

An Article-Oriented Framework for Automatic Semantic Analysis of COVID-19 Researches

Antonio Pedro[1(✉)], Antônio Pereira[1], Pablo Cecilio[1], Nayara Pena[1], Felipe Viegas[2], Elisa Tuler[1], Diego Dias[1], and Leonardo Rocha[1]

[1] Universidade Federal de São João del-Rei, São João del Rei, Brazil
{antoniopedro,antoniopereira,cecilio,nayara.p.pena}@aluno.ufsj.edu.br,
{etuler,diegodias,lcrocha}@ufsj.edu.br
[2] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
frviegas@dcc.ufmg.br

Abstract. In this work, we propose an article-oriented framework for automatically extract semantic topics from scientific articles related to COVID-19 researches. Our framework has four key building blocks (i) pre-processing, (ii) topic modeling, (iii) correlating topics, authors, and institutions, and (iv) a summarization interface. The first one corresponds to apply traditional textual pre-processing strategies in the texts extracted from articles and constructing their data representations. The topic modeling block aims at finding semantic topics from articles based on the most relevant words of each discovered topic. The third block correlates these discovered topics with the articles, authors (researchers), and institutions. The summarization interface provides an intuitive visualization for these results. Our evaluation shows that our framework can automatically extract relevant features from the articles, identifying the key topics covered by them, as well as the contribution of researchers, institutions, and countries to the topics. Our framework can help research institutions and companies to form multidisciplinary teams and funding agencies to identify more promising research approaches regarding COVID-19.

Keywords: COVID-19 · Machine leaning · Topic modeling

1 Introduction

Technological advances in recent decades have considerably increased the processing power and storage capacity of computers. Associated with these advances, the Internet has presented a significant expansion and popularization in recent years. Every day new applications are created, generating and using a higher amount of data (the expectation of reaching 44 ZiB within five years [22]) of the most diverse types. As stated in [2], "too much information is as bad as none", and one of the major challenges in this scenario is to provide tools that can perform intelligent and automatic analyzes on these data to find the information resources to meet the needs of users practically.

Supported by CAPES, CNPq, Finep, Fapesp and Fapemig.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 172–187, 2021.
https://doi.org/10.1007/978-3-030-86970-0_13

We can mention different scenarios where new techniques are being proposed to extract relevant information effectively. One example that illustrates these scenarios has to do with the so-called App Stores, such as Google Store and Apple Store, which provide over 2 million applications (each) so that users can search, buy, share and develop applications for mobile devices. As one could expect, both platforms produce useful data as textual comments and star ratings. Recent studies have shown that we can explore such information to leverage application development, with a clearer and more user-centered definition of requirements and improvements [19]. Currently, an important scenario that affects all segments of society is scientific production, which has been presenting an exponential expansion in terms of the number of articles, the number of researchers, and ongoing researches. The last detailed study on the *Google Scholar*[1] platform, in 2014, estimated an amount of 160 million documents in its database [37]. More recently, this scenario has been receiving even more prominence in consequence of the intense work of researchers to stop the advance of COVID-19, caused by a virus of the coronavirus family. Organizing semantically the information provided by the scientific articles published for this family of viruses can be extremely important to point out more promising research directions, to identify approaches that are still little explored, and to suggest possible collaborations of researchers with prominence in different areas, which corresponds to the application scenario of the present work.

In this work, we propose a general framework that allows filtering, summarizing, and analyzing articles written by researchers and published on different digital platforms. Our framework extracts relevant features from the articles automatically, identifying the key topics covered by them. The main goal of our proposal is to allow researchers and funding agencies to identify well-studied topics and little-explored approaches. Based on this information, take a more promising research direction to not only manage financial resources more efficiently but also to achieve even more innovative results. Based on the topics covered by articles, it is possible to establish the key topics studied by the researchers, authors of articles, and their institutions. This information can encourage researchers that are dealing with the same problem by different approaches to join efforts.

Our framework has four main building blocks (i) preprocessing, (ii) a topic modeling strategy, (iii) correlating topics, authors, and institutions, and (iv) a summarization interface. Our goal is to deal with the texts extracted from scientific articles, a proper data preprocessing allows the framework to achieve better effective results. From the raw textual content of articles, it is possible to apply different preprocessing strategies, such as lower-casing conversion, punctuation, and stopwords removal. The topic modeling strategy aims at finding semantic topics from articles. It is the ability to represent the set of articles as a bag-of-words matrix and decompose the represented dataset matrix properly (design matrix) into matrices that capture the latent relationships between terms and articles. It is also responsible for extracting the most relevant words of each discovered topic. Several strategies exist for this matter, such as non-negative matrix factorization

[1] https://scholar.google.com.

