electronics

MDPI

Article

Content Analysis Using Specific Natural Language Processing Methods for Big Data

Mironela Pirnau [1], Mihai Alexandru Botezatu [2], Iustin Priescu [1], Alexandra Hosszu [3], Alexandru Tabusca [2], Cristina Coculescu [2] and Ionica Oncioiu [4,5,*]

[1] Department of Informatics, Faculty of Informatics, Titu Maiorescu University, 040051 Bucharest, Romania; mironela.pirnau@prof.utm.ro (M.P.); iustin.priescu@prof.utm.ro (I.P.)
[2] Department of Informatics, Statistics and Mathematics, School of Computer Science for Business Management, Romanian American University, 012101 Bucharest, Romania; mihai.botezatu@rau.ro (M.A.B.); alex.tabusca@rau.ro (A.T.); cristina.coculescu@rau.ro (C.C.)
[3] Department of Sociology, Faculty of Sociology and Social Work, University of Bucharest, 030018 Bucharest, Romania; alexandra.hosszu@sas.unibuc.ro
[4] Faculty of Economic Sciences, Titu Maiorescu University, 040051 Bucharest, Romania
[5] Faculty of Economics and Business Administration, "Eugeniu Carada" Doctoral School of Economic Sciences, University of Craiova, 200585 Craiova, Romania
* Correspondence: ionica.oncioiu@prof.utm.ro

Abstract: Researchers from different fields have studied the effects of the COVID-19 pandemic and published their results in peer-reviewed journals indexed in international databases such as Web of Science (WoS), Scopus, PubMed. Focusing on efficient methods for navigating the extensive literature on COVID-19 pandemic research, our study conducts a content analysis of the top 1000 cited papers in WoS that delve into the subject by using elements of natural language processing (NLP). Knowing that in WoS, a scientific paper is described by the group Paper = {Abstract, Keyword, Title}; we obtained via NLP methods the word dictionaries with their frequencies of use and the word cloud for the 100 most used words, and we investigated if there is a degree of similarity between the titles of the papers and their abstracts, respectively. Using the Python packages NLTK, TextBlob, VADER, we computed sentiment scores for paper titles and abstracts, analyzed the results, and then, using Azure Machine Learning-Sentiment analysis, extended the range of comparison of sentiment scores. Our proposed analysis method can be applied to any research topic or theme from papers, articles, or projects in various fields of specialization to create a minimal dictionary of terms based on frequency of use, with visual representation by word cloud. Complementing the content analysis in our research with sentiment and similarity analysis highlights the different or similar treatment of the topics addressed in the research, as well as the opinions and feelings conveyed by the authors in relation to the researched issue.

Keywords: natural language processing; big data; sentiment analysis; similarity analysis; word dictionary; word cloud

Citation: Pirnau, M.; Botezatu, M.A.; Priescu, I.; Hosszu, A.; Tabusca, A.; Coculescu, C.; Oncioiu, I. Content Analysis Using Specific Natural Language Processing Methods for Big Data. *Electronics* **2024**, *13*, 584. https://doi.org/10.3390/electronics13030584

Academic Editors: Ioannis Yiannis Kompatsiaris, Stefanos Vrochidis, Giuseppe Amato and Sotiris Diplaris

Received: 22 December 2023
Revised: 26 January 2024
Accepted: 29 January 2024
Published: 31 January 2024

1. Introduction

The COVID-19 pandemic status, officially declared by the World Health Organization (WHO) on 11 March 2020 [1], has caused a global crisis deeper than all previous epidemics or pandemics, with a much stronger impact on the life of the population worldwide in all its aspects: micro- and macroeconomic, health, social, cultural, educational, emotional.

The World Health Organization [2] has recommended, through appeals to scientists, researchers, and specialists in various fields, to conduct research during the pandemic and to publish the results immediately in order to help patients and limit the catastrophic global consequences of the spread of the virus. Moreover, this research, conducted individually or in collaboration, was considered an ethical obligation of specialists during the pandemic and is practically a public health measure.

Bibliometric and scientometric analyses were carried out at various short intervals (3–6 months), which tried to capture and assess, even during the pandemic, the main issues that were the focus of scientific research activities of specialists in different fields of activity, which were published in various international journals or databases. These analyses, using specific research tools and methods, had the immediate aim of transmitting, as quickly as possible, the results of the latest findings in the field of prevention, treatment, and convalescence of those infected with COVID-19, disseminating them at all levels of age, occupation, environment, to all decision-makers involved in the prevention and treatment of the population. Therefore, for that period, all these research and publication efforts were extremely beneficial.

Research conducted in medicine, health, and virusology were presented [3–6] in the fields of social sciences [7,8], arts and humanities [9], accounting and COVID-19 published during the pandemic [10,11], business and economics [12–14], business and management [15], the online shopping behavior of the population [16–18], policies to prevent and stop the pandemic by applying innovative scientific approaches [19]. The use of textiles for protection in pandemics has been another topic of interest [20]; the study [21] can provide valuable information for the treatment and eradication of COVID-19, and other papers present future research perspectives [5,22] and education perspectives for public health [23], training local medical specialists, and bringing clinical services online [24].

The OA (open access) publication of studies during this period [25] has been extremely beneficial, ensuring free access to all this information, rapid and wide dissemination of research results in order to facilitate knowledge transfer, good collaboration between specialists, and thus an increase in their chances of blocking the devastating effects of the spread and resistance of the COVID-19 virus.

Under these circumstances, we consider that a content analysis of scientific publications on the COVID-19 pandemic and its effects is very useful and beneficial now after more than one year after the end of the pandemic, as it can provide comprehensive information and more accurate and rigorous knowledge in various topics related to the pandemic, including for the adoption of integrated strategies and initiatives on economic development, education, labor, recreation, public health treatments and measures, necessary equipment for hospitals and ambulance services, intensive care.

Nowadays, the interest of researchers and specialists for a specific event or crisis situation is materialized by a multitude of scientific papers published in numerous international journals or databases. Many of these are indexed by Clarivate Analytics Web of Science (WoS) or SCOPUS, have high visibility, ensure online access for those interested, and have the knowledge, best practices, and technology transfer that can be achieved at a rapid pace and with an extremely high volume of information. The exploitation of large volumes of digital data from various domains through natural language processing methods contributes to finding relevant information aiming to know, in a timely manner, the society's diagnosis, its reaction, the responses of the business environment, health, culture, education, and all other activities related to the topic, the event, or the crisis.

Natural language processing (NLP) includes a content analysis of texts and is complemented by automatic evaluation methods of opinions and feelings [26], of states experienced and transmitted by authors in relation to the event/subject/crisis analyzed. NLP is a subfield of artificial intelligence and linguistics dedicated to making computers understand texts and words written in various languages to manage human–computer language interaction.

The large amounts of data available to us in the digital age make access to all of them beyond our ability to understand them, and it is necessary to communicate with computers in natural language. In this context, the components of NLU (Natural Language Understanding or Linguistics and Natural Language Generation), a series of computational tools and methods developed, and neural networks introduced in the field of NLP, are extremely useful for users in various fields, such as spam removal, information extraction, summarization, automatic translation, answering questions.

This study focuses on publications related to the topic "COVID-19" during the pandemic period until now, aiming to identify and extract the most used words by using code snippets written by authors in the open source language Python 3.12 and to provide the user with data collections and tools for further analysis. A content analysis of the selected papers was performed, followed by analyses of the approach and general opinion conveyed by the authors in order to better understand the priorities and concerns of the research environment for different domains during the COVID-19 pandemic. The foundation of our research is based on the analysis of large-sized and well-structured WoS content, including publications of high visibility and appreciated especially by the academic and research environments, specialists, and recognized scientists. In this context, our research objectives are to identify the top 1000 most cited papers (publications) in the WoS database on the subject "COVID-19" (basically, these are the most relevant papers of high interest for the topic.); to draw up three dictionaries of words ordered by their frequency of use for the three components of a scientific paper: "Title", "Abstract", "Keyword"; to represent each dictionary suggestively with a word cloud (based on a set number of words); to carry out a sentiment analysis in order to understand the general opinion conveyed by the authors on the subject; to draw up a similarity analysis of the selected papers to see if the approaches to the topic in the research were unique or, on the contrary, similar. In the present study, using Python packages such as NLTK, TextBlob, and VADER, we performed sentiment analysis on the titles and abstracts of the research papers.

After analyzing the results, we extended the comparison range for sentiment scores using Azure Machine Learning-Sentiment analysis. Our innovative approach to analysis can be applied to any research topic or topic in different areas of specialization. This allowed us to create a concise dictionary of frequently used terms, visually represented with a word cloud. What sets our study apart is the integration of sentiment analysis, which enriches content analysis by highlighting the varied treatments of the research subjects. In addition, it provides insights into the opinions and emotions conveyed by the authors in relation to the topic.

This paper is structured as follows: Section 2 discusses a summary and comparative analysis of scientific publications reviewed during pandemics. Section 3 presents aspects related to the literature review and research objectives. Section 4 outlines the research method and covers findings from the empirical research. Finally, this paper ends with a summary of the research conclusions and presents the limitations of this study.

2. A Summary and Comparative Analysis of Scientific Publications Reviewed during Pandemics

Before COVID-19, mankind has faced many pandemics such as bubonic plague, cholera, Spanish flu, HIV/AIDS. Being concerned about the volume of publications in WoS on these pandemics, we made a brief analysis of the interest and papers published by researchers and specialists in the field, taking into consideration the number of scientific papers published on these subjects. Thus, until 15 August 2023, we queried the WoS database [27] and extracted the scientific publications that have the name of the pandemic in the content of the published papers, forming a database for each type of pandemic.

