3.5. The early diagnosis of catastrophic events: weak signals detection based on information field monitoring

Oleksii MINTS

Department of Finance and Banking

SHEI Pryazovskyi State Technical University

Dnipro, Ukraine

Mykolas Romeris University

Ateities g. 20, Vilnius, 08303, Lithuania

E-mail: mints_a_y@pstu.edu

ORCID ID: 0000-0002-8032-005X

Zaneta SIMANAVICIENE

Public Security Academy

Mykolas Romeris University

Vilnius, Lithuania

E-mail: zasiman@mruni.eu

ORCID ID: 0000-0001-6008-2405

Kateryna POLUPANOVA

Public Security Academy

Mykolas Romeris University

Vilnius, Lithuania

E-mail: polupanova@mruni.eu

ORCID ID: 0000-0001-7310-0596

***Abstract**. Objective: This article explores the possibility of the early diagnosis of catastrophic events based on weak signal detection and information field monitoring. Data and Methods: Publications in The New York Times and The Guardian are*

used as a source of data. To analyze the dynamics of the information field, the concept of weak signals is applied. Text mining methods (text preprocessing, frequency analysis, word clouds) are used. Statistical methods for analyzing distributions and testing hypotheses are used to determine anomalies in keyword frequencies.

Results: The study makes it possible to confirm the assumption that the information field begins to change some time before the real onset of catastrophic events. The changes in the information field that preceded the two selected catastrophic events – the COVID-19 pandemic and Russia's military invasion of Ukraine – are analyzed. Statistically significant changes in the information field in both cases were observed long before the start of these events. It was revealed that deviations in the frequency of keyword mentions more than 4σ are anomalous, and can signal impending disasters. The assumption that it is possible to track the state of the information field through open media resources is also confirmed.

Conclusions: The results of the study confirm the possibility of using the analysis of the information field for the early diagnosis of global catastrophes. The creation of a system for monitoring the information field using the methods of machine linguistics can contribute to the timely preparation and minimization of threats to economic security.

Keywords: *catastrophes, economic security, text mining, detection, weak signals, black swan.*

Introduction

The modern world is characterized by a high degree of variability, and the number of factors that affect global processes is constantly increasing. The influence of unpredictable or difficult-to-predict factors – "black swan events" that directly determine economic security at the state level – is also growing. Over the past few years, among the most significant examples of such factors are the COVID-19 pandemic (2020) and Russia's military aggression against Ukraine (2022).

The originator of the "black swan" concept, N. Taleb, defines such events as fundamentally unpredictable (Taleb, 2012). On the other hand, I. Ansoff (1975) substantiated the need for and possibility of tracking "weak signals" in the context of increasing economic security. That is, he assumed that signals of catastrophic events can be detected some time before their actual occurrence (at least in some cases).

Despite the general similarity of views on the importance of studying global crises and catastrophes, there is no established opinion in the scientific community about the reaction to them. N. Taleb (2012) observed that it is possible to mitigate the consequences of a crisis by following the concept of "antifragility." J. Stiglitz (2018) believes that crises create new dynamics in the global economic system. For example, the 2008 financial crisis exposed flaws in the rules necessary for a stable and prosperous global economy, including the management of inevitable cross-border externalities. Jiuping Xu & Yi Lu (2018) note that it is necessary to rethink disaster management in terms of complex systems engineering. The same authors emphasize the importance of interdisciplinary research and systems thinking for disaster preparedness and relief. At the same time, it is emphasized that it is necessary to change the paradigm of the study of natural disasters towards their prevention.

Some authors, in particular Voronkova et al. (2022), put forward the idea of strategic forecasting, which is possible in the digital economy through an understanding of its global trends and processes. The same idea was independently developed by Štitilis and Laurinaitis et al. (2020). An important area of disaster management is research in the field of economic security, which makes it possible to minimize the consequences of disasters as well as to combat their occurrence. In this direction, we can also note the works of Malnar and Malnar (2015), Quinn and Cahill (2016), Huet et al. (2017), Kiseľáková et al. (2018), and Khadzhynova and Simanavičienė et al. (2022).

The problem of global catastrophes is undoubtedly important for ensuring economic security. The early identification of forthcoming transformations in the geopolitical situation will make it possible to better prepare for them, and the principal condition for ensuring economic security in these conditions is the formation of a combination of economic, social, political, cultural, and other factors that would ensure a sufficient level of stability. This corresponds to the concept of I. Ansoff (1975), but is expanded to the level of countries and regions.

At the same time, the issues of the early diagnosis of global catastrophes are poorly studied. In some cases, these events arise unpredictably, as noted by N. Taleb (2012) – for example, the crisis caused by the Evergiven container ship in the Suez Canal. However, a number of studies show that individual signals of approaching changes can be detected in advance (Welz et al., 2012; Zhan, 2022). It has also been found that social networks have the potential to monitor and detect signs of catastrophes (G. Amoudi et al., 2022; S. Mukherjee et al., 2022; K. Carley et al., 2016; Yin et al.; 2012; Štitilis et al., 2023). Moreover, Charitonidis et al. (2015) directly point to the possibility of using social media as a source of weak signals to predict adverse events in the real world.

T. Ahlqvist compared weak signals in social media with "hardly discernible cracks anticipating such earthquakes" (Ahlqvist et al., 2007). However, the aforementioned work describes only general ideas for the identification of such signals. Charitonidis et al. (2015) offered a more formal approach, but applied it to the prediction of adverse events in the social environment. Other works on diagnosing weak signals were analyzed in the preparation of this article, but they also distanced themselves from the problem of economic security.

It should be noted that Charitonidis et al. (2015) analyzed social media to predict adverse events using machine linguistics methods. The same methods were used in the works of Rajput et al. (2020) and Altmann et al. (2013) for similar studies on trends in social media (although without predicting their development).

