Revealing the Reflections of the Pandemic by Investigating COVID-19 Related News Articles Using Machine Learning and Network Analysis

Araştırma Makalesi/Research Article

Ulya BAYRAM

Department of Electrical and Electronics Engineering, Çanakkale Onsekiz Mart University, Çanakkale, Turkey
ulya.bayram@comu.edu.tr

(Geliş/Received:11.06.2021; Kabul/Accepted:12.04.2022)
DOI: 10.17671/gazibtd.949599

***Abstract*—** Social media data can provide a general idea of people's response towards the COVID-19 outbreak and its reflections, but it cannot be as objective as the news articles as a source of information. They are valuable sources of data for natural language processing research as they can reveal various paradigms about different phenomena related to the pandemic. This study uses a news collection spanning nine months from 2019 to 2020, containing COVID-19 related articles from various organizations around the world. The investigation conducted on the collection aims at revealing the repercussions of the pandemic at multiple levels. The first investigation discloses the most mentioned problems covered during the pandemic using statistics. Meanwhile, the second investigation utilizes machine learning to determine the most prevalent topics present within the articles to provide a better picture of the pandemic-induced issues. The results show that the economy was among the most prevalent problems. The third investigation constructs lexical networks from the articles, and reveals how every problem is related through nodes and weighted connections. The findings exhibit the need for more research using machine learning and natural language processing techniques on similar data collections to unveil the full repercussions of the pandemic.

***Keywords*—**LDA, BERT, machine learning, news articles, natural language processing, network analysis

Pandeminin Yansımalarını Ortaya Çıkarmak için COVID-19 ile İlgili Gazete Makalelerinin Makine Öğrenimi ve Ağ Analizi Yöntemleri ile İncelenmesi

***Özet*—** Sosyal medya platformlarından elde edilmiş veriler, insanların COVID-19 pandemisine karşı gösterdiği tepkiler hakkında bilgi verse de gazete makaleleri kadar objektif bir şekilde bilgi içeremezler. Pandemi sürecinde yayınlanmış makaleler genel halkın yaşanılan krizden nasıl etkilendiği hakkında bilgi vermekle birlikte, aynı zamanda siyasi ve daha farklı alanlardaki etkilerden de bahsederler. Bu makaleler, pandemiyle ilgili çok farklı paradigmaları içermeleri sebebiyle doğal dil işleme araştırmaları için faydalı veri kaynaklarıdır. Bu çalışmada, 2019 ve 2020 yıllarındaki COVID-19 ile ilgili uluslararası haber organizasyonları tarafından dokuz ay boyunca yayınlanmış gazete makaleleri koleksiyonunu kullanmaktadır. Bu koleksiyon üzerine üç kademeli bir inceleme çalışması uygulayarak pandeminin sebep olduğu sonuçları farklı derecelerde açığa çıkarmayı amaçlamaktadır. İlk çalışma, kelime istatistiklerini kullanarak pandemi sürecinde makalelerde en çok bahsedilen problemleri ortaya çıkarır. İkinci çalışma ise, makalelerden pandeminin sebep olduğu problemleri daha iyi ortaya çıkarmak için makine öğrenimi yöntemleriyle konu modelleme yapar. Sonuçlara göre en sık bahsedilen pandemi sebepli problemlerden biri ekonomik olanlardır. Üçüncü çalışma da gazete makalelerinden sözlüksel ağ oluşturarak düğüm ve ağırlıklı bağlantılar üzerinden pandemi sürecinde birçok problemin nasıl bağlantılı olduğunu gösterir. Buluntulara göre makine öğrenimi ve doğal dil işleme yöntemleri ile benzer veri setleri üzerinde pandeminin tüm etkilerinin daha çok araştırılması gerektiği görülmektedir.

***Anahtar Kelimeler*—** LDA, BERT, makine öğrenimi, gazete makaleleri, doğal dil işleme, ağ analizi

1. INTRODUCTION

The COVID-19 pandemic affected many paradigms related to health care and social life. The extent of affected real-life dynamics, however, is much larger than it appears. Many researchers have analyzed social media posts to understand the reflections of the pandemic using machine learning and natural language processing (NLP) techniques [1, 2, 3]. However, social media cannot be treated as a source of balanced and objective information; it can only provide a general idea of people's responses towards the pandemic, the related government measures and the global consequences. Those who seek a relatively unbiased source of information should look no further than the news articles covering various news related to the pandemic. They can provide insights about the pandemic, how the governments around the world reacted to it, how the daily life of the public was affected globally, among various other perspectives. To enable researchers to study this rich source of data, in 2020, a service provider rendered a data collection containing international news articles from global news organizations, spanning November 2019 to July 2020, and made it available for the use of researchers [4]. The present study investigates this collection through NLP, machine learning, and network analysis techniques, where the main aim is to reveal the wide-range reflections of the pandemic.

The types of investigative methods one can apply on an unlabeled collection of news articles are naturally limited to unsupervised techniques. The experiments conducted in this study span a wide range of unsupervised methodologies in parallel with the aim of unveiling the consequences of the pandemic. The first analysis employs statistical methods to demonstrate what a simple approach can reveal from the news articles. The use of statistics for performing NLP tasks is quite common in the literature because they can reveal intriguing phenomena from the texts [5, 6]. Meanwhile, the second investigation increases the methodological complexity and attempts to determine the general topics present within the collection through unsupervised machine learning methods, which are commonly used for discovering hidden semantic structures in texts [7]. In terms of unveiling the repercussions of the pandemic, topic modeling can reveal the hidden semantic structures associated with many global pandemic-related problems. To validate the findings, two different machine learning techniques are utilized for topic modeling and the results are compared. Additionally, an alternative technique is applied to further investigate the reflections of the pandemic to test whether a different approach can reveal other latent problems that were not detected through statistics or machine learning. For many data domains, the network theory is commonly found to be successful at yielding explanations for various phenomena [8]. Thus, the third investigation involves construction of a network from the news articles to capture semantically associated words through mathematical computations. Such networks constructed from word associations can capture information the other methods cannot produce [6, 9]. Using networks for this task can also reveal whether the pandemic-induced issues are related to each other, which can be detected through the existence of strong network connections. All these investigation techniques reveal significant details regarding the obvious as well as the latent consequences of the pandemic. Such revelation is significant to inform the literature and the public regarding which issues were present during the pandemic, since the popular presence of certain subjects in the news can overlay serious problems that deserve more spotlight. These research studies can encourage positive steps to be taken towards resolving the under-studied issues.

In the following sections, a brief literature survey, the utilized news data collection and the selected analysis methods are described in detail. Subsequently, the conducted experiments and the analysis results are displayed and interpreted in the light of the pandemic. Finally, the conclusions are drawn regarding the presented results and the interpretations along with the reveal of future directions.