2.1. "Bubonic Plague" Pandemic

The "bubonic plague" pandemic was recorded during the years 1347–1351, then 1665–1666, 1894–1898, and reappeared in a weaker form in 1907 and after 1970. A total of 926 papers are published in WoS [27] for the topic "bubonic plague", out of which only 3 scientific papers were published in 1976 and 164 scientific papers were published between 2020–2023. Figure 1 shows the distribution of articles published on the "bubonic plague" pandemic by year.

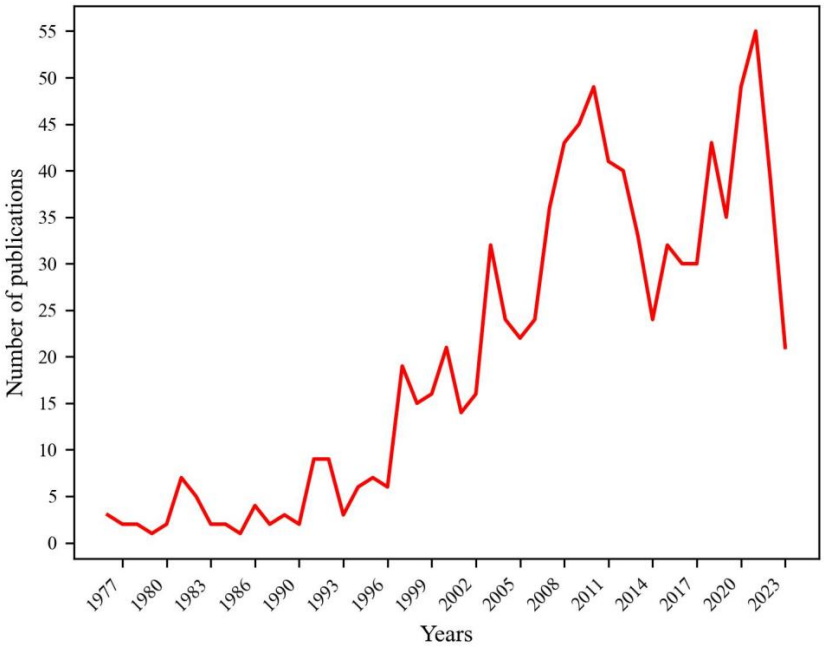

Figure 1. Number of publications in WoS related to the "bubonic plague" pandemic.

It was found that of the total scientific publications related to the "bubonic plague" pandemic existing in WoS, between 2020–2023, there are about 18%, which shows that during the COVID-19 pandemic period, research on the bubonic plague pandemic has been highlighted.

2.2. "Cholera" Pandemic

The "Cholera" pandemic was registered in the 19th century, 1892. A total of 25,010 papers are published in WoS for the topic "cholera", most of them after 1990. The distribution of articles published on the "cholera" pandemic by year is shown in Figure 2.

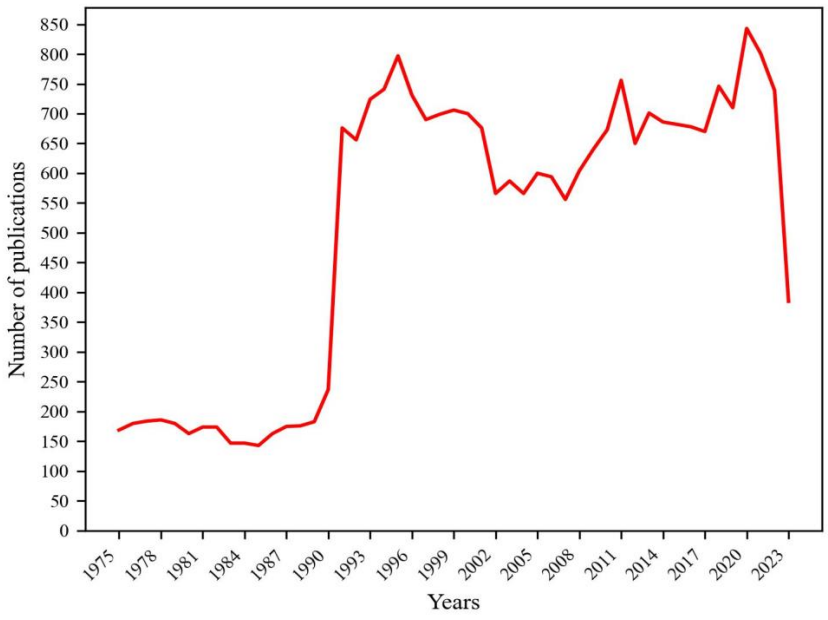

Figure 2. Number of publications in WoS related to the "cholera" pandemic.

It can be seen that since 2020, the number of publications related to the "cholera" pandemic has been decreasing, given the increased interest in the COVID-19 pandemic.

2.3. "Spanish Flu" Pandemic

The "Spanish flu" Pandemic was recorded between 1918 and 1920. A total of 765 publications are listed in WoS on the topic "Spanish flu", most of them published after 2004. The distribution of articles published on the "Spanish flu" pandemic by year is shown in Figure 3. Over the period under review, the empirical points curve has a shape that can be approximated by a straight line; thus, the model that can be used to approximate the evolution of the number of publications in WoS related to the "Spanish flu" pandemic is of the form

$$y_t = f(t) + u_t \quad (1)$$

where:

y_t—values of the dependent variable.

$f(t)$ = the trend component that can be described using a linear function:

$$f(t) = a + b \cdot t \quad (2)$$

u_t—the residual variable, representing the influences of the other factors of variable y not specified in the model considered as random factors with insignificant influences on the dependent variable y.

Estimating the parameters of the linear regression model led to function 3:

$$f(t) = 1.5842 \cdot t - 3153.7 \quad (3)$$

The value of the determination coefficient R^2 (in percentages) expresses how much of the variation in the dependent variable (y_t) can be explained by its linear relationship with the independent variable (t). $R^2 = 0.35$. The 35% of the variation in the total number of papers published in WoS on the "Spanish flu" pandemic can be explained by its linear relationship with time.

The regression coefficient $\hat{b} = 1.5842$ (slope of the line) indicates that for each increase in the independent variable (t = time) by one unit, the number of papers published increases on average by 1.5842 units. It can be seen that since the onset of the COVID-19 pandemic, in 2020, a total of 138 papers on the "Spanish flu" pandemic appeared in WoS; in 2021, the number increased by 19 papers, and from 2022 onwards, it decreased considerably.

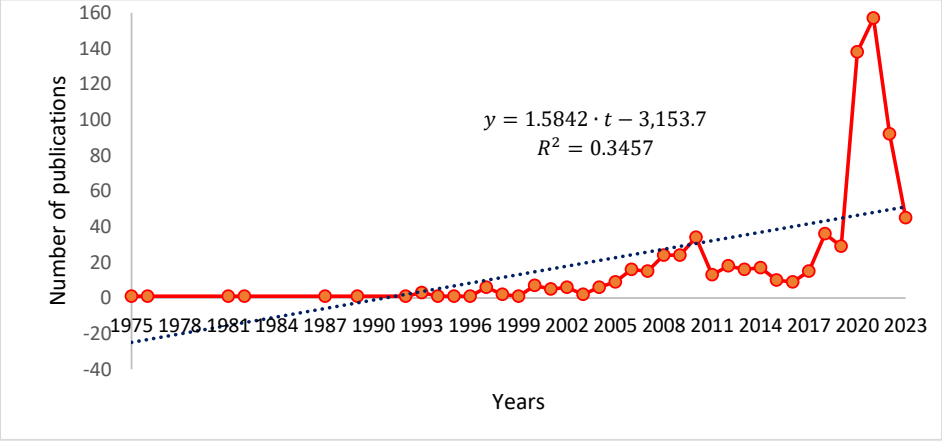

Figure 3. Number of publications in WoS related to the "Spanish flu" pandemic.

In Figure 3, the evolution of the number of publications related to the "Spanish flu" pandemic in WoS is depicted in red, and the trend of this evolution is represented by the dotted line.

2.4. Pandemia HIV/SIDA

For the topic "HIV/AIDS", there are a total of 426,261 published papers on WoS. The number of scientific papers published after 1987 has increased to several thousand, and since 2004, the number of papers published has been increasing to more than ten thousand. The distribution of the publication of articles per year on the HIV/AIDS pandemic is shown in Figure 4.

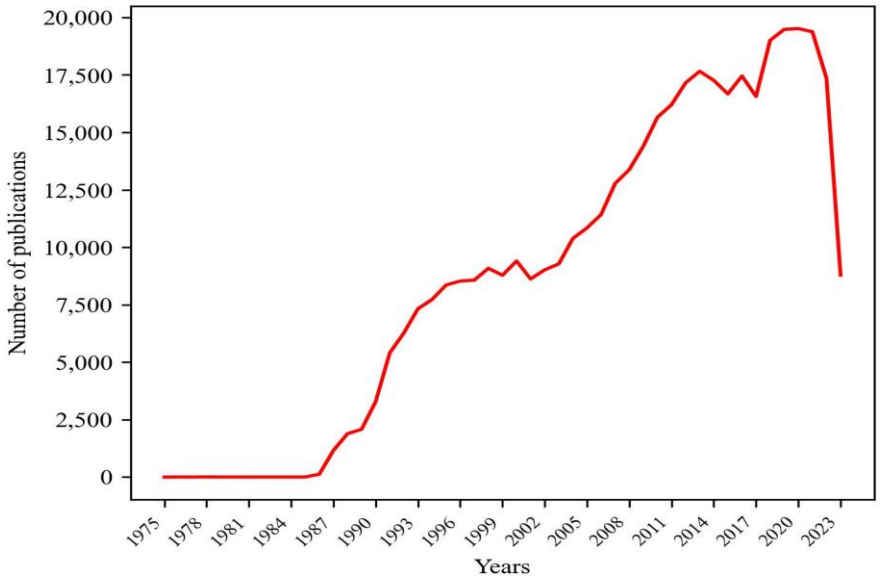

Figure 4. Number of publications in WoS related to the "HIV AIDS" pandemic.