However, there are no studies that clearly identify the sources of weak signals and prove that they can be used to improve economic security.

The main goal of this study is to test the fundamental possibilities of diagnosing catastrophic events on the basis of the study of the information field. To achieve this, we will substantiate the use of the information field as a data source for weak signals. Based on an information base from publications in The New York Times and The Guardian, two daily newspapers, we will develop a model for analyzing the information field using methods and tools of applied machine linguistics. Statistical hypotheses will then be formulated and tested regarding the characteristics of the information field in calm periods and before the onset of catastrophic events.

Methods

In this study, the concept of "weak signals" proposed by I. Ansoff (1975) is used to analyze the dynamics of the information field. According to this concept, global changes are preceded by weak, elusive signals that can be detected during the analysis of the information field. To interpret the concept of the information field, we start from the definition given in Morrill & Pitts (1967), developing it and adapting it to the context of the study.

To study the phenomenon of the early diagnosis of global catastrophes, this article uses a mixed approach, which combines qualitative and quantitative analysis. The research methodology consists of the following components.

Data Collection. As the main source of data, the archives of the publications of two authoritative daily newspapers specializing in international news were used: The New York Times (USA) and The Guardian (UK). This choice was determined by factors such as representativeness, regularity, openness, and accessibility. Both publications provide extensive coverage of world events, providing information on various regions and areas. The daily release of these publications ensures a

high frequency of data updates. Both publications provide access to publication archives via API, which simplifies data collection and analysis.

Machine Linguistics and Text Mining. The main tool for data analysis in this study is the application of machine linguistics and text analysis methods. We use statistical techniques for text data processing, including text preprocessing, frequency analysis, and natural language processing models, to extract insights from large amounts of data (Mints, 2017). Word Clouds were used to visualize the main topics and terms found in publications and to identify related events. The adequacy of the use of these methods in the context of the article is confirmed by the studies of Charitonidis et al. (2015), Rajput et al. (2020), and Altmann et al. (2013).

Statistical Analysis. Data analysis includes the use of statistical tests to identify anomalies in keyword frequencies that may indicate upcoming crisis events. This approach is based on the assumption that the prior frequency of word use in texts is subject to the empirical Zipf's Distribution Law (Baayen, 2001). To test statistical hypotheses, the Shapiro-Wilk test, the Two-sample T-test, the consequence of the Chebyshev inequality, and other tools of mathematical statistics were used. The principles of statistical studies of the distribution of words in large texts and their dynamics are described in the works of Baayen (2001), Rajput et al. (2020), and Altmann et al. (2013).

Ethical Considerations. This study took into account the ethical aspects of data collection and analysis. Particular attention was paid to the protection of personal data and the confidentiality of information.

Theoretical results

In the process of conducting the study, we proceeded from the fact that the object of the application of methods for analyzing text data and extracting information is *the information field.* Morrill & Pitts (1967) define the information field

as a carrier of the properties of the surrounding world. In contrast to the passive category of information space, the information field contains quantitative and qualitative characteristics that can be investigated.

For the sake of clarity, in this study we understand the information field as: *the space of the Internet, where multifaceted media communication between users is carried out, and where hypertext links between information elements exist and are created.* Such a definition makes it possible to specify the object of the research and its methodology.

For further investigation, let us make the following assumptions:

1. The information field begins to change some time before the real onset of catastrophic events.
2. Quantitative and qualitative characteristics of the information field can be tracked through open media resources.

The creation of a system for collecting, analyzing, and monitoring the information field based on open information, using the methods of machine linguistics, will make it possible to understand the processes in the digital world and predict sharp changes in trends. Monitoring current discussions will allow an observer to highlight which topics are becoming relevant and conduct additional research on their impact on economic security. Additional contextual analysis will reveal people's views on various topics discussed in the community.

It is important to consider the location of the information field in relation to the "data-information-knowledge-wisdom" pyramid, also known as the DIKW pyramid. This concept was first proposed in the context of information management and information technology, and describes a hierarchical approach to transforming raw data into information, then into knowledge, and finally into wisdom.

However, one of the first mentions of the pyramid is considered to be the work of Zeleny (1987) in the context of management systems. R. Ackoff (1989)

also made a significant contribution to the development and popularization of this idea. In particular, he observed that: data is transformed into information when it is organized for a purpose; information becomes knowledge when it is used to understand something; and knowledge becomes wisdom when it is used to improve human well-being (Ackoff 1989). The concept of information refinement, similar to the DIKW pyramid, was proposed by Dave Campbell (*Lankow et al., 2012*) to describe big data transformations. In this concept, the data layer is preceded by a signal, and the final stage of information enrichment is "insight."

Since its inception, the DIKW pyramid concept has gained widespread use in analyzing and understanding knowledge management, organizational decision-making, and individual thinking. However, in the context of this study, "data" can already be considered to represent a certain level of abstraction, since they already represent a certain sample set that characterizes the information field. Therefore, it is the information field that should be considered as the fundamental level of the DIKW pyramid, and as the basis for collecting data and their subsequent transformation (Fig. 1).

