Information & Security

D. Lande, I. Subach & A. Puchkov,
vol. 47, no. 1 (2020): 44-61
https://doi.org/10.11610/isij.4703

Published since 1998

ISSN 0861-5160 (print), ISSN 1314-2119 (online)

Research Article

# A System for Analysis of Big Data from Social Media

***Dmytro Lande*** [a,b] (✉), ***Igor Subach*** [b], ***Alexander Puchkov*** [b]

[a] *Institute for Information Recording, National Academy of Sciences of Ukraine, Kyiv, Ukraine, http://www.ipri.kiev.ua*

[b] *Institute of Special Communications and Information Protection, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine, https://iszzi.kpi.ua*

ABSTRACT:

The article presents the basic principles of building and using a monitoring and analysis system of social media on cybersecurity, based on the concepts of Big Data, Data/Text Mining, Information Extraction, Complex Networks. The authors substantiate information technologies for creating a system of content monitoring, selection of relevant information from social networks, implementation of search engines for their refinement by users, saving queries as RSS feeds, and maintaining personal databases in client applications.

The described OSINT system is based on collection of information from open sources, its analysis, preparation and timely delivery of the final product to the customer in order to solve certain intelligence tasks. Hence, the system is the result of a systematic collection, processing and analysis of the necessary publicly available information. It is based on the application of methods and tools of information retrieval, data analysis and aggregation of information flows, and is used for social media content monitoring as a component of decision support systems for information and cybersecurity.

ARTICLE INFO:

RECEIVED: 03 JUN 2020
REVISED: 30 JUL 2020
ONLINE: 21 AUG 2020

KEYWORDS:

social media monitoring, cyber security, OSINT, big data, cyber aggregator



✉ Corresponding Author: Tel.: +380674668917; E-mail: dwlande@gmail.com

## Introduction

Currently, in a hybrid war with a developed information component at many levels of government, there is a need to take into account information that appears in social media. It is known that information flows are sometimes a component of information confrontations, the content of which is aimed at implementing pre-planned psychological impacts on the audience to achieve predetermined goals. Such information, on the one hand, contains a lot of “noise,” misinformation, and, on the other hand, is the most operational. A significant amount of information resources in global networks contains various expert assessments, some of which are associated with the implementation of information threats, in particular, cyber threats. Based on this, accounting for information on web feeds is of great importance for solving problems in the field of ensuring cyber security, but until now there were no affordable budgetary solutions to the problem of targeted information, providing analytical generalizations to corporate users based on information from social networks. The paper suggests and substantiates approaches to building a corporate system for monitoring and analysing social media, which are based on the concept of processing large amounts of data (Big Data) from social media on cybersecurity issues, extracting knowledge from text arrays (Text Mining, Information Extraction).

One of the most important tools for ensuring cybersecurity today is Open Source INTelligence (OSINT) – one of the areas of intelligence, which includes the search, selection and collection of intelligence information obtained from public sources, and analysis of this information. The essence of the OSINT process is to search and analyze information obtained from open sources, collecting information and its further analysis, generating reports on the object of observation.

When creating systems for analyzing information from social media, it is necessary to solve the problem of large amounts of data obtained for analysis, their dynamics and a tendency to constant changes. This problem and ways to overcome it are called Big Data today. In this case, the implementation of the collection, cleaning, storage, retrieval, access, transfer, analysis, and visualization functions of such sets as a holistic entity rather than local fragments is problematic.[1] “Three V” are noted as defining characteristics for big data: (Volume, volume of physical volume), speed (Velocity, growth rate and the need for high-speed processing and obtaining results), diversity (Variety, the ability to simultaneously process different types of structured and weak structured data).

Today, according to research by Gartner, the term Big Data has already exceeded the peak of the famous Hype Cycle. In Fig. 1 shows the statistics of user queries to the Google system for the phrase "Big Data" (Google Trends service, https://trends.google.ru/).

Big data is a term that means a lot of data sets so voluminous and complex that makes it impossible to process them using the existing traditional database and application management tools.[2]

The work goal – a description of the technological foundations and tools for analyzing the content of social networks on cybersecurity, the corporate system with the maximum use of open access components to automate the processes of searching, monitoring, collecting, processing, analyzing, accumulating and storing information and then providing it to users and interacting decision support subsystems.