(NMF), latent dirichlet allocation (LDA), and its probabilistic counterpart, to name a few. Once the key topics of the articles have been identified, the third block consists of associating these topics with the articles and their authors and, its respective institutions. The summarization interface provides an intuitive visualization of the research topics and their associations with articles, authors, and institutions. Thus, the desiderata for our proposed framework is to offer, for each of its fundamental building blocks, methods that can perform their tasks losing no useful information, allowing fully automatic operation.

To evaluate our proposal, we instantiated the framework and provided an extensive analysis considering a dataset composed of articles published regarding COVID-19, SARS-CoV-2 and other coronaviruses, available at Kaggle [1]. Although this dataset contains complete works written in several languages and dated at least 20 years ago, we considered only the title and abstract of the 26 thousand articles published in 2020. We instantiated the first block of our framework applying some traditional preprocessing strategies: lower-casing conversion, punctuation, stopwords removal and named entity recognition (NER) approaches [13,25]. In the second block, we adopted the topic modeling strategy NMF [17] to infer different topics. For the third block, we proposed a strategy that consists to manipulate the matrices provide by NMF that correlates topics and articles, introducing in those the information of the articles' authors and their institutions and countries. Finally, we presented a visual interface proposal that summarizes all the gathered information, highlighting the main topics related to each article and author of our dataset. This interface is available at `labpi.ufsj.edu.br/covpapers/`[2]. Analyzing the results achieved by the instantiation of our framework, we were able to answer several questions such as: (i) *Which authors contribute the most to a certain topic?*; (ii) *Which lines of research are most explored by researchers at an institution?*; (iii) *Which countries lead research in a given area?*; and, (iv) *Which topics are most explored by a given author?*. Besides, with the definition of topics related to a set of articles, we discuss how it is possible to assist: (i) authors in the definition of descriptive tags about their research; (ii) companies to form multidisciplinary teams based on the researchers who most contribute to the desired area; (iii) government agencies, focused on Education and Research, to provide support in determining areas which are more, and less, studied, allowing a more equal redirection of financial resources.

2 Related Work

The proposal of strategies that aim to provide mechanisms for filtering, summarizing, and analyzing a specific area's scientific production is not new [7,8,29,36]. However, all these works concentrate on considering just metadata to provide this type of analysis, such as title, keywords, and tags, arguing that they are enough to provide an efficient data summarization. In this article, we raised some questions about these works: *To what extent are only metadata good descriptors*

[2] username: user, password: covid-19.

broadly? Are the keywords, alone, good descriptors about a large repository of articles? Can a researcher who wants to keep up to date on the scientific production in their fields be able to rely only on keywords to get this information? We believe that the answer to all these questions is "**No!**", motivating us to present the framework described and evaluated in this article, providing all these analyses considering the text that compose the articles.

Although potentially able to provide rich details information, using textual information is also more challenging and essentially requires preprocessing strategies application to minimize textual noise, selecting only data able to contribute to further steps of a machine learning processing, such as Topic Modeling (TM) [35]. In [5] the authors provided a complete analysis regarding different combinations of preprocessing strategies for text classification tasks. Regarding TM, in [30] the authors show that despite there is no correct, generic and unique sequence of preprocessing techniques, lower-casing conversion, punctuation and stopwords removal can achieve better effective results.

Another step that is considered in the textual preprocessing process is the data representation. It is necessary to build a $m \times n$ article-term matrix representation (or some latent term encoding) of the dataset. The representation widely adopted of this matrix in text mining exploits the term frequency-inverse document frequency (TF-IDF) paradigm (and its variants) [28]. The major problems with the TF-IDF representation have to do with its high dimensionality, sparseness and its lack of useful information such as context. Other alternative representations that aim at producing a more compact space in terms of latent dimensions (e.g., distributional word embeddings) exist, such as [24,27]. These strategies are based on co-occurrence statistics of textual datasets, representing words as vectors such that their similarities correlate with semantic relatedness. They usually use contextual information, such as terms adjacent to a target word. There is a consensus on literature regarding the contribution of these representations to TM tasks [34], especially the one presented in [32,33], called by the authors CluWords. CluWords correspond to clusters of semantically related word embeddings [24] built employing distance functions and filtering mechanisms. More than simple groups of (filtered) related words, CluWords are coupled with specific weighting schemes to capture their importance to a specific task.