2. BACKGROUND

News collections became an attractive research domain during the pandemic. One of the prevalent research problems the researchers focused on was detecting fake news. It became necessary to separate the false information during the pandemic due to the political climate and the misinformation spread. They released datasets for training machine learning models for automatically detecting fake news from social media and news articles [10]. Subsequently, they utilized the power of machine learning and deep learning for fake news detection [11]. For example, a study proposes a web application that combines machine learning and crowd-sourcing to validate the credibility of the news content to avoid the spread of misinformation during the pandemic [12].

In addition to fake news detection, some studies focus on capturing general information from news articles. One of such studies evaluates the effects of the pandemic on the stock market using sentiment analysis on the news data using the articles published in MarketWatch.com, Reuters.com, and NYtimes.com [13]. They report a positive correlation between the sentiment scores in news articles and the stock market reactions. Meanwhile, another study employs Latent Dirichlet Allocation (LDA) on news articles published in Brazil [14]. They acquire topics about confirmed cases, economic influences, entertainment, medical supplies, treatment and research, politics, stories, and pandemic prevention, covering the reaction in Brazil towards the pandemic and its consequences. Likewise, another study applies topic modeling on the Croatian news articles published during the COVID-19 pandemic using LDA [15]. They detect a topic related to the earthquake that occurred in Croatia during the first year of the pandemic. Further, they find topics about online education, elections in Croatia, crime, anti-epidemic-measures protests, and travel. Another study utilizes the news collected from World Health Organization and the Global Public Health Intelligence Network and conducts an embedded topic modeling method to analyze the data [16]. A study similar to the current one proposes a new outtake on the Latent Dirichlet Allocation (LDA) method, called PAN-LDA, a new feature extraction technique for conducting machine learning studies on news articles [17]. They use a shorter time-spanning version of the same news articles collection from AYLIEN that ends in May 2020. They detect topics spanning sports, finance, business, entertainment, countries, politics, and health-related terms. Then, they propose ways of utilizing the new features on other machine learning tasks.

In the current study, in addition to the longer-time-spanning dataset, three different topic detection techniques are utilized to analyze the main issues covered by the media during the first year of the pandemic. Additionally, this study provides a statistical temporal analysis of the selected terms and discusses the observations from the social science perspective. (Scripts available in https://github.com/ubayram/covid_news_analysis)

3. METHODS

3.1. Data and Pre-processing

The news collection used in this study was provided by AYLIEN news API [4], publicly available for researchers. It spans larger than 25 Gigabytes in size after decompression, and the file is in “JSONL” format. This uncompressed file contains a collection of news articles written in English language, collected from November 2019 to the end of July 2020, which spans the beginning period of the global pandemic. The articles are assembled from around 440 international news agencies by the API, therefore they contain global news coverage. Along with the complete news in the textual format, it also contains details such as the date of publication, author, and even additional metadata such as the sentiment analysis results. The sentiment analysis of the articles were obtained by the methods AYLIEN researchers used, providing whether the news article has a positive or negative connotation, and an associated score.

In this study, all the metadata available are removed and only the text bodies of the news articles are kept. Next, in order to make sure only the news articles related to the pandemic are used in the experiments, texts that do not mention “Covid” or “Coronavirus” keywords are eliminated. After these processes, the collection had 1,231,556 remaining news articles. Before applying any machine learning or other computational methods, standard text pre-processing techniques are applied to the texts. These techniques include lowercase conversion and punctuation removal except the contractions. Also, the end-of-sentence marks are kept/used in the network construction but are removed for other processes. Lemmatization or stemming methods are not implemented, because a difference between the plural/singular forms of the words might provide important information. Next, stopwords are removed before the LDA topic modeling and the network construction. In this context, stopwords are the set of frequent words that do not contribute to a meaning alone (e.g. the, a, could, should), which might bias the analyses through their high occurrences. For the implementation, the set of stopwords present in the sklearn-API are used, which contains 318 words [18].

3.2. Latent Dirichlet Allocation

Topic modeling is a sub-field of both machine learning and NLP. As a field, it encloses the detection of “abstract topics” from large text collections in terms of associated words for making tasks like document summarization scalable [19]. Since it is unfeasible to label large collections of texts that require summarization or topic modeling, most methods available for the task are unsupervised. Among the unsupervised topic modeling methods, a popular one is LDA [20, 21]. This method is based on two main assumptions. First, it assumes that documents, i.e. texts, contain multinomial probability distributions of latent topics. Second, it assumes that topics are multinomial probability distributions over words. An LDA implementation requires two main inputs to be provided by the user: the text collection and the parameter “k” that is the number of topics in the text. However, for a user to know or prescient the number of topics available in a large text collection is unfeasible, which makes this one of the disadvantages of the LDA approach. Yet, this obstacle can be handled by experimenting with various numbers of topics and analyzing the results [22]. After the inputs are provided to the model, it implements an unsupervised methodology to compute and return a set of probabilities for each word and topic. Specifically, every topic in the set of “k” topics has the same vocabulary of words, but every word in every topic has a different probability associated with its prevalence within the context of that topic. Thus, each value represents a word’s probability of relevance to the detected topic. It is necessary to mention that LDA ignores any correlation between the topics. Meanwhile, its success in the literature undeniably validates its use despite the assumptions and the parameter tuning. In this study, the sklearn API’s LDA implementation with online variational Bayes algorithm is used [18].

3.3. BERT-based Topic Modeling

Following the development of models like ELMo and GPT-1, Bidirectional Encoder Representations from Transformers (BERT), a bidirectional transformer-based machine learning model has changed the course of NLP research significantly since its release in 2018. Through bidirectionality and self-attention, it can capture the context regardless of its local distance, which provided success on various benchmarks [23], and influenced many new models to be constructed. Among these new models are RoBERTa [24], ALBERT [25], DistilBERT [26], and the race for releasing different/better BERT-based models is still ongoing. Initially pre-trained on large text collections on expensive TPUs for many days, the released models are available to fine-tune and use for various NLP tasks. Influenced by all the exaltation about BERT-based models, this study includes a BERT-based topic modeling approach for comparison.

