Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Search queries related to COVID-19 based on keyword extraction

Livia Kelebercová, Michal Munk

Abstract

Background: Pandemic COVID-19 caused an infodemic – massive spread of true and fake information about novel coronavirus. This study aims to present the possibility of using Keyword Extraction as a tool to obtain the most trending search queries related to COVID-19 and analyze the possibility of including their search volume in models for the prediction of fake news.

Methods: The study used Python implementation of the machine learning-based technique KeyBERT to extract keywords from true and fake news. These keywords were used in the next step to obtain related search queries with Google Trends API.

Results: Non-parametric Spearman Rank Order Correlation was identified as a statistically positive correlation ($p < 0.001$) between the occurrence of false news and top query / rising query metrics provided by Google Trends of queries related to extracted keywords pandemic, HIV, lockdown, plague, Michigan, and protest, which proves that search volume can identify fake news.

Conclusions: Experiments done in this research proved that Keyword Extraction from false news is useful for obtaining related search queries and the top query and rising query metrics can be used to increase the accuracy of fake news prediction models.

© 2022 The Authors. Published by Elsevier B.V.

Keywords: Natural Language Processing; Keyword Extraction; Google Trends; Fake News Detection

1. Introduction

Infodemic is a term that refers to the increasing volume of information in which it is hard for people to find trustworthy information sources [1]. Research done on the Italian population [2] proves that 82% of people are not...
able to decide if the information is a hoax or not. In the context of the pandemic COVID-19 lot of misinformation appeared in the media. According to multiple studies, misinformation can negatively affect mental health. A brief by WHO revealed that anxiety and depression increased by 25% in the first year of pandemic COVID-19 [3]. A recent study [4] was discovering if media has an impact on quality of life. Authors found an association between hours spent on the news during a pandemic and overall quality of life. According to [5] many print media were stopped due to the pandemic which caused an increase in interest in other sources of information such as the Internet. Pandemics had a significant impact all around the world and we have to deal with the amount of inaccurate and false information spread on social media, blogs, or Internet news.

To discover the level of interest in fake news or conspiracy theories it is possible to use Google Trends. It is an online tool provided by Google that can return related searches and interest in them from any time period and country. An interesting study [6] used Google Trends to obtain search volumes for words associated with COVID-19 and evaluated their Granger-causality to weekly positivity in multiple countries. Another study [7] compared the government statements and trending search queries and they found out that these statements were the reason behind different information-seeking trends.

Research by Timoneda and Vera [8] compares data from Google Trends with epidemiological data and election results in 2016. They found that Google Trends data close matches epidemiological data and differences exist between states that supported Clinton or Trump.

An interesting study [9] claims that Google Trends can be used to build the predictive model for case numbers in a pandemic outbreak. They found a positive correlation between the trend volume of keywords and cases in countries ($p < 0.05$). Another similar study [10] investigated the relationship between search trends related to COVID-19 over countries to predict a number of cases. They found that Deep Learning techniques outperform other techniques for predicting the number of confirmed cases.

Google Trends has been widely used in studies connected to COVID-19 but they were rather concentrated on the identification of positive cases than on the research of interest in true and fake news and the possibility of integration of Google Trends data in fake news prediction models.

The remainder of this paper is structured as follows: Section 2 describes the necessary existing theory on Keywords Extraction – a method of Natural Language Processing which we used to extract the keywords from a dataset [11] of articles about COVID-19, which are marked as true, false or partially false. In section 3 we present the methodology. We describe the pre-processing techniques we needed to do before the keyword extraction process and the implementation of the KeyBERT technique and how we obtain related searches from Google Trends, interpret their metrics and process these data for evaluation. The fourth section deals with the results of our work. We describe statistical tests we did on our data to verify if the search volume of Google Trends can be used to identify fake news or not.

2. Keyword Extraction

Keyword Extraction (KE) is described as a technique for automatically obtaining the most important words from the text. This extraction process can identify words or phrases that summarize the document. KE approaches can be divided into four basic groups: statistical and machine learning approaches [12].