A set of strategies commonly used to summarize, organize and analyze large volumes of textual data are those related to TM [12,14], which addresses the problem of discovering relationships between documents (D) and topics (Z), and relationships between the words (W) that compose the documents and topics. Each topic $z_i \in D$ comprises a probabilistic distribution between words that co-occur frequently, and the documents are represented by probabilistic distributions among the topics. They can be divided into two approaches: supervised and unsupervised. In supervised approaches, prior knowledge about the documents previously organized into semantically related groups is used so that the assignment of new documents to these groups can be performed [10]. With the ever-increasing and extremely dynamic data volume, the adoption of such strategies in real-world scenarios has become practically infeasible. Unsupervised

approaches have been increasingly applied in the literature [19]. The unsupervised techniques are classified into probabilistic and non-probabilistic models. One of the most cited probabilistic techniques in the literature is LDA, which is based on a three-level hierarchical Bayesian model [3]. To use it, it is necessary to adjust and refine some parameters, among them the number of topics.

From the applicability point of view, we found in the literature several works that make use of probabilistic TM strategies to summarize and understand a large volume of textual data. In [16], the authors proposed the LDA as a topic modeling technique to analyze large journalistic repositories and check for trends and patterns in news across time. Whereas, in [4], the authors analyzed the topics raised in the posts of super users of Internet Support Groups (ISGs), concluding that a large part of the posts is related to mental health and recovery. In [31], meanwhile, the authors aimed to understand through LDA the various meanings that the term "Big Data" takes on in the biomedical literature. Tourism is also an area of great possibilities for using topic modeling techniques, especially with LDA, as presented in the works of [15] and [20]. LDA summarizes potential attractions content for tourist recommendation systems, assisting the used operators in recommending and deciding tourists. In the second, LDA has applied to reviews made at Disney parks, returning a series of topics discussed and commented on by park visitors. This allows the parks' management to have a better understanding of visitor perception, driving marketing campaigns.

Despite the diverse use of LDA, the work of [26] showed that, although topic modeling by LDA is an excellent technique for qualitative text analysis, the content obtained in these topics may not be understandable to humans due to multiple parameter settings. In this sense, non-probabilistic TM algorithms stand out, especially those based on matrix decomposition, which has good interpretability and few parameter settings. Despite the popularity of (probabilistic) LDA, the TM study began with an article defining a non-probabilistic algorithm for the TM task. In [6], the authors proposed the latent semantic analysis (LSA), an algorithm that decomposes the A word document matrix into three other matrices employing the singular value decomposition (SVD) technique. Despite being pioneering, the resulting SVD matrices are still uninterpretable, mainly by returning negative values. Thus, another matrix decomposition technique has gained prominence. In [18], the authors proposed the Non-Negative Matrix factorization strategy (NMF), which is a non-probabilistic method to retrieve topics from documents. This technique, which needs only the number of topics to work, decomposes $m \times n$ article-term matrix into two other matrices, W (of documents by topics) and H (of topics by terms). From the point of view of the work proposed here, the decomposed matrices are very valuable for further analysis. In [19], for example, the authors used these decomposed matrices to perform a sentiment analysis on topics obtained from reviews performed on mobile apps in the Google PlayStore. In [11], on the other hand, the authors exploited the NMF to evaluate the political agenda and topics of speeches of European parliamentarians through time (1999–2014).

3 Proposed Work

It is a fact that the huge amount of data present on the Internet makes extracting information a challenging task. The difficulty increases when the data to be considered is textual. However, with the help of data mining techniques, especially Topic Modeling, it is possible to combine automation and formalization with the quality of consistent results in the understanding about researchers. Currently, it is common to use metadata for papers published in scientific databases, such as e *tags*, to summarize the lines of research in which a researcher works, or even, the affinity with certain topics. However, these metadata may not be as efficient in the task of summarizing and categorizing researchers, either because of their limitations or because they are filled in by the researcher himself. To ensure more assertive analyzes of these researchers, in this work, we propose the use of the textual information present in these articles (the text itself). To this end, we present a general framework that allows filtering, summarize and analyze articles written by researchers and published on different digital platforms. More specifically, our framework can automatically extract relevant features from the articles, identifying the main topics covered by them, as well as, the main topics related to the authors and research institutions. In Fig. 1 we present an overview of our framework and in the following sections, we detail how we instantiate each one of the blocks that composed it.

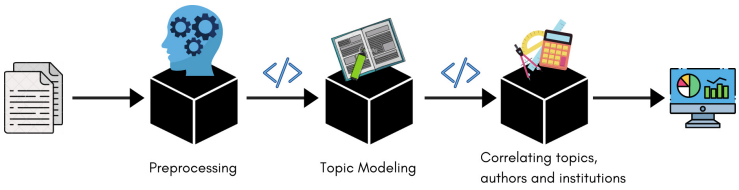

Fig. 1. Proposed framework overview

3.1 Preprocessing

Adopting preprocessing techniques is indispensable to give data consistency. We propose the instantiation of the preprocessing block through the use of four strategies: (i) uppercase to lowercase conversion, (ii) stopwords removal, (iii) tokenization, and (iv) removal of some entities identified in the named entity recognition (NER) process.