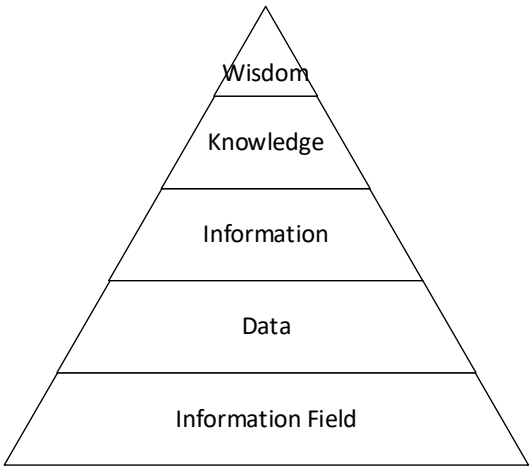


Figure 1. The DIKW pyramid, supplemented by the fundamental level of the information field

Source: authors

Compared to data, a key feature of the information field is that its elements are by definition unstructured and can be presented in text, audio, or visual form. Text form has been the primary way for humanity to exchange information for thousands of years. Compared to audio and visual forms, text allows information to be stored in the most compact way. In addition, modern AI models allow textual descriptions to be created based on content presented in other forms. Therefore, henceforth we will assume that the information field consists of textual information, or can be reduced to it.

Diagnosing global catastrophes in the context of improving economic security should involve extracting data from the information field and gradually transforming it into information, knowledge, and, ultimately, wisdom or understanding.

The first stage of the above, *the transformation of the information field into a compressed form* (data), seems to be the most difficult task due to the fact that it is necessary to simultaneously structure and compress a large amount of content. Taking into account the textual nature of the content to be transformed, methods of machine linguistics and natural language processing were used. These methods included tools which have been developing rapidly over the past few years, from statistical methods and frequency analysis to large language models. Each of these methods has both advantages and disadvantages, so it is important to evaluate them in the context of this study and choose the most suitable. In addition, supporting methods for collecting content and managing databases are necessary at this stage.

Empirical results

The main assumption underlying the concept of this study is that the information field begins to change some time before the actual onset of catastrophic events. Although similar findings can be found in various studies (Ansoff, 1975;

Zhan, 2022; Welz et al., 2012; Charitonidis et al., 2015), the authors were not able to find quantitative studies to support this assumption. Therefore, in order to substantiate the expediency of further research, it was necessary to conduct quantitative experiments to confirm or refute this assumption.

Data sources for the study must meet a number of requirements, including:

- - representativeness – data should describe all regions of the planet as evenly as possible;
- - relevance and regularity – necessary to ensure the prompt accounting of changes in the information field;
- - openness – data should be open and available for research and analysis;
- - compactness – since this study relates to intelligence, the compactness of the data set is also an important requirement.

The analysis of the requirements above shows that publications in daily newspapers that specialize in world news can be considered as an information base of the study. The specification of the requirements and a detailed consideration of various options led to the choice of The New York Times (USA)[3] and The Guardian (UK)[4] newspapers as the information base. Both publications cover global events, are issued daily, and print large volumes of materials. In addition, both publications are open for research and provide access to publication archives via API. Practical experience with these publications has shown that the New York Times API provides more flexible options for building queries, but allows users to process no more than 1000 articles per request. The Guardian API provides fewer options when setting up queries, but it has almost no limit on the number of articles.

The main tool for conducting the study was the Orange Data Mining software package, which includes advanced Text Mining tools. To perform statistical calculations, Matlab was used.

3 https://www.nytimes.com

4 https://www.theguardian.com

A scenario model for studying changes in the information field based on publications in the media is presented in Fig. 2.

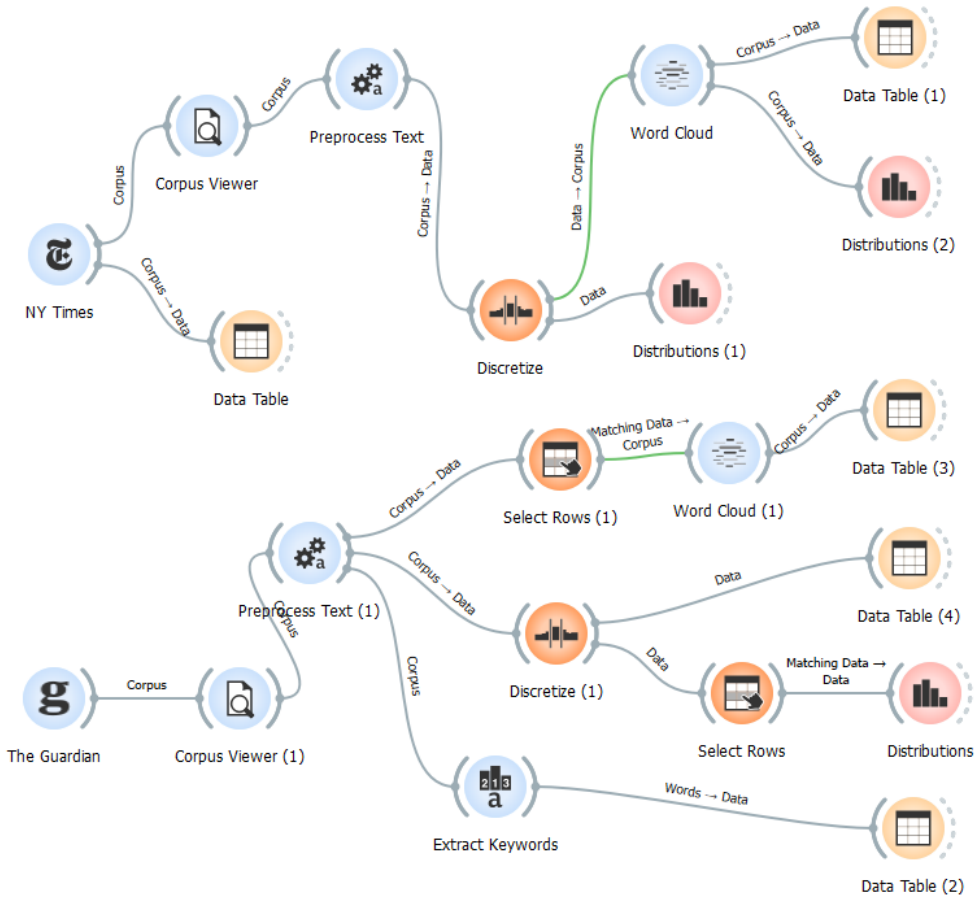


Figure 2. Scenario model for studying changes in the information field based on publications in the media

Source: authors

In the first step, a request is made to retrieve publications from the archive according to the specified keywords and time range. In this case, for further processing, the titles of publications, brief abstracts, or the full text can be used. Already at this stage, information can be obtained regarding the number of publications containing the specified keywords in different periods of time.

