sustainability

MDPI

Article

An Improved Corpus-Based NLP Method for Facilitating Keyword Extraction: An Example of the COVID-19 Vaccine Hesitancy Corpus

Liang-Ching Chen [1,2]

[1] Department of Foreign Languages, R.O.C. Military Academy, Kaohsiung 830, Taiwan; jokiceman@gmail.com
[2] Institute of Education, National Sun Yat-sen University, Kaohsiung 804, Taiwan

Abstract: In the current COVID-19 post-pandemic era, COVID-19 vaccine hesitancy is hindering the herd immunity generated by widespread vaccination. It is critical to identify the factors that may cause COVID-19 vaccine hesitancy, enabling the relevant authorities to propose appropriate interventions for mitigating such a phenomenon. Keyword extraction, a sub-field of natural language processing (NLP) applications, plays a vital role in modern medical informatics. When traditional corpus-based NLP methods are used to conduct keyword extraction, they only consider a word's log-likelihood value to determine whether it is a keyword, which leaves room for concerns about the efficiency and accuracy of this keyword extraction technique. These concerns include the fact that the method is unable to (1) optimize the keyword list by the machine-based approach, (2) effectively evaluate the keyword's importance level, and (3) integrate the variables to conduct data clustering. Thus, to address the aforementioned issues, this study integrated a machine-based word removal technique, the i10-index, and the importance–performance analysis (IPA) technique to develop an improved corpus-based NLP method for facilitating keyword extraction. The top 200 most-cited Science Citation Index (SCI) research articles discussing COVID-19 vaccine hesitancy were adopted as the target corpus for verification. The results showed that the keywords of Quadrant I (n = 98) reached the highest lexical coverage (9.81%), indicating that the proposed method successfully identified and extracted the most important keywords from the target corpus, thus achieving more domain-oriented and accurate keyword extraction results.

Keywords: COVID-19 vaccine hesitancy; keyword extraction; natural language processing (NLP); medical informatics; corpus; i10-index; importance–performance analysis (IPA) method

Citation: Chen, L.-C. An Improved Corpus-Based NLP Method for Facilitating Keyword Extraction: An Example of the COVID-19 Vaccine Hesitancy Corpus. *Sustainability* **2023**, *15*, 3402. https://doi.org/10.3390/su15043402

Academic Editors: Andreas Kanavos and Xuesong (Andy) Gao

Received: 13 January 2023
Revised: 28 January 2023
Accepted: 6 February 2023
Published: 13 February 2023

1. Introduction

With the advancement of information and communication technology (ICT), natural language processing (NLP) has come to play an important role in modern medical informatics, because it is used to determine whether artificial intelligence (AI) can be integrated into the medical informatics field [1]. NLP usually involves the application of algorithms and computational techniques to convert structured or unstructured big textual data into knowledge by extracting and analyzing specific signals, which usually point to keywords. Keyword extraction is a sub-field of NLP that has attracted increasing attention in this information explosion era because it involves the identification of the most relevant terms and representations from a given big textual dataset in a timely manner [2–4].

When it comes to corpus-based research, knowledge acquisition from a specific domain is usually connected to keyword extraction, as keywords act as a pipeline for extracting key information from the target corpus. Traditional corpus-based NLP methods usually rely on the log-likelihood algorithm, which was first proposed by Dunning in 1993 [5] as a statistic-based information retrieval technique to distill keywords from a target corpus. However, although Dunning's log-likelihood algorithm has become a solid foundation

for keyword extraction in modern corpus software, e.g., [6,7], the traditional methods of keyword extraction present several limitations, whose solutions require novel algorithms or computational techniques. This study consulted prior research, e.g., [8–11], to identify the following limitations of traditional methods. Firstly, function words and general-purpose words exist on the keyword list, which interfere with the accuracy of keyword analysis, as such words are usually irrelevant to domain-specific knowledge; moreover, the removal process often relies on manual tasks, which makes the optimization process time-consuming and inefficient. Second, when evaluating the keyword's importance level, the traditional methods only take account of a single variable (i.e., log-likelihood, frequency, or range) to sort the keywords, thus failing to achieve integration among the variables. However, it is controversial to consider high-frequency keywords as important keywords without taking the keywords' range variables into consideration. As Chen and Chang (2021) [12] stated, if a high-frequency word is too concentrated in a small number of sub-corpora (i.e., it has a low range value), its importance level should not be proportional to its high frequency value. Thirdly, as already mentioned, the traditional methods only take account of a single variable to rank the keywords, and so they neither consider multiple variables simultaneously nor conduct data clustering. When determining the so-called important keywords, the traditional methods usually rely on the researchers' arbitrary decisions to decide on the number of important keywords for conducting knowledge acquisition, which makes the analytical results variable and hard to replicate, e.g., [13–15]. The purpose of this study was to develop an improved corpus-based NLP method to address the aforementioned issues and enhance the efficiency and accuracy of keyword extraction. Furthermore, when emerging diseases are encountered, highly efficient and accurate keyword extraction and knowledge acquisition would also directly benefit epidemic prevention and control.

Since the outbreak of the COVID-19 pandemic in 2019, controlling epidemics has not been limited to medical or public health issues but has also included issues such as the efficiency of acquiring information and knowledge about the disease and the identification of its authenticity [16–19]. In the current COVID-19 post-pandemic era, many epidemiologists and public health experts believe that implementing widespread vaccination to generate herd immunity is the best way to fight and control epidemics and allow people to return to their normal pre-pandemic lives [20–23]. However, the COVID-19 vaccine hesitancy phenomenon is hindering this ideal anti-epidemic mechanism [21,24–27]. As the World Health Organization (WHO) stated, vaccine hesitancy refers to a situation wherein vaccination services are available but people still delay or refuse their vaccination. Moreover, the WHO also marked vaccine hesitancy as one of the top 10 global health problems [28]. Thus, identifying the potential factors that cause this phenomenon is critical, allowing the relevant authorities to react earlier and propose proper interventions to mitigate this issue. In Khairat et al.'s (2022) [25] research, the most common reasons behind vaccine hesitancy were found to include a mistrust of COVID-19 vaccines (55%), worries about the adverse side effects of vaccines (48%), and distrust in the government (46%). The authors also suggested that the existence of knowledge gaps should be closed by policymakers to mitigate the public's negative perceptions of vaccines. Accordingly, addressing COVID-19 vaccine hesitancy may not be entirely related to medical issues; in fact, it may be more related to the efficiency of information transmission and the authenticity of the received information [29,30]. That is, efficient and accurate knowledge acquisition pertaining to the disease could also be an important mechanism to facilitate the adoption of epidemic prevention and control measures by the public [31,32].

COVID-19 vaccine hesitation has made us aware of the importance or potential influence of keyword extraction. Although the traditional corpus-based NLP methods have been successfully used for processing medical-related corpora in the past (e.g., [33–35]), because of the above-mentioned limitations, their NLP and keyword extraction results show room for optimization and improvement. Thus, to address the aforementioned limitations, in this pioneering research, the author adopted the function word list [36] and a new general service list (NGSL) [37] as the exclusion baseline and used the machine-based word removal

technique to exclude unnecessary words. Next, two algorithms from the field of bibliometrics and information management, the i10-index [38,39] and the importance-performance analysis (IPA) method [40], were introduced to evaluate the keywords' substantial importance level and conduct keyword data clustering, with the goal of developing an improved corpus-based NLP method for facilitating keyword extraction. The top 200 most cited Science Citation Index (SCI) research articles related to COVID-19 vaccine hesitancy from the Web of Science (WOS) database were collected as the target corpus for verifying the proposed method.

The contributions of the present research are as follows: (1) developing a machine-based word removal technique to exclude meaningless and general-purpose words for optimizing the keyword list; (2) introducing the i10-index to integrate a keyword's frequency and range values for evaluating its importance level; (3) introducing the revised IPA method to integrate variables (i.e., log-likelihood and the i10-index) and conduct data clustering for replacing researchers' arbitrary decisions on determining the number of important keywords in traditional methods; and (4) extracting the highly important keywords from Quadrant I that have both a high keyness level and lexical coverage, which would make the results of keyword extraction more domain-oriented and accurate.

The remainder of the present article is as follows. Section 2 reviews the related works. Section 3 describes each step of the proposed method in detail. Section 4 uses a corpus of COVID-19 vaccine hesitancy literature as the target corpus and real-world natural language data to verify the proposed approach. Section 5 shows the verification of the lexical coverage and a comparative analysis. Section 6 is the conclusion.