As shown by [30], converting all characters to lowercase ensures better results for text classification algorithms, corresponding to our first task. Then, we remove the stopwords from the text, which are words that contribute little to relevant textual analysis [9]. Since such a list of words varies for each context, our proposal consists of using the list of stopwords from the spaCy library[3], adding some other noisy words specific to the dataset context to be analyzed.

Based on a more concise text, it is possible to move on to the third strategy: tokenization. This is the process of transforming words in the text into single tokens

[3] https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py.

- the conversion from uppercase to lowercase makes it easier to group words that refer to the same token, such as "Experiment" and "experiment". These tokens can be generated in several ways, e.g., using whitespaces that appear between words. Although efficient, for large volumes of texts, where contractions and compound words are present in excess, more sophisticated tokenization strategies are important. In this sense, we made use of a package from the NLTK library for Python, which performs tokenization based on Twitter data[4], leaving words closer to their original meaning, i.e., it doesn't decompose compound words like "infectious disease" into two distinct tokens, but into a single token.

The next step in this preprocessing process is to employ a strategy widely used in the literature to categorize people, places, organizations, and other entities of interest in the text [13]. These are recognized in the process of NER. Our work aims to deal with academic databases, composed by published scientific articles, so it is common to use numerical expressions, periods, descriptions of people or physical spaces, such as "in the **second week** the results were better" or "the experiment was performed in the **Turing lab** all week". Therefore, we apply NER to identify textual entities which add little semantic information to algorithms for automatic text manipulation and remove them.

Finally, the last step of this block consists into represent the set of articles as a bag-of-words (BOW) matrix $m \times n$ (article-term matrix). Thus, we adopt the CluWords presented in [32]. Basically, the CluWords transform the traditional BOW representation of documents to include semantic information related to the terms present in the documents. The semantic context is obtained by means of a pre-trained word representation, such as Word2Vec [23], FastText [24]. Each document has a new representation where original words are replaced by the CluWords representation [33]. This transformation is composed of two phases: (1) for each term t of the dataset is computed its corresponding CluWord. A CluWord for a term t is a set of terms in the vocabulary which word vectors are most similar to term t; (2) a modified version of the TF-IDF weighting scheme is calculated for the CluWords on which it is possible to exploit these new terms as a richer representation of documents of the collection. The best results reported in the literature point out the use of CluWords as the state-of-the-art for text representations for TM tasks [32,33].

3.2 Topic Modeling

The TM block is based on NMF, a strategy that simultaneously performs dimension reduction and clustering, with successful application in topic modeling. NMF [17] produces a "part-based" decomposition of latent relationships of a non-negative design matrix $A \in \mathbb{R}^{n \times m}$, where n is the number of documents and m the number of terms. The goal is to find a k-dimensional approximation of A ($k \ll m$) in terms of non-negative factors $H \in \mathbb{R}^{n \times k}$ and $W \in \mathbb{R}^{k \times m}$. H encodes the relationship between documents and topics while W encodes the relationship between terms and topics. The key idea behind NMF is to approximate the

[4] https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.casual.

column vectors of A by non-negative linear combinations of non-negative basis vectors (columns of H) and the coefficients given by the columns of W. The Fig. 2 illlustrate this process.

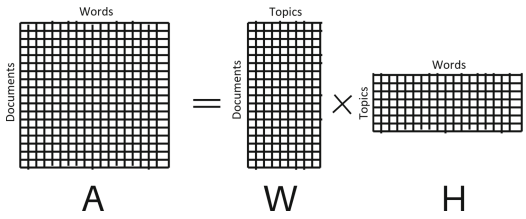

Fig. 2. Decomposition process performs by NMF

3.3 Correlating Topics, Authors and Institutions

Finding out which actors (authors and institutions) contribute more to each topic is a key point to enrich our analysis. Let's consider that the two previous steps of our framework have been applied to a database containing scientific articles related to the Computer Science area. Thus, at the end of the first two steps, we have the matrices H and W, that relate topics to articles and topics by words, respectively. Following the example, consider that the $i-th$ article in the H matrix deals mostly with the topic "Machine Learning", while the $j-th$ article deals with the topic "Sensor Networks". A pertinent question that can be asked is: *Who are the authors of the article that deal mostly with the topic "Machine Learning"?* It is evident that, for this example, each article has one or more authors, so it is possible, through the relations between articles and found topics, to highlight which authors most contribute to the topics obtained by TM. Similarly, as the authors of an article belong to research institutions, it is also possible to highlight the institutions by topics, also considering the relationship between articles and topics. Therefore, for the third block, we propose a strategy that consists to manipulate the matrices provided by NMF that correlates topics and articles, introducing the information of the articles' authors and their institutions. We will demonstrate below how this manipulation happens through an example.

