the need for direct comparison of the short texts. Despite the use of such clustering and other external information [9, 21], there is a lack of comprehensive studies comparing the compatibility of keywords extracted from abstracts and full texts. The accurate representation of the semantic information present in scientific papers is crucial in many areas, as it forms the foundation for many scientometric studies. Thus, this study aims to address the following research questions:

1. To what extent are keywords extracted from abstracts similar to those extracted from the corresponding full papers?
2. Is there consistency in the set of keywords extracted by distinct community detection methods?
3. Does using citations result in superior performance as compared to directly assessing abstract similarity via textual information?

We employed clustering methods to extract keywords from abstracts and compared them with keywords extracted from the corresponding full texts. Using a citation network, we evaluated the performance of various established community detection methods in identifying groups of related papers for the purpose of keyword extraction. We also evaluated clustering approaches that do not rely on citation information, including techniques based on neural embeddings.

The study revealed several interesting results. All evaluated methods were found to have a considerable discrepancy with keywords found in the full texts. We observed that clustering methods that rely solely on textual information outperformed those based on citation networks, indicating that citations may not be an optimal proxy for semantic similarity. Furthermore, our results indicate that the various community detection strategies evaluated yielded similar performance levels, despite the observed differences in the set of keywords identified by each approach.

In summary, our findings suggest that the quantity of information used to extract keywords can strongly impact the performance of the task. Therefore, studies using similarity networks should consider the use of full texts, when available, to provide more robust information regarding topics extracted from paper networks.

The structure of this paper is as follows: In Section II, we present a comprehensive review of the most pertinent studies in the area of keyword extraction. The proposed methodology for extracting keywords from both short and long texts is outlined in Section III, which also includes information regarding the adopted datasets. The main results are presented and discussed in Section IV. Finally, in Section V, we summarize with conclusions and suggest potential perspectives for future research.

II. RELATED WORKS

The early works that addressed the keyword extraction problem focused on statistical methods. The spatial distribution of words along the text is used to gauge words' relevance [7]. The most simple approach is based on word frequency, where words with higher frequency values are considered keywords. However, these methods do not consider word order, therefore, if the text is shuffled, a meaningless version of the text would generate the same set of keywords. The combination of frequency and spatial distribution was then proposed to address this issue via word clustering and entropy [7, 29]. The idea behind these methods is that important words are commonly concentrated in certain parts of the text, where the main topics are located. In this sense, irrelevant words are distributed regularly along texts, while keywords present an uneven distribution and tend to form semantic groups. Another improvement to frequency-based methods is the tf-idf approach, which weights the importance of a word according to its frequency within a text and the frequency along the dataset. The main advantage of these methods is that they are simple and do not require an external corpus or knowledge of the language.

Graph-based methods have also been used approaches to model texts [23]. Several works addressed the keyword extraction problem representing documents as word co-occurrence networks, where two words are connected if they co-occur in a given context [40]. Centrality metrics are then used to assign an importance value to each word. In [20], the authors concluded that network metrics are able to successfully extract relevant words for the keyword extraction task. They also highlighted that network-based approaches do not need the use of external corpora and they are language independent. The use of word embeddings and large contexts has also been useful in improving the quality of co-occurrence networks when extracting keywords [40]. A different approach was proposed by [16], where community detection methods were applied to a network of semantic relationships. The authors used Wikipedia to establish the semantic relatedness between the words of the document. According to [16], important words tend to be grouped into highly connected communities, which are related to the main topics of the document.

In order to address the keyword extraction problem in short texts, several works rely on the use of semantics and background knowledge. According to to [9], extracting only basic or straightforward features from the words is insufficient for finding keywords from short texts. In [21], the authors remarked that text clustering approaches could also be useful in addressing the semantic sparseness of short texts. These techniques enable the clustering of related texts, thereby allowing for the extraction of more semantic information by aggregating texts in the same cluster. [44] employed clustering algorithms to identify the most relevant words for each cluster, operating under the hypothesis that texts with similar topics contain similar keywords. Then, a graph-based approach was applied to each text cluster; and the PageRank algorithm was used to extract keywords.

In [39], the authors proposed a method for selecting keywords based on the informativeness value of each word. This score was calculated at the corpus, cluster, and document levels. At the corpus level, the informativeness was computed taking into account all the documents, while at the cluster level, the word importance was calculated within a group of related texts. The results from the previous steps were then used to compute the informativeness at the document level. This approach yielded a good performance for extracting keywords.

Regarding graph-based techniques, several studies have enhanced TextRank [24] by incorporating different semantic relationships between words as node weights for the word ranking algorithm. For instance, [22] used Wikipedia as an external knowledge base, while [21] employed the Word2Vec and Doc2Vec embedding models to compute the semantic similarity between words.

While many works focus on extracting keywords either from short- and long-texts, here we conduct a comparative analysis of well-established methods for extracting keywords from both short and long texts. We focus on determining the compatibility of keywords extracted from abstracts and those extracted from the full content of research papers.

III. MATERIAL AND METHODS

The framework proposed to extract keywords comprises the following main steps: i) text pre-processing; ii) network construction; iii) community detection; iv) short texts keyword extraction; and v) long texts keyword extraction. The steps are summarized below and illustrated in Figure 1.

1. *Text pre-processing*: this phase comprises the text-processing and vectorization steps. The first step includes the removal of stopwords. The remaining words are stemmed and the tf-idf approach is employed to obtain the vectorized form of the pre-processed texts. Additional details on the pre-processing steps applied can be found in Section III B.

2. *Network creation*: we first constructed a paper citation network, which is used for short text keyword extraction. We also modeled the complete content of each paper as word co-occurrence networks, which were used to extract keywords from long texts. In Section III C, we describe the required steps for the creation of both network models.

3. *Community detection*: we applied community detection methods to the citation networks in order to find clusters of related papers (see Section III D).

4. *Short texts keyword extraction*: this phase is responsible for the extraction of keywords from short texts (paper abstracts). The clusters obtained in the previous step are used in this phase. The relevance of each word is computed inside and outside communities. We also proposed two methods for keyword extraction based on tf-idf and the K-Means algorithm. The methods for short texts keyword extraction are described in Section III E.

5. *Long texts keyword extraction*: to identify reference keywords, we used the complete content of the papers as input from several well-known keyword extraction methods. We evaluated methods based on word frequency, tf-idf, entropy, intermittency, BERT, Yake and TextRank [1, 5, 7, 14, 24]. We also used a network approach based on co-occurrence networks and centrality metrics to find keywords for long texts. These networks were characterized using centrality metrics. A detailed explanation of the adopted methodology is shown in Section III F.

A. Datasets

The following two datasets were used:

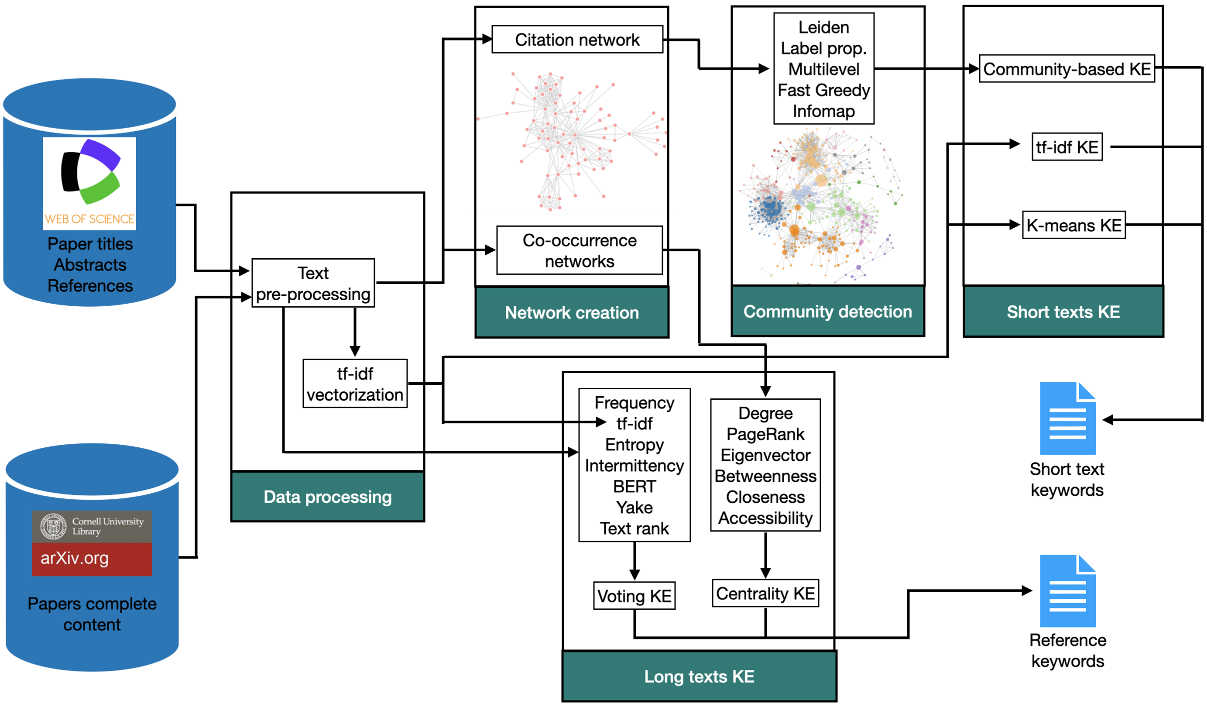


FIG. 1: The workflow of the adopted keyword extraction system involves several stages. The initial step is pre-processing of the text. Next, a citation network is constructed and community detection algorithms are employed to extract keywords from short texts. For comparative evaluation, we implemented both a tf-idf and K-Means approach. Both approaches do not consider citations to cluster papers. In order to generate reference keywords, we employed a combination of statistical and traditional keyword extraction methods on the full texts. Additionally, we also evaluated a co-occurrence network approach as an alternative method for keyword extraction.