Recent research and publications analysis. When analyzing the literature related to the topic of this article, one can identify several important areas related to working with open data for solving cyber security problems, namely, open source intelligence (OSINT), working with big data (Big Data), and social networks analysis (Social Networks Analysis), extracting knowledge from texts, deep analysis of texts (Information Extraction, Text Mining).

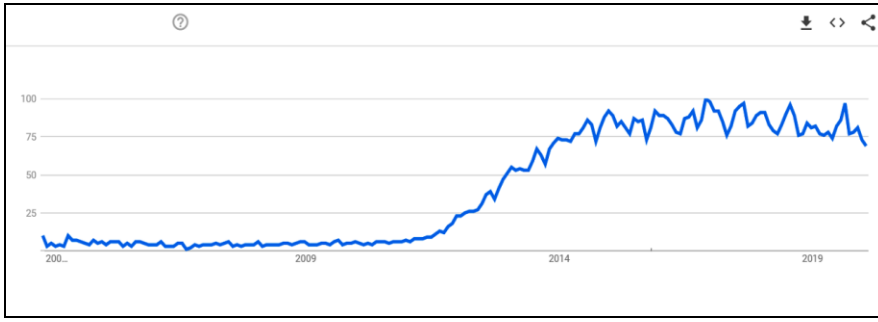


**Figure 1: – Queries "Big Data" dynamics (on the horizontal axis - time, on the vertical axis - popularity points from 0 to 100).**

The most significant of modern research, which belong to the direction of OSINT, include works of Layton and Watters [3] (Issues of automatic data collection, APIs and tools, machine learning algorithms, application of geographic information methods), Akhgar, Bayerl, and Sampson [4] (An open source operational research methodology for addressing security, combating organized crime and counter-terrorism issues – from planning to deployment), Memon and Alhajj [5] (Open source intelligence in the context of counter-terrorism, including models, tools, methods and case studies), Appel [6] (trends in the use of OSINT in the field of cybersecurity, methodology of Internet analytics in the legal framework) and many others.

As part of the analysis of Big Data are currently known studies by authors such as J.W. Foreman[7] (Mathematical optimization methods and genetic algorithms, data clustering methods, forecasting for processing large amounts of data), N. Marz, J. Warren[8] (Approach to the organization of data storage and processing, methodology of using open tools: Hadoop, Cassandra, Cascalog, ElephantDB and Storm with Trident), D. Cielen, A. Meysman, M. Ali[9] (Theoretical foundations, machine learning algorithms, NoSQL type DBMS, stream data processing, text mining, information visualization), K. Krishnan[10] (Strategies, archi-

tectures and implementation of high-performance solutions for data warehouses and unstructured data) and other.

Social network analysis (SNA) is the process of studying social structures using networks and graph theory. SNA characterizes network structures in terms of nodes (individual actors, people, or things) and the edges or connections (relationships or interactions) that connect them. Examples of social structures that are usually visualized using special tools are social networks, meme distribution networks, information networks, friendship and dating networks, knowledge networks, working relationships, cooperation, business networks, social networks, family ties, infection, and the like. These networks are often visualized as sociograms, in which nodes are represented as points, and connections are represented as lines.

The works of such modern scientists as J. Kleinberg and D. Easley[11] are devoted to the analysis of social networks (Strong and weak connections in social networks, structural balance in networks, modeling of network traffic using game theory, strategic interaction in networks, information networks), G. Ragozini, M. P. Vitale[12] (Creating association rules based on network analysis. Application in medicine, formation and research of a disease codes network, analysis of such networks by means of clustering algorithms), M. Kaya, J. Kawash, S. Khoury, M.Y. Day[13] (Analysis of social networks, time activity of publishing pages, identification of harmful sources of information in networks with the use of artificial neural networks, machine learning), M. Kaya, Ö Erdogan[14] (Memes on social networks studying and clustering, similarities between texts, emerging events identifying, forecast), K.A. Zweig[15] (Network analytical activities, methods of presenting data in the form of complex networks, models of random graphs, centrality indices and their use in network analysis, the humanitarian aspect of the analysis of social networks, etc). A large number of works are currently devoted to the in-depth analysis of information from social networks, among which the monographs by M. Russell[16,17] (Information extraction from various social networks, research of application program interfaces of various networks, analysis of text files, determination of text similarity, classification, pattern recognition, neural networks in the analysis of social networks).