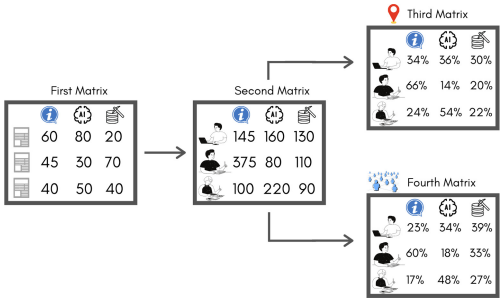

Fig. 3. Calculating author contributions to topics

Let us follow the same example above, considering that we generated the matrices H and W for three topics. First, you can identify what each topic is about by analyzing the H matrix and finding which words are most strongly associated with each topic. Let's assume the example where the first topic is mostly associated with "Information Retrieval", the second topic is related to "Artificial Intelligence", and the third one is related to "Data Mining". The next step is to analyze the W matrix that relates documents and topics. For this, let's take as an example the first matrix in Fig. 3, containing three articles, where each position presents the "relevance" of the topic for the document. From this matrix, we can group and sum the topic values achieved for articles that belong to the same author, leading us to the second matrix in Fig. 3. Assuming that the three articles in the first matrix belong to the first author in the second matrix, we can infer the "relevance" of this author for each topic: 145 (60+45+40) for the "Information Retrieval" topic, 160 (80+30+50) for the "Artificial Intelligence" topic, and 130 (20+70+40) for the "Data Mining" topic. Considering the complete W matrix, this same process can be applied to all collection authors.

From these data, we can perform two different analyses: (i) calculate for each author the distribution among the topics by which their works are related; or (ii) evaluate among all authors those who have their works most correlated to each topic. The first analysis is local information, so it is based only on what has been written by the same author. Thus, by performing a normalization on the rows representing each author in the second matrix of Fig. 3, it is possible to know how much the research of each author is related to each topic. We illustrate the result in the third matrix of Fig. 3. Analyzing the first row that represents the first author, we can observe that his articles are 34% related to the topic "Information Retrieval", 36% to the topic "Artificial Intelligence", and 30% to the topic "Data Mining". We can say that this author's articles deal more with the "Artificial Intelligence" topic. We based the second analysis on global information and consider everything that has been written for the same topic to distinguish those authors who have their articles more aligned with the topics got. In the fourth matrix of Fig. 3 we normalized each of the columns, which represents the topics got from the second matrix. Thus, following our example, we can observe that the first author, despite having his articles dealing mostly with "Artificial Intelligence" is not the one who has the greatest alignment with this topic, the third author has the greatest alignment. Even though the first author has their articles less related to the topic "Data Mining", among the authors evaluated, they are presenting articles more aligned with this topic. Considering that each author registers in their articles the research institution to which they are affiliated and that each institution is in a country, we can do an analogous analysis to the one presented above for institutions and countries.

Following the strategy presented by us, it is possible to perform several analyses, such as identifying the topics that have been most explored by researchers or by a specific researcher, evaluate how institutions and/or countries have been exploring each of the topics identified, and many other analyses. These analyses can help research institutions and companies to form multidisciplinary teams, and funding agencies to identify areas that need more resources.

3.4 Summarization Interface

We have developed a visualization interface that intuitively summarizes the topics found and their associations with researchers and institutions. This way, all the analyses proposed in Sect. 3.2 can be performed quickly and efficiently. We present the data related to topics and their correlations with articles, researchers, and institutions through a large set of visual metaphors in a web application. Charts, heatmaps, tables, and interactive search methods are available, allowing the same result to be viewed from different perspectives. This Web App is available at our website[5]. In Fig. 4 we present our application with one of the visual metaphors provided, the one detailing the topics related to a researcher. Other visual metaphors will be presented in the Sect. 4.1 when we will also discuss their potential to perform the possible analyses discussed in the previous section.

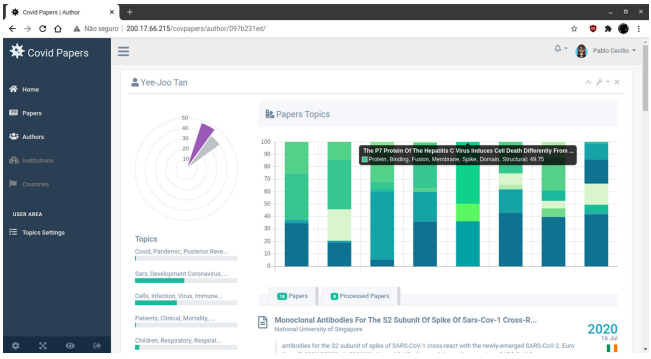

Fig. 4. Visualization of author in COVID's base

4 Experiments and Results

4.1 Dataset

To validate the proposed framework, we considered a dataset of published articles related to COVID-19, SARS-CoV-2 and, other viruses of the coronaviruses group. It is available at Kaggle [1], by the Allen Institute for AI in partnership with Chan Zuckerberg Initiative, Georgetown University Center for Security and Emerging Technologies, Microsoft Research, IBM, US National Library of Medicine, and in coordination with the White House Office of Science and Technology Policy. The dataset contains articles published in several languages, such as, Portuguese, English, French, German and, others. However, in our experimental evaluation, we only considered articles written in English. The dataset has several articles distributed in dates that mark various moments in history

[5] `labpi.ufsj.edu.br/covpapers/`
username: user
password: covid-19.