1. *Short texts KE dataset*: we used the dataset collected in [37]. The authors retrieved the information from 11,063 papers on the complex networks field. The data was obtained from the *Web of Science* (WOS) database [27]. The selected papers were published from 1991 to 2013. For each paper, the authors extracted the title, abstract, and list of references. The latter was used to construct a citation network. The title and abstract of each paper were used as input for the application of keyword extraction techniques.

2. *Long texts KE dataset*: In order to generate a list of reference keywords for each abstract, we collected the full content of each paper, including the introduction, methodology, results and discussion, conclusions, and appendix sections of each research article. We used the API of the arXiv database [28] to extract the complete content of the papers. We performed an automatic search using both the title and the abstract of each paper as keywords for the arXiv API.

Table I presents a summary of the statistical information for the datasets. The information provided was calculated from the pre-processed versions.

TABLE I: Statistical information from datasets. $|D|$ stands for the number of documents. We also show the average number of tokens (W_{avg}), and the first (W_{q1}) and third (W_{q3}) quartiles of the distribution of the number of tokens. Similarly, U_{avg} represents the average vocabulary size, while U_{q1}, and U_{q3} are the first and third quartiles of the vocabulary size distribution, respectively.

Dataset	**Description**	$\|D\|$	W_{avg}	W_{q1}	W_{q3}	U_{avg}	U_{q1}	U_{q3}
Short texts	Paper abstracts	11,063	79.12	59.00	95.00	55.97	43.00	66.00
Long texts	Full paper content	1,982	2,020.58	1,170.25	2,405.00	517.64	394.00	595.00

B. Data processing

This phase comprises three steps: data preparation, text pre-processing, and tf-idf vectorization. The data preparation step consisted of processing the recovered papers from the arXiv database (for the full content of the papers). We obtained a LaTeX version of each paper, so we had to remove all LaTeX tags. We also removed the authors list, institutions, and acknowledgments from the cleaned text. The following sections were included in the analysis of full papers: introduction, related works, methodology (or materials and methods), results, discussion, conclusion, and appendix sections.

Text pre-processing transformations were applied to all texts of the dataset. We first removed stopwords and punctuation marks. Then, a stemming step was applied to the remaining words. This step is required in order to map each word into its root or stem [30]. The tf-idf technique was used to transform the pre-processed text into a sparse vector representation. To compute the importance of a word w, the technique considers the internal frequency of w in a single document. Moreover, the internal frequency is compared with the relative frequency of w in all documents of the dataset [36]. The tf-idf representation of w in a document d is computed as

$$\text{tf-idf}(w, d) = \frac{f(w, d)}{n_d}.\frac{\log N}{\log(N_w)}, \tag{1}$$

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

We used the tf-idf vector representations of each abstract as input values of a K-Means based method for short texts KE. We also used the tf-idf weight of each word from the full content of the papers in order to give an importance value for a long text KE method (LKE).

C. Network creation

Given the short size of paper abstracts, it is infeasible to extract statistically significant information from individual texts. As such, we employed network representation techniques to extract supplementary information that would enhance the keyword extraction process. Upon analyzing the topological and structural properties of the networks, we are able to infer attributes of the texts that enable us to determine the relative importance of each word.

Two different network models are used in our study. In order to cluster *short texts* into groups of related papers, we used a citation network for the *short texts* keyword extraction task. In this case, the network structure represent the whole dataset of documents. Conversely, when extracting keywords from long texts, each text is modeled as a word co-occurrence network [8, 19, 38].

The unweighted paper citation network was built following the methodology described in [37]. The citation networks are intended to represent a semantical similarity structure that do not use textual information to establish links between papers. The resulting network was composed of $11,063$ nodes and $94,472$ edges. The community structure of this network and the information of title and abstract are then used to detect the most important words in each network community.

The *full content* of a paper is modeled as a word co-occurrence network. In this graph model, each node represents a word, and the edges between two nodes are based on the neighborhood relationship of two words. We used the approach that can include virtual links, so that similar words can be linked. This model and its variations have been used in many different scenarios [11, 13, 15, 31]. The networks were characterized using well-known centrality measurements to rank the words according to their structural importance in the networks [26].

D. Community detection

This phase is responsible for detecting communities, i.e. clusters of papers linked via citation links. Communities are groups of nodes that are more densely interconnected with each other in comparison with the rest of the nodes from the network [32]. The identification of communities in large networks is quite a useful task. For example, the nodes that belong to the same community likely share several common properties. Also, the number of found communities and their respective features could help to identify the category of a network for classification tasks [45]. The identification of communities is also useful to understand the dynamic evolution and organization of a network [12]. In this paper, we evaluated the following methods: Multilevel, Label Propagation, Infomap, Fast Greedy, and Leiden method [4, 10, 33, 35, 42]. In the Appendix, we provide a brief description of each method.

In the context of community detection methods, we investigated if community-based methods are consistent in the sense that they generate well-defined, large communities. This is an important step in our analysis because small communities can lead to low performance [37]. In the paper citation network, most of the community detection methods found between 23 and 39 paper communities, which leads to communities comprising more than 100 papers, typically. The infomap, however, generated more than 400 communities, and most of them comprised less than 10 papers. Before the computation of the relevance of each word, we decided to filter out those communities that contain few papers.

E. Short texts keyword extraction

This step consists of the extraction of keywords from the pre-processed paper abstracts. We evaluated a network community-based approach that generates a word importance index to rank each word from the paper abstracts. For comparison purposes, we also evaluated tf-idf and K-Means-based methods for the short texts KE task.

1. *Community-based approach:* we used the community structure found from papers citation networks to detect the word importance index of each word from paper abstracts. The adopted index quantifies the relative frequency of a word appearing inside a community against its frequency in the remaining documents of the citation network [37]. To compute the word importance index I for a word w, we first compute the frequency

of the word inside a community α. This quantity is the relative internal frequency $F_\alpha^{(\text{in})}(w)$, given by

$$F_\alpha^{(\text{in})}(w) = \frac{n_\alpha(w)}{|\alpha|}, \tag{2}$$

where $n_\alpha(w)$ is the total number of papers containing w appears within a community α, and $|\alpha|$ represents the number of papers associated with a community α. We also compute the relative frequency of w outside α, $F_\alpha^{(\text{out})}(w)$, which is computed as:

$$F_\alpha^{(\text{out})}(w) = \sum_{\gamma \neq \alpha} \frac{n_\gamma(w)}{N - |\alpha|}, \tag{3}$$

where N is the total number of papers in the network. Then, the importance index $I(w)$ is calculated as the highest difference between the relative in-community and out-community frequencies, i.e.:

$$I(w) = \max_\alpha \left[F_\alpha^{(\text{in})}(w) - F_\alpha^{(\text{out})}(w) \right]. \tag{4}$$

The word importance index was computed for all words from paper abstracts, and then the best-ranked words were considered as relevant keywords for each abstract.

2. *tf-idf based approach:* the tf-idf values considering all paper abstracts from the dataset are computed. For each abstract, we considered the tf-idf weights of the words comprising the abstract. The words with the highest tf-idf values were selected as relevant keywords for each abstract.

3. *K-Means based approach:* this method is equivalent to the *community-based approach*. The difference is that clusters are obtained via the K-Means algorithm [34]. To obtain the cluster, we first obtained the embedding of each abstract. Then we evaluated several values of K to find the optimal number of clusters.