2. Related Works

2.1. Log-Likelihood

Since the advent of ICT, the machine-based information retrieval and NLP tasks have gradually received more attention. The log-likelihood algorithm proposed by Dunning in 1993 [5] has become a solid foundation for keyword extraction for modern corpus software (e.g., [6,7]). Dunning's (1993) [5] log-likelihood algorithm is a practical measure which was inspired by statistical considerations and can be used in various situations. This statistics-based method is applicable to both large and small textual data and can directly compare the significance of rare and common phenomena. Empirically, Dunning's (1993) [5] log-likelihood algorithm can find a word with higher frequency in a target corpus but a significantly lower frequency in a benchmark corpus, although this word has a high log-likelihood value (i.e., keyness) and is considered to be a keyword of the target corpus. To date, the log-likelihood method is reliable and commonly used. A detailed definition of the log-likelihood method is as follows:

Definition 1 ([5,41]). *If A is a word's frequency in Corpus 1, B is the word's frequency in Corpus 2, C is the total number of words of Corpus 1, and D is the total number of words of Corpus 2, then we can establish the contingency table to clarify the interrelationships between the variables (see Table 1).*

Table 1. The contingency table.

	Corpus 1 (Target Corpus)	**Corpus 2 (Benchmark Corpus)**	**Total**
Frequency of word	A	B	A + B
Frequency of other words	C − A	D − B	C + D − A − B
Total words	C	D	C + D

The log-likelihood value can then be calculated by the following equations:

$$Z_i = \frac{X_i \sum_i Y_i}{\sum_i X_i} \tag{1}$$

$$-2\ln\lambda = 2\sum_i Y_i \ln\left(\frac{Y_i}{Z_i}\right) \tag{2}$$

$$(i = \text{Corpus 1 or Corpus 2})$$

where X_i represents the total words of the ith corpus's data, indicating that $X_1 = C$ and $X_2 = D$; Y_i represents a word's frequency in the ith corpus, indicating that $Y_1 = A$ and $Y_2 = B$.

2.2. The i10-Index

To measure global researchers' academic performance, Google Scholar invented a bibliometric method called the i10-index, which is similar to Hirsch's (2005) [42] h-index, following the general principle that repeats the computation of the index for the same publication set [43]. The i10-index measures an author's number of publications with 10 or more citations. It is a very simple measure and is only used by Google Scholar currently, although it helps us to gauge the substantial contributions of a researcher [44]. A detailed definition of the i10-index is as follows:

Definition 2 ([38,39]). *If the value of the function f is the number of times each article has been cited, and the articles are sorted on the basis of their citations in descending order (see Equation (3)), then we can find how many values of $f(n)$ are equal to or larger than 10 (see Equation (4)), which will be the i10-index value.*

$$\text{i10-index}(f) = \max_n \min(f(1), \ldots, f(n-1)) \tag{3}$$

$$f(n) \geq 10 \tag{4}$$

Here, n is the number of an author's published articles, $\max_n \min(f(1), \ldots, f(n-1))$ indicates that the articles are ranked by their number of citations from maximum to minimum, and $f(n)$ is the number of citations of the nth article.

2.3. IPA Method

The IPA method was firstly proposed by Martilla and James in 1977 [40]. It is a technique used to evaluate the main strengths and weaknesses of an organization's services or products (or value offered) from the perspective of customer satisfaction. Martilla and James (1977) [40] defined customer satisfaction as being about customers' perceptions of whether the quality of an organization's products or services can meet the customers' expectations. Hence, the IPA method was based on two components of the product's or service's attributes, including the product's or service's importance to the customer and the performance of the product or service provided by the organization. This method was used to measure the quantitative results of customer satisfaction surveys.

To date, the IPA method is still considered to be a useful technique for assessing customer satisfaction and developing management strategies, and has been applied in many fields including management and information management e.g., [45–47]. Notably, for facing multiple criteria decision making (MCDM) problems, the IPA method is also used as a decision support tool to integrate multiple variables of the attributes and cluster them on the basis of the relative importance and performance levels among them e.g., [48–50].

A detailed definition of the IPA method is as follows: First, define the x-axis as an attribute's importance, while the y-axis is the attribute's performance. We can then calculate the grand mean of the attributes' importance and the grand mean of the attributes'

performance. These grand mean values are used to draw two referential lines, and the intersection of these two referential lines creates a two-dimensional graph that clusters and allocates all the attributes into four quadrants (see Figure 1).

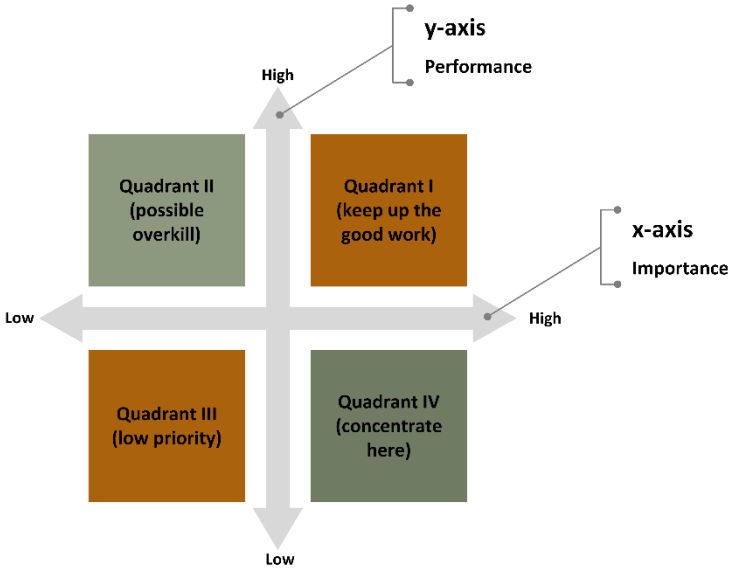


Figure 1. The typical IPA graph.

Depending on the different quadrants, the attributes' major or minor strengths and weaknesses are described as follows [40]:

Quadrant I (keep up the good work): In this quadrant, the attributes are important to customers and the performance provided by an organization is high; hence, the attributes are considered as major strengths and opportunities for achieving or maintaining a competitive advantage.

Quadrant II (possible overkill): In this quadrant, the attributes are less important to customers, but an organization provides high levels of performance; hence, the organization should relocate its resources to other quadrants in need of improved performance.

Quadrant III (low priority): In this quadrant, the attributes are unimportant to customers and the performance provided by the organization is low; hence, they do not need to be improved instantly.

Quadrant IV (concentrate here): In this quadrant, the attributes are important to customers, but the performance provided by the organization are low; hence, these attributes are considered as major weaknesses, and it is necessary to improve them immediately.

3. The Proposed Method

The field of NLP or computational linguistics has been propelled forward by advanced algorithms [51]. NLP industries involve the engineering of information or computational models for solving practical problems for computers in understanding and processing human languages. For keyword extraction, the traditional corpus-based NLP methods are only based on the words' log-likelihood values to determine the keywords and sort them, which leaves some potential concerns about the accuracy and efficiency of NLP (e.g., [8–11]). Such concerns include (1) the inability to optimize the keyword list in a machine-based way, (2) the inability to evaluate the keyword's importance level effectively, and (3) the inability to integrate the variables to conduct data clustering. To solve these issues, this study proposed an improved corpus-based NLP method to remove unnecessary words in a machine-based way, use the i10-index to integrate the keywords' frequency and range values for evaluating their importance level, use a revised IPA method to integrate the keywords' log-likelihood and i10-index values. Based on these two variables, the IPA method was used to cluster and allocate the keywords to different quadrants, then distill

the highly important keywords from Quadrant I for enhancing the efficiency and accuracy of keyword extraction (see Figure 2). Detailed descriptions of each step are given below.

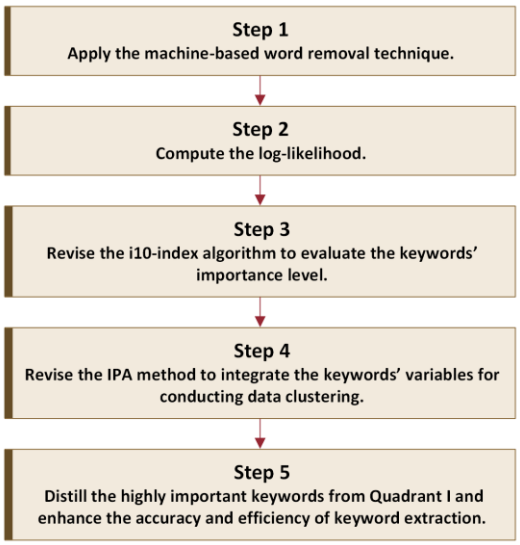


Figure 2. The flow chart of the proposed method.

Step 1. Apply the machine-based word removal technique.

Although algorithms can quantify natural languages, from the perspective of English linguistics, sentences and texts are often filled with large amounts of function words and general-purpose words. Such words usually have extremely high frequency values, which, in turn, affect the calculation results of the log-likelihood method. Pojanapunya and Todd (2018) [52] mentioned that the calculation of the log-likelihood requires four parameters, including the respective frequency of a word in the target corpus and in the benchmark corpus, the total number of running words of the target corpus, and the total number of running words of the benchmark corpus. Therefore, when function words or general-purpose words have extremely high frequency values, they will have stronger statistical significance and a larger effective size after conducting statistical tests, which will make the NLP tool identify such words as keywords in the process of keyword extraction. This phenomenon can be defined as the misjudgment of keywords because keywords should be the words that reflect the characteristics or patterns of a certain context rather than function words and general-purpose words.