## Open Sources Intelligence

According to [18], OSINT is based on the collection of information from open sources, its analysis, preparation and timely delivery of the final product to the customer in order to solve certain intelligence tasks, that is, OSINT is the result of a systematic collection, processing and analysis of the necessary publicly available information.

OSINT in cybersecurity is determined by a number of aspects, including promptness of receipt, volume, quality, reliability, ease of further use, cost of receipt, etc. The process of planning and preparing for OSINT management is influenced by the following factors: [19]

- the effectiveness of information support is achieved by collecting information from the Internet, user content, hashtags, etc.

- relevance. The availability, depth and scope of publicly available information allows you to find the necessary information without involving other specialized intelligence tools;
- simplification of data collection processes. OSINT provides the necessary information, eliminating the need to attract unnecessary technical and human resources;
- depth of data analysis. As part of the intelligence process, OSINT enables in-depth analysis of publicly available information in order to make appropriate decisions.;
- efficiency. A sharp reduction in access time to information on the Internet. Quick receipt of valuable operational information. The situation, which is changing rapidly during crises, is most fully reflected in the current news;
- volumes. Ability to mass monitor certain information sources in order to find targeted content, people and events;
- quality. Compared to special agents' reports, open source information is devoid of subjectivity;
- reliability;
- ease of use. OSINT data can be easily transferred to any interested instances, they are open;
- cost. The cost of obtaining data in OSINT is minimal.

## Technologies for Working with Big Data

When collecting, analyzing open data from the Internet, problems arise in processing large amounts of data, the need for search and navigation in dynamic information flows. A huge number of multilingual dynamic information resources, the dominance of information noise makes it difficult to find the necessary information, operational analysis, and hence the use of open sources in information and analytical work.

Most of the above problems are topical issues of semantic processing of large dynamic information text arrays.

Currently, technological concepts such as Big Data, Complex Networks, Cloud Computing, Data/Text Mining are used to solve these problems.

Increasingly, an ontological approach is used to build subject areas models, in particular, in the cybersecurity field.

Over the past few years, various systems for storing and processing large amounts of data have appeared. Among them are Hadoop ecosystem projects, NoSQL non-relational databases (DBs), and search and analytic systems like Elasticsearch. Hadoop and any NoSQL database have their advantages and applications.

Elastic Stack is an ecosystem of components that are used to search and process data. The main components of Elastic Stack are Kibana, as well as Logstash, Beats, X-Pack and Elasticsearch. The core of the Elastic Stack is the Elasticsearch search engine, which provides capabilities for storing, searching and processing

data. The Kibana utility, also called a window in Elastic Stack, is a visualization tool and user interface for Elastic Stack. Logstash and Beats components allow you to transfer data to Elastic Stack. X-Pack provides powerful auxiliary functionality.

Elasticsearch is a fast, distributed real-time search engine for full-text data search and analysis. Typically, Elasticsearch is used as the base search engine and is the main component of Elastic Stack.

Elasticsearch is built on Apache Lucene technology, and therefore differs from traditional relational database or NoSQL solutions. The main features of using Elasticsearch are listed below:

- unstructured data that are processed;
- search capability;
- data analysis capability;
- support for custom libraries and REST API;
- easy management and scaling;
- high speed.

The Elasticsearch system can be used separately, without any other Elastic Stack components.

Kibana – visualization tool for Elastic Stack, which provides a visual representation of data in Elasticsearch. Kibana offers many options for visualizing information, such as a bar graph, map, line graphs, etc. Kibana allows you to create visualizations and examine data in an interactive view in real time, as well as generate high-quality reports.

Elastic Stack is a flexible platform with an expanded set of tools that allows developers to create their own programs thanks to the great support of programming languages and REST API.

Today, Elastic Stack components is a promising main building system for big data from social networks, in particular, on cybersecurity issues.