BERTopic is a BERT-based topic modeling framework developed by Grootendorst [27]. In the first step, it transforms the texts into embeddings. By default, it uses the "all-MiniLM-L6-v2" BERT sentence-transformer, but the users can provide their preferred embedding models that could be outside the range of BERT models. Since "all-MiniLM-L6-v2" is among the fastest available BERT sentence-transformers and is trained on a large collection of various styles of texts, it is employed in this study. These embedding models are applicable to paragraphs as well as sentences. Yet, very long texts can cause a computational burden. Therefore, in the implementation, the long news articles are shortened to contain only the first 125 words, which should be enough to capture the context of the news article's subject of interest. At the end of the first step, a 384-dimensional dense vector space is obtained from the news articles. In the next step, the framework applies the Uniform Manifold Approximation and Projection (UMAP) dimension reduction method on embedding space to reduce the dimensions from 384 to 5 [28], then utilizes HDBSCAN to cluster the news articles based on their embedding vectors in the new embedding space [29]. Finally, it employs a class-based variant of tf-idf (c-tf-idf) that merges all documents in the same cluster and applies tf-idf to find the statistically influential words in that cluster. Thus, at the end of the topic modeling process, these leading words are utilized to make sense of the detected topics.

3.4. Lexical Networks

Constructing a network from the mathematically computed associations between the words is a beneficial approach to model the contents of textual data [6, 9, 30]. Therefore, as the next step of investigating the news collection, a network is constructed based on the word associations. In this study, the association of two words corresponds to their co-occurrence within the sentences. The reason behind using the sentence-level co-occurrence is the goal of capturing the semantic relevance of words, i.e. two words co-existing in the same sentences would be related to the same sentiment expressed in the sentence. However, in a direct co-occurrence frequency count scenario, the most common words would contaminate the potential for capturing semantic relevance and simply would dominate the associations. Therefore, the mathematical equation that computes the association strength between the words should reduce the effect of the occurrence frequencies of the words, which resulted in selecting pointwise mutual information (PMI) as the metric. PMI is one of such measures that attempts to capture the actual statistical dependence between the word pairs rather than just the linear dependence [31]. The equation is as follows:

$$pmi(i,j) = \frac{log\big(p(i,j)\big) - log\big(p(i)p(j)\big)}{log\big(p(i,j)\big)}$$

In this equation, the association between two words, i and j, are computed, where p(i) and p(j) are the individual probability of occurrences of these words, and p(i, j) is the probability of observing them together in the same sentence. In the implementation, these probabilities are computed by counting the number of sentences they were present together and individually, divided by the complete number of sentences present within the collection. In the equation, through the subtraction operation, the effect of their singular frequencies is eliminated from the associations, therefore commonly occurring words would not dominate the network connections.

After the PMI values are obtained for every word pair, the network is constructed where words are nodes and the words associated with a PMI value higher than a threshold are connected with undirectional, weighted edges. In the experiments, initially the value 0.60 is used as the first threshold. Although this value may appear arbitrary, it corresponds to an acceptable value among the PMI values computed according to the distribution of values. The goal is to restrict the network to contain only the strong connections (having a high connection weight/PMI value). More details about this network construction method and alternative association metrics are available in the literature [9].

4. EXPERIMENTS AND RESULTS

4.1. Statistical Analysis

The first analysis focuses on applying a simple analysis to look for clues reflecting the impact of COVID-19 over the collection. A simple form of statistically analyzing the contents of the dataset is to construct a word cloud as it is a feasible way for visually overviewing prevalent words in the collection [6]. In a word cloud, the words or phrases have font sizes proportional to their frequencies in the dataset, i.e. less common words have small fonts. Meanwhile, different colors improve readability. Figure 1 has the word cloud generated from the complete news data collection after the application of standard text pre-processing methods and stopwords removal. In the figure, the largest three words are "coronavirus," "people," and "covid19" as expected. There are many words related to hospitalizations, masks, lockdowns, and deaths, all intuitively known reflections of the pandemic. Meanwhile, there are also mentions of economy and work-related words in the cloud. The diversity of words highlight how the repercussions of the pandemic have been multidimensional.

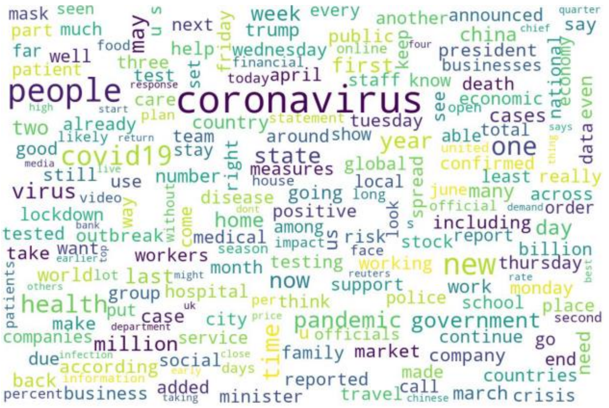


Figure 1. The word cloud constructed from the word frequencies in the complete collection of news articles.

Table 1. The most frequent words excluding stopwords and their relative ranks.

Word	*Rank*	*Word*	*Rank*	*Word*	*Rank*
coronavirus	1	people	2	covid19	4
health	6	pandemic	10	state	11
government	12	virus	13	home	17
million	21	china	23	cases	24
work	28	world	31	march	33
trump	34	country	35	company	37
market	38	outbreak	39	public	42
president	43	help	44	spread	49
hospital	54	lockdown	55	business	57
disease	67	stock	68	death	69
positive	72	crisis	78	medical	80
support	81	test	83	global	85
billion	86	school	87	national	92
testing	94	confirmed	95	workers	103
risk	106	staff	108	minister	109
u s	110	travel	112	countries	117
family	118	economic	119	measures	120

To further detect the most occurring words in the news articles, Table 1 has the words and their ranks in the ordered list of statistical occurrences. The set of words that are not reflective of a situation or issue, such as the month names, days, or numbers are removed from the table, thus the ranks can have large gaps in between due to the excluded words. The words present in Table 1 include politics-related words like "trump," in addition to the words directly related to the COVID-19 like "virus," "hospital," "medical." It also has words about the measures taken to stop the spreading of the virus such as "lockdown," and "school" possibly related to their temporary shutting-down. Among the most prevalent words in the table are related to the financial reflections of the pandemic and the related measures taken by the governments on the workforce as well as on the companies, expressed through the high-ranking words like "company," "economic," "business," and "market."

In Figure 2, the number of news articles mentioning "death" and "lockdown" is very high and at a peak in the month of April 2020. Although this month did not coincide with the most COVID-19 related deaths, it could be the month the public started to worry. Many countries declared lockdowns to prevent the spread and further deaths during that month. Meanwhile, in the following months, the number of news articles mentioning these two words decreased as the public focus started moving towards other issues, such as the BLM movement and the protests. Among the longitudinally analyzed subjects, the most prevalent one is the economy throughout the collection. When counting these news articles, the existence of one of the three words "economy," "economic" and "financial" are considered. How the economy surpasses the other subjects expresses the amplitude of the pandemic on the working-class, companies, stock market, and businesses. It is known that the pandemic caused many job losses and created an economic burden globally [32]. However, the economy-related words being uttered more often compared to the rest of the pandemic-related words, even surpassing the word "death" is a significant finding.