F. Long texts keyword extraction

The keywords obtained from full texts are considered reference keywords when evaluating the quality of keywords extracted from short texts. Here we considered as input texts the complete content of the research papers. We adopted several methods found in the literature to extract keywords documents. The methods can be classified into two approaches: statistical and graph-based approaches:

- *Statistical and traditional keyword extraction methods:* In this step we employed statistical techniques that are commonly used for keyword extraction tasks. These methods perform an appropriate analysis of the statistical distribution of words along documents. The main goal of statistical methods is to detect and rank relevant words of documents without any *a priori* or external information [7]. The methods we adopted are based on frequency, word tf-idf, word entropy, word intermittency, and Yake. We also evaluated a graph-based approach named TextRank, and a method that uses word embeddings based on BERT. The methods are described in the Appendix.

- *Network-based methods:* a comprehensive set of centrality measures were used to analyze the word co-occurrence networks derived from the full content of the papers. The network measurements are useful to identify the most relevant nodes in a network [26]. Therefore, they allow ranking the nodes according to their topological importance so that they can find the most important words for each text [41]. We selected as keywords for each text the best-ranked nodes (words) according to the following centrality metrics: degree, PageRank, betweenness, eigenvector centrality, closeness and accessibility computed at the first two levels [43]. We also employed a methodology that combines the results of each centrality metric. In the methodology henceforth referred to as voting system, the keywords found by the majority of the network measurements were selected as relevant keywords for each text.

IV. RESULTS AND DISCUSSION

Our analysis is divided into two sections. Section IV A describes a statistical analysis of the datasets. Section IV B provides a comparison of keywords extracted from short and full-text sources. We also analyze the performance of distinct network community methods for the task.

A. Dataset analysis and selection of reference keywords

In this section, we first perform a statistical evaluation of the datasets through the analysis of the number of common words between the paper abstracts (short-size texts) and the full content of the research paper (long-size texts). This analysis is an initial step to the generation of a set of gold standard keywords for each paper abstract. Because many datasets comprising full-text papers lack keywords selected by human experts, we used as a starting point the full content (including all sections except the abstract) of each paper. We employed keyword extraction methods to extract reference keywords from the complete content of each paper. However, we first analyzed the number of mutual words existing between each abstract and the full content. In some cases, it is possible that the paper authors use specific words to express their main ideas in the abstract and they could change to other words using synonyms or similar expressions in the rest of the paper. Therefore, it becomes important to analyze whether the information extracted from full texts is compatible with the content of abstracts.

We computed how many words (w) in the abstract are also present in the full content of the papers. The cumulative distribution (i.e. $P(x \geq w)$) of this quantity in the dataset is shown in Figure 2. A significant number of research papers (80%) present a high number of common words (40) between the abstracts and the full content of the papers. 50% of the papers have at least 50 common words. As expected, this means that most of the information in the abstract is also available in the remainder of the paper.

Now we evaluate how many *keywords* found in the *full content analysis* are also present in the abstract. We used two approaches to extract reference keywords considering each paper's full content: statistical and graph-based KE methods. We evaluated these approaches by counting the number of mutual words between the keywords generated by each method and the words composing the paper abstracts. Figure 3 depicts the obtained results for each approach. According to the size of the abstracts and the full content (see Table I), we considered recovering between 5 and 50 keywords generated by each KE method. Then, we count the number of these keywords that are part of the abstracts.

In relation to traditional and statistical methods, the results displayed in Figure 3(a) show that the methods Yake, word frequency, and word entropy outperformed the other KE techniques. These methods were able to find the largest number of common words with

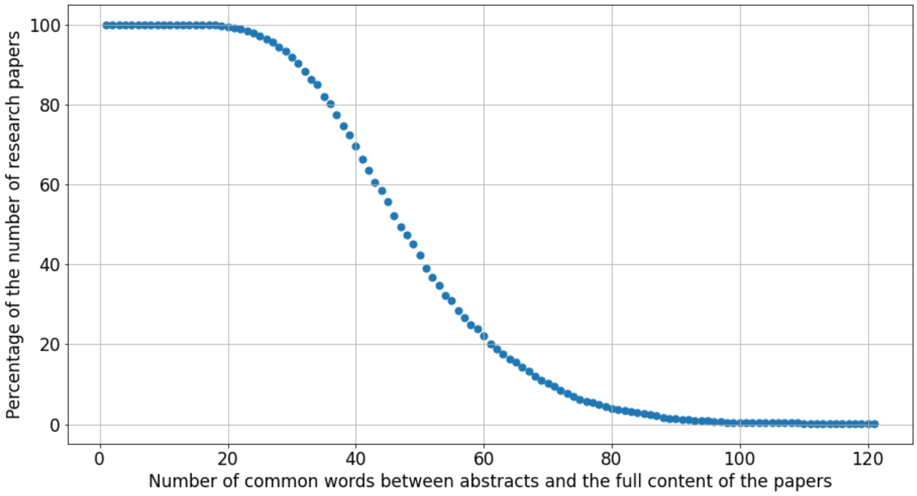
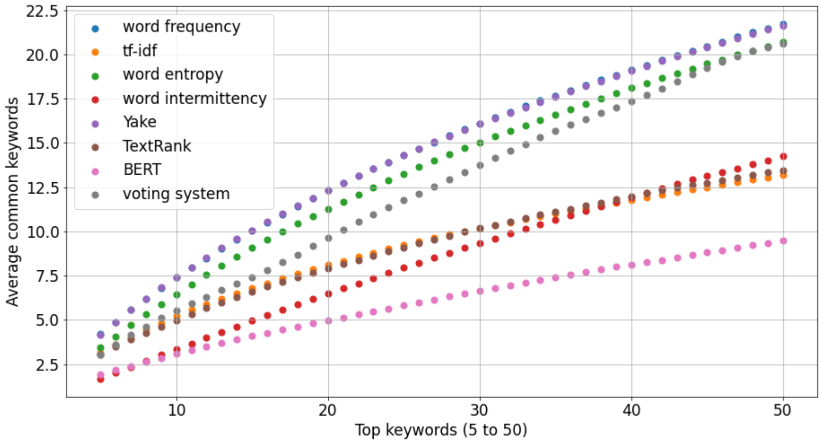


FIG. 2: Analysis of the number of common words between the abstracts and the complete content of the research papers. The x-axis represents the number of mutual words between paper abstracts and their corresponding full content. The y-axis is the representation of $P(x \geq w)$, i.e. the fraction of papers comprising at least w common words between abstracts and full texts.

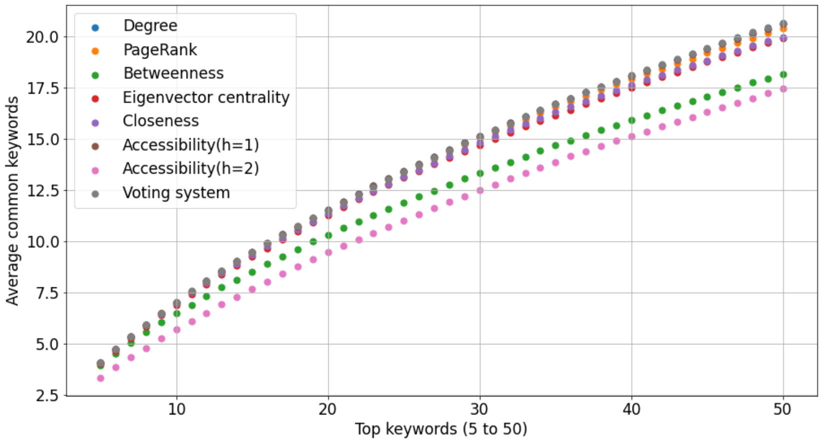
the abstracts. The voting system approach did not achieve the best results. The methods based on word intermittency and BERT also displayed a low number of mutual words with the paper abstracts. We also evaluated the methods based on word co-occurrence networks and centrality measurements. The results depicted in Figure 3(b) suggest that almost all centrality metrics performed similarly. We observed that the voting system, node degree, PageRank, and accessibility (computed at the first hierarchical level) outperformed the other network-based methods. However, the difference in terms of performance with the other network metrics is not significant.

B. Extracting keywords from abstracts

In this section, we analyze if the methods adopted to extract *keywords from abstracts* are able to capture keywords that are found when the full-text content is analyzed. We used accuracy as a performance evaluation measure. The performance of the methods is measured in terms of the number of common words between the reference keywords and the keywords generated by the short-text KE methods, divided by the total number of reference



(a)Analysis of traditional and statistical keyword extraction methods for the full content of research papers.



(b)Analysis of network-based keyword extraction methods for the full content of research papers.

FIG. 3: Analysis of the overlap between the common words found in the paper abstracts and those identified by keyword extraction methods for longer texts (i.e., the full content of the papers). The x-axis represents the number of keywords recovered in the full content, while the y-axis indicates the average number of retrieved keywords that also appear in the paper abstract.