To make the keyword calculation results of the log-likelihood algorithm align more with practical needs, after inputting the target corpus into AntConc 4.1.3 [6], this step adopted the function word list [36] and NGSL [37] as the word exclusion baseline, then utilized Chen et al.'s (2020) [36] lexical filtering method in collaboration with AntConc 4.1.3's lexical filtering function to remove the high-frequency function words and general-purpose words from the target corpus in a machine-based and systematic way.

Step 2. Compute the log-likelihood.

This study adopted the sample data released by the Corpus of Contemporary American English (COCA) in 2021 as the benchmark corpus. In order to effectively highlight the word domain features of the target corpus from the keyword list, the researcher used COCA's nonacademic contexts including blogs, fiction, magazines, news, spoken texts, TV/movies, and web pages, which contain 8,366,198 running words. After the author inputted the benchmark corpus into AntConc 4.1.3, the software used Equations (1) and (2) to compute the words' log-likelihood value, then sorted the words based on their log-likelihood value and generated a keyword list.

Step 3. Revise the i10-index algorithm to evaluate the keywords' importance level.

The traditional i10-index is used to evaluate a researcher's academic contributions by calculating number of publications with 10 or more citations, which takes both the volume

of articles and the number of citations of the respective article into account. To make traditional i10-index suitable for assessing the keywords' importance level, its variables were revised but the original logic was not changed.

From the perspective of corpus linguistics, lexical coverage is an important referential indicator for determining a word's importance level. A word's lexical coverage refers to its proportion within the target corpus, which involves the computation of a word's frequency and range variables [53–55]. Therefore, a revised i-10 index was utilized to integrate these two variables (i.e., frequency and range) for obtaining a value that can represent a keyword's lexical coverage, and this integrated value was taken as the benchmark for evaluating the importance level of the keyword. The detailed definition of the method of computing a keyword's i10-index value is as follows:

Definition 3 ([39]). *If the value of function k is a keyword's frequency in a text (i.e., the sub-corpus data) and the keyword's respective frequency is sorted in descending order (see Equation (5)), then the number of values of $k(n)$ that are equal to or larger than 10 (see Equation (6)), namely, how many texts contain at least 10 occurrences or more of a keyword, is the keyword's i10-index value.*

$$\text{i10-index}(k) = \max_n \min(k(1), \ldots, k(n-1)) \tag{5}$$

$$k(n) \geq 10 \tag{6}$$

Here, n is the number of texts, $\max_n \min(k(1), \ldots, k(n-1))$ indicates that a keyword's respective frequency is ranked from maximum to minimum, and $k(n)$ is the keyword's frequency in the nth text.

After a keyword list was generated from AntConc 4.1.3, this step then used Equations (5) and (6) to compute each keyword's i10-index value.

Step 4. Revise the IPA method to integrate the keywords' variables for conducting data clustering.

According to the original definition of IPA method, the *importance* (on the x-axis) represents the importance of the product or service provided by an organization to customers, while the *performance* (on the y-axis) represents the performance of the product or service provided by an organization. To make the IPA method suitable for use in distilling the highly important keywords, this step modified the x-axis to the i10-index because the i10-index can represent a keyword's lexical coverage, and the lexical coverage points to its importance level; in addition, the y-axis was revised to the log-likelihood because the log-likelihood represents a keyword's keyness level, which is a performance indicator for confirming whether a word can be defined as a keyword.

Before we constructed the revised IPA graph, because the scales of the i10-index and the log-likelihood values are quite different, a normalization process needed to be carried out. Let us assume that ω represents a variable value of a keyword, while ω_{ix} and ω_{iy} are the i10-index and the log-likelihood value of the ith keyword, respectively. The values of ψ_{ix} and ψ_{iy} are the normalizations of ω_{ix} and ω_{iy}, defined as follows [40]:

$$\psi_{ix} = \frac{\omega_{ix}}{\omega_x^{max}} \text{ and } \psi_{iy} = \frac{\omega_{iy}}{\omega_y^{max}} \tag{7}$$

where x = i10-index and y = log-likelihood, $i = 1, 2, 3, \ldots n$.

Once the normalization process has finished, the respective grand mean scores of the keywords' i10-index and log-likelihood values are computed. The grand mean scores of the keywords' i10-index and the log-likelihood values play the roles of the x and y axes' base lines for effectively dividing the revised IPA graph into four quadrants, and thus, the keywords' i10-index and the log-likelihood values need to be standardized. For example, the ith keyword's standardized value of the i10-index (ϕ_{ix}) is the normalized value of its i10-index (ψ_{ix}) minus the grand mean of the normalized i10-index values ($\overline{\psi_x}$).

The process is as follows [56]:

$$\phi_{ix} = \psi_{ix} - \overline{\psi_x} \text{ and } \phi_{iy} = \psi_{iy} - \overline{\psi_y} \quad (8)$$

where x = i10-index and y = log-likelihood, $i = 1, 2, 3, \ldots n$.

Once the standardized values of ϕ_{ix} and ϕ_{iy} have been calculated, all the keywords are marked on the revised IPA graph on the basis of their ϕ_{ix} and ϕ_{iy} values (see Figure 3); in other words, the process is based on the results of the keywords' importance level and their keyness (performance) for clustering the data.

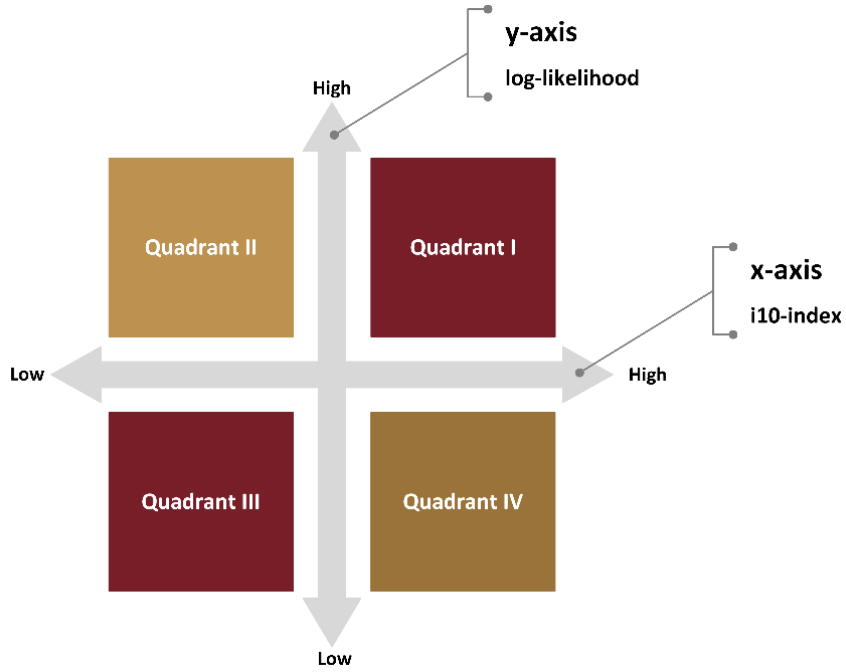


Figure 3. The revised IPA graph.

Step 5. Distill the highly important keywords from Quadrant I and enhance the accuracy and efficiency of keyword extraction.

The revised IPA graph is divided into four quadrants by the intersection of the x-axis and y-axis; moreover, all keywords are clustered and allocated into four quadrants. Based on the original definitions of the IPA method, attributes in Quadrant I have high performances and are considered to be highly important by customers, and are the major strengths for a achieving competitive advantage. Accordingly, the keywords of Quadrant I have high i10-index and log-likelihood values, indicating that the keywords not only reflect the domain features of the target corpus but also have high lexical coverage. Thus, in this step, the so-called highly important keywords are extracted from Quadrant I (see Figure 4). Furthermore, unlike the traditional corpus-based NLP methods, when extracting the keywords of Quadrant I, the proposed method does not need to rely on the researcher's arbitrary decisions to determine the number of important keywords from the keyword list. To summarize, the proposed method optimizes the keyword list in a machine-based way, evaluates the keywords' importance level, and integrates the variables to conduct data clustering, which enhances the accuracy and efficiency of future keyword analysis tasks.

