axioms

MDPI

Article

An Extended AHP-Based Corpus Assessment Approach for Handling Keyword Ranking of NLP: An Example of COVID-19 Corpus Data

Liang-Ching Chen [1] and Kuei-Hu Chang [2,*]

[1] Department of Foreign Languages, R.O.C. Military Academy, Kaohsiung 830, Taiwan
[2] Department of Management Sciences, R.O.C. Military Academy, Kaohsiung 830, Taiwan
* Correspondence: evenken2002@yahoo.com.tw

Abstract: The use of corpus assessment approaches to determine and rank keywords for corpus data is critical due to the issues of information retrieval (IR) in Natural Language Processing (NLP), such as when encountering COVID-19, as it can determine whether people can rapidly obtain knowledge of the disease. The algorithms used for corpus assessment have to consider multiple parameters and integrate individuals' subjective evaluation information simultaneously to meet real-world needs. However, traditional keyword-list-generating approaches are based on only one parameter (i.e., the keyness value) to determine and rank keywords, which is insufficient. To improve the evaluation benefit of the traditional keyword-list-generating approach, this paper proposed an extended analytic hierarchy process (AHP)-based corpus assessment approach to, firstly, refine the corpus data and then use the AHP method to compute the relative weights of three parameters (keyness, frequency, and range). To verify the proposed approach, this paper adopted 53 COVID-19-related research environmental science research articles from the Web of Science (WOS) as an empirical example. After comparing with the traditional keyword-list-generating approach and the equal weights (EW) method, the significant contributions are: (1) using the machine-based technique to remove function and meaningless words for optimizing the corpus data; (2) being able to consider multiple parameters simultaneously; and (3) being able to integrate the experts' evaluation results to determine the relative weights of the parameters.

Keywords: corpus assessment approach; natural language processing (NLP); COVID-19; analytic hierarchy process (AHP); environmental science

Citation: Chen, L.-C.; Chang, K.-H. An Extended AHP-Based Corpus Assessment Approach for Handling Keyword Ranking of NLP: An Example of COVID-19 Corpus Data. *Axioms* **2023**, *12*, 740. https://doi.org/10.3390/axioms12080740

Academic Editors: Nuno Bastos, Touria Karite and Amir Khan

Received: 2 June 2023
Revised: 14 July 2023
Accepted: 17 July 2023
Published: 28 July 2023

1. Introduction

The corpus assessment approach has been applied in the Natural Language Processing (NLP) field for a long time, and it is seen as a critical technique for identifying linguistic patterns [1–3]. Since the end of 2019, the emergence of the novel coronavirus disease COVID-19 has caused serious impacts on global political and economic systems, and even endangered people's lives [4–6]. Diseases always do more harm than good to humans; nevertheless, during the pandemic, scientists discovered that a series of public health policies, such as city lockdowns, as well as decreasing unnecessary commercial activities and travel, can mitigate global environmental pollution issues that we have been helpless to address in the past, especially the air quality index (AQI), which has been shown to have significantly decreased in many modern cities [7–9]. COVID-19 does not seem to be completely eradicated so far; thus, to keep mining knowledge of the disease, it is critical to effectively integrate, process, and reproduce its corpus data.

Corpus assessment approaches have been utilized to process the corpus data of various domains to discover domain-oriented tokens and define linguistic patterns. For example, Poole [3] used the corpus-based approach to process the collected published judicial opinions from 12 geographic distribution areas of the U.S. Federal Court of Appeals

(i.e., the target corpus), for analyzing stance adverbs in its target domain. The contributions of the research defined the linguistic patterns of legal writing styles and provided pedagogical suggestions for legal purposes in English. Otto [2] proposed a three-phase corpus-based data driven learning (DDL) approach to identify special-purpose tokens in the civil engineering domain. The results disclosed that the approach was able to unveil the tokens' functions and improve the efficiency of defining the linguistic patterns in the specialized context of civil engineering. However, when the traditional corpus assessment approach [10] encountered function words and meaningless letters in the keyword list, it could not automatically remove them to conduct corpus optimization, which decreased the efficiency of the corpus assessment. Moreover, the keyword list only adopted the likelihood ratio method [11] as an information retrieval (IR) mean to rank keywords. This caused inaccurate results, because other potential parameters such as frequency and range were not taken into consideration, which made the traditional approach unable to truly define the keywords' level of importance.

The equal weights (EW) method is a classic approach used to process multiple parameters simultaneously when the relative importance of the parameters is unknown. However, the EW method assumes that the relative weights of each parameter are equal, which ignores the relative importance between different parameters. Saaty [12] firstly proposed the analytic hierarchy process (AHP) method to handle the relative importance between different parameters in decision-making problems. The AHP method uses the pairwise comparison between different parameters to compute the eigenvector and eigenvalue and then obtains the relative weights of the parameters. Since then, the AHP method has been adopted in a wide range of applications. For example, Rezaei and Tahsili [13] adopted the AHP method to conduct urban and crisis management, for accessing the vulnerability and immunization parts to decrease the effects of earthquakes. In addition, Ristanovic et al. [14] demonstrated that the AHP method can obtain the best solutions in processing the operational risk management of banks. Prior studies have shown that the AHP method is usually applied in the fields of management and operational research (OR) [12–20]; nevertheless, properly modifying the AHP method can allow it to be used in NLP fields for the computer processing of natural languages, by considering the relative weights of multiple parameters simultaneously.

Corpus assessment approaches have been widely used as an NLP tool in the fields of social sciences and the sciences to explore the linguistic patterns of specific domains [1–3,10,21–23]. The traditional keyword-list-generating approach [10] is based on the likelihood ratio method, which is an IR approach utilized in many types of corpus software [1,23] to calculate a token's keyness value and rank tokens to form a keyword list. Many corpus-based approaches also adopt these types of corpus software to handle corpus analysis tasks [24,25]. However, for traditional keyword ranking, it is difficult to determine the actual importance of each keyword when the program only uses their keyness values for ranking. Namely, the traditional keyword-list-generating approach is only based on one parameter (i.e., the keyness values) to determine and rank keywords, which is insufficient. In the advanced information, communication, and technology (ICT) era, people have developed many algorithms for machine learning and optimizing prior algorithms or machines, with the expectation of machines being able to make more complete and accurate judgments and evaluation results. Thus, the corpus assessment approach should integrate with machine-based corpus optimization and consider multiple parameters (or vectors) simultaneously, to make the evaluation results more accurate. To optimize the deficiency of the traditional keyword-list-generating approach, this paper proposed an extended AHP-based corpus assessment approach to integrate the likelihood ratio method, the corpus optimization approach, and the AHP method, to improve the accuracy of keyword ranking in corpus assessments. The proposed approach firstly optimizes the likelihood ratio method results by removing function words and meaningless letters, and then simultaneously takes three parameters (i.e., the keyness, frequency, and range) into consideration to rank keywords while considering multiple parameters. More importantly,

the relative importance of these parameters is evaluated and determined by experts. That is, the proposed approach not only conducts a complete assessment on the issue but also enables expert evaluation results to be integrated and transformed qualitatively and quantitatively, thereby further making the keyword ranking more complete, precise, and able to satisfy individuals' intentions. To verify the proposed extended AHP-based corpus assessment approach, this paper adopted 53 research articles from the Web of Science (WOS) as empirical examples of natural language data.