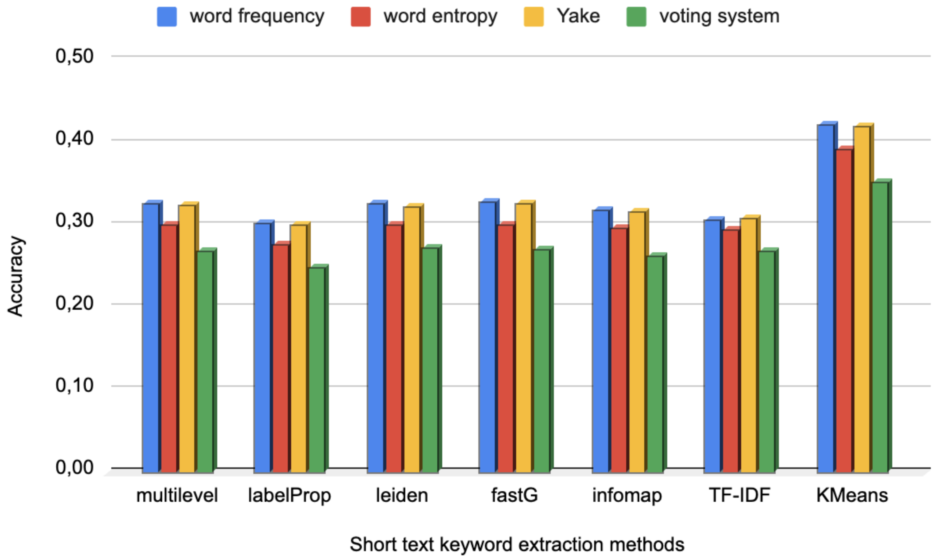
keywords. We established a parameter N to represent the number of reference keywords to be considered in the evaluation. As reference keywords (i.e. keywords obtained from full texts), we used the methods with the highest performance observed in the previous section.

Figure 4(a) shows the performance analysis considering as reference keywords the ones obtained from statistical methods. The results in Figure 4(a) suggest that community-based approaches obtained similar accuracy values since no method clearly outperformed the others. The label propagation method achieved a slightly lower performance than the other methods. We also found that the tf-idf method displayed a performance that is similar to the other network community-based methods. Surprisingly, when citation information is disregarded and only the textual information is used, the performance is improved. The K-Means method is significantly better than all other approaches, with a gain of 25% in performance, in some cases. The complete analysis considering different values for the parameter N is shown in the Appendix.

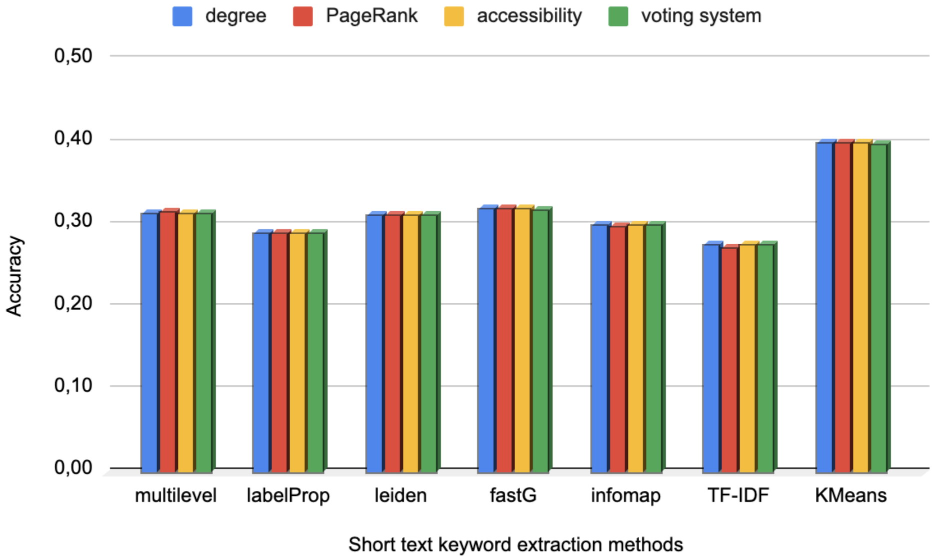
Figure 4(b) shows the performance analysis considering as reference keywords the ones obtained from (co-occurrence) network-based methods. Here we see that all considered network centrality metrics provide almost the same performance. This result is consistent with the literature in text network analysis since there is a correlation of centrality metrics when analyzing written texts. Concerning the formation of reference keywords via community detection methods, it is also worth noting that all community-based methods provided similar performance for the task. Conversely, choosing reference keywords via tf-idf yielded the worst performance. Once again, the best performance was obtained with the KMeans method.

One possible explanation for the similar performance achieved by the community-based detection methods could be the fact that all methods are generating similar partitions and, consequently, they are selecting the same set of keywords for each paper abstract. In order to evaluate this hypothesis, we computed the Spearman correlation coefficient of the ranking of words' relevance generated by different community detection methods. The results (not shown) revealed that the methods are actually selecting different sets of keywords. If one considers the full rank of words, the Spearman correlations are typically lower than 0.10. In a similar fashion, when considering only the top 30 ranked words, all correlations were below 0.27. As expected, in both scenarios, the highest correlation was found for the rankings generated by the Multilevel and Leiden methods [4, 42].

The performance results revealed interesting insights. First, we found that identifying keywords from citation information alone does not provide the highest match between keywords found in abstract and full-text. While citations have been used in numerous contexts [2], one possible reason for the observed low performance is that citations may not



(a) Performance analysis considering the traditional and statistical methods as reference keywords.



(b) Accuracy analysis considering the network-based methods as reference keywords.

FIG. 4: Comparative analysis of the accuracy obtained from the evaluation of KE methods for short texts (paper abstracts). We extracted $N = 30$ keywords extracted from full texts.

reflect the semantic similarity of texts, which may hinder the performance of the community detection methods. In fact, some studies have pointed out a discrepancy between citation and content similarity. For example, [2] found that citation and content similarity are not consistent since the most similar papers are oftentimes disregarded when selecting references for papers. In a similar fashion, the differences in the content have been used to improve models reproducing the growth of citation networks, since content similarity has also been used as an important feature to model the growth of citation networks [46].

While the use of textual information was able to provide a better performance in recovering keywords from full-text content, the obtained accuracy is still below 50%. This means that using cluster information from papers abstracts is not enough to recover the full content of papers. This may have implications in many studies that are based on recovering text content based on keywords. For example, when studying the properties of citation networks, the selection of papers via keywords may affect the stability of citation network metrics. A different number of communities depicting subfields of a major area can be found if the keywords terms are not well-defined to select the relevant papers.

The differences in content extracted from abstracts and full texts can also potentially lead to distinct interpretations in the context of *Science of Science*. In a document similarity network, for example, the centrality of a paper may strongly depend on the use of abstracts or full texts. If such networks are studied in other contexts, this may lead to less robust conclusions. For example, comparing the semantic similarity between papers linked by citations may lead to different results depending on how much text is used to gauge semantic similarity [2]. Therefore, it remains relevant to consider full-text content to draw conclusions relying upon the analysis of papers' semantic similarity.

V. CONCLUSION

The identification of keywords from short texts poses a significant challenge. In this paper, we evaluated whether well-known approaches are able to extract keywords from abstracts that are compatible with a full-content analysis. Due to the limited context provided in abstracts, we employed methods that leverage the citation context to cluster papers into semantically similar groups. Additionally, we used strategies based on statistics and the K-Means algorithm. Reference keywords were obtained from the full content of papers through the use of multiple techniques.

The findings indicate that a simple approach such as the K-Means algorithm outperforms methods that rely on communities derived from citation networks. Additionally, the results demonstrate a similar performance among the various community detection methods applied to citation networks, with no clear superiority demonstrated by any particular method.

All in all, citation networks and alternative methods that do not rely on citations demonstrated suboptimal performance. This result implies that the keywords obtained from abstracts are not consistent with those obtained from a comprehensive content analysis. Consequently, further research is necessary to investigate whether the observed variations may lead to discrepancies in the analysis of document similarity networks [25].

One way to potentially enhance the performance of clustering methods for keyword extraction is through the use of alternative methods for text vector representation. The incorporation of text embeddings, such as those generated by the BERT model [3], may assist in effectively representing the documents. Additionally, implementing synonym handling during the generation of reference keywords could also prove beneficial. The performance could also be improved by integrating citation and text-based information when creating paper networks.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Diego R. Amancio acknowledges financial support from CNPq-Brazil (Grant no. 304026/2018-2, 311074/2021-9) and FAPESP (Grant no. 20/06271-0). Thiago C. Silva (Grant no. 308171/2019-5, 408546/2018-2) gratefully acknowledges financial support from the CNPq foundation.