- Simple Statistical Approaches are methods that use statistical information such as TF-IDF, word frequency, Patricia Tree to identify keywords. These approaches are language-independent and do not require any training data [12] [13] [14].
- Linguistic Approaches use syntactic, semantic, and lexical features of words or sentences. With these algorithms, it is possible to find meaningful key phrases matching the Part of Speech (POS) patterns, for example, nouns + verbs. The main disadvantage of Linguistic Approaches is that the algorithms depend on language [12] [13].
- Machine Learning Approaches are based on supervised learning [12]. These approaches extract keywords using classification or regression. At first, they select candidate words by statistical methods such as TF-IDF and then these algorithms try to determine whether the candidate word in the text is a key phrase or not [13] [14]. As an example, the KeyBERT algorithm itself begins by sending a document to the BERT model, which creates a representation of the document by dividing the text into fixed-size vectors representing the semantics of the document. In the second step, the candidate phrase generator extracts candidate phrases from the document using...
simple techniques such as several occurrences, TF-IDF, and the like. In the next step, these data are sent back to the BERT model and a phrase-level representation is obtained. Subsequently, the cosine similarity is calculated between the document-level representation and the phrase-level representation to obtain the most similar words to the document representation, which are the resulting keywords. [15] Machine Learning Approaches take more computing time and might depend on the language but on the other hand, they can extract keywords with high semantic relevance.

- Other Approaches combine the above methods. Some of them may use HTML tags, the position of words in the text, length of words, and more [12].

3. Methods

For realizing our research, we used Python programming language, Natural Language Toolkit (NLTK) for natural language processing and PyTrends API for Google Trends.

3.1. Keyword Extraction

We used a dataset [7] (Fig. 1) of 3119 articles about COVID-19 during the period of December 2019 to July 2020, which is available in xlsx format. This dataset contains 659 false news, 2061 true news, and 399 partially false news. Columns of the dataset are ID, the title of an article, the text of an article, subcategory, and label.

We used NLTK library to get basic information about dataset such as number of tokens (words and punctuation together), sentences, words, nouns, the average sentence (the number of words divided by the number of sentences) and word length (the total number of letters divided by the number of words). This information is presented in Tab. 1. The title column contains 34617 tokens. The total number of sentences in the column is 335 and the average length of sentence is 94.776. These headings contain 31750 words with an average length of word 5.495. The total number of nouns in the title column is 13412.

The text column contains 1779974 tokens. The total number of sentences in the columns is 66800 and the average length of a sentence is 23.061. This column contains 1540478 words with an average length of word 5.17 and the total number of nouns in all articles is 493682.

| Tab. 1. Information about dataset |
|----------------------------------|
| **titles** | **texts** | **titles and texts** |
| Number of tokens | 34617 | 1779974 | 1814591 |
| Number of sentences | 334 | 66800 | 67134 |
| The average length of sentence | 94.776 | 23.061 | 117.837 |
| Number of words | 31750 | 1540478 | 1572228 |
| The average length of word | 5.495 | 5.17 | 10.665 |
| Number of nouns | 13412 | 493682 | 507094 |
This dataset was collected using Webhose.io and was manually labeled. Subcategory marks each article as true, false news or partially false. The label is a binary value (0/1) which adds 0 value to false news and 1 to true and partially false.

To obtain search queries associated with false or true news, we extracted keywords for each category. Pre-processing is necessary for the KE process. Without appropriate text pre-processing most of the algorithms will select words that are frequent but not meaningful such as “a”, “or”, “and”, and others. We used a list of stop-words from NLTK library. One of the basic pre-processing methods is to clean the text from these stop-words. We also needed to identify and remove other frequent words, such as coronavirus, covid-19, and others. For obtaining related searches we needed to extract only units composed of one word and these words are too general to extract them as keywords. To identify them we used FreqDist function from NLTK.

For KE we used a KeyBERT package. As a sentence transformer we used all-MiniLM-L6-v2 which is 5 times faster than the most common all-mpnet-base-v2 model. We implemented Wordnet lexical database to our KE function to obtain only nouns as keywords. We decided to exclude some of the keywords we obtain such as xenophobia because we were not interested in finding related keywords to them. Fig. 2 shows extracted keywords for each category. Some of them like a pandemic, epidemic, and pathogen KeyBERT extracted for true and also for partially true.