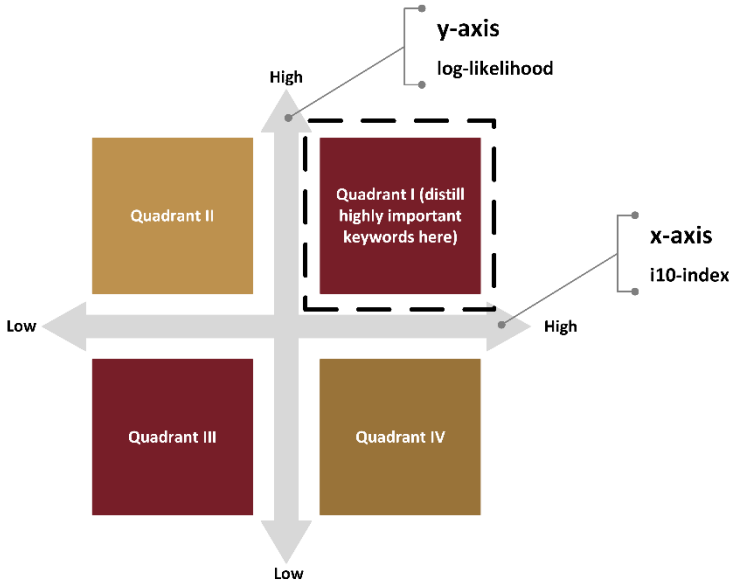


Figure 4. Extracting highly important keywords from Quadrant I.

4. Results

4.1. Overview of the Target Corpus

Since the outbreak of COVID-19 in late 2019, people are still under the shadow of the COVID-19 pandemic. According to the statistical results of WHO, in the fourth season of 2022, the number of confirmed cases worldwide had exceeded 600 million and the number of deaths had reached over 6 million (https://covid19.who.int/ accessed on 15 November 2022). In this long-term battle against COVID-19, many immunologists strongly believe that high vaccine coverage is the most effective way to control the epidemic [20,21]. However, people's vaccine hesitancy may hinder the achievement of herd immunity accompanying high vaccine coverage [25,27]. WHO states that vaccine hesitancy occurs when, even if vaccination services are available, people still delay acceptance of or refuse the vaccination. Moreover, the vaccine hesitancy phenomenon has also been highlighted as one of the top 10 global health problems. As Anakpo and Mishi (2022) [57] claimed, anti-vaccine sentiment and disinformation are major factors causing vaccine hesitancy, while vaccine hesitancy is a type of global health threat and the primary threat handicapping the extinction of COVID-19 through public vaccination. Accordingly, vaccine hesitancy is a critical global public health issue that will impact greatly on most regions. Whether this huge amount of natural language information can be effectively processed will also affect the efficiency of information retrieval, information updates, and public health responses [18,29]. Hence, this study collected textual data related to this topic as the target corpus to verify the proposed method and highlight its contributions in the field of corpus-based NLP.

This study collected the top 200 highly cited research articles related to vaccine hesitancy topic from the Web of Science (WOS) database as the target corpus. WOS is an international pioneering academic database containing high-quality research articles with the Science Citation Index (SCI) and the Social Science Citation Index (SSCI) for each. With today's well-developed internet information, it is sometimes difficult to distinguish genuine and fake information; thus, the main reason for selecting this source was to avoid disputed about the information's authenticity. The target corpus is composed of 16,209 word types and 756,541 tokens (i.e., the total number of running words).

4.2. The Proposed Method

Step 1. Apply the machine-based word removal technique.

After removing the function words and NGSL words, the target corpus decreased its number of word types from 16,209 to 10,922 (−32.62%) and the number of tokens

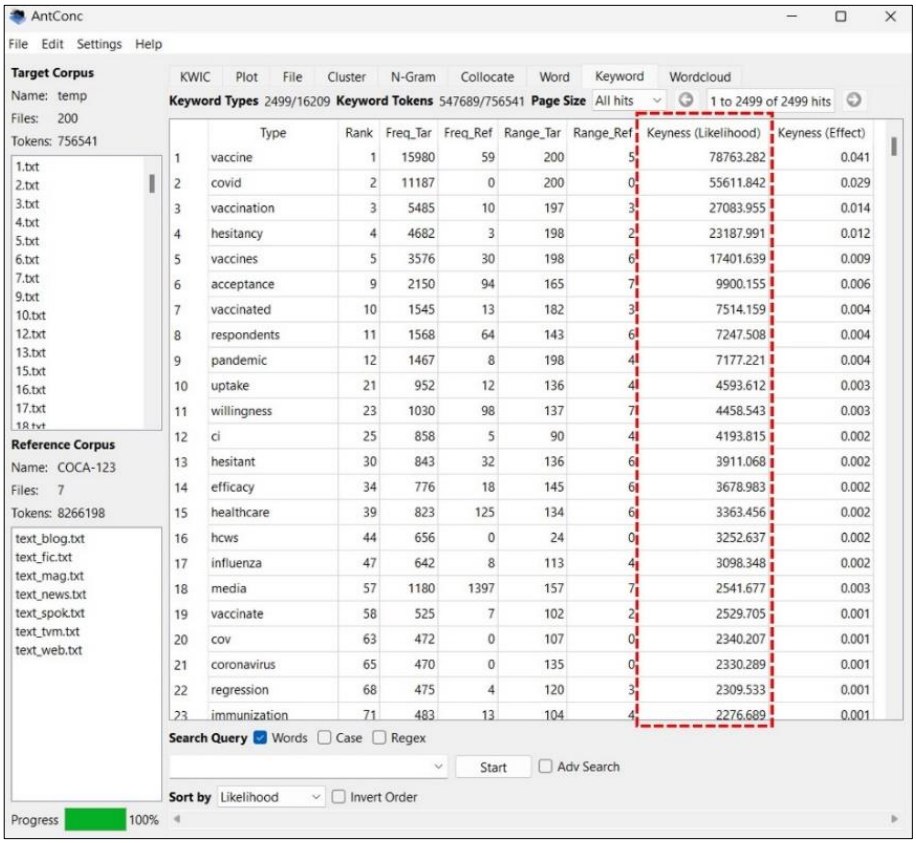
from 756,541 to 142,274 (−81.19%) (see Table 2). Because Chen et al.'s (2020) [36] function word list and Browne et al.'s (2013) [37] NGSL comprise the most common high-frequency general-purpose words, the remaining 10,922 word types were more domain-oriented; in addition, without the interference of extremely high frequency values, the calculated results of log-likelihood were more accurate for extracting the keywords.

Table 2. Data discrepancy after removing the function words and NGSL words.

	Target Corpus	Refined Target Corpus	Data Discrepancy
Word types	16,209	10,922	5287 (−32.62%)
Tokens	756,541	142,274	614,267 (−81.19%)

Step 2. Compute the log-likelihood.

After AntConc 4.1.3 had calculated the log-likelihood values of all words in the refined target corpus, only 1335-word types were considered as keywords and included in the keyword list of AntConc 4.1.3. It can be found from the keyword list interface of AntConc 4.1.3 that the default setting used for sorting keywords is based on the log-likelihood value (see Figure 5). The interface can also sort keywords based on other different variables such as the frequency, range, and so on; however, AntConc 4.1.3 could not simultaneously take all variables into consideration, which caused a lack of integration among the variables. Moreover, it was difficult to effectively evaluate the importance level of the keyword when the software used only a single variable to determine and sort the keywords.



	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	vaccine	1	15980	59	200	5	78763.282	0.041
2	covid	2	11187	0	200	0	55611.842	0.029
3	vaccination	3	5485	10	197	3	27083.955	0.014
4	hesitancy	4	4682	3	198	2	23187.991	0.012
5	vaccines	5	3576	30	198	6	17401.639	0.009
6	acceptance	9	2150	94	165	7	9900.155	0.006
7	vaccinated	10	1545	13	182	3	7514.159	0.004
8	respondents	11	1568	64	143	6	7247.508	0.004
9	pandemic	12	1467	8	198	4	7177.221	0.004
10	uptake	21	952	12	136	4	4593.612	0.003
11	willingness	23	1030	98	137	7	4458.543	0.003
12	ci	25	858	5	90	4	4193.815	0.002
13	hesitant	30	843	32	136	6	3911.068	0.002
14	efficacy	34	776	18	145	6	3678.983	0.002
15	healthcare	39	823	125	134	6	3363.456	0.002
16	hcws	44	656	0	24	0	3252.637	0.002
17	influenza	47	642	8	113	4	3098.348	0.002
18	media	57	1180	1397	157	7	2541.677	0.003
19	vaccinate	58	525	7	102	2	2529.705	0.001
20	cov	63	472	0	107	0	2340.207	0.001
21	coronavirus	65	470	0	135	0	2330.289	0.001
22	regression	68	475	4	120	3	2309.533	0.001
23	immunization	71	483	13	104	4	2276.689	0.001

Figure 5. The keyword list interface of AntConc 4.1.3.

Step 3. Revise the i10-index algorithm to evaluate the keywords' importance level.

Calculating the i10-index value involved simultaneously taking each keyword's frequency and range variables into consideration for evaluating its overall lexical coverage, because lexical coverage brings out its importance level. With the assistance of AntConc 4.1.3, the keywords' frequency values for each text can be displayed through the plot interface and sorted in descending order. The author then used Equations (5) and (6) to compute all the keywords' (n = 1335) i10-index values.