APPENDIX

A. Community detection methods

In this section, we provide a brief description of the network community methods employed in this paper:

1. *Multilevel*: in this algorithm, each node is assigned to a different community. Then nodes are moved to the communities of their corresponding neighbors that yields the highest positive contribution to modularity [4]. This process is repeated until the local contribution of nodes to the modularity is no longer improved. Each community from

the original network is reduced into a single node (maintaining the total weight of the adjacent edges) and the method continues to the next level. The algorithm ends when there is no longer any possibility of increasing the modularity score after reducing communities to nodes.

2. *Label propagation:* The method presented in [33] is based on the principles of neighborhood connectivity and information diffusion in networks. The approach begins by assigning unique community labels to each node in the network. These labels are subsequently propagated throughout the network. During each iteration, each node adopts the most prevalent label within its immediate neighborhood. The edges within the network are then randomly removed, and the nodes are updated in a random order before the next iteration commences. The algorithm stops when the nodes reach a consensus, which is defined as a state in which each node holds the majority label among its neighboring nodes.

3. *Leiden:* The Leiden algorithm, which was proposed in [42], represents an improvement to the widely-used multilevel method [4]. The latter is known to have a weakness of often discovering communities that are weakly connected. In contrast, the Leiden method aims at ensuring that communities are well-connected through the implementation of the following three phases: (i) local moving of nodes (as in the multilevel method); (ii) refinement of partitions; and (iii) aggregation of the network. By incorporating these three phases, the Leiden algorithm is able to uncover higher-quality clusters in significantly less time when compared to the multilevel method.

4. *Fast Greedy:* this algorithm is based on hierarchical agglomerative clustering and aims to optimize the modularity score [10]. The method begins by considering a subnetwork composed exclusively of edges between highly-connected nodes. This methodology subsequently evaluates randomly selected edges that improve the modularity of the subnetwork and aggregates them. This process is repeated until the incremental improvement in modularity becomes negligible. Finally, the communities are obtained by identifying the connected components within the subnetwork.

5. *Infomap:* the algorithm was introduced by [35] and is based on information theory. This method begins by encoding the network into modules in a manner that maximizes

the amount of information retained from the original network. The encoded network is then transmitted through a channel with limited capacity. The goal of the decoder is to attempt to decode the message and construct a set of possible candidates for the original network. The fewer the number of candidates, the more information about the original network has been transmitted. The algorithm also uses random walks to analyze the flow of information through the network.

B. Statistical keyword detection

- *Word frequency and tf-idf methods (Freq. and tf-idf):* one of the simplest techniques for keyword extraction is the frequency-based approach, which assigns relevance to words that occur at a high frequency. The words that rank the highest in terms of frequency are therefore considered as keywords. In order to mitigate the limitations of the frequency-based methods, we also evaluated the tf-idf method. Unlike the frequency-based approach, the tf-idf method assigns a weight to the frequency of each word based on its number of occurrences within the document as well as throughout the entire dataset. In this approach, the words with the highest tf-idf values are considered as keywords.

- *Word entropy (W.E.):* This method leverages Shannon's entropy to analyze the information content of the sequence of occurrences of each word in a given text [7]. This technique requires partitioning the texts into N segments to calculate the entropy of each word. In this study, we partitioned the paper texts according to the number of sentences that make up each text. According to this method, the higher the value of entropy of a word, the greater the heterogeneity of the distribution of that word within the text, and thus the greater its relevance. One of the key advantages of this method is that it does not require a large text corpus for training; it only requires the input text.

- *Word intermittency (W.I.):* This metric takes into account the relationship between the significance of a word and its spatial distribution [1]. Previous research has found that important words are closely related to the main topics of the text and display a highly heterogeneous distribution. Such words tend to be located in specific regions of the text, exhibit large frequency fluctuations and often form clusters [7]. In contrast, common words such as stopwords are distributed randomly throughout the document and exhibit a relatively homogeneous distribution. Thus, as proposed by [7], a statistical analysis of the distribution of word occurrences can be employed to identify relevant keywords within a given text. Similar to the frequency and entropy methods, this technique identifies important words solely based on the target text and does not require external information.

- *Yake*: The Yake method extracts statistical features from the source text to identify the most relevant keywords [5]. Five features are computed for each individual term: (i) casing, (ii) word positional, which assigns greater importance to words that appear at the beginning of a text, (iii) word frequency, which assigns relevance to words that occur more frequently, (iv) word relatedness to context, which measures the number of different terms that appear to the left and right of the target word, and (v) word *difSentence*, which measures how often a word appears across different sentences. These features are then combined into a single measure to assign an importance weight to each word. According to this score, words with the lower values are considered as relevant keywords [6].

- *TextRank (TextR):* The TextRank method, proposed by [24], is a graph-based approach that employs the PageRank algorithm and is widely used for text summarization and keyword extraction tasks. In this method, texts are modeled as word co-occurrence networks, where the nodes are represented by words and edges are established between two nodes if they co-occur within a window size. In the original paper, the window size was set between 2 and 10 words. The PageRank algorithm is employed to rank each word, and the top-ranked words are selected as relevant keywords.

- *BERT-based method*: the Bidirectional Encoder Representations from Transformers (BERT) technique is a state-of-the-art embedding model that captures the semantic content of documents through dense vector representations [14]. The BERT-based method generates word embeddings for each n-gram in the text. Subsequently, the cosine similarity metric is applied to identify the words that are most similar to the original document. The top-ranking similar words are then considered as relevant keywords for each document.

- *Voting system (V.S.):* In order to improve the accuracy of the long text keyword extractor, we combined the results of the above proposed methods. We used a voting system based on the keywords found by most keyword extraction methods.

C. Complete results based on accuracy analysis

TABLE II: Accuracy obtained from the evaluation of keyword extraction methods for short texts (paper abstracts). Here we considered as reference keywords the relevant words found by the *traditional and statistical* methods for the full content of the papers. N represents the number of top keywords we recovered for both short and long texts keyword extraction methods.

	Word Frequency				**tf-idf**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1578	0.2559	0.3269	0.3678	0.0740	0.1066	0.1448	0.1793
labelProp	0.1375	0.2312	0.3022	0.3504	0.0787	0.1033	0.1363	0.1724
leiden	0.1496	0.2533	0.3260	0.3664	0.0819	0.1153	0.1463	0.1785
fastG	0.1465	0.2518	0.3290	0.3684	0.0713	0.1023	0.1419	0.1767
infomap	0.1375	0.2320	0.3175	0.3647	0.1458	0.1561	0.1726	0.1951
tf-idf	0.2858	0.2889	0.3071	0.3350	0.3596	0.3223	0.2984	0.2755
KMeans	0.4253	0.4213	0.4220	0.4146	0.2129	0.2106	0.2198	0.2279
	Word Entropy				**Word Intermittency**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1341	0.2294	0.3004	0.3460	0.0679	0.1291	0.1828	0.2233
labelProp	0.1165	0.2052	0.2769	0.3285	0.0608	0.1159	0.1663	0.2110
leiden	0.1279	0.2293	0.3007	0.3455	0.0651	0.1288	0.1829	0.2228
fastG	0.1237	0.2232	0.3012	0.3456	0.0658	0.1241	0.1807	0.2218
infomap	0.1259	0.2166	0.2967	0.3441	0.0715	0.1274	0.1811	0.2225
tf-idf	0.2477	0.2708	0.2945	0.3217	0.1093	0.1539	0.1868	0.2180
KMeans	0.3596	0.3835	0.3929	0.3925	0.1335	0.1921	0.2280	0.2499
	Yake				**TextRank**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1553	0.2536	0.3242	0.3650	0.1138	0.1689	0.2101	0.2326
labelProp	0.1362	0.2290	0.2997	0.3481	0.0967	0.1594	0.2012	0.2260
leiden	0.1509	0.2516	0.3226	0.3633	0.0986	0.1673	0.2090	0.2319
fastG	0.1448	0.2485	0.3256	0.3653	0.1172	0.1818	0.2184	0.2356
infomap	0.1400	0.2309	0.3153	0.3616	0.0829	0.1503	0.2034	0.2302
tf-idf	0.2905	0.2937	0.3091	0.3353	0.1381	0.1492	0.1703	0.1951
KMeans	0.4221	0.4190	0.4194	0.4121	0.2132	0.2337	0.2459	0.2479
	BERT				**Voting System**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.0573	0.0924	0.1254	0.1501	0.1231	0.1925	0.2688	0.3283
labelProp	0.0710	0.0936	0.1221	0.1462	0.1108	0.1758	0.2485	0.3123
leiden	0.0757	0.1092	0.1325	0.1518	0.1325	0.2003	0.2716	0.3274
fastG	0.0739	0.0965	0.1257	0.1494	0.1300	0.1941	0.2702	0.3280
infomap	0.0596	0.0938	0.1261	0.1483	0.1115	0.1826	0.2633	0.3251
tf-idf	0.0899	0.1043	0.1207	0.1394	0.1891	0.2280	0.2687	0.3070
KMeans	0.1804	0.1668	0.1688	0.1726	0.2761	0.3110	0.3531	0.3746

TABLE III: Accuracy obtained from the evaluation of keyword extraction methods for short texts (paper abstracts). Here we considered as reference keywords the most important words found by the *network-based* methods for the full content of the papers. N is the number of top keywords we recovered for both short and long texts keyword extraction methods.