| Keywords      | True | False | Partially |
|---------------|------|-------|-----------|
| pandemic      | 1    | 0     | 1         |
| outbreak      | 1    | 0     | 0         |
| epidemic      | 1    | 0     | 1         |
| plague        | 1    | 0     | 0         |
| mortality     | 1    | 0     | 0         |
| pathogen      | 1    | 0     | 1         |
| protest       | 0    | 1     | 0         |
| quarantine    | 0    | 1     | 0         |
| lockdown      | 0    | 1     | 0         |
| michigan      | 0    | 1     | 0         |
| gridlock      | 0    | 1     | 0         |
| risk          | 0    | 0     | 1         |
| virulence     | 0    | 0     | 1         |
| danger        | 0    | 0     | 1         |
| hiv           | 0    | 0     | 1         |
| threat        | 0    | 0     | 1         |

Fig. 2 Extracted keywords for each category

3.2. Related search queries

PyTrends library provides an unofficial API for Google Trends. We used it to automate obtaining related search queries. We connected to Google Trends with TrendReq class with host language as a parameter.

To obtain any data from Google Trends it is needed to call the build_payload method with a list of keywords, geographic location, and the period of time as parameters. After the previous step, we created a function to retrieve a Python DataFrame of related trending keywords with top query (TQ) and rising query (RQ) metrics, which we will describe later. Google Trends can analyze only 5 keywords at once, so we had to call our function multiple times. In the end, we exported our DataFrames to xlsx format (Fig.3).
The data obtained by Google Trends reflects popular search queries people make on Google. This tool filters out searches made by only a few people and duplicate searches from the same person in a short period of time.

The most popular searches are represented by the TQ on the relative scale from 0 to 100, where the most frequently searched terms have a value of 100. For example, if we enter in the Google Trends keyword pandemic, the TQ displays the most searched terms related to the keyword pandemic in our selected region and time range.

The second metric is called rising query (RQ). It is expressed in percentage, and it represents the queries with the largest increase in search frequency in a selected region and time range [16]. Rising queries usually represent terms, which were not searched in the past but had the biggest increase since in search volume in the last period of time.

For testing verifying if these data can be used for fake news detection, we needed to process them. Some of our exported DataFames did not contain searches related to COVID-19. That’s why some of the extracted keywords such as gridlock or danger which are present in Tab.1 are not in our results.

We created a variable sum_count, which represents the number of occurrences of each keyword in the article. In the next step, we calculated the mean and median for top queries (TQ) and rising queries (RQ) and as a weight, we used sum count to create total mean TQ, total median TQ, total mean RQ, and total median RQ. The last variable we created a binary variable false news, which is similar to the label column in our dataset but we added value 0 to partially true and true news and 1 to false news.

4. Results

As a part of data exploration, we examined the relationship between binary false news variable (0/1) and aggregated (mean, median) TQ and RQ values from queries that relate to our extracted keywords. With non-parametric Spearman Rank Order Correlation we identified a statistically positive correlation between the false news variables and top query / rising query metrics of queries related to extracted keywords pandemic (p < 0.001), HIV (p < 0.001), lockdown (p < 0.001), plague (p < 0.001), Michigan (p < 0.001) and protest (p < 0.01), regardless of the method of aggregation of the relevant measures demand (total mean TQ, total median TQ, total mean RQ, total median RQ). These variables can help identify false news.
In the next part, we will take a closer look at the results in differences between true and fake news in the previously described variable total mean TQ. For testing differences, we use the non-parametric Mann-Whitney U Test (Tab. 2a, Tab. 2b) and robust descriptive statistics (Tab. 4a, Tab. 4b), given the presence of extreme cases (outliers) (Tab. 3a, Tab. 3b). We can see that the Grubbs Test Statistics are $> 11.6$ (Tab. 3a), respectively $> 9.8$ (Tab. 3b). Their p-values are $< 0.001$. These small p-values are evidence that there are outliers in the investigated variables.