An analysis of the dynamics of changes in the information field before the outbreak of the SARS-CoV-2 (COVID-19) pandemic.

We studied the dynamics of changes in the number of publications containing the words *virus*, *pandemic*, *covid*, and *sars* in the period from November 1, 2019 to March 15, 2020, with an interval of 15 days. The start date of the study period was selected as it was earlier than the first documented case of COVID-19 symptoms in patient zero (Huang et al., 2020). The final interval of the study period includes the date of the pandemic announcement by the World Health Organization (March 11, 2020).

Table 1 and Table 2 contain information on the number of articles in the New York Times and The Guardian mentioning the relevant keywords for the period under study. Table 3 provides similar information for the reference period, which is taken from a similar time interval one year before the events under study.

Table 1. Number of articles with keywords in The New York Times, for the period from 01.11.2019 to 15.03.2020

Source: authors

keywords	2019				2020				
	01-15 Nov	16-30 Nov	01-15 Dec	16-30 Dec	01-15 Jan	16-30 Jan	01-15 Feb	16-29 Feb	01-15 Mar
virus	7	19	10	16	18	138	194	249	650
pandemic	2	1	3	7	8	88	164	208	702
covid	0	0	0	0	2	79	156	193	664
sars	2	1	1	0	7	73	69	41	41

Table 2. Number of articles with keywords in The Guardian, for the period from 01.11.2019 to 15.03.2020

Source: authors

keywords	2019				2020				
	01-15 Nov	16-30 Nov	01-15 Dec	16-30 Dec	01-15 Jan	16-30 Jan	01-15 Feb	16-29 Feb	01-15 Mar
virus	10	15	16	20	16	144	231	278	840
pandemic	3	1	4	0	5	21	33	121	454
covid	0	0	0	0	0	0	31	177	638
sars	0	1	1	0	3	73	72	45	80

Table 3. Number of articles with keywords in The New York Times, for the period from 01.11.2018 to 15.03.2019

Source: authors

keywords	2018				2019				
	01-15 Nov	16-30 Nov	01-15 Dec	16-30 Dec	01-15 Jan	16-30 Jan	01-15 Feb	16-28 Feb	01-15 Mar
virus	16	23	10	17	16	14	16	10	27
pandemic	7	7	3	4	5	7	5	1	7
covid	0	0	0	0	0	0	0	0	0
sars	2	1	3	1	1	1	3	2	3

A simple visual analysis and comparison of the data from Tables 1–3 makes it easy to notice a sharp increase in posts selected for even the most common of the keywords studied – the word "virus."

A similar conclusion can be reached by examining the distribution of publication frequencies over time graphically. Fig. 3 shows a histogram of the frequency of publications in The Guardian for January 2020, taken at intervals of 5 days. It can be seen that the most noticeable increase in the frequency of publications was observed after January 21 (that is, almost 50 days before the official announcement of the pandemic). However, in the previous period (January 16–20), significant changes were noticeable.

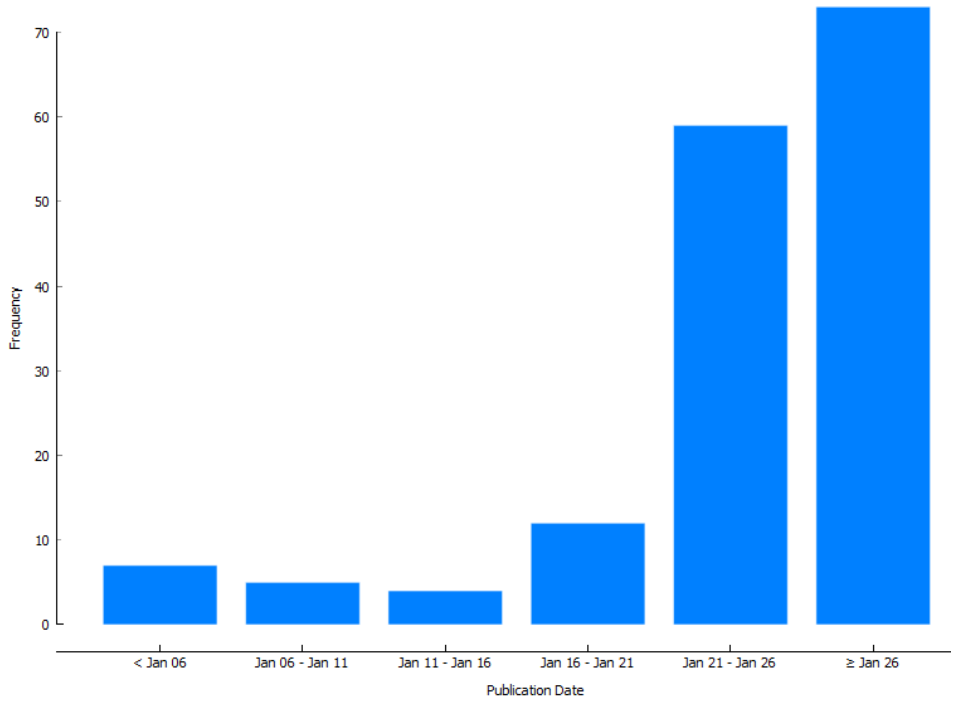


Figure 3. Distribution of publications containing keyword "virus" from Jan 1–30, 2020, in The Guardian