The remainder of this paper is organized as follows. Section 2 presents the background information of related methods and the COVID-19 impacts on environmental sciences. Section 3 describes each step of the proposed extended AHP-based corpus assessment approach. Section 4 uses COVID-19-related research articles as empirical examples to verify the proposed approach and compare it with the other two methods, and highlight the contributions. Section 5 is the concluding section.

2. Background

2.1. Likelihood Ratio Method

With the rise of ICT, people have started to rely on computers to process big natural language data. Dunning [11] first introduced the likelihood ratio method for computing the keyness values of tokens for keyword retrieval in corpus analysis tasks, and it is now considered a critical algorithm that is embedded in many types of corpus software. The logic behind the algorithm is that it compares a token's frequency values in two corpora (i.e., the target corpus and the benchmark corpus). When it finds a token with high frequency values in the target corpus and relatively low frequency values in the benchmark corpus, it will calculate the token's keyness values, after which the computation results of the tokens' keyness values will be ranked for generating a keyword list.

The definition of likelihood ratio method is described as follows:

Definition 1 ([11,21]). *Assume that two random variables, X_1 and X_2, follow the binomial distributions $B(N_1, p_1)$ and $B(N_2, p_2)$; p_1 and p_2 are a single trial's success probability, and n_1 and n_2 represent the number of successes that can occur anywhere among the N_1 and N_2 trials, respectively. The logarithm of the likelihood ratio (λ) can be defined as:*

$$-2log\lambda = 2[logL(p_1, n_1, N_1) + logL(p_2, n_2, N_2) - logL(p, n_1, N_1) - logL(p, n_2, N_2)]$$

where

$$L(p, n, N) = p^n(1-p)^{N-n}$$
$$p_1 = \frac{n_1}{N_1}, p_2 = \frac{n_2}{N_2}, \text{ and } p = \frac{n_1+n_2}{N_1+N_2}$$

2.2. Environmental Science Perspective of COVID-19

The earth is the only planet that humans have detected so far in the vast universe to cultivate life [26]. Creatures on the earth depend on a pleasant environment to survive and grow from generation to generation. However, due to the rapid development of human civilization, people have caused serious damage to the earth's environmental and ecological systems. The emission of large amounts of carbon and toxic pollutants (e.g., PM2.5 and PM10 particulate matter, carbon monoxide (CO), ground-level ozone (O3), sulfur dioxide (SO2), and nitrogen dioxide (NO2)) has caused serious air pollution and global warming, leading to the emergence of extreme climates or weather events, and ultimately damaging the survival of organisms [26–29]. Many countries are continuously advocating pro-environmental behaviors to create sustainable development of the ecosystem and the environment. However, people may believe that environmental impacts are a future matter and that even vigorous efforts to promote environmental protection cannot achieve immediate mitigations [30].

Since 2019, the COVID-19 pandemic has impacted economic and political systems globally [31,32]. The COVID-19 virus has been classified as severe acute respiratory syn-

drome coronavirus 2 (SARS-CoV-2). It is related to SARS-CoV and Middle East Respiratory Syndrome (MERS-CoV), but it has a much higher infectious capability and a lower fatality rate than the former two coronavirus types [31,33–35]. In the middle of 2023, the WHO declared that there were over 765 million confirmed cases, with over 6 million deaths during the COVID-19 pandemic [36]. The genetic formation of the spike protein in SARS-CoV-2 has mutated and caused difficulties for the human immune system to resist the virus, hence causing the virus to have a have rapid infection rate [32,33,37]. Moreover, because of its low fatality rate, the virus can parasitize and remain in its hosts for an extended period, thus giving the virus opportunities to mutate and evolve [38]. Until now, many countries are still suffering from COVID-19 variants (such as the Alpha, Beta, Gamma, and Delta variants), which have caused this anti-virus battle to become endless [39]. Current measurements for fighting the COVID-19 pandemic rely on expanding viral detection, enhancing vaccination rates, and following public health policies [34,35]. In addition, the development and introduction of vaccines and specific medicines indicate that people are gradually gaining the dominant position in this anti-virus battle [40].

From the perspective of environmental science, the series of quarantine policies such as travel limitations, city lockdowns, prohibiting non-essential commercial activities, shutting down unnecessary industries, and banning large gatherings has unexpectedly and significantly mitigated pollution levels and the AQI [26,27,41–44]. Prior studies have taught an important lesson—do not think that the self-contribution of pro-environmental behaviors are insignificant—and the improved AQI has proved that restoration of the environment can be an immediate improvement as long as people are willing to strike a balance between economic development and the environment [27,42,43,45].

3. Methodology

Keyword ranking in the corpus assessment approach is an important technique for handling big natural language data and assisting humans in IR and language pattern recognition. For example, information about COVID-19 continuously spreads in our daily life. Although the vaccine has been invented and people are being vaccinated gradually, the SARS-CoV-2 variants keep mutating and causing the anti-virus war to become endless. To enhance our understanding and awareness of COVID-19, the algorithms used for NLP in corpus analysis must be optimized. Hence, this paper proposes an extended AHP-based corpus assessment approach to integrate the likelihood ratio method, the corpus optimization approach, and the AHP method to improve the accuracy of keyword ranking in corpus assessments. The proposed approach is mainly divided into 11 steps, and a detailed description is described as follows (see Figure 1):

Step 1. Create the target corpus.

Compile the natural language data as the target corpus, and convert the file format of the target corpus from the .docx or .pdf format into the .txt (UTF-8) format.

Step 2. Import the target corpus and the benchmark the corpus to the program.

Input the compiled target corpus to AntConc 3.5.8 [1] (the corpus software adopted in this paper) to compute the frequency of each lexical unit's occurrence. In addition, before generating the keyword list, input the benchmark corpus data. English for general purposes (EGP) genres such as blogs, fictional works, magazines, and news of the Corpus of Contemporary American English (COCA) is adopted as the benchmark corpus.

Step 3. Optimize the target corpus.