	Degree				**PageRank**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1605	0.2490	0.3149	0.3539	0.1622	0.2498	0.3153	0.3514
labelProp	0.1324	0.2234	0.2903	0.3374	0.1329	0.2247	0.2899	0.3349
leiden	0.1417	0.2421	0.3125	0.3522	0.1428	0.2429	0.3120	0.3494
fastG	0.1466	0.2496	0.3195	0.3549	0.1481	0.2510	0.3194	0.3523
infomap	0.1256	0.2152	0.3002	0.3476	0.1247	0.2146	0.2979	0.3445
tf-idf	0.2529	0.2546	0.2757	0.3085	0.2516	0.2521	0.2725	0.3043
KMeans	0.4047	0.3989	0.4010	0.3948	0.4070	0.3998	0.3999	0.3920
	Betweenness				**Eigenvector**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1563	0.2281	0.2809	0.3127	0.1484	0.2385	0.3008	0.3385
labelProp	0.1255	0.2021	0.2571	0.2973	0.1253	0.2136	0.2787	0.3238
leiden	0.1356	0.2202	0.2774	0.3099	0.1343	0.2315	0.2994	0.3370
fastG	0.1392	0.2272	0.2836	0.3134	0.1387	0.2373	0.3031	0.3393
infomap	0.1125	0.1855	0.2596	0.3028	0.1239	0.2120	0.2921	0.3346
tf-idf	0.2291	0.2255	0.2387	0.2661	0.2490	0.2522	0.2726	0.3021
KMeans	0.3828	0.3656	0.3574	0.3503	0.3865	0.3844	0.3847	0.3797
	Closeness				**Accessibility (h=1)**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1585	0.2438	0.3066	0.3426	0.1605	0.2490	0.3149	0.3539
labelProp	0.1317	0.2198	0.2855	0.3278	0.1324	0.2234	0.2903	0.3374
leiden	0.1390	0.2373	0.3052	0.3411	0.1417	0.2421	0.3125	0.3522
fastG	0.1461	0.2446	0.3096	0.3431	0.1466	0.2496	0.3195	0.3549
infomap	0.1209	0.2127	0.2932	0.3367	0.1256	0.2152	0.3002	0.3476
tf-idf	0.2478	0.2499	0.2717	0.3019	0.2529	0.2546	0.2757	0.3085
KMeans	0.3936	0.3883	0.3891	0.3828	0.4047	0.3989	0.4010	0.3948
	Accessibility(h=2)				**Voting System**			
method	**N=10**	**N=20**	**N=30**	**N=40**	**N=10**	**N=20**	**N=30**	**N=40**
multilevel	0.1196	0.1943	0.2526	0.2904	0.1605	0.2481	0.3141	0.3535
labelProp	0.1048	0.1744	0.2338	0.2777	0.1324	0.2226	0.2909	0.3377
leiden	0.1088	0.1869	0.2508	0.2894	0.1426	0.2408	0.3119	0.3522
fastG	0.1108	0.1909	0.2519	0.2899	0.1464	0.2487	0.3175	0.3548
infomap	0.0997	0.1740	0.2419	0.2855	0.1255	0.2147	0.3001	0.3479
tf-idf	0.1977	0.2125	0.2344	0.2638	0.2536	0.2546	0.2759	0.3093
KMeans	0.3162	0.3167	0.3236	0.3264	0.4039	0.3963	0.3985	0.3952

[1] D. R. Amancio. Probing the topological properties of complex networks modeling short written texts. *PloS one*, 10(2):e0118394, 2015.

[2] D. R. Amancio, M. d. G. V. Nunes, O. N. Oliveira Jr, and L. da F. Costa. Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics*, 91(3):827–842, 2012.

[3] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[6] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer, 2018.

[7] C. Carretero-Campos, P. Bernaola-Galván, A. Coronado, and P. Carpena. Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications*, 392(6):1481–1492, 2013.

[8] N. Castro and M. Stella. The multiplex structure of the mental lexicon influences picture naming in people with aphasia. *Journal of Complex Networks*, 7(6):913–931, 2019.

[9] J. Chen, H. Hou, and J. Gao. Inside importance factors of graph-based keyword extraction on chinese short text. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–15, 2020.

[10] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[11] E. A. Corrêa Jr and D. R. Amancio. Word sense induction using word embeddings and community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 523:180–190, 2019.

[12] N. Dakiche, F. B.-S. Tayeb, Y. Slimani, and K. Benatchba. Tracking community evolution in social networks: A survey. *Information Processing & Management*, 56(3):1084–1102, 2019.

[13] H. F. de Arruda, L. d. F. Costa, and D. R. Amancio. Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063120, 2016.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] H. Ferraz de Arruda, F. Nascimento Silva, V. Queiroz Marinho, D. Raphael Amancio, and L. da Fontoura Costa. Representation of texts as complex networks: a mesoscopic approach. *Journal of Complex Networks*, 6(1):125–144, 2018.

[16] M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, pages 661–670, 2009.

[17] A.-W. Harzing. Two new kids on the block: How do crossref and dimensions compare with google scholar, microsoft academic, scopus and the web of science? *Scientometrics*, 120(1):341–349, 2019.

[18] X. Jiang, Y. Hu, and H. Li. A ranking approach to keyphrase extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 756–757, 2009.

[19] S. M. Joseph, S. Citraro, V. Morini, G. Rossetti, and M. Stella. Cognitive network neighbourhoods quantify feelings expressed in suicide notes and reddit mental health communities. *Physica A: Statistical Mechanics and its Applications*, page 128336, 2022.

[20] S. Lahiri, S. ray choudhury, and C. Caragea. Keyword and keyphrase extraction using centrality measures on collocation networks. 01 2014.

[21] J. Li, G. Huang, C. Fan, Z. Sun, and H. Zhu. Key word extraction for short text via word2vec, doc2vec, and textrank. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(3):1794–1805, 2019.

[22] W. Li and J. Zhao. Textrank algorithm by exploiting wikipedia for short text keywords extraction. In *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*, pages 683–686. IEEE, 2016.

[23] J. Machicao, E. A. Corrêa Jr, G. H. Miranda, D. R. Amancio, and O. M. Bruno. Authorship

attribution based on life-like network automata. *PloS one*, 13(3):e0193703, 2018.

[24] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[25] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

[26] M. Newman. *Networks*. Oxford university press, 2018.

[27] https://clarivate.com/webofsciencegroup/solutions/web-of-science/.

[28] https://arxiv.org/.

[29] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Munoz, and A. M. Somoza. Keyword detection in natural languages and dna. *EPL (Europhysics Letters)*, 57(5):759, 2002.

[30] R. Pramana, J. J. Subroto, A. A. S. Gunawan, et al. Systematic literature review of stemming and lemmatization performance for sentence similarity. In *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, pages 1–6. IEEE, 2022.

[31] L. V. Quispe, J. A. Tohalino, and D. R. Amancio. Using virtual edges to improve the discriminability of co-occurrence text networks. *Physica A: Statistical Mechanics and its Applications*, 562:125344, 2021.

[32] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the national academy of sciences*, 101(9):2658–2663, 2004.

[33] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007.

[34] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PloS one*, 14(1):e0210236, 2019.

[35] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008.

[36] G. Salton and C.-S. Yang. On the specification of term values in automatic indexing. *Journal of documentation*, 1973.

[37] 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.

[38] M. Stella. Multiplex networks quantify robustness of the mental lexicon to catastrophic concept failures, aphasic degradation and ageing. *Physica A: Statistical Mechanics and its Applications*, 554:124382, 2020.

[39] M. Timonen, T. Toivanen, Y. Teng, C. Chen, and L. He. Informativeness-based keyword extraction from short documents. In *KDIR*, pages 411–421, 2012.

[40] J. A. Tohalino, T. C. Silva, and D. R. Amancio. Using virtual edges to extract keywords from texts modeled as complex networks. *arXiv preprint arXiv:2205.02172*, 2022.

[41] J. V. Tohalino and D. R. Amancio. Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539, 2018.

[42] V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[43] B. A. N. Travençolo and L. d. F. Costa. Accessibility in complex networks. *Physics Letters A*, 373(1):89–95, 2008.