Source: authors

In order to quantify the magnitude of changes necessary to treat them as anomalous, it is necessary to conduct statistical studies. First of all, the distribution over time of the frequency of the appearance of keywords in the publications must be determined. Similar questions have repeatedly been raised in the scientific literature. The most frequently studied question is the distribution of words in large texts (Baayen, 2001; Baron et al., 2009; Rajput et al., 2020). At the same time, most authors confirm the relevance of the empirical Zipf's Distribution Law. According to this law, in the list of words ranked in descending order of popularity, the frequency of use of the n_{th} word is inversely proportional to its number n:

$$\text{Word frequency} \propto \frac{1}{word\ rank} \quad (1)$$

It follows from this law that the frequency of the use of individual words will be constant on average, and its deviations from the average value in individual texts will be influenced by random factors. If we assume that the probabilities of frequency deviation up and down are the same, then we get a normal distribution law. It should be noted that the frequency of word usage can vary with time, as noted by Altmann et al. (2013) and Rajput et al. (2020). Therefore, it is necessary to be careful when choosing the time interval for studying the parameters of distributions.

The data in Tables 1–3 represent the frequencies of keywords in publications distributed over time. The statistical characteristics of the frequency of the word "virus" in different periods of time can then be calculated. This word was chosen because the frequency of its appearance in the period preceding the pandemic, according to Tables 1–3, is the highest. This, according to the findings of Baayen (2001), increases the total number of observations and reduces the impact of random fluctuations on statistical outcomes.

Thus, the following data can be analyzed:

V1 – a sample of the frequency of the word "virus" in newspaper publications in the study period (from 1.11.2019 to 15.03.2020).

V0 – a sample of the frequency of the word "virus" in newspaper publications in the base period (from 1.11.2018 to 15.03.2019).

V19 – a sample of the frequency of the word "virus" in newspaper publications during 2019 (from 1.01.2019 to 26.12.2019).

As can be seen from this list, in comparison with the data in Tables 1–3, frequency sampling for the whole of 2019 is added. This sample was created with the same discreteness (15 days) as the others, and its analysis is necessary to check the stationarity of the distribution parameters.

To check the data against the normal distribution law, the Shapiro-Wilk test was used. The H0 hypothesis is formulated as follows: "V is normal with

unspecified mean and variance."

At a significance level of α=0.05, the Shapiro-Wilk test produced the following results:

For V19: sw=0.9441; p=0.2197. There is no reason to reject the H0 hypothesis.

For V0: sw=0.899; p=0.2463. There is no reason to reject the H0 hypothesis.

For V1: sw=0.7002; p=0.0036. The hypothesis of a normal distribution must be rejected.

Note that the test values (sw) and p-value for V19 and V0 are fairly close. This indirectly indicates the proximity of the distribution parameters. To assess this more accurately, the mean (μ) and standard deviation (σ) values can be calculated.

V19: μ = 15.0435, σ = 6.0671

V0: μ = 16.5556, σ = 5.5252

The confidence interval for estimating the mathematical expectation of the V19 sample can be found using the following formula:

$$\Delta\mu = \frac{\sigma}{\sqrt{n}} \cdot t(p, f) = \frac{6.0671}{\sqrt{24}} \cdot t(0.95,23) = 2.56 \qquad (2)$$

Since |16.5556-15.0435| = 1.5124 < 2.56, it can be observed that the statistical distribution of V0 corresponds to the distribution of V19.

Similar conclusions can be reached based on the calculation of the Two-sample T-test. The equality of the mean distributions can be taken as the null hypothesis, H0, alongside a significance level of α=0.05.

p-value(ttest2(V19,V0)) = 0.4352 > 0.05

Therefore, we cannot reject the H0 hypothesis regarding the equality of the mean distributions in the V19 and V0 samples.

Thus, in the absence of external disturbances, the frequency of mentions of a keyword in publications is approximately the same, and its variation corresponds to the normal distribution law.

Testing the hypothesis of the equality of the means of V1 and V19 using the Two-sample T-test gives the following value:

p-value (ttest2 (V19,V1)) = 0.005 < 0.05 μ.

The assumption of a normal distribution of words in publications makes it possible to statistically determine the probability of the observed deviation from the mean values. If the probability falls below critical values, then this may mean the emergence of new factors that change external conditions.

Thus, in Table 1, on the basis of which the V1 sample was compiled, in the period from January 16 to January 30, the number of publications mentioning the word "virus" increased to 138. The probability of such an observation can be calculated using the formula:

$$P(X \geq 138)=1-P(X \leq 137)=1-\Phi(z), \quad (3)$$

$$z = \frac{137 - \mu}{\sigma} = \frac{137 - 15.0435}{6.0671} = 20.1$$

Thus, the observed deviation was more than 20σ. With a normal distribution, 99.7% of observations fall within a much narrower interval of ±3σ. Therefore, the deviation of 20σ should be unequivocally interpreted as anomalous. Calculations show that if the previously observed distribution parameters are preserved, the probability of a deviation of 20σ is approximately 1.249×10^{-90}.

In practice, events that deviate from the expected value by more than 3σ are considered unlikely. However, in dynamics, the distribution of words in newspaper publications is most likely quasi-stationary. This follows from the work of Altmann et al. (2013). Therefore, additional research is needed to determine the limits of permissible deviations of frequencies from average values.

Nevertheless, in the context of the observations made, it is obvious that already in the second half of January 2020, the number of publications mentioning the word "virus" went far beyond the statistically possible dispersion. This could

be interpreted as a signal of future catastrophic events.