Before initializing the likelihood ratio calculation, from a linguistic perspective, function words will decrease the accuracy of high frequency words and the keyword-generating process [21]. Therefore, to increase the accuracy and efficiency of soft computing in NLP tasks, this optimization process is inevitable. This step adopts the corpus optimization process of Chen et al. [21], which uses a machine-based processing approach to eliminate function words.

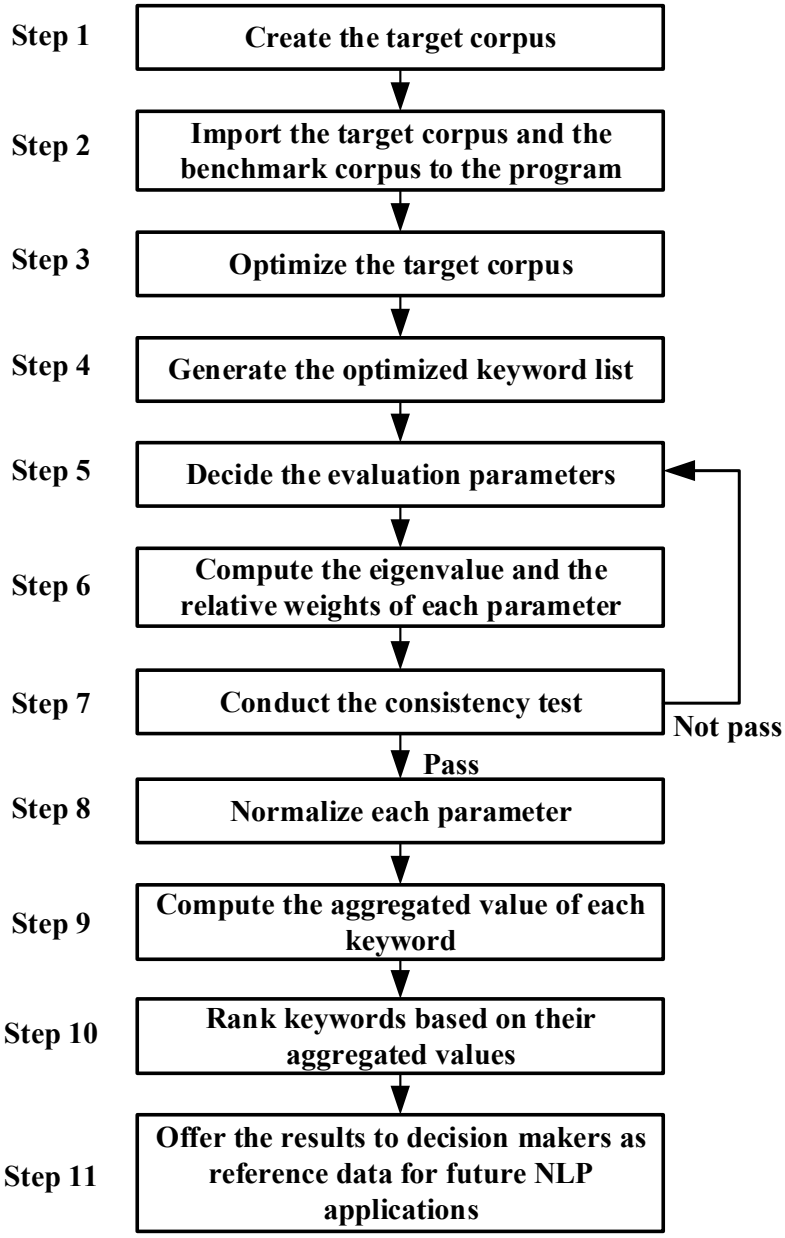

Figure 1. Flowchart of the proposed approach.

Step 4. Generate the optimized keyword list.

After all corpus data is inputted, Dunning's [11] likelihood ratio method will compute and extract words that appear highly frequently in the target corpus in comparison with the words in the benchmark corpus (i.e., computing words' keyness values and ranking them). These words can be considered to be characteristic of the target corpus. Namely, keywords of the target corpus will be retrieved and ranked on the keyword list.

Step 5. Decide the evaluation parameters.

Give experts questionnaires with a paired comparison based on Table 1 to conduct a pairwise comparison of each parameter, in order to, respectively, evaluate the two criteria's relative contribution or importance.

Table 1. Pairwise comparison scale [12].

Relative Importance Scale	Definition of Relatively Important Level	Explanation
1	Equal importance	Two indicators contribute equally to the objective
3	Moderate importance of one over another	From experience and judgment, a certain indicator is slightly important
5	Essential or strong importance	From experience and judgment, a certain indicator is quite important
7	Demonstrated or very strong importance	Practical aspects show that a certain indicator is extremely important
9	Absolute importance	The evidence indicates that a certain indicator is absolutely important
2, 4, 6, 8	The median value of adjacent measures	When a compromise is needed

Then, use Equation (1) to establish the pairwise comparison matrix and proceed with the computation process. If there are n influencing elements, an $\frac{n(n-1)}{2}$ pairwise comparisons must be conducted.

$$\begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ 1/a_{12} & 1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1/a_{1n} & 1/a_{2n} & \cdots & 1 \end{bmatrix} \tag{1}$$

Step 6. Compute the eigenvalue and the relative weights of each parameter.

The eigenvalues and eigenvectors are computed by Equation (2), in which A is the $n \times n$ pairwise comparison matrix, λ is the eigenvalue of matrix A, and X is the eigenvector of matrix A.

$$A \cdot X = \lambda \cdot X \tag{2}$$

After obtaining the maximal eigenvalue λ_{max}, use Equation (3) to calculate the relative weights, W, of each parameter.

$$A \cdot W = \lambda_{max} \cdot W \tag{3}$$

where $W = [w_1, w_2, \ldots, w_n]^T$, and $\sum_{i=1}^{n} w_i = 1$.

Step 7. Conduct the consistency test.

When conducting an expert questionnaire survey, relatively important level scores are usually given by the experts' subjective comments. In other words, the objective and ideal framework should satisfy the transitivity. To inspect whether the pairwise comparison matrix created by the experts' questionnaires is consistent, the consistency index (CI) must be computed by Equation (4) and the consistency ratio (CR) must be calculated by Equation (5) for verification. If the CR value is less than 0.1, the pairwise comparison matrix is consistent.

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{4}$$

$$CR = \frac{CI}{RI} \tag{5}$$

where n is the dimension of the pairwise comparison matrix, λ_{max} is the maximal eigenvalue of the matrix, and RI is the random index (see Table 2).

Table 2. Random index (RI) table [12].

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	N/A	N/A	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48	1.56	1.57	1.59

Step 8. Normalize each parameter.

This paper used three parameters, including the keyness, frequency, and range, to calculate the normalized value of each parameter.