If we take "vaccinate" as an example, the word counts for "vaccinate" was 525 (i.e., frequency = 525) in the target corpus, and "vaccinate" appeared in 102 texts (i.e., range = 102). According to the plot interface of AntConc 4.1.3, the respective frequency values of "vaccinate" in each text were sorted in descending order, which satisfied the conditions of Equation (5). It was then discovered that the word "vaccinate" occurred over 10 times in each of 17 texts, which satisfied the conditions of Equation (6). Thus, the i10-index of "vaccinate" was 17 (see Figure 6).

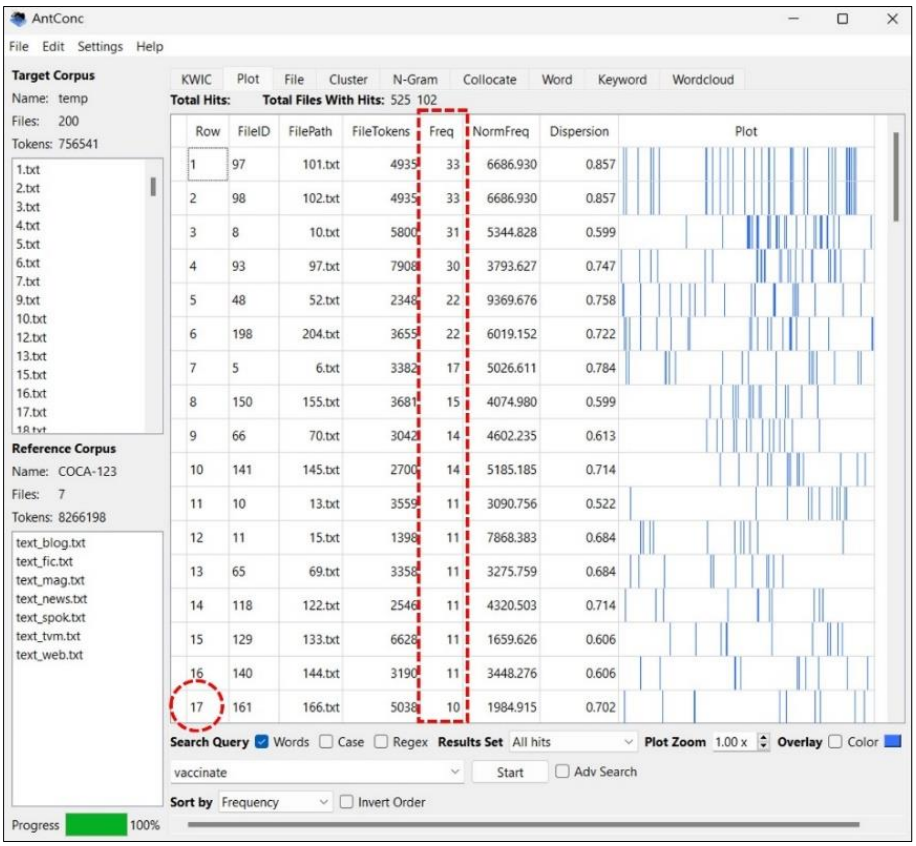


Figure 6. Plot interface of AntConc 4.1.3.

Step 4. Revise the IPA method to integrate the keywords' variables for conducting data clustering.

After the standardized values of the i10-index (ϕ_{ix}) and log-likelihood (ϕ_{iy}) for each keyword had been calculated by Equations (7) and (8), all keywords were marked on the revised IPA graph on the basis of their ϕ_{ix} and ϕ_{iy} values (see Figure 7). The results showed that 98 keywords were clustered in Quadrant I, 70 keywords were clustered in Quadrant II, 1127 keywords were clustered in Quadrant III, and 40 keywords were clustered in Quadrant IV.

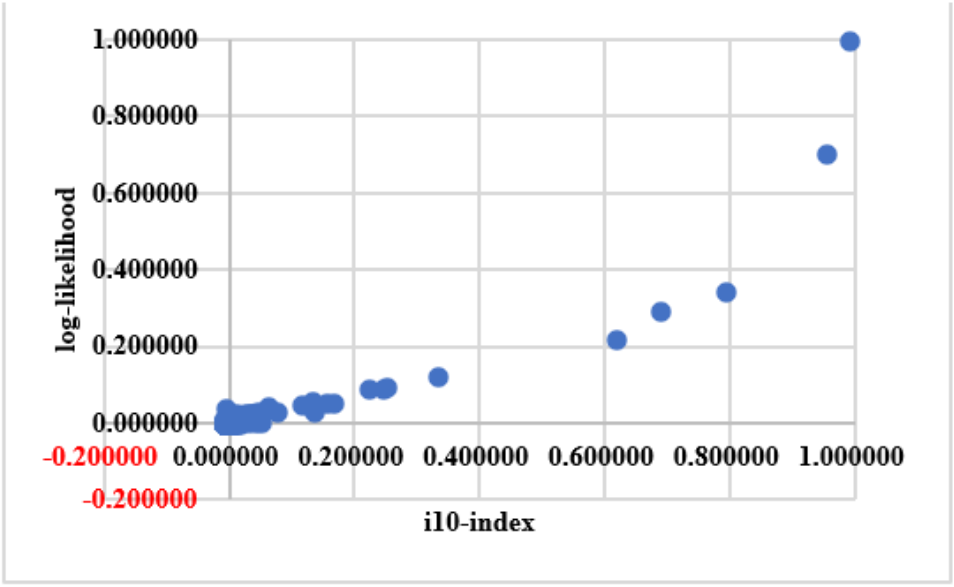


Figure 7. The keywords marked on the revised IPA graph.

Step 5. Distill the highly important keywords from Quadrant I and enhance the accuracy and efficiency of keyword extraction.

The proposed method considered 98 keywords that were clustered in Quadrant I as the highly important keywords for further keyword analysis (see Table 3). Keywords are important channels that help to reveal the key information of a particular big corpus of data. In this case, the 98 highly important keywords can be categorized into five major groups, including medical common nouns (e.g., "*vaccine*", "*COVID*", "*vaccination*", "*vaccinated*", "*pandemic*", etc.), vaccine brands (e.g., "*Pfizer*", "*BNT*", "*AstraZeneca*"), statistics-related words (e.g., "*confidence* interval" (*CI*), "*adjusted odds ratio*" (*AOR*), "*likelihood*", "*regression*", "*health belief model*" (*HBM*), etc.), regions (e.g., "*UK*", "*China*", "*Saudi Arabia*", "*USA*", "*Canada*"), and others (e.g., "*hesitancy*", "*acceptance*", "*respondents*", "*willingness*", "*media*", etc.). Notably, the keywords such as "*willingness*", "*media*", "*conspiracy*", "*misinformation*", "*refusal*", "*mistrust*", "*distrust*", "*undecided*", and "*unsure*" can guide us to find the potential factors that cause vaccine hesitancy in the data of the big target corpus, as such keywords are often found in the abstracta of prior research into vaccine hesitancy (e.g., [58–62]), indicating that the results of distilling keywords by the proposed method are more domain-oriented and precisely reflect the domain's key information.

If we review the whole process of the proposed method, it is machine-based and computed by algorithms, and thus involves fewer human-based tasks and makes the keyword extraction process more efficient. Furthermore, the automated keyword extraction method and the results that aligned more with the practical needs enable future medical professionals to reduce the manual correction tasks in the process of extracting keywords from big corpus data, thereby enhancing the efficiency of keyword analysis.

Table 3. The 98 highly important keywords in Quadrant I.