Sphinx (from SQL Phrase Index) is a full-text search engine for big data - Sphinx. It is used as a search engine in the currently built existing system, which is described in this paper. Sphinx is distributed under the terms of the GNU GPL or, for version 3.0 + without source code. A distinctive feature of Sphinx is its high indexing and search speed, as well as integration with existing DBMSs (MySQL, PostgreSQL) and APIs for common web programming languages (officially supported by PHP, Python, Java; there are implemented for Perl, Ruby, .NET and C + +). The official site of the system is http://sphinxsearch.com/. Sphinx system has the following features:

- high indexing speed (up to 10-15 MB / s for each processor core);
- high speed full-text search (up to 150-250 queries per second for each processor core with 1,000,000 documents);
- large scalability;
- distributed search support;

- support for single-byte encodings and UTF-8;
- support for morphological search - there are built-in modules for several languages;
- support for existing databases (PostgreSQL and MySQL), as well as ODBC-compatible databases (MS SQL, Oracle, etc).

Neo4j, a graph open-source Java database management system with transaction support (ACID), is promising for the processing of conceptual networks of semantic search within the framework of analyzing big data from social media. To date, this system is considered the most common graph DBMS.

Neo4j saves the data in a proprietary format specially adapted for the presentation of graph information. This approach, in comparison with the relational DBMS, allows additional optimization, while the graph does not need to be placed entirely in the server's RAM, which allows processing large network structures. Queries in Neo4j can be done directly through the Java API or in the Gremlin languages created in the open-source project TinkerPop and Cypher, which is the query and data manipulation language for graph storage.

## System Functionality

Relevant approach to solving the problem of creating such a corporate system under consideration is the simultaneous use of methods and tools of information retrieval, data analysis and aggregation of information flows. Created system of the social media monitoring and analysis, automatic processing of full texts from social media for a certain period related to the topic of "cybersecurity."

Information is collected in the search mode on social media (websites, social networks, instant messengers, blogs, etc.). The queries (the key phrase for searching the relevant social network, if possible, otherwise an account) are read by the program from special configuration tables. Next, the search and output records matching queries come. After that, unique records are written to the server database.

Analysis of existing approaches to the aggregation of thematic news has led to the need and possibility of creating a set of tools for social networks content monitoring on selected issues, in particular, cybersecurity. [20]

The system that is described includes means of personalization, providing online access to databases, including from mobile devices, for which the possibilities of RSS formats are widely used. The choice of "ready-made" software components is justified, self-developed tools are described (scanners of social networks, means of forming dynamic RSS feeds), and the results of their integration into a single complex are presented.

For example, we dwell in detail on the methodology for collecting information from the selected channels of the Telegram messenger. There are three known ways of accessing automated data collection to this resource. The first is direct access to the messenger at an address like https://tlgrm.ru/channels/@kpilive. In this case, kpilive is the name of the channel (specifically in this

case – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" channel – KPI live). The second, in another format, at the address of the redirect type: https://tlg.repair/s/kpilive. The third is access to channel information in Atom format through an external aggregator RSSHub: https://rsshub.app/telegram/channel/kpilive.

To create a text corpus based on the content of the channels of the Telegram messenger, the first step is to create a list of channels that may interest the researcher. To do this, you can refer to the numerous directories of these channels located on the network. Choose, for example, located at https://ru.telegram-store.com/catalog/product-category/channels/ (more than 50 thousand channels). Entering in the search mode the word we are interested in, for example "Ukraine", we get two pages of links to the address: https://ru.telegram-store.com/?s=Ukraine. We form the list of channels in a format:

@nowasteukraine
@onlineukraine
@sos_ua
@ua_rozvynuta
@ukraine_novosti
…

We scan the resources of the selected channels using the freely available program for obtaining information from network resources. It is known that there are several software agents (spider programs) that scan content from network resources via HTTP / HTTPS. We select the wget program, which is part of many systems. We use the channel list given in clause 1, form the address lists for the three access paths to the resource:

a:
https://tlgrm.ru/channels/@nowasteukraine
https://tlgrm.ru/channels/@onlineukraine
https://tlgrm.ru/channels/@sos_ua
…
b:
https://tlg.repair/s/nowasteukraine
https://tlg.repair/s/onlineukraine
https://tlg.repair/s/sos_ua
…
c:
https://rsshub.app/telegram/channel/nowasteukraine
https://rsshub.app/telegram/channel/onlineukraine
https://rsshub.app/telegram/channel/sos_ua
…

After starting the spider program with the command:

wget –i addr_list –O file,

we get 3 files corresponding to the given approaches. The parameters of the given program specify the file name with the list of addresses (–i) and the name of the source file (–O). Approach c) gives the fullest completeness; moreover, its Atom format is standard, but considering that it corresponds to an external service with respect to Telegram, approach b) is accepted for implementation, the second in completeness.