Assume that the p_{ij} is the value of the ith item of keyword data and the jth parameter. The value of r_{ij} is the normalization of p_{ij}, defined as follows.

$$r_{ij} = \frac{p_{ij}}{p_j^{max}},\ j = 1, 2, \ldots, 3 \tag{6}$$

Step 9. Compute the aggregated value of each keyword.

The aggregated value of each keyword is computed by the multiplication of the relative weights of the results for the three parameters from step 8 (shown as Equation (7)).

$$\text{aggregated value}_i = \sum_{j=1}^{3} w_j \times r_{ij} \tag{7}$$

Step 10. Rank keywords based on their aggregated values.

Re-rank the keywords based on their aggregated values from step 9, and generate the ultimate optimized keyword list.

Step 11. Offer the results to decision makers as reference data for future NLP applications.

The optimized keyword list can be provided as critical reference data for decision makers in future NLP applications, such as corpus analysis, keyword analysis, or key information extraction.

4. Empirical Analysis

4.1. Overviews of the Target Corpus

This paper adopted 53 research articles published in 2020–2021 from WOS, which is an internationally well-known academic database. These research articles were under the categorization of environmental science as defined by journal citation reports (JCR), and the topics were all centered on COVID-19. The selection of the research articles had to satisfy the following criteria: (1) the research article needed to correlate with COVID-19; (2) the research article needed to belong to the environmental science discipline; (3) the research article needed to be highly cited; and (4) the research article needed to have a science citation index (SCI) or a social science citation index (SSCI). The main reason to set these criteria was that there is bounteous fake news (information) about COVID-19. After the researchers used the above criteria to search for the relevant research articles from the WOS database, during that moment, there were 53 highly cited research articles showing in the search results. Thus, to verify and highlight the contributions of the proposed approach, the 53 research articles were selected as the target corpus for being the rigorous and non-controversial natural language data.

4.2. Traditional Keyword-List-Generating Approach for Ranking Keywords

The traditional keyword-list-generating approach [10] adopted by this study used Dunning's [11] likelihood ratio method as the main algorithm to determine the keywords of the target corpus. However, some deficiencies occurred in the traditional keyword-list-generating approach. First, without the corpus data optimization process, function words and meaningless letters would affect and reduce the tokens' keyness computation accuracy and cause the keyword list to contain unrelated or meaningless tokens; second, if the keyness value was the only parameter used to determine and rank keywords, it would be impossible to define which keyword was the most commonly used or the most widely dispersed. In other words, the tokens' keyness value needed to be computed with other parameters (e.g., frequency and range) to become a multiple-parameter calculation result that could be used to rank keywords.

4.3. The EW Method for Ranking Keywords

The EW method [46,47] assumes that each criterion has the same importance. If the problem to be solved contains n parameters, P_1, $P_2, \ldots, P_n$, the weight of the EW method is $\frac{1}{n}$. Let a_i be the assessment value of criterion P_i. The weights of the aggregated values for the EW method are shown in Equation (8).

$$\text{EW value} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{8}$$

When the EW method was adopted for computing the parameters of this paper (i.e., the keyness, frequency, and range) for ranking keywords, several deficiencies emerged. First, from the linguistic perspective, under the circumstance that the target corpus was not optimized, the keyness calculation results would have interference from function words and meaningless letters, causing the keyness values to be biased at the beginning. Second, although the EW method can simultaneously consider all parameters, the relative importance level of each parameter should not be the same; hence, it was difficult to meet the experts' expectations.

4.4. The Proposed Extended AHP-Based Corpus Assessment Approach

To optimize and address the deficiencies of the two aforementioned methods, this paper adopted the target corpus as the empirical case, to demonstrate and verify the efficacy and practicality of the proposed approach. Detailed descriptions of each step were as follows.

Step 1. Create the target corpus.

The target corpus in this paper was based on 53 research articles with SCI from WOS. The lexical features included 10,595 word types, 189,680 tokens, and a type–token ratio (TTR) of 0.05586 (representing the lexical diversity).

Step 2. Import the target corpus and the benchmark corpus to the program.

To retrieve the keywords, the algorithm of the software will calculate a word's keyness value to determine whether it is the domain-oriented word, by finding the word that has high frequency in the target corpus but has low frequency in the benchmark corpus. From the perspective of linguistic analysis, when the target corpus is the textual data of professional fields, then the benchmark corpus should select more general-purpose-use data (i.e., EGP). In addition, COCA is considered as the biggest and genre-balanced EGP corpus data, and is widely adopted by many corpus-based researchers as the benchmark corpus [11,21], and so did this paper. After processing by the software, the lexical features of the benchmark corpus (i.e., COCA) included 109,306 word types, 8,266,198 tokens, and a TTR of 0.01322.

Step 3. Optimize the target corpus.

To increase the accuracy of keyword extraction, this step adopted the corpus-based machine optimization approach to eliminate function words and meaningless letters [21]. Table 3 shows the refined target corpus, which eliminated 217 word types and 81,097 tokens, and downsized the target corpus by 43%. Without the interference of function words and meaningless letters, the keyword generator could retrieve more domain-oriented or content words to form a more accurate keyword list.

Table 3. Data discrepancy between the original data and the refined data.

Lexical Feature	Original Data	Refined Data	Data Discrepancy
Word Types	10,595	10,378	−217 (decreasing 2%)
Tokens	189,680	108,583	−81,097 (decreasing 43%)
TTR	0.05586	0.09558	

Step 4. Generate the optimized keyword list.

Once the target corpus, the benchmark corpus, and the stop wordlist are input into AntConc 3.5.8 [1], the traditional keyword-list-generating approach is used to exclude function words and meaningless letters to calculate each token's keyness value and determine the keyword list (see Figure 2). However, during this step, the keyword list still remains at the single-parameter evaluation stage.