2.5. COVID-19 Pandemic

The number of publications in WoS [27] for COVID-19 between 2020–2023 is 443,878. A graphical representation of this number, by year of publication, is shown in Figure 5.

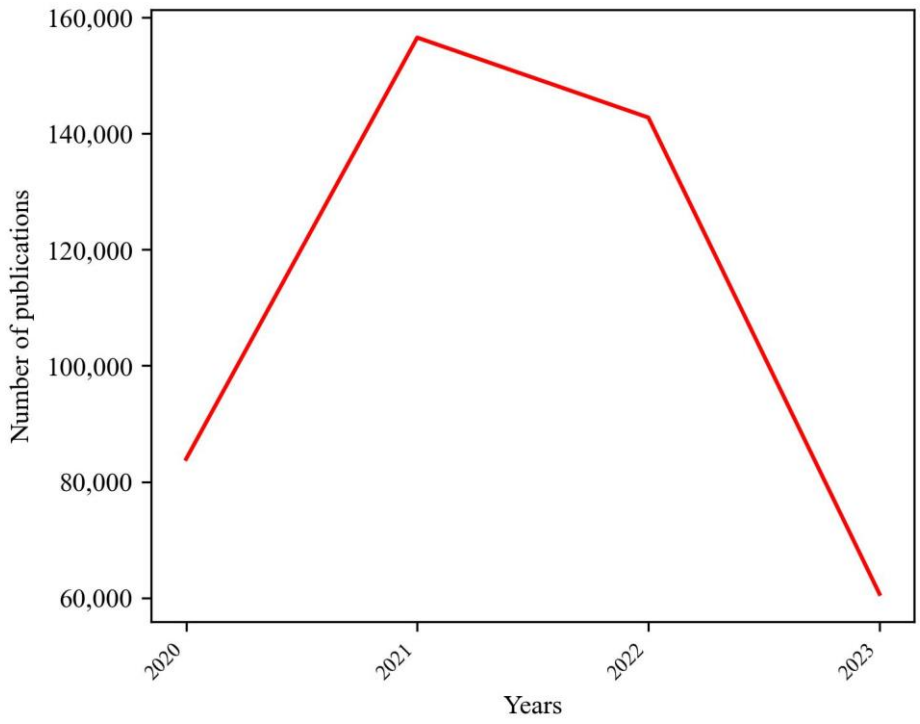

Figure 5. Number of publications in WoS related to COVID-19.

Based on a comparative analysis of the trends shown in Figures 3–5, a similarity in the profiles of the curves for publications from 2020 to 2023 can be noticed due to the fact that publications on the COVID-19 pandemic have included references to previous pandemics in the literature review and the distribution of the number of publications is of a Gaussian type.

3. Literature Review

The performance and accuracy of natural language processing models are dependent on the quality of the NLP dataset [28]. Extending these services to the cloud by building a cloud platform for NLP services [29] will ensure easier management, automatic updating, global accessibility, and allow greater diversity through the use of interactive maps [30]. Prakash, Ohno-Machado, and Chapman [31] detail the incorporation of IR (Information Retrieval) technology in relational database engines, while Jiang, Wang, and Zhou [32] complete the NLP analysis with the need to reduce the linguistic complexity of the original text by capturing and classifying the general sentiment and the opinions conveyed by it using the Abstract Meaning Representation (AMR) model. Kolbe and Burnett [33] analyzed 128 published studies, in which the authors used methods to analyze the content of large databases, suggesting the need to improve the objectivity of these research studies. In a highly industrialized, automated, and competitive economy, some authors [34] analyze AI-integrated NLP and IoT by companies that use the written text or speech of customers to determine their opinion and loyalty towards the company's products to allow them to be customized according to their own wishes. Another paper [35] analyzes an NLP process made in Python to obtain a keyword dictionary, a minimum number of terms used in a paper. The authors calculate the Coherence Scores used as an indicator of the importance of each word and note the qualitative increase in research by incorporating NLP methods simultaneously with the analysis of larger amounts and much more diverse types of data.

Thus, in research paper [36], a bibliometric analysis for 146 scientific publications that use NLP methods, content analysis, and the "word cloud" to promote the field of Smart Agriculture was conducted, while the papers [37,38] present the importance of performing text mining, in order to identify, extract, and understand, from large databases; the most used words on the policies and decisions are applied at the local or central level. Dicle [39] performs and explains text mining using wordfreq and wordcloud (from Stata), which provides the individual researcher with a list, a dictionary of used words, and their frequency of use. In [40], the authors analzsed the importance of this dictionary of frequently used single words for English module students for the specific field of psychology, using the digital tools Wordcloud and Quizlet. The authors showed that by using Wordcloud, through the creation of the "word cloud", students retained a minimal specialized vocabulary in English psychology more quickly.

A multitude of scientific papers published in connection with an event, an earthquake [41], energy crisis [42], a worldwide crisis situation such as the COVID-19 pandemic, has conveyed a series of manifestations, states of distress, emotion, joy or hope, criticism or discouragement, recommendations, opinions, and suggestions of the researchers involved and society.

Because the Web has become a source of information at a global level, knowing these moods, feelings, opinions is necessary for the foundation of the decision-making process at the micro- and macroeconomic, political, social level. The process that automatically analyses utterances expressed in natural language, identifying essential feelings or opinions, which it classifies according to the emotions conveyed, is called sentiment or opinion analysis [43].

The study of opinions and their evaluation can be conducted using Natural Language Processing (NLP) through different specific algorithms of sentiment analysis or opinion mining. The result of sentiment analysis is the contextual polarity of the text [44], which can be positive, negative, or neutral. There are different methods and algorithms for evaluating opinions and determining the corresponding sentiment [44–48]. Sentiment analysis can be

performed per document (Document Analysis), which determines the opinion expressed by a document; per sentence (Sentence Analysis) to evaluate the opinion conveyed by the sentence; per entity or feature (Entity and Aspect or Feature Analysis), which evaluate the opinion conveyed on the entity. It can be realized by training a neural network (Supervised Learning) using SVMs (Support Vector Machines) and NaiveBayesClassifier algorithms (NB), SentiWordNet, social network analysis (SNA) [43,49–55]. Stine [56] went further with the analysis and managed to detect the sarcasm transmitted in the text.

Researchers and specialists in communication and various fields of specialization continued the analysis of the opinions expressed by researchers in scientific publications by analyzing the opinions expressed on social networks. Thus, the aggressiveness manifested by COVID-19, especially at the beginning, caused states of worry, emotions that were explored and evaluated [57] and thus were known by decision-makers in all affected areas. In study [58], the spiritual factor of religious faith is analyzed, which conveyed to the population words such as peace, trust, hope for healing, explaining the positive evaluation of feelings. A number of studies have analyzed the feelings conveyed by the opinions of all categories of the population during the COVID-19 pandemic on social networks in Greece or China [59,60] or in epidemics, pandemics, viruses, or outbreaks in the last 10 years [61].

4. Description of Research Method

4.1. Content Analysis. Frequency of Use and Word Cloud

4.1.1. Logic Flow of Natural Language Processing

The content analysis of natural language text uses terms such as word dictionary, word cloud, word frequency.

The frequency of use of a word is given by the number of occurrences of the word within the volume of data in the papers analyzed. We use the word cloud to graphically represent the words identified in the analyzed papers, in a visual form, so that their size is proportional to their frequency of use. The dictionary represents, in this context, a totality of words used in the analyzed papers, representative of the domain or subject analyzed. Considering that during the COVID-19 pandemic, the number of scientific research studies addressing this topic increased a lot, we searched the WoS database. There were found 443,878 papers that contained the word "COVID-19". We extracted from the WoS database the first 1000 publications (this number depended on the limitation imposed by WoS) that were the most cited, and we formed three data collections containing the fields "title", "abstract", and "keywords", respectively. For each of the three data collections, we took the most used words, identified by their frequency of use (sorted in descending order). In this way, we took into account, in our analysis, the most used words in the respective field/subject used in the "title", "abstract", and "keyword", respectively. We continued the analysis with the word cloud representation using natural language processing in Python. The general description of the logical processing flow to obtain the results in our study is as follows:

- Importing the libraries to perform various operations, including pandas for data manipulation, NLTK for natural language processing, matplotlib for visualization, WordCloud for word cloud generation.
- Downloading the necessary resources for the NLTK library, such as the stopwords list and the tokenizer for tokenizing words [62].
- Loading data from a CSV file ("Abstract.csv"') into a DataFrame (df).
- Concatenating all the texts in the "text" column of the df DataFrame into a single string (text) and converting the text to lowercase.
- Cleaning data—Remove special characters (punctuation) from the text using a list of punctuation characters in the string library.
- Applying the text tokenization operation to individual words using word_tokenize.
- Removing Stopwords—i.e., stopwords in the text, using the English stopwords list.
- Removing specific words or character sequences listed in the words_to_remove list.

- Creating a word_freq dictionary to calculate the frequency of the remaining words in the text.
- Creating a new filtered_word_freq dictionary containing only words with a frequency of at least 1.
- Using the WordCloud library to generate a word cloud (WordCloud) based on the frequency of words in word_freq and using options such as sizes, maximum number of words, and colors.
- Displaying the word cloud generated using Matplotlib. The word cloud is a visual representation of words.
- Creating a DataFrame (filtered_word_freq_df) from the filtered_word_freq dictionary and saving it in a CSV file ("Abstract_frequency_word.csv").
- Ending—We display a message to confirm saving the filtered dictionary.