In the third and final stage, the collected complete data is converted into the text corpus format required for the study (up to XML 1.0 format), which can then be loaded into the Sphinx information retrieval system, a fragment of which is such:

```
<?xml version="1.0" encoding="utf-8"?>
<sphinx:docset>
<sphinx:schema>
<sphinx:field name="subject"/>
<sphinx:field name="content"/>
<sphinx:field name="source"/>
<sphinx:field name="datatime"/>
<sphinx:attr name="url"/>
</sphinx:schema>
 <sphinx:document id="1_tg">
<subject> Festivals, concerts and literary readings </subject>
<content> Festivals, concerts and literary readings - these events are not paused at all due to quarantine...</content>
<source>Telegram: novoe_vremya</source>
<datatime>20200425 14:20</datatime>
<url>https://tlgrm.ru/channels/@novoe_vremya/11263</url>
</sphinx:document>
```

The above technique can be used to extract data, analyze texts published by users of the Telegram messenger. A similar technique with some adaptations was used to form text arrays from other social media.

A model of a system for monitoring information from external systems - social media - was created by the authors in the Institute of Special Communications and Information Protection of National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute.”

Monitoring is carried out continuously in time from such systems as Twitter, Youtube, Reddit, Telegram etc. Monitoring results are available in various modes – search and analytical. The user accesses the system through a web browser or RSS aggregator.

The system of analysis of big data from social media provides the implementation of such functions:

1) Formation of the information fund by collecting according to certain criteria and accounts of the information given in national codings from information resources of the Internet: web-sites; blogs: Twitter, Livejournal; social net-

works: Facebook, Instagram, Reddit, Medium; video hosting: YouTube, RuTube; scientific communities: Academia.edu, ArXiv.org; messengers: Telegram.

2) Adjustment by the system administrator of modules for automatic scanning and primary processing of information from websites and social networks. If necessary, create accounts through which the system will be able to access certain social networks.

3) Maintaining retrospective full-text databases from information collected from the Internet, creating, rotating databases and ensuring the formation of internal vocabulary data sets in different languages (message indexing) using a universal coding system (UTF-8).

4) Identification of duplicates and similar information messages (including in different languages), grouping of duplicates and similar information messages in the search engine results.

5) Implementing a full-text search using queries in different languages.

6) Initial analysis of text messages stored in the system's databases: automatic detection of named entities (faces, company names, brands, geographical names, etc.), determination of tonality, identification of support words by statistical algorithms in information materials in different languages.

7) Formation of analytical reports, including information portraits and plots, which are based on the use of key words in different languages, thematic rubricating of documents.

8) Integration with geographic information system.

9) Data analysis and visualization; visualization of statistical data: according to certain sources, the number of downloaded messages for a period of time; graphs (histograms) of the distribution of the number of information messages, indicating the distribution of quantitative indicators by sources, types of sources, date.

10) The use of wavelet analysis to study thematic information flows. The technology of using wavelets makes it possible to identify single and irregular "bursts," sharp changes in the values of quantitative indicators at different time periods, in particular, the volume of thematic publications on social networks. In this case, the moments of the occurrence of cycles can be determined, as well as the moments when chaotic oscillations occur in periods of regular dynamics. Well-known wavelets such as the Mexican hat and Morlet wavelet accurately reflect the dynamics of information operations.[21]

11) Prediction of the development of events based on the analysis of the dynamics of publications in social media. For this, the fractal theory and nonlinear analysis methods are used.[22] The time series generated by thematic information flows also have fractal properties and can be considered as stochastic fractals. Wave methods can be used for forecasting.[23]

12) Providing access to many users to the system, delimiting access to system resources.

The basis of the hardware platform of systems for analyzing big data from social media is still the servers:

- information proxy server (a rented virtual server providing anonymous collection of information located on an external data center. With the development of the system, there may be several such servers. This server, on the one hand, is designed to provide reliable services to users of corporate networks, and, on the other hand, it can provide data exchange with similar external proxies);
- data collection server (server for collecting data from Internet resources. It can extract data according to scenarios defined by the administrator directly from Internet resources, or through information proxies);
- analytics server (the server carries out analytical processing of information and information retrieval. Using the server, databases of historical information are supported. Analytical processing of information includes: extraction of concepts; geoinformation support; definition of tonality of messages; formation of information; analysis of the dynamics of messages; forecasting; analysis of an array of information sources, etc.);
- front-end server (a web server from which end users can access through web browsers, RSS aggregators, or through application APIs to system resources).