Analysis of the information field using text mining tools not only allows critical signals to be identified, but also permits additional data that is useful for clarifying the directions of danger to be obtained. Figure 4 shows a visualization of a Word Cloud generated from the Guardian based on the keyword "virus" from January 1 to January 15, 2020. It should be noted that within the framework of statistical analysis, this time period is not yet manifested as dangerous or abnormal.



Figure 4. Word Cloud generated from the Guardian based on the keyword "virus," 1–15 Jan, 2020

Source: authors

At the center of the cloud are words such as China, Flu, Illness, and SAR, which specify danger. The analysis of these keywords can provide the researcher with additional information about the state of the information field and potential threats to economic security.

Figure 5 shows a similar chart, which is based on publications with the keyword

SARS (Severe Acute Respiratory Syndrome) for the entire month of January 2020 in the New York Times. An analysis of this diagram already provides much more information not only about the causes, but also about the possible consequences of the spread of coronavirus. In particular, the word "quarantine" is quite clearly visible. It was the announcement of quarantine that became the main reason for economic losses around the world. Obviously, an additional 40 days to prepare for quarantine could have significantly reduced the losses of many companies.



Figure 5. Word Cloud generated from the New York Times based on the keyword "SARS," 1-30 Jan, 2020

Source: authors

Analysis of the dynamics of changes in the information field before the start of the war in Ukraine.

This method can also be used to consider the background of the beginning of another critically important event for the whole world: the war in Ukraine. Figure 6 shows a histogram of the frequency of publications in the Guardian with the

keyword "war" for the period from 1 Jan 2021 to 28 Feb 2022. To reduce information noise, only publications in the "World news" category were selected when building a histogram.

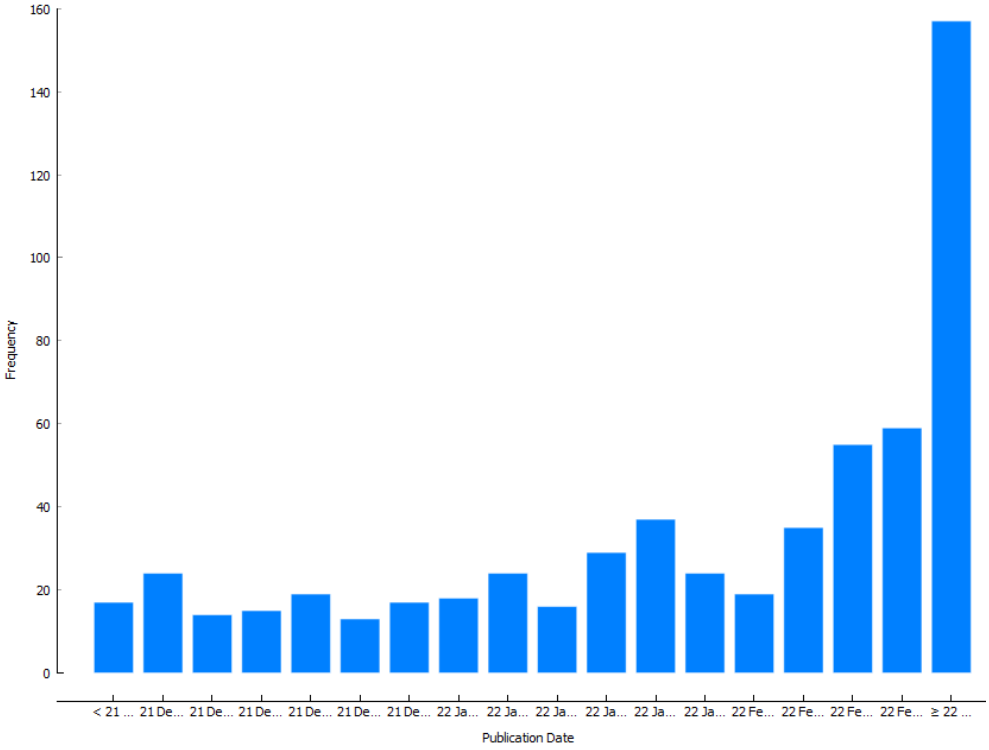


Figure 6. Distribution of publications containing keyword "war" in the Guardian from 1 Dec 2021 to 28 Feb 2022

Source: authors

The analysis of Figure 6 shows that there are no sharp frequency jumps on the histogram until the final pocket, which includes the date of the start of the Russian invasion: Feb 2, 2022.

The analysis of a similar sample of publications based on 2021 data shows that $\mu(war21) = 20.5352$; $\sigma(war21) = 6.4472$. The maximum frequency values were 45 (August 2021) and 42 (May 2021), which is 3.8σ and 3.33σ, respectively. At the same time, the histogram in Figure 6 shows that 10 days before the start of the war in Ukraine, the deviation from the mean value was more than 6σ. The chance

that such deviations are caused by random factors is only 9.87×10^{-10}. A comparison of distribution parameters also shows statistically significant discrepancies as early as January 2022.

The analysis of a Word Cloud generated by the keyword "war" (Fig. 7), as in the case described above, allows the threat and the parties associated with it to be clarified. In this Word Cloud of publications in the Guardian from 1 Dec 2021 to 15 Feb 2022, the central place is occupied by the words Ukraine, Russia, US, and UK. The same conclusion is confirmed by a study of a sample of publications from The New York Times for the keyword "war" for the period of December 15–30, 2021 (that is, almost 2 months before the start of the invasion). There, Russia and Ukraine are among the three most mentioned countries, along with Afghanistan.



Figure 7. Word Cloud generated from the Guardian based on the keyword "war," 1 Dec 2021–15 Feb 2022

Source: authors

The results obtained confirm the possibility of analyzing the information field

for the early detection of potential catastrophes, which can contribute to timely preparation and the minimization of threats to economic security.