It can be seen that the program reads the data in a CSV file, processes them to remove words of no interest, calculates the frequency of the remaining words, and generates a word cloud to visualize the frequency of words in the given text.

4.1.2. Analysis of Input Data for the Title.csv File

The Title.csv input data have 1000 records, containing 2055 different words, of which 1038 are unique. For clarity, we made a cloud of the top 100 most frequently used words in the titles of scientific papers, represented in Figure 6.

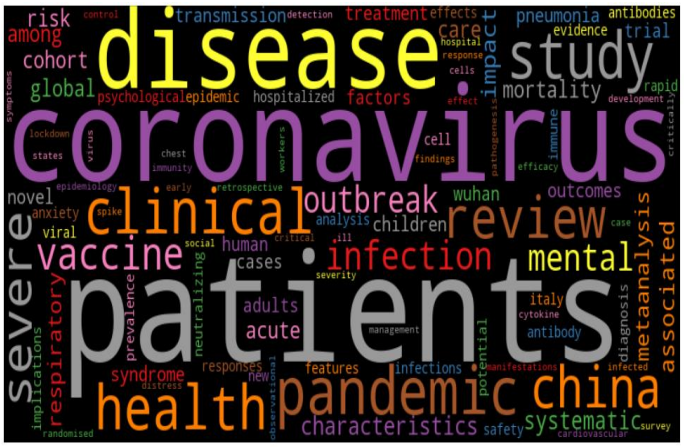

Figure 6. Word cloud for the 100 most used words in the titles of scientific papers.

The frequency of occurrence of the top 10 words is represented in Figure 7 where the word "patients" is the most used at 8.76%.

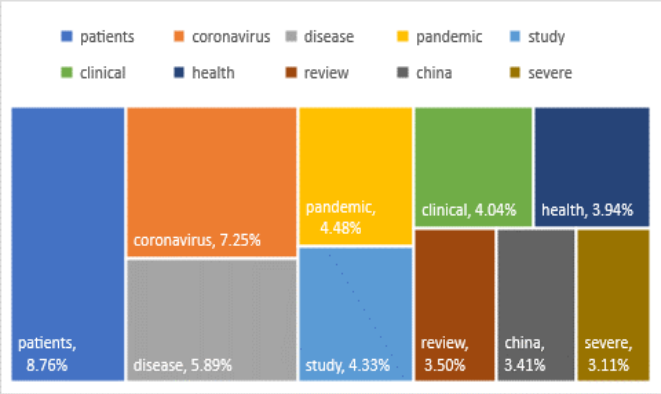

Figure 7. Top 10 frequent words in "Title".

4.1.3. Analysis of Input Data for the Keyword.csv File

The Keyword.csv input data have only 722 records (as not all papers mention keywords), containing 1304 words. We made a cloud of the top 100 most frequently used words in Keyword.csv for scientific papers, represented in Figure 8. If we analyze the obtained dictionary, the unique words in the dictionary represent about 60% of them, which leads to the idea that the analyzed papers deal with a significant number of unique topics.

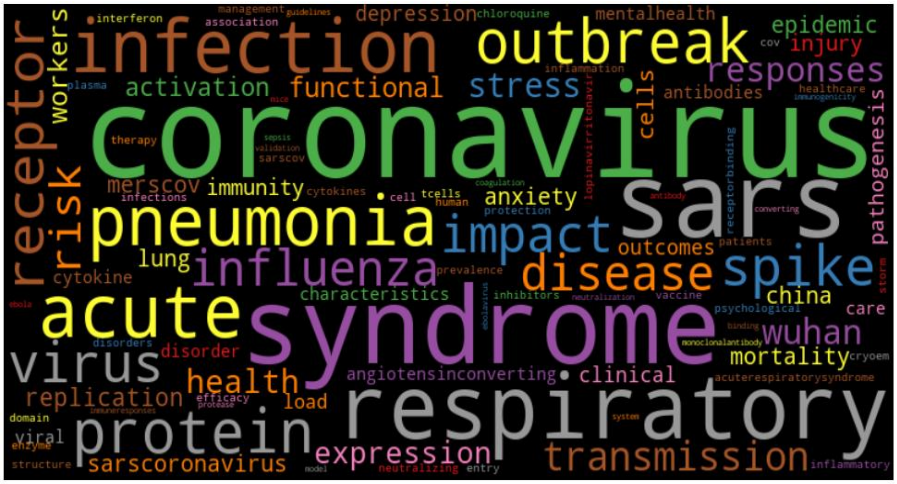

Figure 8. Word cloud for the 100 most used words in publications keywords.

4.1.4. Analysis of Input Data for the Abstract.csv File

The Abstract.csv input data contain 876 records, as there are papers without abstracts among the 1000 papers retrieved from WoS.

A total number of 10,276 words in the dictionary were obtained, of which 4149 are single words, representing about 40% of the total words. The cloud of the top 100 most frequently used words in the abstracts was made for the scientific papers analyzed, represented in Figure 9.

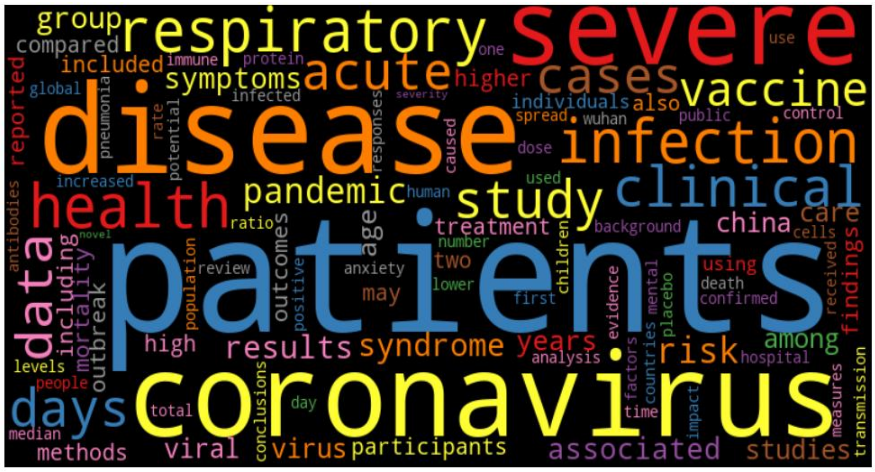

Figure 9. Word cloud for the 100 most used words in the abstracts of publications.

We have basically identified, for the group G = {Title, Abstract, and Keyword}, the most used words of importance for those interested in the topic "COVID-19". These could constitute a minimum vocabulary necessary for those interested in the topic under analysis. In summary, the above data are presented in Table 1.

Table 1. Content of the three dictionaries.

Document Type	Number of Records	Word Dictionary Size
Keyword	722	1304
Title	1000	1038
Abstract	876	10,276

Source: Authors' processing.

4.2. Sentiment Analysis

4.2.1. Sentiment Analysis Using Microsoft Azure AI Language

Sentiment analysis using Microsoft Azure AI Language is achieved by using the natural language processing services offered by the Azure platform. The Azure AI Language service uses pre-trained machine learning models to evaluate the sentiment of a text and to return a sentiment score. This sentiment analysis method also relies on advanced language models and machine learning in order to understand the context and tone of a text and provides accurate results for varied texts and complex contexts. However, this method has limitations due to the fact that access to Azure services involves costs that, depending on the volume of data, are limited. Microsoft Azure AI Language [63,64] has the ability to analyze, evaluate a text, and return sentiment scores for each sentence in a range from 0 to 1. Values close to 1 represent a positive sentiment, those close to 0 a negative sentiment, and those in the middle of the range (0.5) are considered neutral or indeterminate. The sentiment analysis function provides sentiment labels with "negative", "neutral", and "positive" values based on the confidence score obtained. Using Azure AI Language services for sentiment analysis provides a high level of accuracy and flexibility, but the cost associated with using Azure resources must also be considered.

Using Azure Machine Learning-Sentiment analysis in Excel, we calculated for the "Abstract.csv" document, the fields "Sentiment" and "Score". We used the same processing for the "Title.csv" document. The results are shown in Figure 10.

TITLE tweet_text	Abstract tweet_text	Publication Year	Titlle Sentiment	Title Score	Abstract Sentimen	Abstract Score
Presumed Asymptomatic Carrier Transmission of COVID-19	This study describes possible transmission of novel coronavirus disease 2019 (COVID-19) from an asymptomatic Wuhan resident to 5 family members in Anyang, a Chinese city in the neighboring province of Hubei.	2020	positive	0.655316	neutral	0.53463
Detection of SARS-CoV-2 in Different Types of Clinical Specimens	This study describes results of PCR and viral RNA testing for SARS-CoV-2 in bronchoalveolar fluid, sputum, feces, blood, and urine specimens from patients with COVID-19 infection in China to identify possible means of non-respiratory transmission.	2020	positive	0.627734	neutral	0.55133
Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention	This Viewpoint summarizes key epidemiologic and clinical findings from all cases of coronavirus disease 2019 (COVID-19) reported through February 11, 2020, in mainland China, and case trends in response to government attempts to control and contain the infection.	2020	positive	0.737637	positive	0.62089

Figure 10. Section in the window with the results obtained in the sentiment analysis.

Knowing the calculated sentiment for the abstract of some papers helps in choosing the types of papers consulted or analyzed by researchers in the field of COVID-19. It is observed that many of the papers analyzed are "negative", as they deal with topics related to worry, severity, deaths, response to treatment. "Positive" abstracts convey information about

the favorable condition of patients, how they responded favorably to treatment schemes, hope of a cure. Figures 11 and 12 show the weight of the three sentiment categories in "Abstract.csv" and "Title.csv", respectively.