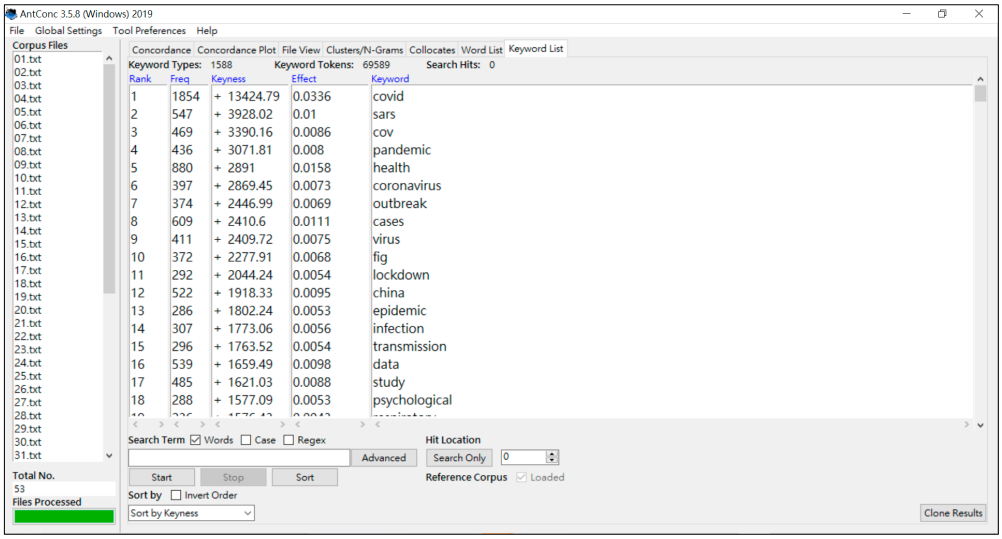

Figure 2. The optimized keyword list on AntConc 3.5.8 [1].

Step 5. Decide the evaluation parameters.

In this step, the evaluation parameters decided by experts are determined as the tokens' keyness, frequency, and range values for the following evaluation processes. The evaluation team in this study included three experts with academic specialties including NLP, corpus linguistics, teachers of English to speakers of other languages (TESOL), performance evaluation, and fuzzy logic. Based on Table 1, the three experts determined the pairwise comparison results of the evaluation parameters, respectively. The results are shown in Table 4.

Table 4. Pairwise comparison results of the parameters.

Criteria	Experts	Experts' Comments		
		Keyness	Frequency	Range
Keyness	Expert 1	1	1/2	1/3
	Expert 2	1	1	1/2
	Expert 3	1	1/2	1/3
Frequency	Expert 1	2	1	1/2
	Expert 2	1	1	1/2
	Expert 3	2	1	1/2
Range	Expert 1	3	2	1
	Expert 2	2	2	1
	Expert 3	3	2	1

Next, the researchers arithmetically averaged each element in the matrix given by the experts and summarized the results as shown in Table 5, and then used Equation (1) to create the matrix for computation in the following steps.

Table 5. The aggregated pairwise comparison matrix.

Criteria	Keyness	Frequency	Range
Keyness	1.000	0.667	0.389
Frequency	1.499	1.000	0.500
Range	2.571	2.000	1.000

Step 6. Compute the eigenvalue and the relative weights of each parameter.

After computing the aggregated pairwise comparison matrix (see Table 5) using Equations (2) and (3), the maximum of the eigenvalue, λ_{max} was 3.003, and the relative weights for the keyness, frequency, and range were 0.195, 0.278, and 0.527, respectively. The relative weights were given by the experts' evaluation and calculated through the AHP computing process, which indicated the relative importance between each vector. Based on the priority vector that range (0.527) > frequency (0.278) > keyness (0.195), we reasoned that the experts' overall assessments indicated that the so-called keywords should also occur widely and frequently in the corpus data.

Step 7. Conduct the consistency test.

To verify the reliability and validity of the relative weights, use Equations (4) and (5), and Table 2 to compute the CI and CR values. The CR value is 0.003, which is less than 0.1, which expressed that the results were acceptable.

Step 8. Normalize each parameter.

Use Equation (6) to normalize each parameter for further aggregated value computation.

Step 9. Compute the aggregated value of each keyword.

Once all parameters were nominalized, the researchers used Equation (7) to compute the aggregated value of the keywords. The partial results of the keywords' aggregated values are presented in Table 6.

Table 6. Keyword list results of the three compared approaches (partial data).

The Traditional Keyword List Generator [10]			The EW Method [47]			The Proposed Method		
Rank	**Keyness Value**	**Token**	**Rank**	**EW Value**	**Token**	**Rank**	**AHP-Based Value**	**Token**
1	14,098.08	COVID-19	1	0.717	COVID-19	1	1.000	COVID-19
2	6008.24	et	2	0.695	the	2	0.699	health
3	4803.88	al	3	0.592	of	3	0.608	coronavirus
4	4129.4	SARS	4	0.552	and	4	0.608	study
5	3562.98	CoV	5	0.488	in	5	0.598	cases
6	3232.45	pandemic	6	0.426	health	6	0.598	China
7	3195.11	health	7	0.406	et	7	0.591	disease
8	3015.85	coronavirus	8	0.403	coronavirus	8	0.587	data
9	2626.35	cases	9	0.377	pandemic	9	0.568	pandemic
10	2584.59	outbreak	10	0.377	al	10	0.557	SARS
11	2560.15	virus	11	0.377	SARS	11	0.555	public
12	2414.4	fig	12	0.376	study	12	0.548	reported
13	2358.18	of	13	0.374	cases	13	0.537	high
14	2151.97	lockdown	14	0.372	china	14	0.531	used
15	2101.9	china	15	0.372	disease	15	0.531	number
16	1907.4	epidemic	16	0.370	by	16	0.527	due
17	1885.46	infection	17	0.360	data	17	0.526	virus
18	1872.07	transmission	18	0.343	were	18	0.515	confirmed
19	1844.72	data	19	0.342	virus	19	0.513	countries
20	1789.67	study	20	0.342	reported	20	0.508	spread
21	1708.46	disease	21	0.336	during	21	0.503	analysis
22	1682.23	psychological	22	0.333	public	22	0.502	outbreak