This study has demonstrated the validity of the assumptions made in the theoretical results section. In both situations considered, the information field began to change significantly before the onset of major catastrophes. Statistically significant changes in the information field were observed around 50 days before the official announcement of the COVID-19 pandemic, and around 10 days before the start of Russia's military invasion of Ukraine.

The method of analyzing the frequency of keyword mentions in different time periods allows researchers to identify anomalous changes in the information field. This method is compatible with the apparatus of mathematical statistics, which is necessary for validating results. Statistical tests confirmed the importance of changes in the information field before the onset of the catastrophes under study. According to the results of the study, deviations in the frequency of mentions of words more than 4σ are abnormal. However, this conclusion requires additional verification.

Discussion and conclusions

Based on the analysis of literary sources, this article confirms the relevance of the scientific problem of the early diagnosis of global catastrophes in the context of ensuring economic security, and highlights promising methods for its solution. The possibility of diagnosing catastrophic events based on the detection of weak signals and monitoring the information field is shown.

This involved the creation of a system for collecting, analyzing, and monitoring the information field based on open information, using machine linguistics methods. This allows the processes in the digital world to be understood, and enables sharp changes in trends to be predicted. The article discusses the theoretical prerequisites for diagnosing global catastrophes in the context of increasing

economic security. The sequence of extracting data from the information field and gradually transforming them into information, knowledge, and, ultimately, wisdom or understanding is substantiated. Critical assumptions describing the relationship between catastrophic events, the information field, and its manifestations are formulated.

This study makes it possible to statistically confirm the assumption that the information field begins to change some time before the real onset of catastrophic events. Changes are manifested in the form of a sharp increase in the frequency of mentioning keywords and related terms, titles, surnames, etc. The assumption that it is possible to track the state of the information field through open media resources is also confirmed.

At the same time, it should be noted that there are some differences between the results of the article and the results obtained by other authors. One of the important results of this article is the conclusion that in the absence of external perturbations, the frequency of keyword mentions in publications is approximately the same, and its variation corresponds to the normal distribution law. However, it may turn out that under other conditions for the formation of the corpus of texts, other laws of distribution will be more suitable. This conclusion can also be reached if we take into account that the decrease in the frequency of words on the left is limited by the frequency of 0, and there are no such restrictions on the right. Rajput et al. (2020) Baron et al. (2009), and Altman et al. (2012) indicate log-normal, or Poisson distribution as possible options. However, it should be taken into account that the conditions for the formation of the data sample in these studies were significantly different from ours. Baron et al. (2009) examined a historical corpus of texts spanning an interval of 400 years; Altman et al. (2012) investigated the evolution of words in the word niche at an interval of around 12 years while Rajput et al., (2020) covered the period of the onset of the COVID-19 pandemic. Although the latter period is short, according to our results, it is in

times of crisis that significant changes in the information field occur. Therefore, further research is needed to confirm our findings, or to establish boundary conditions for them.

It is important to note that the results obtained are based on a relatively limited dataset and simplified analysis models. Since the main purpose of this article is to test the fundamental possibility of diagnosing catastrophic events on the basis of the study of the information field, this approach can be considered justified. Further research towards the creation of a system for the early diagnosis of catastrophes should involve the use of more representative data and sophisticated machine learning algorithms. This is necessary for the development of reliable forecasting methods.

This study shows that the use of text mining tools allows the direction of a threat and the parties associated with it to be clarified. The results obtained demonstrate the potential of the proposed approach for the early diagnosis of global catastrophes. Further research on this topic would involve for experimental work on the basis of open media and participatory media databases in order to analyze quantitative characteristics characterizing changes in the information field during major catastrophes.

References

1. Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16, pp. 3–9.
2. Ahlqvist, T., Halonen, M., & Heinonen, S. (2007). Weak signals in social media: Report on two workshop experiments in Futures monitoring. *SOMED Foresight Report.* Advance online publication. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5a263d161fedbfdde04c7c2e46481cc750856f27
3. Altmann, E. G., Whichard, Z. L., & Motter, A. E. (2013). Identifying trends in

word frequency dynamics. *Journal of Statistical Physics*, 151, pp. 277-288. doi 10.1007/s10955-013-0699-7

4. Amoudi, G., Almansour, A., Watters, C., Alahmadi, D., Alruwaili, F. & Alzahrani S. (2022) Tweet for help: the role of social media in disaster events and the case of the 2015 *Mina stampede, Digital Creativity*, 33:4, pp. 329-348, doi: 10.1080/14626268.2022.2141262
5. Ansoff, I. (1975). Managing strategic surprise by response to weak signals. *California Management Review*, 18(2), pp. 21–33.
6. Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). *Springer Science & Business Media*. Books.
7. Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. Anglistik: *International Journal of English Studies*, 20(1), pp. 41-67.
8. Carley, K.M., Malik, M., Landwehr, P.M., Pfeffer, J., & Kowalchuck, M. (2016). Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. *Safety Science*, 90, pp. 48–61. doi: 10.1016/j.ssci.2016.04.002
9. Charitonidis, C., Rashid, A., & Taylor, P. J. (2015, August). Weak signals as predictors of real-world phenomena in social media. *In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 2015 (pp. 864-871). https://core.ac.uk/download/pdf/76960803.pdf
10. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 395 (10223): pp. 497–506. doi:10.1016/S0140-6736(20)30183-5
11. Huet, C., Ford, J.D., Edge, V.L. et al. (2017) Food insecurity and food consumption by season in households with children in an Arctic city: a cross-sectional study. *BMC Public Health* 17, pp. 578. doi: 10.1186/s12889-017-4393-6
12. Jiuping Xu & Yi Lu (2018) Towards an earthquake-resilient world: from post-disaster reconstruction to pre-disaster prevention,. *Environmental Hazards*, 17:4, pp. 269-275, doi: 10.1080/17477891.2018.1500878
13. Khadzhynova, O., Simanavičienė, Ž., Mints, O., Burak, P., Khachatrian, V.