(avian flu in 2001, swine flu in 2009, and COVID-19 in 2020). Thus, we separated only articles published in the year 2020, resulting in more than 16,000 documents, for which we extracted the title and abstracts for a specific analysis on COVID-19.

4.2 Preprocessing and Topic Modeling

We apply the preprocessing steps highlighted in Sect. 3.1 (including the text representation using CluWords, the state-of-the-art in semantic representation [32,33]). The pre-processed dataset is then forwarded to the step our tool which performs the modeling of topics (as detailed in Sect. 3.2, we use NMF [17]). In the summarization interface, it is possible to define how many topics it should generate (**10, 20 or 30 topics**) and how many words should characterize each of the topics (**10, 20 or 30 words**). In our experimental analysis, we exploit 10 topics with 10 words, described in Table 1.

Table 1. Topic Modeling extraction using NMF, considering 10 topics and 10 words

Topic	Word 1	Word 2	Word 3	Word 4	Word 5
1	Infection	Viruses	Pathogen	Diseases	Parasite
2	Antigen	Immunogen	Crossreactivity	Hybridomas	Monoclonal
3	Economic	Social	Political	Governmental	Education
4	Cells	Neuroprogenitor	Fibroblast	Keratinocytes	Epithelium
5	Patients	Inpatient	Outpatient	Hospital	Clinic
6	Antimalarial	Drugs	Dosage	Antidepressants	Statins
7	Coatings	Alloy	Metals	Oxides	Conductive
8	Proteins	Cysteine	Enzyme	Nucleotides	Ubiquitins
9	Approximation	Parameters	Simulations	Algorithm	Equations
10	Venous	Catheterization	Carotid	Artery	Endovascular
Topic	Word 6	Word 7	Word 8	Word 9	Word 10
1	Contagions	Bacterial	Rhinovirus	Transmissible	Parvovirus
2	Epitope	Toxoid	Autoantibodies	Polyclonal	Sera
3	Cultural	Initiatives	Financial	Institutions	Personal
4	Macrophages	Tissues	Apoptosis	Intercellular	Osteoblasts
5	Doctors	Diagnoses	Nurses	Physicians	Oncologists
6	Immunosuppressant	Aspirin	Methadone	Quinine	Psychoactive
7	Superalloy	Nanocomposite	Thermoplastics	Electrodeposition	Polymer
8	Ribosome	Genes	Residues	Aminoacids	Peptide
9	Probabilistic	Submodel	Methodology	Computation	Model
10	Embolectomy	Thoracotomy	Aneurysm	Sternotomy	Endarterectomy

4.3 Analysing the Correlation Between Topics, Authors and Institutions

A researcher could identify how the proposed framework defines their research areas or which topics their research comes closest to. To illustrate this analysis, we considered the Dutch researcher Marion Koopmans, a virologist who specialized in molecular epidemiology at the Erasmus University Medical Center in Rotterdam, Netherlands. In Fig. 5, we present the topic distribution for her four articles available in the dataset (right graph) and an overall analysis of the researcher (left graph). We can see that her research aimed at 34% on Topic 1 (dealing with virology); 22% on Topic 2 (immune system); and 14% on Topic 3 (socioeconomic effects of the pandemic). This association, automatically identified by our framework, can be corroborated by the author's description in her institutional page[6]. The same association can also be observed in a recent report in the scientific journal Nature [21], which highlights a list of researchers who have been leading the research on the origin of COVID-19.