[44] X. Wan and J. Xiao. Collabrank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969–976, 2008.

[45] X.-G. Wang. A network classification method by using community structure. *Applied Mathematics & Information Sciences*, 9(3):1181, 2015.

[46] Q. Zhao and X. Feng. Utilizing citation network structure to predict paper citation counts: A deep learning approach. *Journal of Informetrics*, 16(1):101235, 2022.

CHAPTER 7

CONCLUSION

Our research focused on developing automated methods to analyze scientific items, specifically in two applications: research grant classification and keyword extraction. For research grant classification, we used machine learning algorithms to predict the productivity of Brazilian research grants based on features extracted from summaries of research projects and researchers. This approach has the potential to facilitate the selection and funding of research projects with higher impact. For keyword extraction, we represented texts using various network models and analyzed their structural properties to identify the most relevant words for each document. This method can assist in summarizing and categorizing large amounts of text data in a more efficient and accurate way.

Our study aimed to analyze various scientific documents, including abstracts of research proposals, paper abstracts, and full research articles, using a combination of different research areas. We had multiple motivations for evaluating these datasets. Firstly, with the exponential growth of scientific literature in online journals, scientometric and bibliometric methods have become crucial in evaluating the quality of this information. Therefore, we utilized methods based on these approaches as external tools that can provide valuable information for the tasks proposed in this work. Additionally, we aimed to address the challenge of characterizing and extracting important patterns in texts of different sizes and types. Hence, considering the type of text documents we used, we were able to complementarily utilize three significant research areas: natural language processing (NLP), complex networks, and scientometrics/bibliometrics.

To achieve the first task, we authored two articles focused on research grant classification, each addressing different feature extraction approaches to characterize each research proposal. In our first article, we evaluated whether the extraction of topical and complexity features solely from the abstracts of research grants is sufficient to predict their future success or productivity. We also extracted bibliometric features related to the academic activity of the researchers, including collaborations with other researchers, publications and citations over time, participation in research projects, and institutions where the researchers worked. Our hypothesis was that a researcher's scientific career excellence is an indicator of their future project success and productivity. In our second paper, we evaluated all of these features.

The results of our previous approaches showed that the productivity of research projects is related to the academic background of the researchers involved. However, using only features based on the abstracts of research grants is insufficient to fully characterize their productivity. For small texts, additional information is needed to extract more robust features. One possible improvement is to combine methods based on text features and bibliometric features, taking the best of each approach. Other productivity criteria, such as the total number of publications, citation patterns, usage counts, or journal reputation, could also be used. However, our methodology is limited to projects funded by a Brazilian grant agency, and to obtain a deeper understanding, larger datasets from other scientific databases such as Scopus or Web of Science should be considered in future research.

Our second goal was to develop methods for extracting keywords from paper abstracts and full content papers using complex network concepts. To achieve this, we devised two different approaches. The first method involved extracting the best-ranked words based on different centrality measurements applied to co-occurrence networks of each text. We varied the number of neighbors considered for co-occurrence edges and also included virtual edges based on semantic similarity between words. We evaluated these approaches on datasets ranging from abstracts to full articles. Our second approach involved modeling the dataset as a citation network, where each paper was represented as a node and two papers were connected if one cited the other. We then used different community extraction methods on this network and extracted groups of related papers. We computed an importance index for each word based on its frequency within and outside these groups. Using this index, we ranked each word based on its relevance. This approach was focused on evaluating small-sized texts such as paper abstracts. Overall, both methods showed promising results in keyword extraction. However, further investigation is needed to evaluate these approaches on larger datasets and to improve their performance.

Our first approach yielded promising results indicating that including virtual edges, which provide semantic information about words, could improve the accuracy of our methodology. However, the number of virtual edges added to the network should be minimal to prevent a decrease in performance. Our findings are consistent with the approach described in Quispe, Tohalino and Amancio (2021), which showed that incorporating edges created via word embeddings improved the performance of methods used for authorship classification. Word embeddings can also manage synonyms, reducing redundancy in co-occurrence networks. Furthermore, we demonstrated that centrality metrics are valuable tools for extracting features from each word, with potential uses for other NLP tasks such as document classification and summarization. Our second approach focused on finding the most significant words in small texts, but the accuracy of our network-based method was lower than that of other simpler methods used in this research. For example, our K-Means algorithm was more effective at finding relevant keywords for our proposed task. However, we believe that community detection methods can still be valuable for other tasks such as paper clustering and visualization of related research. One potential improvement to our methodology is to model the papers of a cluster as co-occurrence networks and then apply centrality metrics to extract the most important words for each group of related articles. It is worth noting that datasets with reference keywords selected by human experts in the area are crucial for further advancements in this field.

Table 1 presents the research papers that were produced during my Ph.D. program. The first two papers focused on research grant classification, while two additional papers on keyword extraction have been submitted to a scientific journal. In collaboration with a colleague from the same Ph.D. program, we also published a paper on authorship attribution using complex network concepts, which although not included in this monograph, is highly relevant to our research.

Table 1 – Research papers produced in the course of this Ph.D. program.

Paper	**State**	**Task**
Analyzing the relationship between text features and grants productivity	Published	Grant classification
On predicting research grants productivity via machine learning	Published	Grant classification
Using virtual edges to extract keywords from texts modeled as complex networks	Submitted	Keyword extraction
Using citation networks to evaluate the impact of text size on the identification of relevant concepts	Submitted	Keyword extraction
Using virtual edges to improve the discriminability of co-occurrence text networks	Published	Authorship attribution

BIBLIOGRAPHY

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citation on page 22.

ALMEIDA, F.; XEXÉO, G. Word embeddings: A survey. **arXiv preprint arXiv:1901.09069**, 2019. Citation on page 28.

AMANCIO, D. R.; ANTIQUEIRA, L.; PARDO, T. A.; COSTA, L. da F.; JR, O. N. O.; NUNES, M. G. Complex networks analysis of manual and machine translations. **International Journal of Modern Physics C**, World Scientific, v. 19, n. 04, p. 583–598, 2008. Citation on page 33.

ANTIQUEIRA, L. **Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2007. Citations on pages 30 and 33.

ANTIQUEIRA, L.; PARDO, T. A. S.; NUNES, M. d. G. V.; JR, O. N. O.; COSTA, L. d. F. Some issues on complex networks for author characterization. **Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial**, v. 11, n. 36, p. 51–58, 2007. Citations on pages 23 and 33.

BELIGA, S. Keyword extraction: a review of methods and approaches. **University of Rijeka, Department of Informatics, Rijeka**, v. 1, n. 9, 2014. Citation on page 30.

BORGATTI, S. P. Centrality and network flow. **Social Networks**, v. 27, n. 1, p. 55–71, 2005. Citation on page 31.

CHEN, J.; HOU, H.; GAO, J. Inside importance factors of graph-based keyword extraction on chinese short text. **ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)**, ACM New York, NY, USA, v. 19, n. 5, p. 1–15, 2020. Citations on pages 21 and 22.

CHERIFI, H. **Complex networks and their applications**. [S.l.]: Cambridge Scholars Publishing, 2014. Citation on page 31.

CHOUDHURY, M.; THOMAS, M.; MUKHERJEE, A.; BASU, A.; GANGULY, N. How difficult is it to develop a perfect spell-checker? a cross-linguistic analysis through complex network approach. **arXiv preprint physics/0703198**, 2007. Citations on pages 23 and 33.

Costa, L. D. F.; Oliveira JR., O.; Travieso, G.; Rodrigues, F. A.; Villas Boas, P.; Antiqueira, L.; Viana, M. P.; Correa Rocha, L. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, p. 329–412, May 2011. Citation on page 33.

Costa, L. D. F.; Rodrigues, F. A.; Travieso, G.; Villas Boas, P. R. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, p. 167–242, Jan. 2007. Citations on pages 30, 31, and 32.

CRAWFORD, M.; KHOSHGOFTAAR, T. M.; PRUSA, J. D.; RICHTER, A. N.; NAJADA, H. A. Survey of review spam detection using machine learning techniques. **Journal of Big Data**, SpringerOpen, v. 2, n. 1, p. 1–24, 2015. Citation on page 29.

EMINAGAOGLU, M. A new similarity measure for vector space models in text classification and information retrieval. **Journal of Information Science**, SAGE Publications Sage UK: London, England, v. 48, n. 4, p. 463–476, 2022. Citation on page 28.

ERDÖS, P.; RÉNYI, A. On random graphs, i. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959. Citation on page 31.

FALAGAS, M. E.; KOURANOS, V. D.; ARENCIBIA-JORGE, R.; KARAGEORGOPOULOS, D. E. Comparison of scimago journal rank indicator with journal impact factor. **The FASEB journal**, Federation of American Societies for Experimental Biology, v. 22, n. 8, p. 2623–2628, 2008. Citation on page 35.