## The interface of the system of analysis of Big Data from social media

The system of analyzing large amounts of data from social media "Cyber aggregator"[24,25] provides the user with a web interface from which functions for searching and analyzing information in social media are available to him (Fig. 2).

The system user is provided with documents upon request both in the retrospective database (Search) and in the current information (Current), as well as for data analysis (Analysis). The centerpiece of the interface is the digest of the most relevant messages. In a separate block (Queries), saved user queries are reflected. Statistical information regarding filling in the system database from individual social media is available in a special section (Sources statistics).

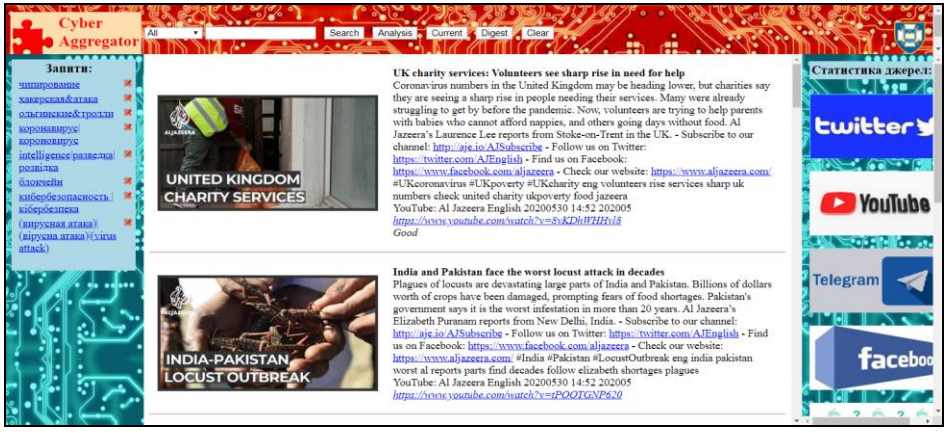


**Figure 2: User interface of Cyber Aggregator system.**

The user can enter and save queries in a special query language that supports searching by words and phrases, using logical operators. Requests can be individual words, for example, “hacking attacks” or “block chain.” Search can be performed within a separate social network or in the entire database. As a result of a search by query “cybersecurity & coronavirus” (Fig. 3), the user is presented with a list of relevant message headers with hyperlinks to the full texts of these messages in the system, as well as to these messages in social media.

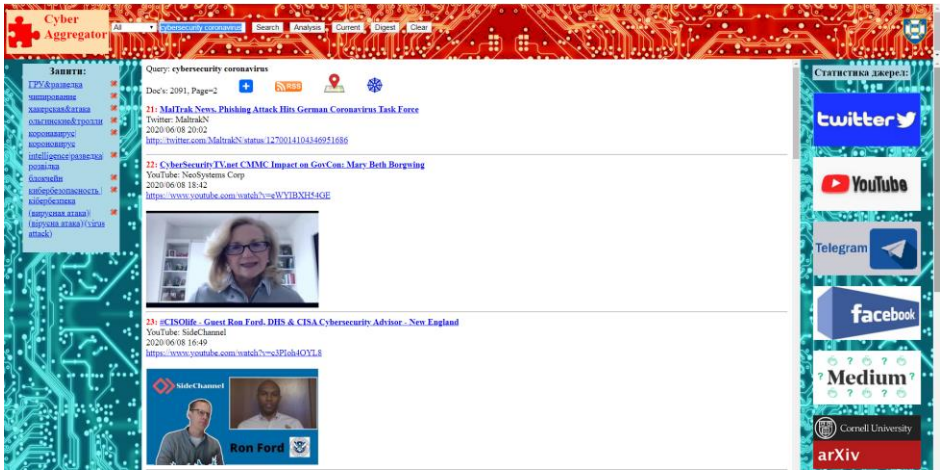


**Figure 3: Fragment of the user interface in search mode (search results for query “cybersecurity & coronavirus”).**

If the query produces documents that meet information needs, then it can be saved for future use (Add Query). It is possible to further output the found messages in RSS format (with the subsequent loading of these results into the so-called RSS aggregators on an ongoing basis), as well as displaying the search results with details on a geographical map, which is scaled both in automatic mode and through settings (Fig. 4).