Table 6. *Cont.*

The Traditional Keyword List Generator [10]			The EW Method [47]			The Proposed Method		
Rank	**Keyness Value**	**Token**	**Rank**	**EW Value**	**Token**	**Rank**	**AHP-Based Value**	**Token**
23	1663.43	respiratory	23	0.329	outbreak	23	0.500	level
24	1639.61	temperature	24	0.328	confirmed	24	0.497	table
25	1602.85	Wuhan	25	0.327	between	25	0.496	results
26	1580.8	confirmed	26	0.326	due	26	0.483	measures
27	1518.98	during	27	0.323	high	27	0.483	significant
28	1504.64	reported	28	0.320	number	28	0.483	period
29	1395.95	anxiety	29	0.319	used	29	0.476	respiratory
30	1307.22	emissions	30	0.318	spread	30	0.475	including
31	1292.84	concentrations	31	0.318	countries	31	0.471	impact
32	1206.96	measures	32	0.317	CoV	32	0.471	infection
33	1205.46	and	33	0.315	analysis	33	0.469	different
34	1182.21	the	34	0.310	respiratory	34	0.468	days
35	1164.24	spread	35	0.307	results	35	0.463	transmission
36	1143.28	march	36	0.307	level	36	0.463	CoV
37	1127.58	pollution	37	0.306	measures	37	0.463	Wuhan
38	1104.61	period	38	0.304	infection	38	0.462	increased
39	1095.2	countries	39	0.304	table	39	0.462	research
40	1079.93	infected	40	0.304	transmission	40	0.459	population
41	1073.05	analysis	41	0.302	Wuhan	41	0.454	March
42	1035.12	CI	42	0.301	significant	42	0.451	related
43	1029.14	emergency	43	0.299	period	43	0.449	studies
44	1022.27	RNA	44	0.296	impact	44	0.448	compared
45	1014.58	impact	45	0.292	e	45	0.448	epidemic
46	1008.79	in	46	0.292	epidemic	46	0.443	using
47	997.92	variables	47	0.289	increased	47	0.437	based
48	991.82	patients	48	0.287	population	48	0.437	associated
49	975.98	PM	49	0.287	march	49	0.434	total
50	958.18	results	50	0.285	related	50	0.433	case
51	942.67	infectious	51	0.283	research	51	0.431	increase
52	935.49	factors	52	0.283	studies	52	0.429	observed
53	933.81	air	53	0.281	compared	53	0.428	low
54	896.63	severe	54	0.278	associated	54	0.428	control
55	894.87	respondents	55	0.273	observed	55	0.426	severe
56	888.35	wastewater	56	0.273	using	56	0.422	February
57	869.62	concentration	57	0.272	based	57	0.421	affected
58	866.76	depression	58	0.271	severe	58	0.416	current
59	856.5	associated	59	0.270	total	59	0.416	patients
60	850.23	stress	60	0.270	affected	60	0.414	higher

※ COVID-19: Corona Virus Disease 2019; CoV: Corona Virus; CI: Confinement Index; PM: Particulate Matter; RNA: Ribonucleic Acid; SARS: Severe Acute Respiratory Syndrome.

Step 10. Rank the keywords based on their aggregated values.

Based on each keyword's aggregated value, the researchers re-ranked the keyword list (see Table 6) to form the ultimate optimized keyword list.

Step 11. Offer the results to decision makers as reference data for future NLP applications.

The results of the ultimate optimized keyword list can be integrated with the complete evaluation results from the experts to provide a more complete benchmark for defining critical lexical units, thereby improving the efficiency and accuracy of NLP.

4.5. Comparison and Discussion

To enhance the accuracy of the corpus evaluation results, a corpus assessment approach must be able to compute multiple parameters at the same time and consider the relative importance between different parameters. However, the traditional keyword-list-

generating approach [10] only uses the likelihood ratio method [11] to determine and rank keywords in the target corpus, which is a deficiency of corpus assessment [2,3,10,22]. Thus, to optimize the aforementioned issues, this paper proposed an extended AHP-based corpus assessment approach that integrated the likelihood ratio method, the corpus optimization approach, and the AHP method to refine corpus data, simultaneously handle multiple parameters, and consider the relative importance between different parameters for accurately evaluating keywords. COVID-19-related research articles (N = 53) from the environmental science discipline were adopted as the target corpus and used as an empirical example to verify the proposed approach.

This paper compared three approaches from three perspectives: (1) corpus optimization; (2) considering multiple parameters simultaneously; and (3) considering the relative importance between different parameters to highlight the contributions of the proposed approach (see Table 7).

Table 7. Comparison of the optimization features between three approaches.

Research Method	Optimization Feature		
	Corpus Optimization	**Considering Multiple Parameters Simultaneously**	**Considering the Relative Importance between Different Parameters**
The traditional keyword-list-generating approach [10]	No	No	No
The EW method [47]	No	Yes	No
The proposed extended AHP-based corpus assessment approach	Yes	Yes	Yes

Firstly, for corpus optimization, Table 6 indicates that function words, such as the, and, of, and in, appeared on the keyword lists generated by the traditional keyword-list-generating approach [10] and the EW method [47]. Due to function words being critical elements to form meaningful sentences, those tokens usually occupy over 40% of the corpus data. If the function words are not eliminated beforehand, the likelihood ratio method [11] will consider them as keywords because their extremely high frequency values will disguise the keyness computation results. Once the function words are included in the keyword list, content words that may be true keywords will be excluded; thus, causing bias in the computation results. Before entering the algorithm computation process, the proposed approach adopted the corpus optimization approach to eliminate function words and meaningless letters, to enhance the computation accuracy.

Secondly, when considering multiple parameters simultaneously, it is insufficient to use the traditional keyword-list-generating approach [10], as it is based on only one parameter (the keyness) to rank keywords. To make the evaluation results approach uncontroversial, the EW method [47] and the proposed approach were used to simultaneously take three parameters (i.e., the keyness, frequency, and range) into consideration, and each keyword's aggregated value was used to re-rank the keyword list.

Finally, in consideration of the relative importance of different parameters, the researchers soon discovered the major problem of the EW method [47]. Although the EW method could consider the three parameters at the same time, the importance between the three parameters would be considered as equal, and the relative importance between the parameters would not be confirmed. To compensate for this deficiency, the proposed approach integrated the AHP method [12] to calculate the relative weights of each parameter and identify the relative importance between parameters. After using the AHP method to calculate the experts' evaluation scores, the researchers discovered that the relative weights of the keyness, frequency, and range were 0.195, 0.278, and 0.528, respectively, which were not equal. The derived implications of the unequal relative weights indicated that, after

generating the keyword list, the experts wanted to identify the most widely- and frequently-used keywords in the target corpus; hence, their assessment results determined the relative importance of the three parameters as range > frequency > keyness.

In summation, to handle the single-parameter evaluation deficiency of keyword ranking and optimize the traditional corpus-based assessment approach, the proposed extended AHP-based corpus assessment approach was able to exclude function words and meaningless letters, simultaneously compute multiple parameters, and consider the relative importance between different parameters.

5. Conclusions

The algorithms used for today's corpus analytical tasks are gradually being used for multiple-parameter and high-precision analysis. Keyword ranking is one of the critical techniques of corpus analysis to extract key information from the target corpus. COVID-19 is no longer limited to medical or public health issues, but also impacts other issues such as ecological systems, environmental science, and economics. High-precision COVID-19 corpus data analysis can enhance the efficiency of knowledge discovery for this novel disease. However, the traditional keyword-list-generating approach [10] is only based on the likelihood ratio method [11] to compute the tokens' keyness values, to determine and rank keywords. Thus, there is still room for optimization, as it does not automatically eliminate function words and meaningless letters or conduct multiple-parameter evaluations. Moreover, when the EW method [47] is adopted as the multiple-parameter evaluation approach to re-rank keywords, it cannot eliminate function words and meaningless letters or confirm the relative importance between each parameter to obtain more accurate results. Hence, this paper proposed an extended AHP-based corpus assessment approach to compensate the aforementioned problems, by optimizing the target corpus and conducting a multiple-parameter evaluation by using the relative weights of the parameters to determine the keywords' actual importance levels.