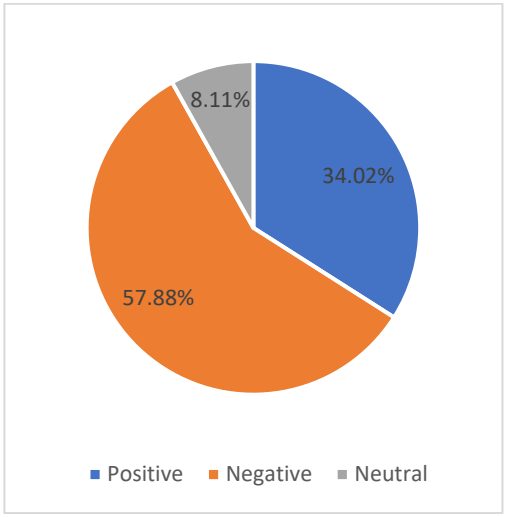

Figure 11. Weight of sentiment categories for "Abstract.csv".

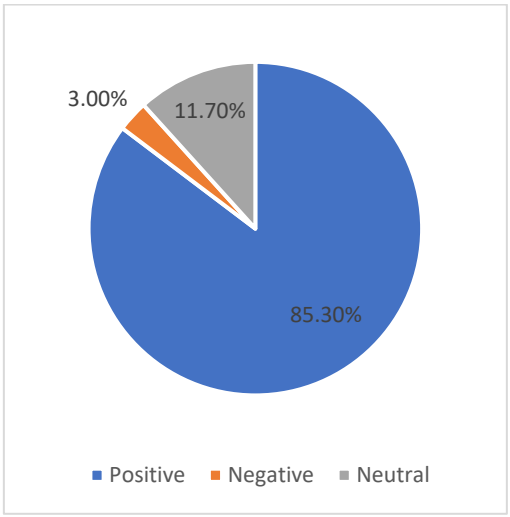

Figure 12. Weighted sentiment categories for "Title.csv".

The comparative analysis of the feelings between "Title" and "Abstract" after removing them from the analysis of papers without abstract is shown in Table 2. It is noticed the variation in the number of papers with corresponding sentiment values for the title of the papers correlated with the sentiment distribution for their abstract.

Table 2. Comparative sentiment analysis between "Abstract" and "Title".

Title	Abstract Positive	Abstract Negative	Abstract Neutral	Total
Positive	257	429	58	744
Negative	5	17	2	24
Neutral	36	61	11	108

Source: Authors' processing.

The sentiment analysis was complemented via a statistical analysis to check if the feelings conveyed in the positive, negative, or neutral abstracts are correlated and if there is a degree of association or dependence between the measured variables.

Microsoft Excel also provides the user with the possibility to calculate the Pearson Function, which calculates and returns the correlation coefficient r, the Pearson coefficient, with values in the range $[-1, 1]$. It measures the degree of association (dependence) between two or more data sets [65].

$$\text{Function syntax: PEARSON } (\text{array}_1, \text{array}_2) \tag{4}$$

where: array_1—a set of independent values; array_2—a set of independent values.

The formula for the Pearson moment correlation coefficient, r, is:

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}} \tag{5}$$

where r is correlation coefficient, n is sample size, X_i represent values of the X variable in a sample, Y_i represent values of the Y variable in a sample, $\overline{X}$ represents the mean of the values of the X variable, and $\overline{Y}$ represents the mean of the values of the Y variable. The Pearson coefficient for the correlation of titles with the corresponding abstracts is approximately 0.9303, which means that "Title" and "Abstract" are strongly correlated.

4.2.2. Sentiment Analysis using Python's Open-Source TextBlob, NLTK, and VADER Libraries

TextBlob uses a pre-trained classifier to assign a polarity (-1 to 1) and subjectivity (0 to 1) to a text so that the sentiment analysis calculated for a text is based on the set of keywords as well as the polarity of these words in a text. Sometimes, this package does not handle complex contexts or subtleties of sentiment in certain texts. NLTK (Natural Language Toolkit) is a dedicated natural language library used for natural language processing (NLP) and computational linguistics. This library provides a rich set of tools and resources for the manipulation and analysis of natural language data. NLTK's main functionalities include tokenization, grammar analysis, sentiment analysis, document classification, language model building. NLTK provides tools for sentiment analysis, including SentimentIntensity-Analyzer, which is a specific module for sentiment analysis in NLTK. It uses the VADER method to assess the intensity of sentiment in a text. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a specialized sentiment analysis system developed to deal with language features, such as expressions and emotions conveyed in social media and beyond. VADER is used by NLTK's SentimentIntensityAnalyzer to evaluate sentiment. VADER is a dictionary of keywords and rules that assigns polarity and intensity scores to words. Basically, the SentimentIntensityAnalyzer in NLTK uses the VADER approach to provide sentiment analysis scores in a text. VADER can also be used independently of NLTK. The choice of one of these packages depends on the specific nature of the text. To use these packages, we have installed them as shown in Figure 13.

```
(base) C:\Users\Mironela Pirnau>pip install textblob nltk vaderSentiment
Collecting textblob
  Using cached textblob-0.17.1-py2.py3-none-any.whl (636 kB)
Requirement already satisfied: nltk in c:\programdata\anaconda3\lib\site-packages (3.8.1)
Collecting vaderSentiment
  Using cached vaderSentiment-3.3.2-py2.py3-none-any.whl (125 kB)
Requirement already satisfied: click in c:\programdata\anaconda3\lib\site-packages (from nltk) (8.0.4)
Requirement already satisfied: joblib in c:\programdata\anaconda3\lib\site-packages (from nltk) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in c:\programdata\anaconda3\lib\site-packages (from nltk) (2022.7.9)
Requirement already satisfied: tqdm in c:\programdata\anaconda3\lib\site-packages (from nltk) (4.65.0)
```

Figure 13. Installing textblob, nltk, and vaderSentiment packages.

Using the Python code in Appendix A, we calculated the sentiment score and its polarity using the textblob, nltk, and vaderSentiment packages for the contents of the "Abstract.csv" and "Title.csv" files.

The results obtained with this code sequence have been centralized in Tables 3 and 4.

Table 3. Sentiment score for Abstract.csv.

Sentiment Type	Sentiment Type TextBlob	Sentiment Type NLTK	Sentiment Type VADER	Sentiment Type Microsoft Azure AI Language
Positive	88.0137%	42.35%	43.04%	34.02%
Negative	9.1189%	56.28%	55.59%	57.88%
Neutral	2.8665%	1.37%	1.37%	8.11%

Source: Authors' processing.

Table 4. Sentiment score for Title.csv.

Sentiment Type	Sentiment Type TextBlob	Sentiment Type NLTK	Sentiment Type VADER	Sentiment Type Microsoft Azure AI Language
Positive	18.00%	19.00%	19.20%	85.30%
Negative	11.30%	25.70%	25.50%	11.70%
Neutral	70.70%	55.30%	55.30%	3%

Source: Authors' processing.

For the results obtained in Table 4, it can be seen that for long texts, the results obtained with NLTK, VADER, and Microsoft Azure AI Language packages are close.

It can be seen from the results obtained in Table 4 that for short texts, the sentiment score values obtained with the NLTK and VADER packages are close. Identifying the sentiment in the text is a complex task and depends on human subjectivity. As shown in Tables 3 and 4, the results may vary depending on the algorithms used, the datasets, and the overall context. Natural language processing combined with machine learning tools has multiple applicability [28,62,66].

For highly accurate sentiment analysis, we continue our study by using the Hugging Face Transformers library, which provides access to numerous pre-trained natural language models, including models for sentiment analysis. We will use the BERT (Bidirectional Encoder Representations from Transformers) model, developed by Google, as it has been pre-trained on large volumes of text. It is known to have achieved outstanding performance for NLP applications, as it uses a transformer architecture to capture the bidirectional meaning of words in context.

4.3. Similarity Analysis for the Works Analysed

Often, in Big Data models, similar entities need to be identified, which can be studies, research on a certain topic or characteristics that define a certain product, services. Examining similar items (sentences) is a fundamental data-mining problem. There are several ways to determine the similarity of these entities. For the huge amount of data existing on the Internet, plagiarism, "mirrored" web pages, or articles coming from the same source can be identified in this way [67]. Jaccard similarity, defined as the ratio between the number of common elements and the total number of elements of the data sets, is described in [67]. It is possible to determine the similarity of two strings using specific functions in Microsoft Excel [68] or VBA code. Variants of saliency-based fuzzy logic are used to model natural language when using linguistic variables [69].

Fuzzy Lookup technology measures the similarity between two data sets using Jaccard similarity, defined as the size of the intersection of the sets of objects (data) divided by the

size of their union [70]. Thus, the Fuzzy Lookup component for Excel identifies matches of text data in Excel.

The matching is robust even in cases of misspellings, abbreviations, added or missing data. The more common data that are identified, the closer the Jaccard similarity will be to a value of 1. The technology allows assigning weights (importance) to some objects, thus obtaining the weighted Jaccard similarity.

An implementation of this algorithm was carried out in the *fuzzywuzzy* package in Python. To calculate the similarity score between each pair of texts (eliminating the situation where a text is compared to itself), we have $\frac{n \cdot (n-1)}{2}$ similarity scores, where n is the number of records. For our case, with 1000 records in the Title.csv file, we obtained a total of 499,500 similarity scores between all possible pairs of texts in the 1000 records/title works. These scores were stored in a new dataset.

The value calculated by the *fuzzywuzzy* algorithm represents the degree of similarity between two strings.