Keywords	ϕ_{ix}	ϕ_{iy}	Keywords	ϕ_{ix}	ϕ_{iy}
Vaccine	0.992119	0.995679	CDC	0.017245	0.003960
COVID	0.956943	0.701742	Authorities	0.017245	0.003372
Vaccination	0.796139	0.339544	Behavioral	0.017245	0.002172
Hesitancy	0.690611	0.290080	Intent	0.017245	0.000509
Vaccines	0.620260	0.216615	Regression	0.012219	0.025001
Acceptance	0.333827	0.121374	Determinants	0.012219	0.011189
Vaccinated	0.253425	0.091080	Infected	0.012219	0.009070
Pandemic	0.248400	0.086803	Prevalence	0.012219	0.008959
Respondents	0.223275	0.087695	Susceptibility	0.012219	0.006187
Willingness	0.167998	0.052285	Ethnicity	0.012219	0.005775
CI	0.157948	0.048924	Pfizer	0.012219	0.005480
Media	0.137848	0.027948	Tweets	0.012219	0.004721
Uptake	0.132822	0.054000	Rollout	0.012219	0.002668
Healthcare	0.132822	0.038382	Pharmaceutical	0.012219	0.002276
Hesitant	0.117747	0.045335	Resistant	0.012219	0.002186
Vaccinate	0.077546	0.027796	BNT	0.012219	0.001973
Efficacy	0.062471	0.042388	Physicians	0.012219	0.001452
HCWs	0.062471	0.036975	HBM	0.012219	0.001092
Conspiracy	0.052420	0.016141	VHS	0.012219	0.000427
CoV	0.047395	0.025391	USA	0.012219	0.000007
Questionnaire	0.047395	0.023626	Herd	0.007194	0.009008
UK	0.047395	0.011462	Distrust	0.007194	0.007144
Misinformation	0.042370	0.020678	Distribution	0.007194	0.006593
Chronic	0.042370	0.003145	Literacy	0.007194	0.005212
China	0.042370	0.002102	Measles	0.007194	0.004178
Immunization	0.037345	0.024584	Subgroups	0.007194	0.003849
SARS	0.037345	0.023872	Dose	0.007194	0.003719
Immunity	0.037345	0.022916	Providers	0.007194	0.002686
Vaccinations	0.037345	0.020081	Canada	0.007194	0.000600
Fig	0.037345	0.008768	Epidemic	0.007194	0.000574
AOR	0.037345	0.006883	Supplementary	0.002169	0.004842
Effectiveness	0.032320	0.021091	Statistically	0.002169	0.004548
Demographic	0.032320	0.020223	Additionally	0.002169	0.004364
Flu	0.032320	0.013837	Preventive	0.002169	0.003756
Sociodemographic	0.032320	0.012548	Administered	0.002169	0.003678
Saudi	0.032320	0.005046	SD	0.002169	0.003633
Anti	0.032320	0.003917	AstraZeneca	0.002169	0.003169
Arabia	0.032320	0.002004	Consent	0.002169	0.002728
Coronavirus	0.027295	0.025265	Adherence	0.002169	0.001786
Predictors	0.027295	0.013092	Correlation	0.002169	0.001591
mRNA	0.027295	0.009397	Propensity	0.002169	0.001341
Severity	0.027295	0.008802	Viral	0.002169	0.001207
Odds	0.027295	0.008023	Unvaccinated	0.002169	0.001146
VH	0.027295	0.005815	Disparities	0.002169	0.001092
Refusal	0.022270	0.015445	WTP	0.002169	0.000531
Mistrust	0.022270	0.014165	Undecided	0.002169	0.000525
Adverse	0.022270	0.010694	Unsure	0.002169	0.000382
Likelihood	0.017245	0.007472	EUA	0.002169	0.000353
Doses	0.017245	0.005757	Contextual	0.002169	0.000336

CI, confidence interval; HCWs, healthcare workers; UK, United Kingdom; SARS, severe acute respiratory syndrome; AOR, adjusted odds ratio; VH, vaccine hesitancy; CDC, Centers for Disease Control and Prevention; BNT, Pfizer-BioNTech; HBM, health belief model; VHS, vaccine hesitancy scale; SD, standard deviation; WTP, willingness to pay; EUA, emergency use authorization.

5. Discussion

5.1. Verification of the Lexical Coverage of Keywords in Different Quadrants

Lexical coverage determines the level of importance of a word as it relates to the word's frequency and range of occurrence. From the point of view of human language

acquisition, there is a strong positive correlation between lexical coverage and literal comprehension [53–55]. Accordingly, the NLP technique should also take lexical coverage into consideration when identifying highly important keywords for humans. This study not only introduced the i10-index algorithm to integrate the keywords' frequency and range values but also introduced the revised IPA method to integrate the keywords' i10-index and log-likelihood values to conduct data clustering, which successfully clustered the highly important keywords in Quadrant I.

Through verification of the lexical coverage (i.e., calculating the substantial of keywords in the target corpus), although highly the important keywords in Quadrant I (n = 98) only accounted for 8.6% of the overall keywords (n = 1335), their lexical coverage was the highest (9.81%) among the full set of keywords (see Table 4), indicating that such keywords although having high log-likelihood values, also appeared frequently and widely in the target corpus. There were fewer keywords in Quadrants II (n = 70; 5.2%) and IV (n = 40; 3%), and their lexical coverage was quite low and even less than 1% (see Table 4). Even though most of the keywords were clustered in Quadrant III (n = 1127; 84.42%), their lexical coverage (3.5%) was still well below the lexical coverage of the highly important keywords in Quadrant I (9.81%) (see Table 4). Even though AntConc 4.1.3 extracted 1335 keywords, there were not many keywords that were important and valuable to analyze because important keywords must be able to lead us to generally understand the key information from in big corpus data. Hence, highly important keywords must have high keyness and lexical coverage values; only the keywords of Quadrant I could satisfy these conditions.

Table 4. Lexical coverage of the keywords in each quadrant.

	Word Types	Tokens	Lexical Coverage
Quadrant I	98	74,225	9.81%
Quadrant II	70	5946	0.79%
Quadrant III	1127	26,460	3.5%
Quadrant IV	40	6217	0.82%
Total keywords	1335	112,848	14.92%

5.2. Comparison of the Proposed Method with the Traditional Corpus-Based NLP Methods

In this section, the proposed method was compared with three traditional corpus-based NLP methods including a corpus software package, namely, AntConc 4.1.3, and two corpus-based NLP methods that have been used for processing real-world natural language data [8,11]. We discuss the results from three aspects, including the optimization of the keyword list in a machine-based way, evaluating the keywords' importance level, and integrating the variables to conduct data clustering (see Table 5) to highlight the advantages and contributions of the proposed method.

Table 5. Comparison of the proposed method with the traditional methods.

	Optimizing the Keyword List in a Machine-Based Way	Evaluating the Keywords' Importance Level	Integrating the Variables to Conduct Data Clustering
AntConc 4.1.3 [6]	No	No	No
Kithulgoda and Mendis's corpus-based NLP method [8]	No	No	No
Zhong et al.'s corpus-based NLP method [11]	No	No	No
The proposed method	Yes	Yes	Yes

For optimizing the keyword list in a machine-based way, AntConc 4.1.3 has long been favored as an NLP tool by corpus-based researchers. Although it could identify keywords from the target corpus, sort them according to different variables, and generate a keyword list, the keyword list still contained some meaningless letters, function words, and general-

purpose words (see the examples in Table 6), and such words are not helpful for mining key information and even decrease the analytical efficiency and accuracy. To handle this problem, in Kithulgod and Mendis's (2020) [8] research, after they used AntConc 3.4.4 (i.e., older version of AntConc 4.1.3) to process a specialized Welcome Address (WA) corpus, they relied on manual tasks to remove the lexical units that had no semantic significance in the list of n-grams before they conducted the keyword analysis. Similar situation also occurred in Zhong et al.'s (2020) [11] corpus-based research. As mentioned above, meaningless letters, function words, and general-purpose words usually have extremely high frequency values that could interfere the calculation of the log-likelihood values [36]. Therefore, removing such words is inevitable; however, it is not recommended to rely on manual tasks. The proposed method set the function words and NGSL words as the baseline for exclusion and used a machine-based word removal technique [12,36] to remove meaningless letters, function words, and general-purpose words, which enhanced the efficiency of removal.

Table 6. Top 50 keywords from the original keyword list (a part of the keyword list).

Rank	Log-Likelihood	Keyword	Rank	Log-Likelihood	Keyword
1	78763.28	Vaccine	26	4159.43	Associated
2	55611.84	Covid	27	4141.98	Public
3	27083.96	Vaccination	28	4132.75	Sample
4	23187.99	Hesitancy	29	4126.79	Studies
5	17401.64	Vaccines	30	3911.07	Hesitant
6	11144.91	Study	31	3893.01	Individuals
7	10707.55	Health	32	3876.22	Perceived
8	10138.47	Participants	33	3717.94	Table
9	9900.16	Acceptance	34	3678.98	Efficacy
10	7514.16	Vaccinated	35	78763.28	Vaccine
11	7247.51	Respondents	36	3620.22	Intention
12	7177.22	Pandemic	37	3511.41	Higher
13	7161.98	Et	38	3453.18	Social
14	6754.69	Survey	39	3453.01	Disease
15	6284.55	Were	40	3363.46	Healthcare
16	6132.97	Among	41	3355.39	Data
17	5717.95	Population	42	3293.09	Attitudes
18	5664.36	Al	43	3263.40	Information
19	5267.72	Of	44	3256.14	Countries
20	4692.43	Factors	45	3252.64	Hcws
21	4593.61	Uptake	46	3138.45	Confidence
22	4489.16	Reported	47	3136.38	Safety
23	4458.54	Willingness	48	3098.35	Influenza
24	4376.07	Risk	49	2941.16	Infection
25	4193.82	CI	50	2927.96	Results

For evaluating the keywords' importance level, the traditional methods [6,8,11] did not consider the lexical coverage of the keywords before analyzing them, while the so-called important keywords they analyzed were either directly obtained from the keyword list of AntConc 4.1.3, or the process additionally set the keywords' frequency values as the filtration thresholds for re-ranking them, then obtained the so-called high-frequency keywords. However, Chen and Chang (2021) [12] stated that if high-frequency words are excessively concentrated in a small sub-corpus, their importance could be challenged; hence, the range of the keywords must also be considered. In the proposed method, the author claimed that a keyword's importance level is determined by its lexical coverage, and the lexical coverage involves the frequency and range values. This research introduced the i10-index to simultaneously take the keywords' frequency and range values into consideration, as it not only represents the keywords' lexical coverage but also evaluate the keywords' importance at a deeper level. In addition, verification of the lexical coverage of the keywords in

Quadrant I with high i10-index values indicated that these keywords also had the highest lexical coverage, which also proved the applicability of the i10-index algorithm.