Tab. 2. (a) Mann-Whitney U Test for related searches to keywords Michigan, pandemic, HIV, plague, protest, lockdown, and mortality.

|                      | U       | Z       | Z adjusted | p-value |
|----------------------|---------|---------|------------|---------|
| Michigan total mean TQ | 1079500 | 0.452258| 3.77427    | 0.0002  |
| pandemic total mean TQ| 916184  | 7.31127 | 12.10033   | 0.00000 |
| HIV total mean TQ     | 1038829 | 2.1604  | 5.98202    | 0.00000 |
| plague total mean TQ  | 1062362 | 1.172025| 4.36049    | 0.00001 |
| protest total mean TQ | 1075063 | 0.638626| 3.19892    | 0.0014  |
| lockdown total mean TQ| 1037731 | 2.206506| 5.11468    | 0.00000 |
| mortality total mean TQ| 1075609| 0.615695| 1.80505    | 0.0711  |

Based on the results of the nonparametric Mann-Whitney U Test (Tab. 2a), we reject the null hypotheses at the 0.001 significance level in the case of related queries to keywords Michigan, pandemic, HIV, plague lockdown, and in the case of the words protest at the 0.01 significance level with reliability 99%, so we demonstrated statistically significant differences between false and true news in terms of the TQ. From Tab. 4a we can see the identified statistically significant differences in favor of false news. The results are in agreement with the initial data exploration (non-parametric correlation). The results for all mean estimates (Tab. 4a) are equally robust.

Tab. 2. (b) Mann-Whitney U Test for keywords threat, risk, virulence, epidemic, and quarantine.

|                      | U       | Z       | Z - adjusted | p-value |
|----------------------|---------|---------|--------------|---------|
| threat total mean TQ | 1079675 | 0.4449  | 0.7095       | 0.4780  |
| risk total mean TQ   | 1085341 | -0.2070 | -0.2809      | 0.7788  |
| virulence total mean TQ | 1090129 | -0.0059 | -0.0489      | 0.9610  |
| epidemic total mean TQ| 1068745 | -0.9040 | -1.2990      | 0.1940  |
| quarantine total mean TQ| 1053948| -1.5254 | -2.0565      | 0.0397  |

Tab. 3. (a) Grubbs Test Statistics for keywords Michigan, pandemic, HIV, plague, protest, lockdown, and mortality.

|                      | Valid N | Grubbs Test Statistic | p-value | Valid N | Grubbs Test Statistic | p-value |
|----------------------|---------|-----------------------|---------|---------|-----------------------|---------|
| Michigan total mean TQ | 1058    | 21.0270               | 0.0000  | 2061    | 38.8920               | 0.0000  |
| pandemic total mean TQ | 1058    | 11.6159               | 0.0000  | 2061    | 30.8075               | 0.0000  |
| HIV total mean TQ     | 1058    | 23.6207               | 0.0000  | 2061    | 28.9532               | 0.0000  |
| plague total mean TQ  | 1058    | 22.2368               | 0.0000  | 2061    | 20.1421               | 0.0000  |
| protest total mean TQ | 1058    | 23.0720               | 0.0000  | 2061    | 20.5421               | 0.0000  |
| lockdown total mean TQ| 1058    | 12.8655               | 0.0000  | 2061    | 20.6365               | 0.0000  |
| mortality total mean TQ| 1058    | 11.9890               | 0.0000  | 2061    | 25.7945               | 0.0000  |
Tab. 3. (b) Grubbs Test Statistics for keywords threat, risk, virulence, epidemic, and quarantine.

| False News = 1 | False News = 0 |
|----------------|----------------|
| Valid N        | Grubbs Test Statistic | p-value | Valid N       | Grubbs Test Statistic | p-value |
| Threat total mean TQ | 1058 | 14.4618 | 0.0000 | 2061 | 10.7463 | 0.0000 |
| Risk total mean TQ | 1058 | 13.0683 | 0.0000 | 2061 | 34.3090 | 0.0000 |
| Virulence total mean TQ | 1058 | 14.5052 | 0.0000 | 2061 | 25.1660 | 0.0000 |
| Epidemic total mean TQ | 1058 | 9.8832 | 0.0000 | 2061 | 14.3379 | 0.0000 |
| Quarantine total mean TQ | 1058 | 12.5336 | 0.0000 | 2061 | 18.5046 | 0.0000 |

The results for the remaining keywords (Tab. 2b) are consistent with the data exploration. In the case of these keywords, there were no statistically significant differences in the (TQ) of their respective queries between false and true news, except for the keyword quarantine, where we reject the null hypothesis at the 0.05 significance. In this case, a statistically significant difference was identified (Tab. 4b) in favor of true news, which is in line with the data exploration, where a statistically significant negative correlation was identified. The results for all mean estimates (Tab. 4b) are equally robust.