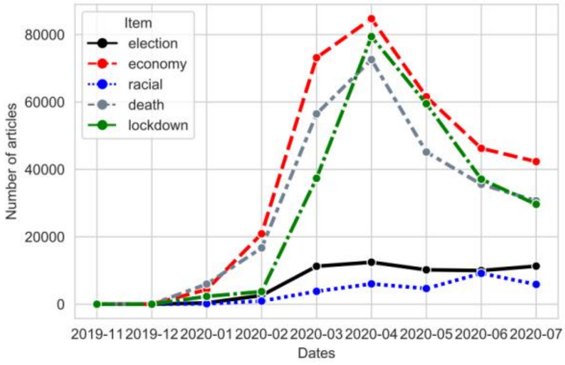


Figure 2. Monthly frequencies of the five subjects in terms of number of articles mentioning the related words.

4.2. Topic Modeling

The subsequent investigation focuses on taking the methodological complexity a little further. This analysis uses two machine learning methods to automatically determine the set of topics present in the collection. Topic modeling provides insightful, condensed information from vast text collections regardless of the language they contain [22]. In this context, topic modeling is expected to reveal the burdens of the pandemic in more detail compared to the previous statistical investigation in addition to the other contexts present within the news articles. Although it is one of the most popular topic modeling methods, LDA requires the exact number of topics, k, to be provided as an input parameter by the user. This is a disadvantage because the selection of a too small or a too high value might return undesirable topic modeling results. For example, if the k value was too high for the collection, some of the returned topics would not be interpretable. To overcome this disadvantage, this study experiments with LDA using two different "k" values. This repetitive experimentation would reveal whether the two LDA results obtained with different numbers of topics will have overlapping topics. Finding some topics that overlap between the low and high k-valued results would validate the presence of these topics in the collection. Therefore, one of the goals with this repetition is the detection of such overlaps. In the implementation, the remaining parameters are kept at defaults: learning method="batch", decay=0.7, offset=10.0, max iter=10, batch size=128.

Table 2. Top 20 words returned by the LDA model with 5 topics.

Topics	*Top 20 words per topic*
1:	demand, new, oil, financial, hedge, fund, global, economic, growth, china, stocks, 2020, stock, quarter, company, coronavirus, billion, year, market, million
2:	years, know, told, year, players, league, family, day, world, pandemic, it, going, season, new, home, people, like, just, time, coronavirus
3:	use, coronavirus, pandemic, covid19, make, company, going, help, think, business, need, it, re, new, work, time, just, we, like, people
4:	lockdown, reported, disease, positive, spread, hospital, minister, outbreak, number, country, patients, china, new, government, virus, health, covid19, cases, people, coronavirus
5:	told, federal, white, week, officials, covid19, city, news, york, public, house, pandemic, health, states, people, new, state, president, trump, coronavirus

Table 2 shows the topics returned from the LDA where five topics are selected for the experiment. Each topic is represented by the top twenty words with the highest probabilities associated with it. The first topic in the table contains the words "oil," "stocks," and "economy," which expresses the financial events reported during the pandemic by the news. However, based on the top words, these financial matters do not appear to be directly related to how the pandemic affected the society financially. Instead, they express how the stock market was affected. The second topic containing the top words "season," "league," and "players" shows it is expressing the news coverage about how the sports games have been handled during the pandemic. The third topic includes words like "company," "business," "people," and "work." It reveals how the society was financially affected by the pandemic. It appears that LDA captured the difference between the stock market-related news and the societal reflections separately as two different topics. Subsequently, the fourth topic in Table 2 expresses words related to the spread of the virus and the last topic reveals how the pandemic was handled in the United States by the president.

The second experiment with LDA using twice the number of topics returns the ten topics in Table 3. Coincidentally, the first topic detected by this LDA model also appears to be capturing the stock market-related words. Meanwhile, the second topic captures the football games. The words in this topic are more explicitly related to sports than those in the second topic of the previous LDA with five topics. This coherency can be explained by the increased number of topics allowed, so the model starts capturing the topics in more fine-grained fashion. Further, the third topic consists of words mentioning "business," "growth," and "vaccine," which might be a bit difficult to intuitively associate to a single topic. Meantime, the fourth topic is similar to the first one, expressing stock market related words. The contents of this and the previous topics returned by this LDA show that ten topics might be too many for this collection. Subsequently, the fifth topic contains words related to traveling during the pandemic and lockdowns, the sixth one expresses how the pandemic affected children, families, and education, and the seventh topic mentions the COVID tests and hospitalizations. Next, the eighth one comprises words associated with politics in the United States and the presidential election. Likewise, the ninth topic contains expressions about the medical workers, and the use of masks, while the last topic is related to the cases in India and how the government officials handled the situation there.

To compare with the LDA results and validate the findings in the previous tables, BERTopic is utilized as an alternative. For easy comparability with the LDA results, in the first experiment the number of topics is selected as ten. Instead of the complete collection, a large subset (hundred thousand news articles) of the corpus is selected at random and is used due to computational memory limitations. As there were no computational resources available with a compatible GPU, GoogleColab environment's free time-limited GPU is employed to obtain the BERTopic results.

Table 3. Top 20 words returned by the LDA model with 10 topics.

Topics	*Top 20 words per topic*
1:	price, earnings, shares, revenue, capital, investment, 2019, sales, 2020, year, investors, quarter, billion, funds, stocks, hedge, million, stock, company, market
2:	premier, training, teams, 2020, pandemic, fans, year, time, play, club, world, sports, game, football, team, games, coronavirus, players, league, season
3:	growth, work, really, good, that, ve, vaccine, company, quarter, time, million, new, going, like, it, just, year, think, business, re
4:	pandemic, cent, million, outbreak, world, new, markets, demand, reuters, prices, market, chinese, bank, economy, economic, year, global, oil, coronavirus, china
5:	work, airlines, open, lockdown, closed, flights, minister, workers, measures, employees, pandemic, march, business, restrictions, businesses, people, new, travel, government, coronavirus
6:	years, video, children, really, life, pandemic, school, know, says, going, it, day, family, new, coronavirus, home, just, time, like, people
7:	hospital, outbreak, positive, tested, disease, number, city, confirmed, patients, state, china, reported, deaths, covid19, virus, new, health, people, cases, coronavirus
8:	washington, administration, money, democratic, election, donald, people, campaign, americans, new, federal, biden, white, pandemic, states, state, house, coronavirus, president, trump
9:	information, home, use, virus, pandemic, help, face, work, medical, care, government, social, need, workers, masks, public, covid19, coronavirus, health, people
10:	prime, chief, indian, ministry, district, hospital, delhi, positive, country, health, police, state, cases, lockdown, people, india, minister, covid19, government, coronavirus

Contrary to the LDA implementation, no text pre-processing techniques are applied initially to the news articles before feeding them into the BERTopic. The pre-trained BERT sentence-transformers are trained on raw data containing stopwords, punctuation, and uppercase letters. BERT algorithm is known to ignore noise and it computationally highlights the contextually relevant words. Likewise, the tf-idf approach should also provide smaller weights to the stopwords. However, the experimental results returned the following top words for a topic with many stopwords: the, of, in, to, and, on, said, covid, 19, for, cases, from, has, have, was, government, minister, state, with, at. Thus, as the next step, the Gensim library is used to remove the stopwords from the news articles without applying any other text pre-processing before applying BERTopic [33]. It is important to note that this library does not include the “The” with an uppercase first letter in its default list of stopwords. Following the stopword removal without performing any other text pre-processing steps, the top twenty words for each of the ten topics returned by the BERTopic model are obtained as in Table 4. The topics are ordered from the most prevalent to the least based on the number of news articles that have fallen under each topic.