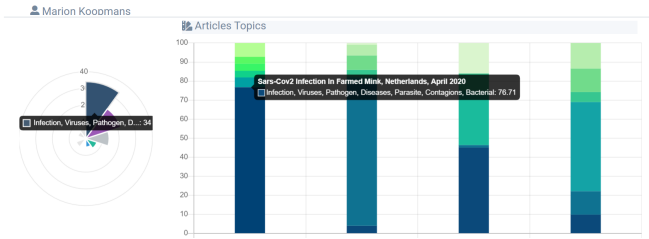

Fig. 5. Relationship between topics and articles by Marion Koopmans

The second analysis evaluates, among all the authors, those whose work is the most correlated to a topic. Returning to our previous example, despite Marion Koopmans' research having Topic 1 as her primary focus, the framework shows another researcher standing out in this area. He is the Japanese researcher Hiroshi Nishiura. In Fig. 6, we illustrate the topic distribution of his article entitled "*Incubation Period And Other Epidemiological Characteristics Of 2019 Novel Coronavirus Infections With Right Truncation: A Statistical Analysis Of Publicly Available Case Data*", which has 52.89% of its content linked to Topic 1. Just one article does not define the association of a researcher for a specific research topic, but several published articles mostly related to one research topic lead the framework to infer the researcher's contribution to this topic - Hiroshi Nishiura. He emerges as the researcher who has the highest correlation with Topic 1 (Fig. 7).

[6] https://www.erasmusmc.nl/en/research/researchers/koopmans-marion.

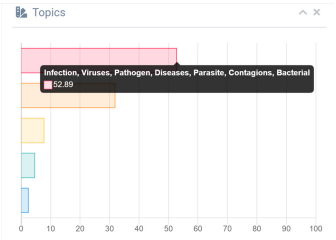

Fig. 6. Topics Related to a article

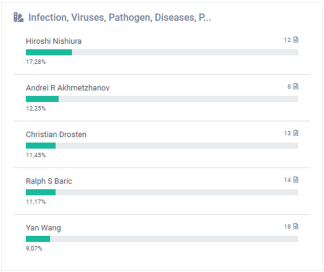

Fig. 7. Authors who contribute the most to a topic

Besides the analyses performed on researchers and articles, research institutions committee might be interested in finding out what their institutions study and how their studies affect the global scenario facing COVID-19. For example, the city of Wuhan, the epicenter of the new coronavirus pandemic, has some educational institutions that have contributed essentially so that the rest of the world could also understand the new virus. Our framework allowed not only to confirm this simple statement but also to highlight that the studies performed at the Huazhong University of Science and Technology, in Wuhan, China, placed them in the research leadership for several topics (5 out of 10 considered topics), as shown in Fig. 8.

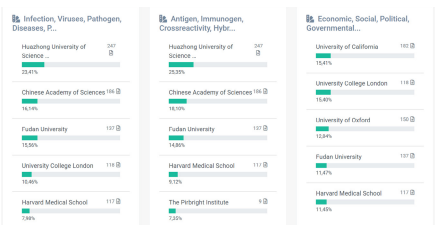

Fig. 8. Correlation per institution

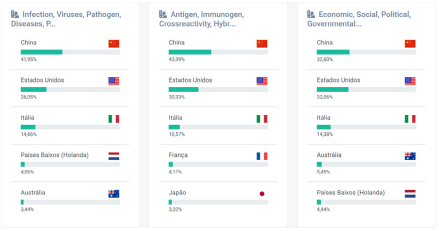

Fig. 9. Correlation per country

Finally, governments and research foundations interested in visualizing how their countries are facing the various topics related to the new coronavirus can use our framework to find out which areas of study have been prioritized and if the studies have had a significant impact when compared to other countries. The countries that have invested the most in fighting the pandemic or have suffered from it are the ones that lead the listed topics. Following the basic principle of science, they report, study, and propose solutions, and highlight potential problems that can surge elsewhere. Thus, the participation of China, the United States, Italy, India, the Netherlands, France, and Spain in the topics got by our method stands out as shown in Fig. 9.

5 Conclusion and Future Works

In this work, we proposed a novel framework that analyzing scientific articles published on different digital platforms and offers to researchers and funding agencies a more detailed and informative view of the well-studied topics and little-explored approaches considered on COVID-19 researches, pointing out more promising research direction. This framework has four key building blocks that automatically extract semantic topics from the papers, identifying the primary subjects covered by them; correlate these topics with the papers, authors (researchers), and institutions, and summarize all information in an intuitive interface. Analyses on a dataset composed of papers published regarding COVID-19 and other corona viruses [1] demonstrated that we can: (1) identify correctly the topics referred to an article, (2) identify the contribution of researchers to different topics, and (3) evaluate how institutions and/or countries have been contributing to each topic. In sum, the main implication is that this new framework can help research institutions and companies to form multidisciplinary teams and funding agencies to identify less and more explored areas to direct resources to more promising research approaches regarding COVID-19. As future work, we plan to instantiate different configurations regarding topic modeling and data representation. We also intend to propose and evaluate dynamic and interactive methods in the visualization interface and apply them to other databases.

Acknowledgements. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and National Council for Scientific and Technological Development – CNPq.