Tab. 4. (a) Robust Statistics for keywords Michigan, pandemic, HIV, plague, protest, lockdown, and mortality.

| False News = 1 | False News = 0 |
|----------------|----------------|
| Mean           | Trimmed mean 1% | Winsorized mean 1% | Mean           | Trimmed mean 1% | Winsorized mean 1% |
| Michigan total mean TQ | 0.93*** | 0.04*** | 0.43*** | 0.20 | 0.00 | 0.00 |
| Pandemic total mean TQ | 13.45*** | 10.28*** | 11.86*** | 3.90 | 2.36 | 2.97 |
| HIV total mean TQ | 2.70*** | 1.57*** | 1.92*** | 0.92 | 0.37 | 0.55 |
| Plague total mean TQ | 1.90*** | 0.90*** | 1.40*** | 0.64 | 0.15 | 0.40 |
| Protest total mean TQ | 2.76** | 0.75** | 1.35** | 0.92 | 0.00 | 0.00 |
| Lockdown total mean TQ | 6.86*** | 4.52*** | 5.90*** | 2.07 | 1.27 | 1.82 |
| Mortality total mean TQ | 1.39 | 0.88 | 1.22 | 1.00 | 0.48 | 0.81 |

Note: *** p < 0.001, ** p < 0.01, * p < 0.05

Tab. 4. (b) Robust Statistics threat, risk, virulence epidemic, and quarantine.

| False News = 1 | False News = 0 |
|----------------|----------------|
| Mean           | Trimmed mean 1% | Winsorized mean 1% | Mean           | Trimmed mean 1% | Winsorized mean 1% |
| Threat total mean TQ | 1.40 | 1.10 | 1.29 | 1.20 | 0.97 | 1.10 |
| Risk total mean TQ | 7.30 | 5.94 | 6.75 | 6.73 | 5.45 | 6.10 |
| Virulence total mean TQ | 0.04 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 |
| Epidemic total mean TQ | 9.05 | 7.35 | 8.46 | 11.04 | 8.90 | 10.44 |
| Quarantine total mean TQ | 12.73 | 10.67 | 12.16 | 16.57* | 13.57* | 14.97* |

Note: *** p < 0.001, ** p < 0.01, * p < 0.05

The graph of point and interval (95% confidence interval) mean estimation (Fig. 4) visualizes the identified statistically significant differences (Tab. 4a, Tab. 4b) in the total mean TQ values between false and true news.

In the case of the remaining variables total median TQ, total mean RQ, and total median RQ, the results are in agreement with the previous analysis as well as with the initial data exploration (correlation analysis). For this reason, we are limited to data visualization in this case.
Tab. 4. (a) Robust Statistics for keywords Michigan, pandemic, HIV, plague, protest, lockdown, and mortality. (Tab. 4b) are equally robust. 

...exploration, where a statistically significant ne...case, a statistically significant difference was identified (Tab. 4b) in favor of true news, which is in line with the data...true news, except for the keyword quarantine, where we reject the...keywords, there were no statistically significant differences in the (TQ) of their respective queries between false and...data visualization in this case.

The results for the remaining keywords (Tab. 4a, Tab. 4b) in favor of false news. In...reason, we are limited to data v...statistically significan...In the case of these variables (Fig. 5, Fig. 6, Fig. 7), statistically, significant differences between false and true news were shown in the values of the variables total median TQ, total means RQ, total median RQ expressing top and rising queries related to keywords pandemic, lockdown, protest, HIV, plague, Michigan in favor of false news.

Tab. 3. (b) Grubbs Test Statistics for keywords threat, risk, virulence, epidemic, and quarantine total mean TQ.

| Keyword | Total Mean TQ (False News) | Total Mean TQ (True News) | Mean | Trimmed Mean 1% Winsorized Mean 1% |
|---------|---------------------------|---------------------------|------|----------------------------------|
| Threat  | 10.74                     | 12.16                     | 1.29 | 1.20                             |
| Risk    | 6.73                      | 5.45                      | 6.10 |                                |
| Virulence | 7.35                   | 8.46                      | 0.00  | 0.00                             |
| Epidemic | 10.67                   | 12.16                     | 16.57*| 13.57*                           |
| Quarantine | 6.73                  | 5.45                      | 6.10 |                                |
| Michigan | 10.74                    | 12.16                     | 16.57*| 13.57*                           |

The results for all mean estimates...variables total median TQ, total mean RQ, total median RQ between false and true news except keywords related to quarantine, where a statistically significant difference in the values of the variables total median TQ, total mean RQ, total median RQ was demonstrated in favor of true news.
5. Conclusion

In this paper, we propose KE for obtaining relevant search queries in Google Trends related to true and fake news. Our experiments proved that KE can be useful for obtaining related queries and their metrics can be used to identify fake news.