For integrating variables to conduct data clustering, the traditional methods [6,8,11] were based on a single variable used to rank the keywords. It can be seen in the results of Kithulgoda and Mendis's (2020) [8] research that the variables of the keyword (i.e., log-likelihood, frequency, and range) were presented separately and lacked integration. Moreover, Zhong et al.'s (2020) [11] method first extracted the keywords then re-ranked the keywords based on their frequency values, which did not integrate these two variables either. This phenomenon also happened in prior corpus-based research (e.g., [9,10]). That is, the integration of the variables was not achieved by the traditional methods; moreover, they used a single variable as a benchmark to rank the keywords, and the top keywords were usually regarded as more important. In addition, the number of important keywords for further keyword analysis was often based on the researchers' arbitrary decisions (e.g., [13–15]), which could cause the analytical results to vary and make them hard to replicate. Nevertheless, the proposed method used the revised IPA method to integrate the keywords' log-likelihood and i10-index values, and used these two variables to cluster and allocate the keywords into the four quadrants. The data analyst then needs to focus only on the keywords in Quadrant I, which makes the process of the corpus-based NLP method more standardized and efficient.

6. Conclusions

The corpus-based NLP method plays an important role in modern medical informatics, as it transforms abundant data from text and discourse transcripts into knowledge by extracting and analyzing the keywords. An efficient and accurate algorithm or data processing method is the foundation of successful NLP used in corpus-based research. Recently, COVID-19 vaccine hesitancy has caught epidemiologists' and public health experts' attention because it hinders the progress of public vaccination [21,24–27]. In the post-epidemic era, implementing widespread vaccination to generate herd immunity seems to be the best way to fight and control the epidemic. Thus, figuring out the potential factors that may cause the COVID-19 vaccine hesitancy phenomenon is critical, as this would enable the relevant authorities to understand the problems and propose appropriate interventions to mitigate such a phenomenon. Although prior corpus-based research had successfully extracted the keywords from medical-related textual data for obtaining domain-specific knowledge or key information (e.g., [33–35]), the traditional corpus-based NLP methods still have room to be optimized for enhancing the efficiency and accuracy of keyword extraction and analysis. To address this issue, this study proposed an improved corpus-based NLP method to distill the highly important keywords from data in the COVID-19 vaccine hesitancy corpus for facilitating the progress of an NLP application in medical informatics.

The proposed method makes three significant contributions: (1) the proposed method uses a machine-based word removal technique to remove meaningless and general-purpose words for optimizing the keyword list; (2) the proposed method introduces the i10-index to integrate a keyword's frequency and range values for evaluating its importance level; (3) the proposed method introduces the revised IPA method to integrate variables (i.e., log-likelihood and i10-index) and cluster the data, thus replacing the researchers' arbitrary decisions for determining the number of important keywords, as in traditional methods; and (4) the proposed method extracts highly important keywords from Quadrant I that have both a high keyness level and high lexical coverage, which makes the results of keyword extraction more domain-oriented and accurate.

The limitation of this study is that the two algorithms introduced here (i.e., the i10-index and the revised IPA method) have not been coded into software, which makes the author need to extract additional quantitative data of the keywords for calculation. Moreover, synonyms, homonyms, and polysemy in the target corpus cannot be considered and handled by the proposed method. Future research could be based on the present study to integrate this method with other NLP methods, such as sentiment analysis or topic modeling, for addressing these

issues and facilitating the progress of keyword extraction in medical informatics. Nevertheless, this pioneering research has verified the feasibility of importing the two algorithms into the corpus-based NLP method, which could effectively extract the highly important keywords and enhance the efficiency and accuracy of keyword analysis.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Meystre, S.M.; Heider, P.M.; Kim, Y.; Davis, M.; Obeid, J.; Madory, J.; Alekseyenko, A.V. Natural language processing enabling COVID-19 predictive analytics to support data-driven patient advising and pooled testing. *J. Am. Med Inf. Assoc.* **2021**, *29*, 12–21. [CrossRef] [PubMed]
2. Garg, M. A survey on different dimensions for graphical keyword extraction techniques issues and challenges. *Artif. Intell. Rev.* **2021**, *54*, 4731–4770. [CrossRef]
3. Mao, K.J.; Xu, J.Y.; Yao, X.D.; Qiu, J.F.; Chi, K.K.; Dai, G.L. A text classification model via multi-level semantic features. *Symmetry* **2022**, *14*, 1938. [CrossRef]
4. Trappey, A.J.C.; Liang, C.P.; Lin, H.J. Using machine learning language models to generate innovation knowledge graphs for patent mining. *Appl. Sci.* **2022**, *12*, 9818. [CrossRef]
5. Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* **1993**, *19*, 61–74.
6. Anthony, L. *AntConc*; Version 4.1.3; Waseda University: Tokyo, Japan, 2022; Available online: https://www.laurenceanthony.net/software (accessed on 1 November 2022).
7. Scott, M. *WordSmith Tools, Version 8.0*; Oxford University Press: Oxford, UK, 2020.
8. Kithulgoda, E.; Mendis, D. From analysis to pedagogy: Developing ESP materials for the welcome address in Sri Lanka. *Engl. Specif. Purp.* **2020**, *60*, 140–158. [CrossRef]
9. 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. Med. Soc.* **2018**, *4*, 2056305118776010. [CrossRef]
10. Todd, R.W. An opaque engineering word list: Which words should a teacher focus on? *Engl. Specif. Purp.* **2017**, *45*, 31–39. [CrossRef]
11. Zhong, Y.P.; Zhu, W.Z.; Zhou, Y.Y. CSR image construction of Chinese construction enterprises in Africa based on data mining and corpus analysis. *Math. Probl. Eng.* **2020**, *2020*, 7259724. [CrossRef]
12. 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]
13. Bi, J. How large a vocabulary do Chinese computer science undergraduates need to read English-medium specialist textbooks? *Engl. Specif. Purp.* **2020**, *58*, 77–89. [CrossRef]
14. Munoz, V.L. The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *Engl. Specif. Purp.* **2015**, *39*, 26–44. [CrossRef]
15. Nation, P. Teaching and learning vocabulary. In *Handbook of Research in Second Language Teaching and Learning*; Hinkel, E., Ed.; Lawrence Erlbaum: Mahwah, NJ, USA, 2005.
16. Hadlington, L.; Harkin, L.J.; Kuss, D.; Newman, K.; Ryding, F.C. Perceptions of fake news, misinformation, and disinformation amid the COVID-19 pandemic: A qualitative exploration. *Psychol. Pop. Media* **2022**, *12*, 40–49. [CrossRef]
17. Luo, Y.F.; Shen, H.Y.; Yang, S.C.; Chen, L.C. The relationships among anxiety, subjective well-being, media consumption, and safety-seeking behaviors during the COVID-19 epidemic. *Int. J. Environ. Res. Public Health* **2021**, *18*, 13189. [CrossRef]
18. Lyu, J.C.; Le Han, E.; Luli, G.K. COVID-19 vaccine-related discussion on Twitter: Topic modeling and sentiment analysis. *J. Med Internet Res.* **2021**, *23*, e24435. [CrossRef]
19. Otegi, A.; San Vicente, I.; Saralegi, X.; Penas, A.; Lozano, B.; Agirre, E. Information retrieval and question answering: A case study on COVID-19 scientific literature. *Knowl.-Based Syst.* **2022**, *240*, 108072. [CrossRef]
20. Haque, A.; Pant, A.B. Mitigating COVID-19 in the face of emerging virus variants, breakthrough infections and vaccine hesitancy. *J. Autoimmun.* **2022**, *127*, 102792. [CrossRef]
21. Pertwee, E.; Simas, C.; Larson, H.J. An epidemic of uncertainty: Rumors, conspiracy theories and vaccine hesitancy. *Nat. Med.* **2022**, *28*, 456–459. [CrossRef]
22. Pfattheicher, S.; Petersen, M.B.; Bohm, R. Information about herd immunity through vaccination and empathy promote COVID-19 vaccination intentions. *Health Psychol.* **2022**, *41*, 85–93. [CrossRef]
23. Yoo, J.H. What we do know and do not yet know about COVID-19 vaccines as of the beginning of the year 2021. *J. Korean Med Sci.* **2021**, *36*, e54. [CrossRef]