The proposed extended AHP-based corpus assessment approach has the following significant contributions. First, the proposed approach uses a machine-based approach to eliminate function words and meaningless letters for optimizing the target corpus, thereby further enhancing the accuracy of the followed algorithms' computations. Second, the proposed approach uses the AHP method to fully consider the relative weights of three parameters to provide calculation results with higher accuracy. Third, the proposed approach is a corpus-based assessment approach based on the perspectives of multiple parameters, which differs from traditional approaches that are based on the perspective of a single parameter. The optimized keyword list represents that each keyword has been fully considered as being truly important, which enhances the accuracy of keyword application. Fourth, the traditional corpus-based assessment approaches that were mentioned in this paper were just special cases of the proposed extended AHP-based corpus assessment approach. In addition to optimizing the traditional approaches, the proposed approach also makes itself more generally applicable. Once the keyword ranking results are optimized and improved by the proposed method, the important and domain-oriented words (i.e., keywords) will be ranked in the ahead ranks, which will improve users' IR efficiency through the corpus software. In other words, without the optimization, the ahead ranks will show the words of grammar, or those which are meaningless, unimportant, or even unrelated to the domain, which will rely on human's tasks to filter the unnecessary information. The target corpus (i.e., COVID-19 corpus data) used in this paper was only a specific case for verification and highlighted the advantages of the proposed approach; namely, any corpus data can be processed and optimized by the proposed approach.

This paper has some limitations for future researchers to overcome. With today's advanced information technology, future studies can be based on the proposed approach to develop other algorithms for optimizing corpus analytical tasks, such as the Term Frequency-Inverse Document Frequency (TF-IDF) method, high-precision NLP techniques e.g., [48,49], multiple-parameter evaluation models, and novel corpus programs.

Author Contributions: Conceptualization, L.-C.C. and K.-H.C.; methodology, L.-C.C.; software, L.-C.C.; validation, K.-H.C.; writing—original draft preparation, L.-C.C.; writing—review and editing, K.-H.C.; funding acquisition, K.-H.C. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank the National Science and Technology Council, Taiwan, for financially supporting this research under Contract Nos. MOST 111-2221-E-145-003 and NSTC 111-2221-E-145-003.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

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

References

1. Anthony, L. AntConc (Version 3.5.8), Corpus Software. 2019. Available online: https://www.laurenceanthony.net/software/antconc/ (accessed on 1 January 2022).
2. Otto, P. Choosing specialized vocabulary to teach with data-driven learning: An example from civil engineering. *Engl. Specif. Purp.* **2021**, *61*, 32–46. [CrossRef]
3. Poole, R. A corpus-aided study of stance adverbs in judicial opinions and the implications for English for legal purposes instruction. *Engl. Specif. Purp.* **2021**, *62*, 117–127. [CrossRef]
4. Akhtaruzzaman, M.; Boubaker, S.; Sensoy, A. Financial contagion during COVID-19 crisis. *Financ. Res. Lett.* **2021**, *38*, 101604. [CrossRef]
5. Antonakis, J. Leadership to defeat COVID-19. *Group Process Intergroup Relat.* **2021**, *24*, 210–215. [CrossRef]
6. Chilamakuri, R.; Agarwal, S. COVID-19: Characteristics and therapeutics. *Cells* **2021**, *10*, 206. [CrossRef]
7. Aydin, S.; Nakiyingi, B.A.; Esmen, C.; Guneysu, S.; Ejjada, M. Environmental impact of coronavirus (COVID-19) from Turkish perceptive. *Environ. Dev. Sustain.* **2021**, *23*, 7573–7580. [CrossRef]
8. Sahraei, M.A.; Kuskapan, E.; Codur, M.Y. Public transit usage and air quality index during the COVID-19 lockdown. *J. Environ. Manag.* **2021**, *286*, 112166. [CrossRef]
9. SanJuan-Reyes, S.; Gomez-Olivan, L.M.; Islas-Flores, H. COVID-19 in the environment. *Chemosphere* **2021**, *263*, 127973. [CrossRef]
10. Ross, A.S.; Rivers, D.J. Discursive Deflection: Accusation of "fake news" and the spread of mis- and disinformation in the Tweets of president Trump. *Soc. Media Soc.* **2018**, *4*, 2056305118776010. [CrossRef]
11. Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* **1993**, *19*, 61–74.
12. Saaty, T.L. *The Analytic Hierarchy Process*; McGraw-Hill: New York, NY, USA, 1980.
13. Rezaei, A.; Tahsili, S. Urban vulnerability assessment using AHP. *Adv. Civ. Eng.* **2018**, *2018*, 2018601. [CrossRef]
14. Ristanovic, V.; Primorac, D.; Kozina, G. Operational risk management using multi-criteria assessment (AHP model). *Teh. Vjesn.* **2021**, *28*, 678–683.
15. Chang, K.H. Generalized multi-attribute failure mode analysis. *Neurocomputing* **2016**, *175*, 90–100. [CrossRef]
16. Chang, K.H.; Chang, Y.C.; Chain, K.; Chung, H.Y. Integrating soft set theory and fuzzy linguistic model to evaluate the performance of training simulation systems. *PLoS ONE* **2016**, *11*, e0162092. [CrossRef] [PubMed]
17. Durao, L.F.C.S.; Carvalho, M.M.; Takey, S.; Cauchick-Miguel, P.A.; Zancul, E. Internet of Things process selection: AHP selection method. *Int. J. Adv. Manuf. Technol.* **2018**, *99*, 2623–2634. [CrossRef]
18. Han, Y.; Wang, Z.H.; Lu, X.M.; Hu, B.W. Application of AHP to road selection. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 86. [CrossRef]
19. Saaty, T.L. Rank from comparisons and from ratings in the analytic hierarchy/network processes. *Eur. J. Oper. Res.* **2006**, *168*, 557–570. [CrossRef]
20. Chang, K.H.; Chain, K.; Wen, T.C.; Yang, G.K. A novel general approach for solving a supplier selection problem. *J. Test. Eval.* **2016**, *44*, 1911–1924. [CrossRef]
21. Chen, L.C.; Chang, K.H.; Chung, H.Y. A novel statistic-based corpus machine processing approach to refine a big textual data: An ESP case of COVID-19 news reports. *Appl. Sci.* **2020**, *10*, 5505. [CrossRef]
22. Chen, L.C.; Chang, K.H. A novel corpus-based computing method for handling critical word ranking issues: An example of COVID-19 research articles. *Int. J. Intell. Syst.* **2021**, *36*, 3190–3216. [CrossRef]
23. Scott, M. PC analysis of key words-and key key words. *System* **1997**, *25*, 233–245. [CrossRef]
24. Brookes, G. 'Lose weight, save the NHS': Discourses of obesity in press coverage of COVID-19. *Crit. Discourse Stud.* **2021**, *19*, 629–647. [CrossRef]
25. Ong, T.T.; McKenzie, R.M. The language of suffering: Media discourse and public attitudes towards the MH17 air tragedy in Malaysia and the UK. *Discourse Commun.* **2019**, *13*, 562–580. [CrossRef]
26. Gautam, S. The influence of COVID-19 on air quality in India: A boon or inutile. *B. Environ. Contam. Tox.* **2020**, *104*, 724–726. [CrossRef]
27. Gope, S.; Dawn, S.; Das, S.S. Effect of COVID-19 pandemic on air quality: A study based on Air Quality Index. *Environ. Sci. Pollut. R.* **2021**, *28*, 35564–35583. [CrossRef]