Table 4. Top 20 words returned by the BERTopic model with 10 topics.

Topic s	*Top 20 words per topic*
1:	league, season, players, the, football, club, premier, coronavirus, team, game, nba, clubs, games, play, manchester, year, said, player, united, training
2:	food, stores, the, coronavirus, said, sales, company, store, restaurants, pandemic, restaurant, customers, amazon, retail, 19, new, business, delivery, covid, it
3:	the, she, old, family, year, coronavirus, hospital, home, instagram, said, star, died, shared, it, prince, daughter, actor, he, mother, time
4:	cases, covid, positive, 19, said, hospital, the, tested, district, patients, state, number, health, reported, coronavirus, total, old, delhi, people, police
5:	trump, president, biden, house, coronavirus, democratic, donald, the, white, campaign, joe, election, senate, said, sanders, washington, presidential, vice, fauci, stimulus
6:	china, wuhan, cases, korea, chinese, coronavirus, virus, new, the, beijing, reported, south, outbreak, said, health, people, city, province, confirmed, hubei
7:	gold, investors, hedge, stocks, market, stock, points, index, funds, the, markets, dow, coronavirus, 500, nasdaq, trading, futures, prices, global, shares
8:	manila, philippines, said, city, covid, 19, the, sa, quarantine, ng, health, disease, duterte, cebu, coronavirus, doh, na, philippine, department, government
9:	airlines, flights, airline, air, said, airport, travel, flight, the, coronavirus, passengers, international, reuters, aircraft, boeing, aviation, india, passenger, delta, carrier
10:	masks, face, mask, wear, wearing, the, coronavirus, public, said, protective, coverings, people, medical, equipment, health, covid, 19, ppe, it, workers

In Table 4, the first topic is related to sports, conceivably about the cancellations of the games during the pandemic.

The second topic is about businesses like restaurants and delivery services, reflecting how they got affected by the lockdowns and the pandemic. The third topic appears to be about the reflections on entertainment with words about death, hospitals, stars, and actors, and the next one mentions Delhi, hospitalizations, and COVID. The fifth topic captures news articles about the reflections of the pandemic on US politics. The sixth one includes consistent words about China and the pandemic. Next, the seventh topic covers stock-market-related news. Similar to the sixth, the eighth topic is about the Philippines and their response to the pandemic. The next one enfolds news articles about air travel during the pandemic, and the last one covers the pandemic precautions like face masks and equipment. Compared to Table 3, these results match with LDA topics to a great extent with some exceptions. For example, BERTopic returns two distinct topics, one about China, another about the Philippines that LDA could not capture. Meanwhile, LDA appears to have captured different topics BERTopic failed to capture, about daily-life and about reflections on the economy.

While the first BERT experiment in Table 4 shows what BERTopic returns when the model is enforced to return specifically only ten topics, the follow-up experiment allows the model to auto-detect the number of topics it can discover through clustering the embeddings after dimensionality reduction. In the experiment, BERTopic automatically detected 72 topics. Due to space limitations, in Table 5, only the top fourteen of these topics are demonstrated. The topics are sorted from the most predominant to the least based on the number of news articles falling under each topic. The first topic in Table 5 appears to be related to India and Pakistan's response to the pandemic. The second and third topics are the same as the first and second topics in Table 4, capturing news articles covering businesses and sports events during the pandemic. It is noteworthy that all the sports-related terms detected for the third topic in Table 5 seem to be about England. Meanwhile, topic number seven also has sports-related words; but mainly American sports. Thus, allowing the model to automatically find the number of topics ended up sub-dividing the first topic in Table 4 into two: American and British. Subsequently, the fourth topic is unequivocally about the unemployment that affected many lives during the pandemic and the debated stimulus packages released by some countries such as the US. Beginning from the sixth up to the tenth topic, together with the twelfth topic in Table 5, found topics appear to be the same as the sixth, third, seventh, eighth, tenth topics in Table 4. Three new topics visible in Table 5 are the eleventh topic about how schools and academia were affected by the pandemic, the thirteenth mentioning Africa, and the fourteenth mentioning Canada during the pandemic.

Table 5. Top 20 words of top 10 topics returned by the BERTopic model with 72 automatically found topics.

Topics	*Top 20 words per topic*
1:	cases, covid, said, 19, delhi, state, minister, government, district, india, positive, the, lockdown, police, hospital, people, health, tested, patients, pakistan
2:	food, stores, sales, store, restaurants, restaurant, company, customers, amazon, retail, delivery, the, said, pandemic, coronavirus, business, meat, workers, employees, online
3:	league, premier, club, football, season, players, clubs, cricket, manchester, england, liverpool, rugby, united, team, matches, cup, match, training, chelsea, arsenal
4:	unemployment, businesses, business, work, million, stimulus, small, jobs, employees, program, claims, loans, benefits, economic, companies, bank, the, workers, rate, week
5:	trump, biden, president, house, democratic, donald, joe, white, campaign, senate, sanders, fauci, vice, presidential, washington, coronavirus, democrats, bernie, republican, pence
6:	she, old, died, star, family, hospital, daughter, year, mother, instagram, baby, the, home, shared, son, husband, her, children, it, coronavirus
7:	season, nba, players, league, nfl, baseball, team, teams, game, games, sports, play, draft, football, the, player, mlb, espn, basketball, nhl
8:	stocks, stock, investors, index, market, points, markets, dow, hedge, shares, rs, 500, nasdaq, crore, trading, quarter, funds, sp, wall, earnings
9:	china, wuhan, chinese, beijing, hubei, virus, cases, outbreak, province, coronavirus, new, reported, city, confirmed, health, people, the, commission, said, mainland
10:	manila, philippines, city, sa, ng, duterte, cebu, doh, na, philippine, covid, quarantine, 19, said, ang, disease, department, filipinos, health, cbn
11:	students, school, schools, education, exams, class, university, board, classes, exam, examinations, online, examination, teachers, campus, parents, the, learning, universities, academic
12:	masks, face, mask, wear, wearing, coverings, protective, public, equipment, medical, ppe, n95, people, surgical, coronavirus, the, personal, said, health, covering
13:	nigeria, africa, state, ghana, lagos, cases, government, 19, covid, president, country, governor, the, said, confirmed, health, african, buhari, south, recorded
14:	canada, province, 19, cases, covid, health, trudeau, ottawa, ontario, cbc, quebec, ca, news, toronto, font, manitoba, canadian, care, says, new

When the two BERTopic results are compared to each other, allowing the model to auto-detect the number of topics have resulted in returning more fine-grained, detailed topics, each focusing on more specific contexts. Especially the division of the sports articles into two based on the countries they are popular at is noteworthy.