24. Hsu, A.L.; Johnson, T.; Phillips, L.; Nelson, T.B. Sources of vaccine hesitancy: Pregnancy, infertility, minority concerns, and general skepticism. *Open Forum Infect. Dis.* **2022**, *9*, ofab433. [CrossRef] [PubMed]
25. Khairat, S.; Zou, B.M.; Adler-Milstein, J. Factors and reasons associated with low COVID-19 vaccine uptake among highly hesitant communities in the US. *Am. J. Infect. Control.* **2022**, *50*, 262–267. [CrossRef]
26. Kiefer, M.K.; Mehl, R.; Costantine, M.M.; Johnson, A.; Cohen, J.; Summerfield, T.L.; Landon, M.B.; Rood, K.M.; Venkatesh, K.K. Characteristics and perceptions associated with COVID-19 vaccination hesitancy among pregnant and postpartum individuals: A cross-sectional study. *BJOG* **2022**, *129*, 1342–1351. [CrossRef]
27. Xiao, J.Y.; Cheung, J.K.; Wu, P.; Ni, M.Y.; Cowling, B.J.; Liao, Q.Y. Temporal changes in factors associated with COVID-19 vaccine hesitancy and uptake among adults in Hong Kong: Serial cross-sectional surveys. *Lancet Reg. Health-W. Pac.* **2022**, *23*, 100441. [CrossRef]
28. Kelkar, A.H.; Blake, J.A.; Cherabuddi, K.; Cornett, H.; McKee, B.L.; Cogle, C.R. Vaccine enthusiasm and hesitancy in cancer patients and the impact of a webinar. *Healthcare* **2021**, *9*, 351. [CrossRef] [PubMed]
29. Griffith, J.; Marani, H.; Monkman, H. COVID-19 vaccine hesitancy in Canada: Content analysis of tweets using the theoretical domains framework. *J. Med Internet Res.* **2021**, *23*, e26874. [CrossRef]
30. Meraya, A.M.; Salami, R.M.; Alqahtani, S.S.; Madkhali, O.A.; Hijri, A.M.; Qassadi, F.A.; Albarrati, A.M. COVID-19 vaccines and restrictions: Concerns and opinions among individuals in Saudi Arabia. *Healthcare* **2022**, *10*, 816. [CrossRef] [PubMed]
31. Luo, Y.F.; Chen, L.C.; Yang, S.C.; Hong, S. Knowledge, attitude, and practice (KAP) toward COVID-19 pandemic among the public in Taiwan: A cross-sectional study. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2784. [CrossRef]
32. Scheiber, A.; Prinster, T.B.; Stecko, H.; Wang, T.N.; Scott, S.; Shah, S.H.; Wyne, K. COVID-19 vaccination rates and vaccine hesitancy among Spanish-speaking free clinic patients. *J. Community Health* **2022**. [CrossRef]
33. Gong, H.; Barlow, M. A corpus-based analysis of research article macrostructure patterns. *J. Engl. Acad. Purp.* **2022**, *58*, 101138. [CrossRef]
34. Shen, Q.; Tao, Y.T. Stance markers in English medical research articles and newspaper opinion columns: A comparative corpus-based study. *PLoS ONE* **2021**, *16*, e0247981. [CrossRef]
35. Sun, X.M.; Chalupnik, M. Sacrificing long hair and the domestic sphere: Reporting on female medical workers in Chinese online news during COVID-19. *Discourse Soc.* **2022**, *33*, 650–670. [CrossRef]
36. 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]
37. Browne, C.; Culligan, B.; Phillips, J. The New General Service List. 2013. Available online: http://www.newgeneralservicelist.org (accessed on 1 November 2022).
38. Chopra, K.; Swanson, E.W.; Susarla, S.; Chang, S.; Stevens, W.G.; Singh, D.P. A comparison of research productivity across plastic surgery fellowship directors. *Aesthet. Surg. J.* **2016**, *36*, 732–736. [CrossRef] [PubMed]
39. da Silva, J.A.T. The i100-index, i1000-index and i10,000-index: Expansion and fortification of the Google Scholar h-index for finer-scale citation descriptions and researcher classification. *Scientometrics* **2021**, *126*, 3667–3672. [CrossRef]
40. Martilla, J.A.; James, J.C. Importance-performance analysis. *J. Mark.* **1977**, *41*, 77–79. [CrossRef]
41. Rayson, P. From key words to key semantic domains. *Int. J. Corpus Linguist.* **2008**, *13*, 519–549. [CrossRef]
42. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. [CrossRef]
43. Kozak, M.; Bornmann, L. A new family of cumulative indexes for measuring scientific performance. *PLoS ONE* **2012**, *7*, e47679. [CrossRef]
44. Sadeghi-Bazargani, H.; Bakhtiary, F.; Golestani, M.; Sadeghi-Bazargani, Y.; Jalilzadeh, N.; Saadati, M. The research performance of Iranian medical academics: A national analyses. *BMC Med. Educ.* **2019**, *19*, 449. [CrossRef]
45. Joung, J.; Kim, H.M. Approach for importance-performance analysis of product attributes from online reviews. *J. Mech. Des.* **2021**, *143*, 081705. [CrossRef]
46. Rasovska, I.; Kubickova, M.; Ryglova, K. Importance-performance analysis approach to destination management. *Tour. Econ.* **2021**, *27*, 777–794. [CrossRef]
47. Wang, Z.L.; Shen, H.C.; Zuo, J. Risks in prefabricated buildings in China: Importance-performance analysis approach. *Sustainability* **2019**, *11*, 3450. [CrossRef]
48. Chang, K.L. A new hybrid MCDM model for esports caster selection. *J. Mult.-Valued Log. Soft Comput.* **2021**, *37*, 573–590.
49. Tsai, J.F.; Wang, C.P.; Chang, K.L.; Hu, Y.C. Selecting bloggers for hotels via an innovative mixed MCDM model. *Mathematics* **2021**, *9*, 1555. [CrossRef]
50. Wen, T.C.; Chang, K.H.; Lai, H.H.; Liu, Y.Y.; Wang, J.C. A novel rugby team player selection method integrating the TOPSIS and IPA methods. *Int. J. Sport Psychol.* **2021**, *52*, 137–158.
51. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 604–624. [CrossRef]
52. Pojanapunya, P.; Todd, R.W. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguist. Linguist. Theo.* **2018**, *14*, 133–167. [CrossRef]
53. Durbahn, M.; Rodgers, M.; Peters, E. The relationship between vocabulary and viewing comprehension. *System* **2020**, *88*, 102166. [CrossRef]

54. Herman, E.; Leeser, M.J. The relationship between lexical coverage and type of reading comprehension in beginning L2 Spanish learners. *Mod. Lang. J.* **2022**, *106*, 284–305. [CrossRef]
55. Xodabande, I.; Ebrahimi, H.; Karimpour, S. How much vocabulary is needed for comprehension of video lectures in MOOCs: A corpus-based study. *Front. Psychol.* **2022**, *13*, 992638. [CrossRef] [PubMed]
56. Phadermrod, B.; Crowder, R.M.; Wills, G.B. Importance-Performance Analysis based SWOT analysis. *Int. J. Inf. Manage.* **2019**, *44*, 194–203. [CrossRef]
57. Anakpo, G.; Mishi, S. Hesitancy of COVID-19 vaccines: Rapid systematic review of the measurement, predictors, and preventive strategies. *Hum. Vaccines Immunother.* **2022**, *18*, 2074716. [CrossRef]
58. Allington, D.; McAndrew, S.; Moxham-Hall, V.; Duffy, B. Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the COVID-19 pandemic. *Psychol. Med.* **2021**, *53*, 236–247. [CrossRef]
59. Mascherini, M.; Nivakoski, S. Social media use and vaccine hesitancy in the European Union. *Vaccine* **2022**, *40*, 2215–2225. [CrossRef] [PubMed]
60. Ouyang, H.; Ma, X.H.; Wu, X. The prevalence and determinants of COVID-19 vaccine hesitancy in the age of infodemic. *Hum. Vaccines Immunother.* **2022**, *18*, 2013694. [CrossRef]
61. Pierri, F.; Perry, B.L.; DeVerna, M.R.; Yang, K.C.; Flammini, A.; Menczer, F.; Bryden, J. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci. Rep.* **2022**, *12*, 5966. [CrossRef]
62. Zhang, X.N.; Guo, Y.Q.; Zhou, Q.; Tan, Z.X.; Cao, J.L. The mediating roles of medical mistrust, knowledge, confidence and complacency of vaccines in the pathways from conspiracy beliefs to vaccine hesitancy. *Vaccines* **2021**, *9*, 1342. [CrossRef]