The ratio method in *fuzzywuzzy* calculates the similarity score using the Levenshtein distance, which measures the minimum number of operations (insertions/deletions/replacements) performed to turn one string into another. Similarity scores are dependent on the type of analysis and the algorithm used. To normalize the similarity scores to a range of 0–100, Python code was used (normalization 1):

$$norm\_score1 = \left(fuzz.\frac{ratio(string1,\ string2)}{len(string1)} \right) \cdot 100 \quad (6)$$

To normalize the similarity score according to the total length of both strings being compared, Python code was used:

$$norm\_score2 = \left(fuzz.\frac{ratio(pair[0],\ pair[1])}{(len(pair[0]) + len(pair[1]))} \right) \cdot 100 \quad (7)$$

Using the code in Appendix B, we calculated the weighted similarity scores for the records in the Title.csv file. Then we performed the calculations again using cosine similarity [71], which is based on term frequency vectors and the scikit-learn library, see Appendix C.

In Table 5, we have centralized the cosine similarity with the scikit-learn library, the number of papers according to the similarity interval for the two normalization methods.

Table 5. Similarity results calculated for "Title.csv".

Similarity Interval for Title	normalized_score1 Number of Papers	normalized_score2 Number of Papers	Cosine Similarity with the Scikit-Learn Library [71]
(0–50%)	471,916	498,123	499,334
(50%–60%)	16,638	984	95
(60%–70%)	6031	258	42
(70%–80%)	2511	83	21
(80%–90%)	1133	30	4
(90%–100%)	1271	22	4

Source: Authors' processing.

It can be seen that, regardless of the method applied, around 94% of the paper titles analyzed have a degree of similarity below 50%.

The study of the similarity of some entities is increasingly used in research from all fields: medical, public health, education, management, environment, business, crisis events/situations in society, communication. Through the computational determination of the similarity of some texts and some news transmitted online regarding a special

event or situation, it is possible to evaluate the effectiveness of communication actions in influencing society's behavior regarding public health [72], to achieve the semantic grouping of documents [73], to analyze the similarity of the weighting methods of the criteria taken into account in Multicriteria Decision Support Systems in management [74], and to study cerebral activity on several channels in medicine [75].

5. Conclusions and Discussion

The usefulness of our work is both theoretical and practical. At a theoretical level, the dictionaries of words with the highest frequency of use in a field or topic of study (selected from the most cited scientific papers, works of great interest) represent an extremely important resource for initiation, information, or additional learning for those interested, such as young researchers who want to know the terminology specific to a field/theme of research; translators who specialize in a particular field and need to have a minimum knowledge of a dictionary (vocabulary) on a particular field, as presented in [40]; trainers (teachers) and trainees (pupils or students) in order to initiate and/or specialize [76] in a vocabulary specific to a field/research topic; researchers/scientists in order to address interdisciplinary themes; politicians and decision-makers who have to own a minimum vocabulary necessary to address and understand complex social problems [37,38]. Recognizing emotions in conversation is one branch of sentiment analysis [77]. The analysis of feelings for the abstract and title of the reviewed papers provides valuable information to both researchers and decision-makers in different fields [78].

Knowing a researcher's opinion, which can be inferred from the abstract of a paper, helps in choosing the types of papers studied by researchers in certain areas of interest, in our case, "COVID-19". The failure to manage the crisis and its negative effects on society (such as feelings of worry and the restrictions imposed) or, on the contrary, feelings of hope, the process of healing, the success of the measures taken, the return to normality, feelings of joy are opinions conveyed by the authors of research papers that represent extremely important signals for the population but also for specialists and researchers interested in analyzing various moments of crisis manifestation in society. In addition, the similarity noticed in the case of the most representative works in the field under discussion showed that the topics approached were different (taken from various areas which were in the attention of the population, such as health, economic, and social life) and targeted both the academic and research environment, that is, specialists and researchers interested in finding solutions to stop and cure the virus and in returning to normal life. The results of this analysis can also be used by those interested in consulting, initiating or deepening their knowledge in the field for a better understanding of the reaction and behavior of those affected.

From a practical point of view, this paper used a method of exploring the informational content of large databases with a great diversity and complexity of the topics addressed in order to identify and extract a dictionary of words with a high frequency of use in a certain field or research topic. We considered the group G = {Title, Abstract, Keyword} for a publication of any type, an article, book, research project, which contains the most important specialized words used in a certain field or research topic. The method uses working tools consisting of code sequences written by authors that use Python functions and allow the creation of data collections (.csv): a dictionary of words based on the frequency of their use, ordered in descending order, in the Abstract, Title, and Keywords belonging to the selected works. The representation of the first "n" most used words by word cloud (we considered 100) in the case of each component taken into account in a research paper (Title, Abstract, Keywords) is suggestive, allowing any researcher to get used to the terminology specific to the respective field faster. Therefore, we can say that we have obtained dictionaries with the most used words from the analyzed works, and they are part of a machine learning method [79,80].

At a practical level, as in the case of learning a foreign language, a minimum vocabulary of words can help the training process of a beginner in the same way the built

vocabulary will be able to help a researcher/specialist in related fields or a beginner/trainee to get used to specific terminology. The terms identified in a vocabulary, without necessarily referring to those related to COVID-19, do not remain unchanged throughout time. They do not remain unchanged even within the vocabulary of the same language. In a field of research (especially in one in which modern technology, IC and T, new discoveries in science revolutionize the world) the terminology, the procedures, the norms as well as the vocabulary are continuously updating. Therefore, the conditions for selecting the works from the accessed database can change, for example, "the most cited works published in connection with... in the period". Thus, we can be sure that they are articles of research that are recognized in the respective field, which are quoted and, at the same time, use the latest terminology.

If this study were completed with similar studies conducted to find out the reaction of the population through social networks [41,42,77,80,81], then the picture of the pandemic, of the event considered, with the states of crisis and pain or recovery of the population, of the areas of economic, social, cultural life would be more complete. Recognizing the problems society faces would become a prerequisite for finding solutions.

Therefore, this study has this limitation since the analysis of scientific papers published by specialists and researchers is carried out without taking into account the many messages and opinions transmitted by community members through social networks. The analysis of these messages and the monitorization of the reactions, opinions, and discussions carried out by citizens and organizations on social networks can be a future stage of development of this paper by analyzing the feelings as well as the similarities existing in the opinions of the world's population irrespective of age, occupation, ethnicity, or religion.

Another limitation is the relatively small number (1000) of papers extracted from WoS and analyzed (compared to the total number of papers published and existing in WoS (443,878 papers). In addition, we have considered English-language publications, but it is known that there are a number of other publications in other languages and in international journals and databases other than WoS.

However, we considered representative the 1000 works extracted and analyzed from WoS, having selected the most cited works that addressed the issue of COVID-19. The analysis method proposed by us can be applied to any research topic in papers, articles, or projects in various fields/specializations in order to identify a minimum dictionary (vocabulary) of the most commonly used terms with visual representation through word clouds. Those interested in a particular research problem can quickly and correctly inform themselves about published papers in the field, can access a minimal dictionary of terms used, can identify some of the feelings conveyed by those papers, thus being able to choose a topic that will be investigated in future research.

Author Contributions: Conceptualization, M.P., M.A.B., I.P., A.H., C.C. and I.O.; methodology M.P., M.A.B., I.P., A.H., A.T., C.C. and I.O.; software, M.P., M.A.B., I.P., A.H., C.C. and I.O.; validation M.P., M.A.B., I.P., A.H., A.T., C.C. and I.O.; formal analysis, M.P., M.A.B., I.P., A.H. and I.O.; investigation, M.P., M.A.B., I.P., A.H. and I.O.; resources, M.P., M.A.B., I.P., A.H. and I.O.; writing—review and editing, M.P., M.A.B., I.P., A.H., A.T., C.C. and I.O; visualization, M.P., M.A.B., I.P., A.H., A.T., C.C. and I.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the first author or corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

```
import pandas as pd
from textblob import TextBlob
from nltk.sentiment import SentimentIntensityAnalyzer
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer as VaderSentimentIntensityAnalyzer
def determina_tip_sentiment(polaritate):
if polaritate > 0:
return 'Positive'
elif polaritate < 0:
return 'Negative'
else:
return 'Neutral'
def calculeaza_sentiment(nume_fisier_csv, nume_fisier_output_csv):
data = pd.read_csv(nume_fisier_csv)
coloana_text = data.columns[0]
# TextBlob
data['Sentiment_TextBlob'] = data[coloana_text].apply(lambda x: TextBlob(str(x)).sentiment.polarity)
data['Tip_Sentiment_TextBlob'] = data['Sentiment_TextBlob'].apply(determina_tip_sentiment)
# NLTK
sia = SentimentIntensityAnalyzer()
data['Sentiment_NLTK'] = data[coloana_text].apply(lambda x: sia.polarity_scores(str(x))['compound'])
data['Tip_Sentiment_NLTK'] = data['Sentiment_NLTK'].apply(determina_tip_sentiment)
# VADER
vader_sia = VaderSentimentIntensityAnalyzer()
data['Sentiment_VADER'] = data[coloana_text].apply(lambda x: vader_sia.polarity_scores(str(x))['compound'])
data['Tip_Sentiment_VADER'] = data['Sentiment_VADER'].apply(determina_tip_sentiment)
data.to_csv(nume_fisier_output_csv, index=False)
nume_fisier_csv = 'c://2023/title.csv'
nume_fisier_output_csv = 'c://2023/title_sen.csv'
calculeaza_sentiment(nume_fisier_csv, nume_fisier_output_csv)
```