(2022) Assessment of the EU Countries' Economic Security based on the Composite Indicators // WSEAS transactions on business and economics. Athens : *WSEAS Press*, vol. 19, pp. 690-700. doi: 10.37394/23207.2022.19.61.

14. Kiseľáková, D., Šofranková, B., Čabinová, V., Onuferová, E. (2018). Competitiveness and sustainable growth analysis of the EU countries with the use of Global Indexes' methodology. *Entrepreneurship and Sustainability* Issues, 2018, 5 (3), pp.581 - 599. doi: 10.9770/jesi.2018.5.3(13)
15. Lankow, J. and Ritchie, J. and Crooks, R. (2012) Chapter 7. Data Visualization Interfaces // Infographics: *The Power of Visual Storytelling. Wiley*, 2012. 264 p.
16. Malnar, D. & Malnar, A. (2015). 'Demographic Security Trends in Southeastern Europe', Croatian International Relations Review, 21(73), pp. 57-87. doi:10.1515/cirr-2015-0011
17. Mints A.Y. (2017) Classification tasks of data mining and data processing in economy. *Baltic Journal of Economic Studies*, 3(3), pp. 47-52. https://doi.org/10.30525/2256-0742/2017-3-3-47-52
18. Morrill R. L., Pitts F. R. Marriage, migration, and the mean information field: a study in uniqueness and generality //*Annals of the Association of American Geographers*. 1967. T. 57. № 2. pp. 401-422.
19. Mukherjee, S., Kumar, R. & Kumar P. B. (2022) Managing a natural disaster: actionable insights from microblog data, Journal of Decision Systems, 31:pp. 1-2, pp. 134-149, doi: 10.1080/12460125.2021.1918045
20. Quinn, J. F., & Cahill, K. E. (2016). The new world of retirement income security in America. American Psychologist, 71(4), pp. 321–333. doi: 10.1037/a0040276
21. Rajput, N. K., Grover, B. A., & Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. arXiv preprint arXiv:2004.03925.
22. Stiglitz, Joseph E. (2018), "Lessons from the financial crisis and their implications for global economic policy", pp. 227-238, doi:10.7916/d8-1b0v-m790
23. Štitilis, Darius; Laurinaitis, Marius; Verenius, Egidijus. (2023) The use of biometric technologies in ensuring the security of critical Infrastructure: the context of the protection of personal data // Entrepreneurship and sustain-

ability issues : *Entrepreneurship and Sustainability Center*., vol. 10, iss. 3, pp. 133-150. doi:10.9770/jesi.2023.10.3(10).

24. Štitilis, Darius; Rotomskis, Irmantas; Laurinaitis, Marius; Nadvynychnyy, Sergiy; Khorunzhak, Nadiya (2020) National cyber security strategies: management, unification and assessment // *Independent journal of management &production: Special Edition (Baltic States)*. São Paulo : Instituto Federal de Educação, Ciência e Tecnologia de SãoPaulo. vol 11, no. 9, November, pp. 2341-2354. doi: 10.14807/ijmp.v11i9.1431.
25. Taleb N. N. (2012). Antifragile: Things That Gain from Disorder. Random House. 525 p.
26. Voronkova, V., Kaganov, Y., & Metelenko, N. (2022). Conceptual basis of "the digital economy forsite model": european experience. *Humanities studies*, pp. 9-19 https://doi.org/10.26661/hst-2022-10-87-01
27. Welz K, Brecht L, Pengl A, Kauffeldt J V, Schallmo D R (2012) Weak signals detection: Criteria for social media monitoring tools. Proc. the 5th ISPIM Innovation Symposium. 2012(International Society for Professional Innovation Management, Seoul, Korea),1. https://www.researchgate.net/publication/326260498
28. Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012)Using social media to enhance emergent situation awareness.. *In IEEE Intelligent Systems*, 27(6), pp. 52–59. Vol. November/, doi: 10.1109/MIS.2012.6
29. Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management, 7*(1), pp. 59–70.
30. Zhan, C. (2022). Early Identification Methods for Emerging Technologies Based on Weak Signals. https://doi.org/10.21203/rs.3.rs-2291140/v1

MYKOLAS ROMERIS UNIVERSITY

Public Security Academy

TRANSFORMATIONS, CHALLENGES AND SECURITY

Collective monograph

Mykolo Romerio universitetas

MYKOLAS ROMERIS UNIVERSITY

Public Security Academy

TRANSFORMATIONS, CHALLENGES AND SECURITY

Collective monograph

Mykolo Romerio universitetas

Vilnius, 2024

Reviewed:
Prof. dr. Badri Gechbaia (Batumi Shota Rustaveli State University)
Prof. dr. Tadeuz Trocikowski (Cuiavian University in Włocławek)

Editor-in-chief:
Prof. habil. dr. Žaneta Simanavičienė

Approved for publication by the Council of the Public Security Academy of Mykolas Romeris University, at the sitting on 13 September 2024. Resolution No. 1VS-8.

All publishing rights are reserved. This book, or any part of it, may not be reproduced, edited, or otherwise distributed without the permission of the publisher.
The bibliographic data of the publication is available on the portal of the Lithuanian Integrated Library Information System (LIBIS) "ibiblioteka.lt".

ISBN: ISBN 978-609-488-096-4 (El. - PDF)
ISBN 978-609-488-097-1 (Print)

© Mykolas Romeris University, 2024