28. Liu, Q.; Harris, J.T.; Chiu, L.S.; Sun, D.L.; Houser, P.R.; Yu, M.Z.; Duffy, D.Q.; Little, M.M.; Yang, C.W. Spatiotemporal impacts of COVID-19 on air pollution in California, USA. *Sci. Total Environ.* **2021**, *750*, 141592. [CrossRef]
29. Yao, Y.; Pan, J.H.; Liu, Z.X.; Meng, X.; Wang, W.D.; Kan, H.D.; Wang, W.B. Ambient nitrogen dioxide pollution and spreadability of COVID-19 in Chinese cities. *Ecotox. Environ. Safe* **2021**, *208*, 111421. [CrossRef] [PubMed]
30. Lee, P.S.; Sung, Y.H.; Wu, C.C.; Ho, L.C.; Chiou, W.B. Using episodic future thinking to pre-experience climate change increases pro-environmental behavior. *Environ. Behav.* **2020**, *52*, 60–81. [CrossRef]
31. Baloch, S.; Baloch, M.A.; Zheng, T.L.; Pei, X.F. The coronavirus disease 2019 (COVID-19) pandemic. *Environ. Dev. Sustain.* **2020**, *250*, 271–278. [CrossRef] [PubMed]
32. Yi, H.S.; Ng, S.T.; Farwin, A.; Low, A.P.T.; Chang, C.M.; Lim, J. Health equity considerations in COVID-19: Geospatial network analysis of the COVID-19 outbreak in the migrant population in Singapore. *J. Travel. Med.* **2021**, *28*, taaa159. [CrossRef] [PubMed]
33. Huang, X.Y.; Wei, F.X.; Hu, L.; Wen, L.J.; Chen, K. Epidemiology and clinical characteristics of COVID-19. *Arch. Iran. Med.* **2020**, *23*, 268–271. [CrossRef] [PubMed]
34. Klopfenstein, T.; Kadiane-Oussou, N.J.; Toko, L.; Royer, P.Y.; Lepiller, Q.; Gendrin, V.; Zayet, S. Features of anosmia in COVID-19. *Med. Maladies Infect.* **2020**, *50*, 436–439. [CrossRef] [PubMed]
35. Pascarella, G.; Strumia, A.; Piliego, C.; Bruno, F.; Del Buono, R.; Costa, F.; Scarlata, S.; Agro, F.E. COVID-19 diagnosis and management: A comprehensive review. *J. Intern. Med.* **2020**, *288*, 192–206. [CrossRef] [PubMed]
36. World Health Organization (WHO). WHO Coronavirus (COVID-19) Dashboard. 2021. Available online: https://covid19.who.int/ (accessed on 1 May 2023).
37. Othman, H.; Bouslama, Z.; Brandenburg, J.T.; da Rocha, J.; Hamdi, Y.; Ghedira, K.; Srairi-Abid, N.; Hazelhurst, S. Interaction of the spike protein RBD from SARS-CoV-2 with ACE2: Similarity with SARS-CoV, hot-spot analysis and effect of the receptor polymorphism. *Biochem. Biophys. Res. Commun.* **2020**, *527*, 702–708. [CrossRef]
38. Wibmer, C.K.; Ayres, F.; Hermanus, T.; Madzivhandila, M.; Kgagudi, P.; Oosthuysen, B.; Lambson, B.E.; de Oliveira, T.; Vermeulen, M.; van der Berg, K.; et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **2021**, *27*, 622–625. [CrossRef]
39. World Health Organization (WHO). SARS-CoV-2 Variants, Working Definitions and Actions Taken. 2021. Available online: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed on 1 May 2023).
40. Forni, G.; Mantovani, A. COVID-19 vaccines: Where we stand and challenges ahead. *Cell Death Differ.* **2021**, *28*, 626–639. [CrossRef]
41. Berman, J.D.; Ebisu, K. Changes in US air pollution during the COVID-19 pandemic. *Sci. Total Environ.* **2020**, *739*, 139864. [CrossRef]
42. Bashir, M.F.; Ma, B.J.; Shahzad, L. A brief review of socio-economic and environmental impact of COVID-19. *Air Qual. Atmos. Health* **2020**, *13*, 1403–1409. [CrossRef]
43. Srivastava, A. COVID-19 and air pollution and meteorology-an intricate relationship: A review. *Chemosphere* **2021**, *263*, 128297. [CrossRef] [PubMed]
44. Travaglio, M.; Yu, Y.Z.; Popovic, R.; Selley, L.; Leal, N.S.; Martins, L.M. Links between air pollution and COVID-19 in England. *Environ. Pollut.* **2021**, *268*, 115859. [CrossRef] [PubMed]
45. Saadat, S.; Rawtani, D.; Hussain, C.M. Environmental perspective of COVID-19. *Sci. Total Environ.* **2020**, *728*, 138870. [CrossRef]
46. Cusmariu, A. A proof of the arithmetic mean geometric mean inequality. *Am. Math. Mon.* **1981**, *88*, 192–194. [CrossRef]
47. Chunaev, P. Interpolation by generalized exponential sums with equal weights. *J. Approx. Theory* **2020**, *254*, 105397. [CrossRef]
48. Stefano, M.; Siino, M.; Garbo, G. Improving Irony and Stereotype Spreaders Detection using Data Augmentation and Convolutional Neural Network. *CEUR Workshop Proc.* **2022**, *3180*, 2585–2593.
49. Siino, M.; Ilenia, T.; Marco, L.C. T100: A modern classic ensemble to profile irony and stereotype spreaders. *CEUR Workshop Proc.* **2022**, *3180*, 2666–2674.