Appendix B

```
import pandas as pd
from fuzzywuzzy import fuzz
from itertools import combinations
import csv
file_path = 'c://2023/title.csv'
df = pd.read_csv(file_path)
if 'text' not in df.columns:
print("error in file CSV.")
else:
with open('c://2023/title_similarity.csv', 'w', newline='') as csv_file:
csv_writer = csv.writer(csv_file)
csv_writer.writerow(['text1', 'text2', 'similaritate_normalizata'])
for pair in combinations(df['text'], 2):
norm_score = (fuzz.ratio(pair[0], pair[1]) / (len(pair[0]) + len(pair[1]))) * 100
csv_writer.writerow([pair[0], pair[1], norm_score])
```

Appendix C

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from itertools import combinations
```

```
import csv
file_path = 'c://2023/title.csv'
df = pd.read_csv(file_path)
if 'text' not in df.columns:
print("Eroare în fișierul CSV.")
else:
with open('c://2023/cosine_similarity.csv', 'w', newline='') as csv_file:
csv_writer = csv.writer(csv_file)
csv_writer.writerow(['text1', 'text2', 'similaritate_cosinus'])
vectorizer = TfidfVectorizer()
vectors = vectorizer.fit_transform(df['text'])
for pair in combinations(range(len(df['text'])), 2):
text1, text2 = pair
cosine_sim = cosine_similarity(vectors[text1], vectors[text2])[0][0]
csv_writer.writerow([df['text'][text1], df['text'][text2], cosine_sim])
```

References

1. Cucinotta, D.; Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Biomed.* **2020**, *91*, 157–160. [CrossRef]
2. World Health Organization. Available online: https://www.who.int/ (accessed on 1 September 2023).
3. Roychowdhury, K.; Bhanja, R.; Biswas, S. Mapping the research landscape of COVID-19 from social sciences perspective: A bibliometric analysis. *Scientometrics* **2022**, *127*, 4547–4568. [CrossRef]
4. Akl, E.A.; Meho, L.I.; Farran, S.H.; Nasrallah, A.A.; Ghandour, B. The Pandemic of the COVID-19 Literature: A Bibliometric Analysis, Running Title: Bibliometric Analysis of the COVID-19 Literature. *Res. Sq.* **2020**, 1–20. [CrossRef]
5. Ageel, M. Pandemic Critical Care Research during the COVID-19 (2020–2022): A Bibliometric Analysis Using VOSviewer. *BioMed. Res. Int.* **2022**, *2022*, 8564649. [CrossRef]
6. Hod, R. Bibliometric Analysis on Medical Education During COVID-19 Pandemic. *Malays. J. Med. Health Sci.* **2022**, *18* (Suppl. 14), 111–119. [CrossRef]
7. Nasir, A.; Shaukat, K.; Hameed, I.A.; Luo, S.; Alam, T.M.; Iqbal, F. A Bibliometric Analysis of Corona Pandemic in Social Sciences: A Review of Influential Aspects and Conceptual Structure. *IEEE Access* **2020**, *8*, 133377–133402. [CrossRef] [PubMed]
8. Hosszu, A.; Rughiniș, C.; Rughiniș, R.; Rosner, D. Webcams and social interaction during online classes: Identity work, presentation of self, and well-being. *Front. Psychol.* **2022**, *12*, 761427. [CrossRef] [PubMed]
9. Shapira, P. Scientific publications and COVID-19 "research pivots" during the pandemic: An initial bibliometric analysis. *bioRxiv* **2020**, 1–42. [CrossRef]
10. Firmansyah, I.; Rusydiana, A.S. Bibliometric Analysis of Articles on Accounting and COVID-19 during the Pandemic. *Libr. Philos. Pract. (E-J.)* **2021**, *5179*, 1–15.
11. Leoni, G.; Lai, A.; Stacchezzini, R.; Steccolini, I. The pervasive role of accounting and accountability during the COVID-19 emergency, Accounting. *Audit. Account. J.* **2022**, *35*, 1–19. [CrossRef]
12. Nagy, A.M.; Konka, B.; Torok, A. The COVID problem reflected by economics—A bibliometric analysis. *Acta Oeconomica* **2021**, *71*, 205–221.
13. Popescu, A.I. Business Formation during the Coronavirus Pandemic. A Regional Analysis Considering Knowledge and Technological Intensity. *Econ. Comput. Econ. Cybern. Stud. Res.* **2021**, *55*, 199–214.
14. Paraschiv, D.M.; Țițan, E.; Manea, D.I.; Bănescu, C.E. Quantifying the Effects of Working from Home on Privacy. An Empirical Analysis in the 2020 Pandemic. *Econ. Comput. Econ. Cybern. Stud. Res.* **2021**, *55*, 21–36. Available online: https://ecocyb.ase.ro/Articles2021_4.htm (accessed on 11 June 2023).
15. Verma, S.; Gustafsson, A. Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. *J. Bus. Res.* **2020**, *118*, 253–261. [CrossRef] [PubMed]
16. Cetina, I.; Vinerean, S.; Opreana, A.; Radulescu, V.; Goldbach, D.; Radulian, A. The Impact of the COVID-19 Pandemic on Consumers' Online Shopping Behaviour—An Empirical Model. *Econ. Comput. Econ. Cybern. Stud. Res.* **2022**, *56*, 41–56.
17. Stancu, A.; Filip, A.; Dumitru, I.; Alniacik, U.; Ionescu, F.T.; Mogoș, O.; Cânda, A. Modelling m-Commerce Adoption among Generation Z in the Pandemic Context. *J. Econ. Comput. Econ. Cybern. Stud. Res.* **2023**, *57*, 187–202.
18. Gherghina, Ș.C.; Botezatu, M.A.; Simionescu, L.N. Exploring the Impact of Electronic Commerce on Employment Rate: Panel Data Evidence from European Union Countries. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 3157–3183. [CrossRef]
19. Aristovnik, A.; Ravšelj, D.; Umek, L.A. Bibliometric Analysis of COVID-19 across Science and Social Science Research Landscape. *Sustainability* **2020**, *12*, 9132. [CrossRef]
20. Paim, A.A.M.; Carneirode Andrade, M.; Steffens, F. Mapping and bibliometric analysis of scientific publications on the use of textile materials for protection in pandemics. *Braz. J. Inf. Sci. Res. Trends* **2022**, *16*, 6. [CrossRef]
21. Shi, Y.; Song, Y.; Guo, Z.; Yu, W.; Zheng, H.; Ding, S.; Zhan, S. COVID-19 pharmacological research trends: A bibliometric analysis. *Intell. Med.* **2023**, *3*, 1–9. [CrossRef]

22. Zyoud, S.H.; Al-Jabi, S.W. Mapping the situation of research on coronavirus disease-19 (COVID-19): A preliminary bibliometric analysis during the early stage of the outbreak. *BMC Infect. Dis.* **2020**, *20*, 561. [CrossRef] [PubMed]
23. Selva-Pareja, L.; Camí, C.; Roca, J.; Espart, A.; Campoy, C.; Botigué, T. Knowledge, attitudes, and practices about COVID-19 pandemic: A bibliometric analysis. *Front. Public Health* **2023**, *11*, 1075729. [CrossRef] [PubMed]
24. DEFelice, F.; Polimeni, A. Coronavirus Disease (COVID-19): A Machine Learning Bibliometric Analysis. *Vivo* **2020**, *34* (Suppl. 3), 1613–1617. [CrossRef] [PubMed]
25. Nair, C.S.; Shivdas, A.; Yasmin, M. Research trends of open access publications in library and information science during COVID-19 pandemic: A bibliometric analysis. *J. Posit. Sch. Psychol.* **2022**, *6*, 1517–1524.
26. Almasoud, A.S.; Alshahrani, H.J.; Hassan, A.Q.A.; Almalki, N.S.; Motwakel, A. Modified Aquila Optimizer with Stacked Deep Learning-Based Sentiment Analysis of COVID-19 Tweets. *Electronics* **2023**, *12*, 4125. [CrossRef]
27. Available online: https://www.webofscience.com (accessed on 1 September 2023).
28. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [CrossRef]
29. Pais, S.; Cordeiro, J.; Luqman Jamil, M. NLP-based platform as a service: A brief review. *J. Big Data* **2022**, *9*, 54. [CrossRef]
30. Bjørnar, T.; Solveig, B.; Weiqin, C.; Lars, N. Word cloud visualisation of locative information. *J. Locat. Based Serv.* **2015**, *9*, 254–272. [CrossRef]
31. Prakash, M.N.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inf. Assoc.* **2011**, *18*, 544–551. [CrossRef]
32. Jiang, X.; Wang, Z.; Zhou, G. Semantic Simplification for Sentiment Classification. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022.
33. Kolbe, R.H.; Burnett, M.S. Content-Analysis Research: An Examination of Applications with Directives for Improving Research Reliability and Objectivity. *J. Consum. Res.* **1991**, *18*, 243–250. Available online: https://www.jstor.org/stable/2489559 (accessed on 20 June 2023). [CrossRef]
34. Mah, P.M.; Skalna, I.; Muzam, J. Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0. *Appl. Sci.* **2022**, *12*, 9207. [CrossRef]
35. Abram, M.D.; Mancini, K.T.; Parker, R.D. Methods to Integrate Natural Language Processing Into Qualitative Research. *Int. J. Qual. Methods* **2020**, *19*, 1609406920984608. [CrossRef]
36. Patil, R.R.; Kumar, S.; Rani, R.; Agrawal, P.; Pippal, S.K. A Bibliometric and Word Cloud Analysis on the Role of the Internet of Things in Agricultural Plant Disease Detection. *Appl. Syst. Innov.* **2023**, *6*, 27. [CrossRef]
37. Laver, M.; Benoit, K.; Garry, J. Extracting policy positions from political texts using words as data. *Am. Political Sci. Rev.* **2003**, *97*, 311–331. [CrossRef]
38. Grimmer, J.; Stewart, B.M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Anal.* **2013**, *21*, 267–297. [CrossRef]