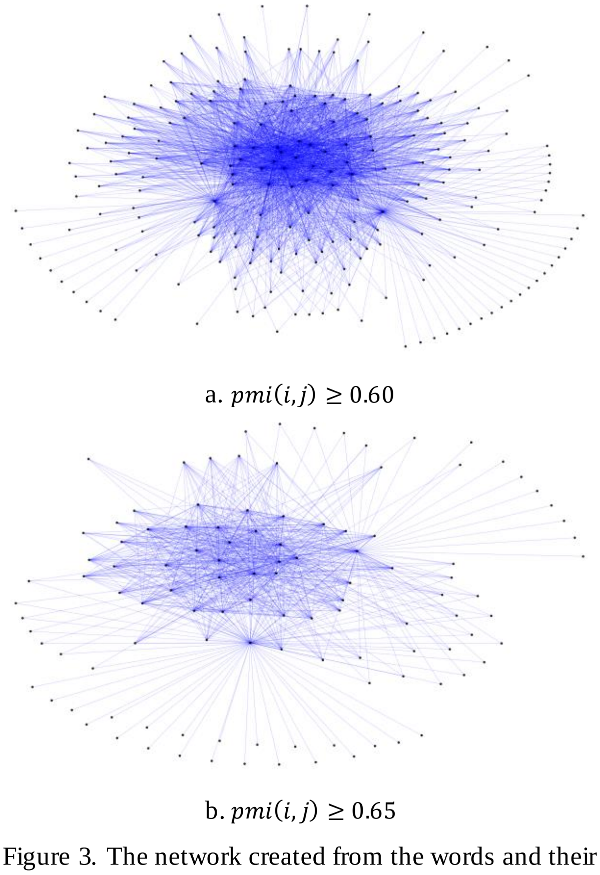
Meantime, the change in the order of topics is also noticeable as in the second experiment the India-related topic becomes the first topic instead of the fourth, possibly due to the fact that the first topic in Table 4 got divided and its cluster shrank in size. In terms of comparing LDA and BERTopic, there is a large amount of overlap between the topics they returned. Therefore selecting one method over the other requires the consideration of other factors in effect. For example, if the goal is to obtain more fine-grained topics where each topic is specifically about a theme, such as British sports, it might be ideal to select the unrestricted BERTopic. However, another important factor that needs consideration is the computational resources these methods require. BERT-based models require GPU for speed, and a compatible GPU is unfortunately expensive and unavailable for many underprivileged researchers. Although environments such as Google Colab make it possible for everyone to experience the power of GPU on these complex models, it provides a time-limited access with a quota that can be increased for an additional cost, which also is difficult for the underprivileged. LDA could be a cheaper option when crude (compared to the fine-grained BERTopic) results are acceptable. In terms of the context, a significant outcome is the prevalence of sports, politics, economy, and business-related topics in the news articles. However, the number of articles expressing the real-life impact of the pandemic on everyday life being less than the sports-related articles is surprising and should be discussed and analyzed by social scientists.

When the previous study with the shorter version of the same dataset ([17]) is considered in comparison to the current results, their method returned a topic related to unemployment while their traditional LDA experiment failed to do so. Meanwhile, in the current study, BERTopic succeeds in returning unemployment in the fourth topic when the number of topics is unrestricted. Also the topics they returned overlap those in the current study, validating their results as well as the current study. Overall, LDA and BERTopic demonstrated their success as unsupervised and semi-supervised machine learning approaches by capturing similar topics regardless of the parameters or methodologies. Nonetheless, conducting another experiment using a completely different technique could reveal further details about the pandemic and how it affected social domains. The next experiment constructs a network from the relations between the words present in the news articles and applies network analysis.

4.3. *Network Analysis*

Following the previously disclosed methodology, an undirected, weighted network is constructed using the word associations present in the news collection. For scalability, the month where the number of news articles was prevalent, April, is selected. Since this month alone had more than 300,000 news articles mentioning the pandemic, a random sampling is conducted among these texts for computational scalability. Hence, 50,000 articles are randomly sampled and used in the network construction.

From the sampled texts, stopwords are removed, and the remaining words are lemmatized. Next, when constructing a vocabulary from the texts, words that occur at least 500 times (minimum occurring in 1% of the set) in the sampled collection are detected. This resulted in a vocabulary of 4,556 unique words. Every word in this vocabulary became a node in the network. Subsequently, the PMI values between the pairs of these words are computed over every sentence present in the sample set. Finally, to contain only the strong associations in the network, PMI values higher than or equal to 0.60 are used to create the edges between the nodes (words). The visualized network is present in Figure 3a. Once the PMI threshold is increased by 0.05, the eliminated connections result in a clearer network as in Figure 3b.



a. $pmi(i,j) \geq 0.60$

b. $pmi(i,j) \geq 0.65$

Figure 3. The network created from the words and their associations within the news articles, visualized after two different weight thresholding scenarios.

There are a few methods available for analyzing and interpreting such networks. For example, networks constructed from original data are commonly found to have dense local connections and sparse outer connections [34]. Such networks are called to possess the small-world attribute [35]. Whether a network has a small-world property can be determined using two topological measures: average path length, and the clustering coefficient. The first measure is the average of the shortest path lengths from one node to another. It provides information regarding how connected the network nodes are, i.e. a small average path length would mean nodes are densely connected in the network. The following measure is the degree to which nodes in the network tend to cluster together. It is computed based on the number of triangles present between the connections. In a small-world network, the average path length is expected to be low, while the clustering coefficient is expected to be high [35]. However, as it is challenging to interpret what value is high or low in the absence of a comparison baseline, researchers developed measures to determine the small-worldness by performing comparisons with random graphs. One of them is the *omega* measure that returns values between -1 and 1 [36]. In this metric, values close to 0 mean that the network has small-world properties, while values close to -1 mean the network has a lattice structure, and values around 1 constitute a random graph. When this metric is applied to the network in Figure 3b, the omega value returned by NetworkX API ([37]) over comparisons with five random graphs and ten iterations is 0.036, which is close to 0. To further validate this small-world finding, the next available measure to experiment with is called *sigma* [35], where networks with values above 1 are considered to have small-worldness. The metric returns 0.996 from the network in Figure 3b. Both these findings then validate that the thresholded network created from the word associations present in the news articles has the small-world property.