References

1. Ai, A.I.F.: Covid-19 open research dataset challenge (cord-19) (2020)
2. Auden, W.H.: The Complete Works of WH Auden: Prose, vol. 2. Princeton University Press, Princeton, April 2002
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(null), 993–1022 (2003)
4. Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., Griffiths, K.M.: What's all the talk about? topic modelling in a mental health internet support group. BMC Psychiatry **16**(1), 367 (2016)
5. Cunha, W., et al.: Extended pre-processing pipeline for text classification: on the role of meta-feature representations, sparsification and selective sampling. Inform. Process. Manag. **57**(4), 102263 (2020)
6. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**, 391–407 (1990)
7. Dutta, P.S., Tahbilder, H., et al.: Prediction of rainfall using data mining technique over Assam. Indian J. Comput. Sci. Eng. (IJCSE) **5**, 85–90 (2014)
8. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Metadata analysis of scholarly events of computer science, physics, engineering, and mathematics. In: TPDL (2018)

9. Gerlach, M., Shi, H., Amaral, L.A.N.: A universal information theoretic approach to the identification of stopwords. Nat. Mach. Intell. **1**(12), 606–612 (2019)
10. Ghosh, S., et al.: Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. Sci. Rep. **7**, 40841 (2017)
11. Greene, D., Cross, J.P.: Exploring the political agenda of the european parliament using a dynamic topic modeling approach (2016)
12. Griffiths, T.L., Steyvers, M.: A probabilistic approach to semantic representation. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 24 (2002)
13. Gudivada, V.N., Arbabifard, K.: Chapter 3 - open-source libraries, application frameworks, and workflow systems for nlp. In: Gudivada, V.N., Rao, C. (eds) Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications, volume 38 of Handbook of Statistics, pp. 31–50. Elsevier (2018)
14. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 289–296. Morgan Kaufmann Publishers Inc. (1999)
15. Huang, C., Wang, Q., Yang, D., Xu, F.: Topic mining of tourist attractions based on a seasonal context aware lda model. Intell. Data Anal. **22**, 383–405 (2018)
16. Jacobi, C., van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit. Journalism **4**(1), 89–106 (2016)
17. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
18. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00, pp. 535–541. MIT Press, Cambridge, MA, USA (2000)
19. Luiz, W., et al.: A feature-oriented sentiment rating for mobile app reviews. In: Proceedings of the 2018 World Wide Web Conference, WWW 2018, pp. 1909–1918. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018)
20. Luo, J.M., Vu, H.Q., Li, G., Law, R.: Topic modelling for theme park online reviews: analysis of disneyland. J. Travel Tourism Market. **37**(2), 272–285 (2020)
21. Mallapaty, S.: Meet the scientists investigating the origins of the covid pandemic (2020)
22. Marr, B.: 20 fatos sobre a internet que você (provavelmente) não sabe (2015)
23. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. CoRR, abs/1712.09405 (2017)
24. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: LREC 2018 (2018)
25. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticæ Investigationes **30**(1), 3–26 (2007)
26. Nikolenko, S.I., Koltcov, S., Koltsova, O.: Topic modelling for qualitative studies. J. Inf. Sci. **43**(1), 88–102 (2017)
27. Pennington, J., Socher, R., Manning, V.: Glove: global vectors for word representation. In: EMNLP (2014)
28. Qaiser, S., Ali, R.: Text mining: use of tf-idf to examine the relevance of words to documents. Int. J. Comput. Appl. **181**(1), 25–29 (2018)
29. Sundar, N.A., Latha, P.P., Chandra, M.R.: Performance analysis of classification data mining techniques over heart disease database, vol. 2, pp. 470–478. Citeseer (2012)
30. Uysal, A.K., Gunal, S.: The impact of preprocessing on text classification. Inform. Process. Manag. **50**(1), 04–112 (2014)

31. van Altena, A.J., Moerland, P., Zwinderman, A., Olabarriaga, S.: Understanding big data themes from scientific biomedical literature through topic modeling. J. Big Data **3**, 1–21 (2016)
32. Viegas, F., et al.: Cluwords: exploiting semantic word clustering representation for enhanced topic modeling, pp. 753–761 (2019)
33. Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., Goncalves, M.: CluHTM - semantic hierarchical topic modeling based on CluWords. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8138–8150. Association for Computational Linguistics, Online, July 2020
34. Viegas, F., et al.: Semantically-enhanced topic modeling. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, pp. 893–902. Association for Computing Machinery, New York, NY, USA (2018)
35. Vijayarani, S., Ilamathi, M.J., Nithya, M.: Preprocessing techniques for text mining-an overview. Int. J. Comput. Sci. Commun. Netw. **5**, 7–16 (2015)
36. Yethiraj, N.G.: Applying data mining techniques in the field of agriculture and allied sciences. Int. J. Bus. **001**, 40–42 (2012)
37. You, J.: Just how big is google scholar? ummm..., (2014)