39. Dicle, M.F.; Dicle, B. Content analysis: Frequency distribution of words. *Stata J.* **2018**, *18*, 379–386. [CrossRef]
40. Belles-Fortuno, B.; Martinez-Hernandez, A.I. English In The Healthcare Setting: The Use Of Wordcloud And Quizlet with Psychological Pathologies. In Proceedings of the 11th International Conference On Education and New Learning Technologies (EDULEARN 19), Palma, Spain, 1–3 July 2019.
41. Teodorescu, H.N. Using analytics and social media for monitoring and mitigation of social disasters. *Procedia Eng.* **2015**, *107*, 325–334. [CrossRef]
42. Teodorescu, H.N.; Pirnau, M. Twitter's Mirroring of the 2022 Energy Crisis: What It Teaches Decision-Makers—A Preliminary Study. *Rom. J. Inf. Sci. Technol.* **2023**, *26*, 312–322. [CrossRef]
43. Mayur, W.; Annavarapu, C.S.R.; Chaitanya, K. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
44. Devika, M.D.; Sunitha, C.; Ganesh, A. Sentiment Analysis: A Comparative Study on Different Approaches. In Proceedings of the Fourth International Conference on Recent Trends in Computer Science & Engineering, Chennai, Tamil Nadu, India, 29–30 April 2016.
45. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
46. Bashri, M.F.A.; Kusumaningrum, R. Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization. In Proceedings of the 5th International Conference on Information and Communication Technology, Melaka, Malaysia, 17–19 May 2017.
47. Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Appl. Sci.* **2023**, *13*, 4550. [CrossRef]
48. Gao, Z.; Li, Z.; Luo, J.; Li, X. Short Text Aspect-Based Sentiment Analysis Based on CNN + BiGRU. *Appl. Sci.* **2022**, *12*, 2707. [CrossRef]
49. Popa, C.G. Analiza Sentimentelor şi Complexitatea Opiniilor Online. *Today Software Magazine* (TSM) 2022, Issue 32. Available online: https://www.todaysoftmag.ro/article/1297/analiza-sentimentelor-si-complexitatea-opiniilor-online (accessed on 22 June 2023).

50. Montiel-Vázquez, E.C.; Ramírez Uresti, J.A.; Loyola-González, O. An Explainable Artificial Intelligence Approach for Detecting Empathy in Textual Communication. *Appl. Sci.* **2022**, *12*, 9407. [CrossRef]
51. Bhadane, C.; Dalal, H.; Doshi, H. Sentiment analysis: Measuring opinions. In Proceedings of the International Conference on Advanced Computing Technologies and Applications, Mumbai, India, 26–27 March 2015.
52. Fang, X.; Zhan, J. Sentiment analysis using product review data. *J. Big Data* **2015**, *2*, 5. [CrossRef]
53. Taherdoost, H.; Madanchian, M. Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. *Computers* **2023**, *12*, 37. [CrossRef]
54. Keramatfar, A.; Amirkhani, H. Bibliometrics of sentiment analysis literature. *J. Inf. Sci.* **2019**, *45*, 3–15. [CrossRef]
55. Lin, J.K.; Chien, T.W.; Yeh, Y.T.; Ho, S.C.; Chou, W. Using sentiment analysis to identify similarities and differences in research topics and medical subject headings (MeSH terms) between Medicine (Baltimore) and the Journal of the Formosan Medical Association (JFMA) in 2020: A bibliometric study. *Medicine* **2022**, *101*, e29029. [CrossRef] [PubMed]
56. Stine, R.A. Sentiment Analysis. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 287–308. [CrossRef]
57. Chakraborty, K.; Bhatia, S.; Bhattacharyya, S.; Platos, J.; Bag, R.; Hassanien, A.E. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers-A study to show how popularity is affecting accuracy in social media. *Appl. Soft Comput. J.* **2020**, *97*, 106754. [CrossRef]
58. Sánchez-Garcés, J.; López-Gonzales, J.L.; Palacio-Farfán, M.; Coronel-Sacón, V.; Ferney-Teheran, Y.; Peñuela-Pineda, J.; Avila-George, H. Exploratory Analysis of Fundamental Spiritual Support Factors to a Positive Attitude in Patients with COVID-19 Using Natural-Language Processing Algorithms. *Appl. Sci.* **2021**, *11*, 9524. [CrossRef]
59. Samaras, L.; García-Barriocanal, E.; Sicilia, M.A. Sentiment analysis of COVID-19 cases in Greece using Twitter data. *Expert Syst. Appl.* **2023**, *230*, 120577. [CrossRef]
60. Luo, H.; Meng, X.; Zhao, Y.; Cai, M. Exploring the impact of sentiment on multi-dimensional information dissemination using COVID-19 data in China. *Comput. Hum. Behav.* **2023**, *144*, 107733. [CrossRef]
61. Alamoodi, A.; Zaidan, B.; Zaidan, A.; Albahri, O.; Mohammed, K.; Malik, R.; Almahdi, E.; Chyad, M.; Tareq, Z.; Albahri, A. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Syst. Appl.* **2020**, *167*, 114155. [CrossRef]
62. Teodorescu, H.M. Machine Learning Methods for Strategy Research. *Harvard Business School Research Paper Series No. 18-011*. 2017. Available online: https://ssrn.com/abstract=3012524 (accessed on 15 June 2023).
63. Available online: https://learn.microsoft.com/en-us/training/modules/analyze-text-with-text-analytics-service/2-get-started-azure (accessed on 15 September 2023).
64. What Is Sentiment Analysis and Opinion Mining in the Language Service? Azure AI Services | Microsoft Learn. Available online: https://learn.microsoft.com/en-us/azure/ai-services/language-service/sentiment-opinion-mining/overview?tabs=prebuilt (accessed on 15 June 2023).
65. Available online: https://support.microsoft.com/ro-ro/office/pearson-func%C8%9Bia-pearson-0c3e30fc-e5af-49c4-808a-3ef66e034c18 (accessed on 10 September 2023).
66. Teodorescu, M.H.; Ordabayeva, N.; Kokkodis, M.; Unnam, A.; Aggarwal, V. Determining Systematic Differences in Human Graders for Machine Learning Based Automated Hiring. *Brookings Working Paper Series*. 2022, pp. 1–36. Available online: https://www.brookings.edu/wp-content/uploads/2022/06/Determining-systematic-differences-in-human-graders-for-machine-learning-based-automated-hiring.pdf (accessed on 10 June 2023).
67. Foysal, A.A.; Böck, R. Who Needs External References?—Text Summarization Evaluation Using Original Documents. *AI* **2023**, *4*, 970–995. [CrossRef]
68. Fry, A.; Gieseck-Ashworth, J.; Seiler, C. Loving statistics & excel fuzzy lookup in the time of COVID-19. *Ser. Libr.* **2022**, *82*, 145–149. [CrossRef]
69. Teodorescu, H.N. Improve the design and testing of fuzzy systems with a set of (almost) simple rules. *Int. J. Comput. Commun. Control* **2022**, *17*, 1–8. [CrossRef]
70. Microsoft 365. Fuzzy Lookup Add-In for Excel. Available online: https://www.microsoft.com/en-US/download/details.aspx?id=15011 (accessed on 12 September 2023).
71. Wang, L.; Luo, J.; Deng, S.; Guo, X. RoCS: Knowledge Graph Embedding Based on Joint Cosine Similarity. *Electronics* **2024**, *13*, 147. [CrossRef]
72. Cezario, S.; Marques, T.; Pinto, R.; Lacerda, J.; Silva, L.; Santos, L.T.; Santana, O.; Ribeiro, A.G.; Cruz, A.; Araújo, A.C. Similarity Analysis in Understanding Online News in Response to Public Health Crisis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 17049. [CrossRef] [PubMed]
73. Haji, S.H.; Jacksi, K.; Salah, R.M. A Semantics-Based Clustering Approach for Online Laboratories Using K-Means and HAC Algorithms. *Mathematics* **2023**, *11*, 548. [CrossRef]
74. Paradowski, B.; Shekhovtsov, A.; Baczkiewicz, A.; Kizielewicz, B.; Sałabun, W. Similarity Analysis of Methods for Objective Determination of Weights in Multi-Criteria Decision Support Systems. *Symmetry* **2021**, *13*, 1874. [CrossRef]
75. Kriegeskorte, N.; Mur, M.; Bandettini, P. Representational similarity analysis—Connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2008**, *2*, 4. [CrossRef] [PubMed]
76. Hosszu, A.; Rughiniș, C. Digital divides in education. An analysis of the Romanian public discourse on distance and online education during the COVID-19 pandemic. *Sociol. Românească* **2020**, *18*, 11–39. [CrossRef]

77. Fu, Y.; Yuan, S.; Zhang, C.; Cao, J. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics* **2023**, *12*, 4714. [CrossRef]
78. Teodorescu, H.N.; Bolea, S.C. Comparative Lexical Analysis of Three Romanian Works–The Etymological Metalepsis Role and Etymological Indices. *Sci. Technol. (Romjist)* **2022**, *25*, 275–289.
79. Macanovic, A. Text mining for social science—The state and the future of computational text analysis in sociology. *Soc. Sci. Res.* **2022**, *108*, 102784. [CrossRef]
80. Patel, R.; Passi, K. Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. *IoT* **2020**, *1*, 218–239. [CrossRef]
81. Nemes, L.; Kiss, A. Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic. *Appl. Sci.* **2021**, *11*, 11017. [CrossRef]