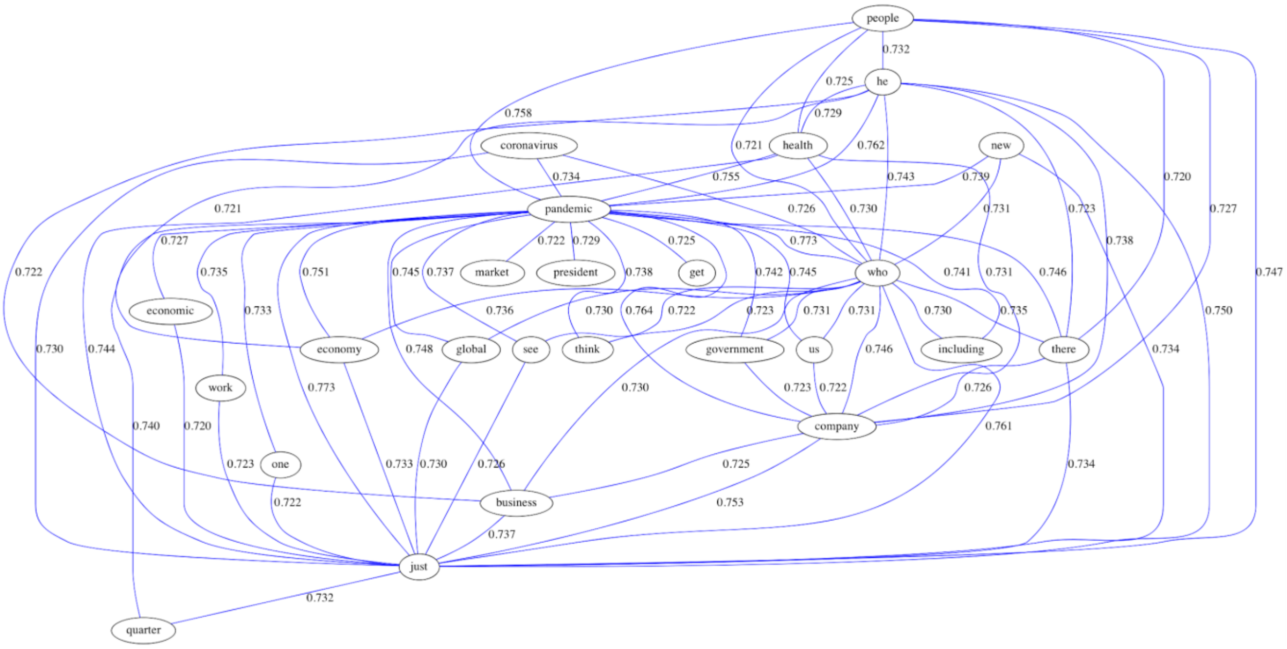


Figure 4. The strongest connections in the network were created from the words and their associations within the news articles, visualized after thresholding the connection weights by 0.72.

Semantic networks in the literature showed small-world properties before [38]. In this data domain, the network showing small-world property expresses a few things about the method, and the data. First, it means that the PMI metric and the way the network was constructed from the data are able to capture the relations between words in a meaningful manner, which should be considered by future studies. Second, the fact that some central nodes are connected to many local ones, creating grouping-like trends (i.e. small-worldness) means that the central connections are worth further analysis, as they are expected to be related to the most influential words in the articles. Thus, the next step should be to visually observe which words are these strongly connected central nodes. Consequently, increasing the PMI threshold to 0.72 returns the network in Figure 4. In this figure, among the most central nodes with influential connections to the others are "who," which is the lowercase version of WHO (World Health Organization) that had a great responsibility during the pandemic, together with "pandemic" as one of the most predictable words along with "just" and "company." The words "economic" and "economy" are both directly connected to the node "pandemic" with strong connections, which validates the significant finding of this study: one of the most important effects of the pandemic has been financial. Meanwhile, the node of WHO connected to "economy," "company," and the node "us" which is the lowercase version of the shortened "United States" are striking findings. The WHO website reveals that since 2020, the organization has been discussing the effects of the pandemic on the economy and since then they have been developing a new council for discussing the economical effects of health crises [39]. Thus, the network captures this effect strongly as it captures the anticipated problems related to the outbreak. Meanwhile, the absence of a strong connection between the nodes "president" and any of the economy-related nodes shows that news coverage was not connecting the two subjects.

5. CONCLUSION

The COVID-19 pandemic affected many paradigms. The health-care systemgot affected by the outbreak, education systems had to be adjusted for lockdowns, politicians debated many aspects of the pandemic, governments declared some measures like banning travel, among many other aspects of life that were affected. In this study, complementary to the existing studies, various different investigative approaches are used to reveal the pandemic-affected paradigms covered by the news during the first nine months of the outbreak. According to the results, the detected paradigms include sports, politics, elections, stock-market, education, and most prominently, the economy. The machine learning and network-based investigative methods all support the findings of each other regarding how the news organizations covered the various repercussions of the pandemic. In fact, how these repercussions were detected along with other more prevalent pandemic-related issues like hospitalizations and deaths, even the ongoing presidential elections in the US and all the news covering the sports-related events during the era could not overshadow the seriousness of the economic problems faced by many. The investigation framework proposed in this study is applicable and extendable to other problem domains, and can be adapted by other researchers and has a lot of potential for revealing latent information. In future work, more research can be conducted to dive deeper into different aspects of the pandemic using more complex, unsupervised and semi-supervised deep learning techniques, while social media data can be introduced for comparing the news coverage with people's opinions.

REFERENCES

[1] A. Khattar, P.R. Jain, S. M. K. Quadri, "Effects of the Disastrous Pandemic COVID-19 on Learning Styles, Activities and Mental Health of Young Indian Students - A Machine Learning Approach," **In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)**, 1190–1195, 2020.

[2] M. Yang, C. Han, "Revealing Industry Challenge and Business Response to COVID-19: A Text Mining Approach," *International Journal of Contemporary Hospitality Management*, 33(4), 1230-1248, 2021.

[3] N. Tuna, A. Sebatlı Sağlam, F. Çavdur, "Covid-19 Salgını ile İlgili Paylaşımlar Üzerinde Veri Analizi", *Bilişim Teknolojileri Dergisi*, 15(1), 13-23, 2022, doi:10.17671/gazibtd.928990.