The limitation of our work is that the algorithm we used for KE depends on the language. KeyBERT works perfectly for English but if we would like to implement this technique for documents written in a less used language such as the Slovak language there would not be as many existing sentence transformers, and we would probably need to create our own.

In our future work, we would like to repeat this experiment with datasets in different languages. The main goal of our future work will be to test if our proposed variables can improve fake news detection in multiple classifiers such as Naïve Bayes, Logistic Regression, Support Vector Machine, Stochastic Gradient, or Random Forest. We would experiment with different ratios of training and testing data. We would also like to experiment with metrics from different analytic tools such as Google Ads or Google Search Console.
Acknowledgments

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-18-0473 and Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of the Slovak Academy of Sciences (SAS) under the contracts No. VEGA-1/0821/21.

References

[1] Pan American Health Organization, World Health. 2020. “Understanding the infodemic and misinformation in the fight against COVID-19.” Washington D.C.: PAHO. <Available at: https://iris.paho.org/bitstream/handle/10665.2/52052/Factsheet-infodemic_eng.pdf>. [Accessed March 1, 2022].
[2] UNISOB. 2018. “Report dei risultati di ricerca 2018.” UNISOB. <Available at: https://www.unisob.na.it/eventi/pdf/20180720.pdf>. [Accessed March 1, 2022].
[3] WHO. 2022. “Mental Health and COVID-19: Early evidence of the pandemic’s impact.” WHO. <Available at: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1>. [Accessed March 3, 2022].
[4] Sharma, P., Gupta, S., Kuswaha, P., Shekhawat, K. 2020. “Impact of Mass Media on Quality of Life during COVID-19 Pandemic among Indian Population.” International Journal of Science and Healthcare Research. 5(3).
[5] Sharma, A., Choudhary, P., Randhawa, R. 2022. “Media as an Infodemic.” International Journal of Advance Nursing Education. 1(2).
[6] Sato, K., Mano, T., Iwata, A., Toda, T. 2021. “Need of care in interpreting Google Trends-based COVID-19 infodemiological sturdy results: potential risk of false positivity.” BMC Med Res Methodol. 21.
[7] Limilia, P., Pratamawaty, B. 2020. “Google Trends and Information Seeking Trends of COVID-19 in Indonesia.” Journal ASPIKOM. 5(2).
[8] Timoneda, J., Vera, S. 2021. “Will I die of coronavirus? Google Trends data reveal that politics determine virus fears.” PLoS ONE. 16(10).
[9] Doganer, A., Zhang, Z. 2021. “An Infodemical Analysis of Google Trends in COVID-19 Outbreak: Predict Case Numbers and Attitudes of Different Societies.” Journal of Database Management (JDM). 32(2).
[10] Pan, Z., Nguyen, H., Abu-gellban, H., Zhang, Y. 2020. “Google Trends Analysis of COVID-19 Pandemic.” IEEE International Conference on Big Data. pp. 2857-2866.
[11] Koirala, A. 2021. “COVID-19 Fake News Dataset.” Mendley Data. <Available at: https://data.mendeley.com/datasets/zwfdmp5syg/1>. [Accessed March 3, 2022].
[12] Beliga, S., Meštrović, A., Martinčić-Ipišić, S. 2015. “An Overview of Graph-Based Keyword Extraction Methods and Approaches.” Journal of information and organizational sciences. 39.
[13] Sharma, P.; Li, Y. 2019. “Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling.” Preprints. pp. 1–6.
[14] Beliga, S., 2014. “Keyword extraction: a review of methods and approaches.” The University of Rijeka, Department of Informatics. <Available at: http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf> [Accessed February 3, 2022].
[15] Devlin, J. et al., 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of the 2019 Conference of the North, pp. 4171–4186.
[16] McCarthy, A.D. & McGoldrick, D., 2021. “Analyzing public interest in metabolic health-related search terms during COVID-19 using google trends.” Cureus, 13(6).