GARFIELD, E. *et al.* The impact factor. **Current contents**, New York, v. 25, n. 20, p. 3–7, 1994. Citation on page 35.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citation on page 32.

GUIDA, G.; MAURI, G. Evaluation of natural language processing systems: Issues and approaches. **Proceedings of the IEEE**, IEEE, v. 74, n. 7, p. 1026–1035, 1986. Citation on page 27.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National academy of Sciences**, National Acad Sciences, v. 102, n. 46, p. 16569–16572, 2005. Citation on page 34.

HOOD, W.; WILSON, C. The literature of bibliometrics, scientometrics, and informetrics. **Scientometrics**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV . . . , v. 52, n. 2, p. 291–314, 2001. Citation on page 35.

HOTHO, A.; NÜRNBERGER, A.; PAASS, G. A brief survey of text mining. In: **Ldv forum**. [S.l.: s.n.], 2005. v. 20, n. 1, p. 19–62. Citation on page 27.

KENTON, J. D. M.-W. C.; TOUTANOVA, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of naacL-HLT**. [S.l.: s.n.], 2019. p. 4171–4186. Citations on pages 22 and 29.

KESSLER, M. M. Bibliographic coupling between scientific papers. **American documentation**, Wiley Online Library, v. 14, n. 1, p. 10–25, 1963. Citation on page 35.

KOSTOFF, R. N. Co-word analysis. In: **Evaluating R&D impacts: Methods and practice**. [S.l.]: Springer, 1993. p. 63–78. Citation on page 36.

KOWSARI, K.; MEIMANDI, K. J.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text classification algorithms: A survey. **Information**, MDPI, v. 10, n. 4, p. 150, 2019. Citations on pages 21 and 29.

LANCASTER, F. W.; GALLUP, E. **Information retrieval on-line**. [S.l.], 1973. Citation on page 28.

LEYDESDORFF, L. Alternatives to the journal impact factor: I3 and the top-10%(or top-25%?) of the most-highly cited papers. **Scientometrics**, Springer, v. 92, n. 2, p. 355–365, 2012. Citation on page 35.

LEYDESDORFF, L.; OPTHOF, T. Scopus's source normalized impact per paper (snip) versus a journal impact factor based on fractional counting of citations. **Journal of the American society for information science and technology**, Wiley Online Library, v. 61, n. 11, p. 2365–2369, 2010. Citation on page 35.

LI, J.; FAN, Q.; ZHANG, K. Keyword extraction based on tf/idf for chinese news document. **Wuhan University Journal of Natural Sciences**, Springer, v. 12, n. 5, p. 917–921, 2007. Citations on pages 22 and 30.

LI, J.; HUANG, G.; FAN, C.; SUN, Z.; ZHU, H. Key word extraction for short text via word2vec, doc2vec, and textrank. **Turkish Journal of Electrical Engineering and Computer Sciences**, v. 27, n. 3, p. 1794–1805, 2019. Citations on pages 21 and 22.

LÜ, J.; CHEN, G.; OGORZALEK, M. J.; TRAJKOVIĆ, L. Theory and applications of complex networks: Advances and challenges. In: IEEE. **2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)**. [S.l.], 2013. p. 2291–2294. Citation on page 30.

MIKOLOV, T.; CHEN, K.; CORRADO, G. S.; DEAN, J. Efficient estimation of word representations in vector space. **CoRR**, abs/1301.3781, 2013. Citations on pages 22 and 29.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119. Citations on pages 22 and 29.

MINGERS, J. Exploring the dynamics of journal citations: modelling with s-curves. **Journal of the Operational Research Society**, Taylor & Francis, v. 59, n. 8, p. 1013–1025, 2008. Citation on page 34.

MINGERS, J.; LEYDESDORFF, L. A review of theory and practice in scientometrics. **European journal of operational research**, Elsevier, v. 246, n. 1, p. 1–19, 2015. Citations on pages 23, 33, and 34.

MOHD, M.; JAN, R.; SHAH, M. Text document summarization using word embedding. **Expert Systems with Applications**, Elsevier, v. 143, p. 112958, 2020. Citation on page 22.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011. Citation on page 27.

NALIMOV, V. V.; MULCHENKO, Z. M. **Measurement of science. Study of the development of science as an information process**. [S.l.], 1971. Citation on page 34.

NEWMAN, M. **Networks: an introduction**. [S.l.]: Oxford university press, 2010. Citations on pages 31 and 32.

NEWMAN, M. E. Coauthorship networks and patterns of scientific collaboration. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 101, n. suppl 1, p. 5200–5205, 2004. Citation on page 36.

PRICE, D. J. D. S. Networks of scientific papers. **Science**, JSTOR, p. 510–515, 1965. Citation on page 35.

PRITCHARD, A. *et al.* Statistical bibliography or bibliometrics. **Journal of documentation**, New York, v. 25, n. 4, p. 348–349, 1969. Citation on page 34.

QUISPE, L. V.; TOHALINO, J. A.; AMANCIO, D. R. Using virtual edges to improve the discriminability of co-occurrence text networks. **Physica A: Statistical Mechanics and its Applications**, v. 562, p. 125344, 2021. ISSN 0378-4371. Available: <https://www.sciencedirect.com/science/article/pii/S037843712030707X>. Citations on pages 23, 29, 107, and 140.

RAAN, A. F. V. Sleeping beauties in science. **Scientometrics**, Springer, v. 59, n. 3, p. 467–472, 2004. Citation on page 34.

RADICCHI, F.; CASTELLANO, C.; CECCONI, F.; LORETO, V.; PARISI, D. Defining and identifying communities in networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 101, n. 9, p. 2658–2663, 2004. Citation on page 32.

REUTERS, T. A guide to evaluating research performance with citation data. **Accessible by: http://ip-science. thomsonreuters. com/m/pdfs/325133_thomson. pdf**, 2008. Citation on page 35.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based methods for multi-document summarization: exploring relationship maps, complex networks and discourse information. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2012. p. 260–271. Citation on page 21.

SALTON, G.; WONG, A.; YANG, C.-S. A vector space model for automatic indexing. **Communications of the ACM**, ACM, v. 18, n. 11, p. 613–620, 1975. Citations on pages 22 and 28.

SILUO, Y.; QINGLI, Y. Are scientometrics, informetrics, and bibliometrics different? In: **16th International Society of Scientometrics and Informetrics Conference ISSI**. [S.l.: s.n.], 2017. Citation on page 34.

SMALL, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. **Journal of the American Society for information Science**, Wiley Online Library, v. 24, n. 4, p. 265–269, 1973. Citation on page 35.

STEIN, R. A.; JAQUES, P. A.; VALIATI, J. F. An analysis of hierarchical text classification using word embeddings. **Information Sciences**, Elsevier, v. 471, p. 216–232, 2019. Citation on page 22.

STROGATZ, S. H. Exploring complex networks. **Nature**, Nature Publishing Group, v. 410, n. 6825, p. 268–276, 2001. Citation on page 31.

TIMONEN, M.; TOIVANEN, T.; TENG, Y.; CHEN, C.; HE, L. Informativeness-based keyword extraction from short documents. In: **KDIR**. [S.l.: s.n.], 2012. p. 411–421. Citation on page 21.

TOHALINO, J. V.; AMANCIO, D. R. Extractive multi-document summarization using multilayer networks. **Physica A: Statistical Mechanics and its Applications**, v. 503, p. 526–539, 2018. ISSN 0378-4371. Available: <https://www.sciencedirect.com/science/article/pii/S0378437118303212>. Citations on pages 23, 30, and 33.

VINKLER, P. Indicators are the essence of scientometrics and bibliometrics. **Scientometrics**, Springer, v. 85, n. 3, p. 861–866, 2010. Citation on page 23.

WAN, X.; YANG, J.; XIAO, J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In: **Proceedings of the 45th annual meeting of the association of computational linguistics**. [S.l.: s.n.], 2007. p. 552–559. Citation on page 30.

WANG, R.; LIU, W.; MCDONALD, C. Using word embeddings to enhance keyword identification for scientific publications. In: SPRINGER. **Australasian Database Conference**. [S.l.], 2015. p. 257–268. Citation on page 22.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citations on pages 22 and 31.

ZHOU, G.; HE, T.; ZHAO, J.; HU, P. Learning continuous word embedding with metadata for question retrieval in community question answering. In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. [S.l.: s.n.], 2015. p. 250–259. Citation on page 22.

ZUPIC, I.; ČATER, T. Bibliometric methods in management and organization. **Organizational Research Methods**, SAGE Publications Sage CA: Los Angeles, CA, v. 18, n. 3, p. 429–472, 2015. Citations on pages 23 and 35.

ICMC USP
SÃO CARLOS