[4] Internet: AYLIEN Coronavirus News Dataset. http://info.aylien.com/coronavirus-dataset, 2020-08-09.

[5] J. Jensen, S. Naidu, E. Kaplan, L. Wilse-Samson, D. Gergen, M. Zuckerman, A. Spirling, "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech", *Brookings Papers on Economic Activity*, 1–81, 2012.

[6] U. Bayram, J. Pestian, D. Santel, A. A. Minai, "What's in a Word? Detecting Partisan Affiliation from Word Use in Congressional Speeches", **In 2019 International Joint Conference on Neural Networks (IJCNN)**, 1–8, 2019.

[7] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, 993–1022, 2003.

[8] S. P. Borgatti, A. Mehra, D. J. Brass, G. Labianca, "Network Analysis in the Social Sciences," *Science*, 323(5916), 892–895, 2009.

[9] U. Bayram, A. A. Minai, J. Pestian, "A Lexical Network Approach for Identifying Suicidal Ideation in Clinical Interview Transcripts", **In International Conference on Complex Systems**, 165–172, 2018.

[10] P. Patwa, S. Sharma, S., S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, "Fighting an Infodemic: Covid-19 Fake News Dataset", **In International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation**, Springer, Cham, 21-29, 2021.

[11] R. Varma, Y. Verma, P. Vijayvargiya, P. P. Churi, "A Systematic Survey on Deep Learning and Machine Learning Approaches of Fake News Detection In The Pre-and Post-COVID-19 Pandemic", *International Journal of Intelligent Computing and Cybernetics*.

[12] N. L. Kolluri, D. Murthy, "CoVerifi: A COVID-19 News Verification System", *Online Social Networks and Media*, 22, 100123, 2021.

[13] M. Costola, M. Nofer, O. Hinz, L. Pelizzon, "Machine Learning Sentiment Analysis, COVID-19 News and Stock Market Reactions", *SAFE Working Paper*, 288, 2020.

[14] T. de Melo, C. M. Figueiredo, "Comparing News Articles and Tweets About COVID-19 In Brazil: Sentiment Analysis and Topic Modeling Approach", *JMIR Public Health and Surveillance*, 7(2), e24585, 2021.

[15] P. K. Bogović, A. Meštrović, S. Beliga, S. Martinčić-Ipšić, "Topic Modelling of Croatian News During COVID-19 Pandemic", **International Convention on Information, Communication and Electronic Technology (MIPRO)**, 1044-1051, IEEE, 2021.

[16] Y. Li, P. Nair, Z. Wen, I. Chafi, A. Okhmatovskaia, G. Powell, Y. Shen, D. Buckeridge, "Global Surveillance of COVID-19 by Mining News Media Using a Multi-Source Dynamic Embedded Topic Model", **In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics**, 1-14, 2020.

[17] A. Gupta, R. Katarya, "PAN-LDA: A Latent Dirichlet Allocation Based Novel Feature Extraction Model for COVID-19 Data Using Machine Learning", *Computers in biology and medicine*, 138, 104920, 2021.

[18] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vander-Plas, A. Joly, B. Holt, G. Varoquaux, "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project", **In ECML PKDD Workshop: Languages for Data Mining and Machine Learning**, 108–122, 2013.

[19] A. Haghighi, L. Vanderwende, "Exploring Content Models for Multi-Document Summarization", **In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)**, 362-370, 2009.

[20] M. Hoffman, F. R. Bach, D. M. Blei, "Online Learning for Latent Dirichlet Allocation", *In Advances in Neural Information Processing Systems*, 856–864, 2010.

[21] M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, "Stochastic Variational Inference", *The Journal of Machine Learning Research*, 14(1), 1303–1347, 2013.

[22] K. Deng, P. K. Bol, K. J. Li, J. S. Liu, "On the Unsupervised Analysis of Domain-Specific Chinese Texts", **Proceedings of the National Academy of Sciences**, 113(22), 6154–6159, 2016.

[23] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", **North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**, 4171-4186, 2019.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, D., …, V. Stoyanov, "ROBERTA: A Robustly Optimized BERT Pretraining Approach", *arXiv preprint*, arXiv:1907.11692, 2019.

[25] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut,. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", *arXiv preprint* arXiv:1909.11942, 2019.

[26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter", *arXiv preprint* arXiv:1910.01108, 2019.

[27] M. Grootendorst, "BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics", Zenodo, Version v0.9.4, 2020.

[28] L. McInnes, J. Healy, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction", *ArXiv* e-prints 1802.03426, 2018.

[29] R. J. Campello, D. Moulavi, J. Sander, J. "Density-based clustering based on hierarchical density estimates", **In Pacific-Asia conference on knowledge discovery and data mining**, Springer, Berlin, Heidelberg, 160-172, 2013.

[30] U. Bayram, R. Roy, A. Assalil, L. BenHiba, "The Unknown Knowns: A Graph-Based Approach for Temporal COVID-19 Literature Mining", *Online Information Review*, 45(4), 687–708, 2021.

[31] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction", **Proceedings of GSCL**, 31–40, 2009.

[32] L. Nassif-Pires, L. L. Xavier, T. Masterson, M. Nikiforos, F. Rios-Avila, **Pandemic of Inequality**, Technical Report, Levy Economics Institute, 2020.

[33] R. Rehurek, P. Sojka, "Software Framework for Topic Modelling with Large Corpora", **In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks**, 45-50, 2010.

[34] M. E. J. Newman, "Modularity and Community Structure in Networks", **Proceedings of the National Academy of Sciences**, 103(23), 8577–8582, 2006.

[35] M. D. Humphries, K. Gurney, "Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence", *PloS One*, 3(4):e0002051, 2008.

[36] Q. K. Telesford, K. E. Joyce, S. Hayasaka, J. H. Burdette, P. J. Laurienti, "The Ubiquity of Small-World Networks", *Brain Connectivity*, 1(5), 367–375, 2011.

[37] D. A. Schult, P. Swart, "Exploring Network Structure, Dynamics, and Function Using NetworkX," **In Proceedings of the 7th Python in Science Conferences (SciPy 2008)**, 11–16, Pasadena, CA, 2008.

[38] Y. N. Kenett, O. Levy, D. Y. Kenett, H. E. Stanley, M. Faust, S. Havlin, "Flexibility of Thought in High Creative Individuals Represented by Percolation Analysis", **Proceedings of the National Academy of Sciences**, 115(5), 867–872, 2018.

[39] Internet: World Health Organization (WHO). Global experts of new WHO Council on the Economics of Health for All Announced. https://www.who.int/news/item/06-05-2021-global-experts-of-new-who-council-on-the-economics-of-health-for-all-announced, 2021-06-05.