In Analytical mode (Analysis), the user is provided with a number of tools, the first of which is a graph (Graph) corresponding to the time series of the number of relevant message requests per day (Fig. 5).

The user is also given the opportunity to view the main plots (Digest) on the topic “blockchain” (Fig. 6), clusters, grouped according to predefined key words.

The system provides modes for forming networks of concepts that correspond to individual messages (persons, brands), information sources (Fig. 7). These modes allow you to rate the concept, to explore the relationships between them.

The "Analytics" mode provides the possibility of forecasting (Forecast) by the method suggested by D. Sornette[26,27] which is based on the analysis of the regularity of market prices in commodity and stock markets before the collapse, crises. "Crises" in the information flows from social media is a sharp increase in

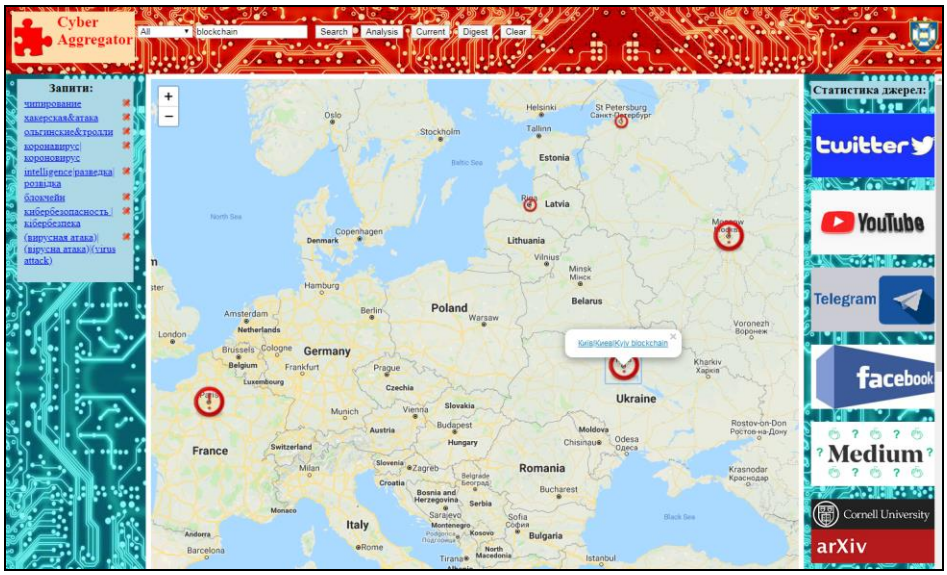


**Figure 4: Fragment of the geographic information system interface.**

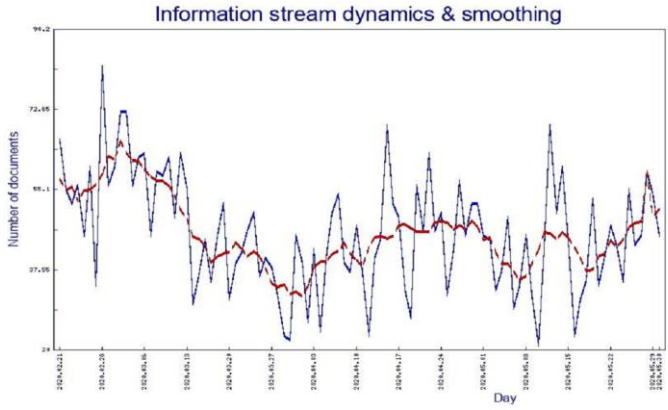


**Figure 5: Dynamics of thematic messages on query «Blockchain».**

the number of publications on a specific topic. Such an increase may signal an information attack, for example. Information flows are very similar in behavior to the processes that Sornette studied. Therefore, its model is selected as the main model for the forecast. It is noted in the Sornette works that before the collapse, the price is characterized by a power-law growth, complicated by log-periodic fluctuations that converge to a critical point, where the probability of collapse reaches its maximum value. The corresponding power model, which takes into account linear log-periodic oscillations, has the following form:

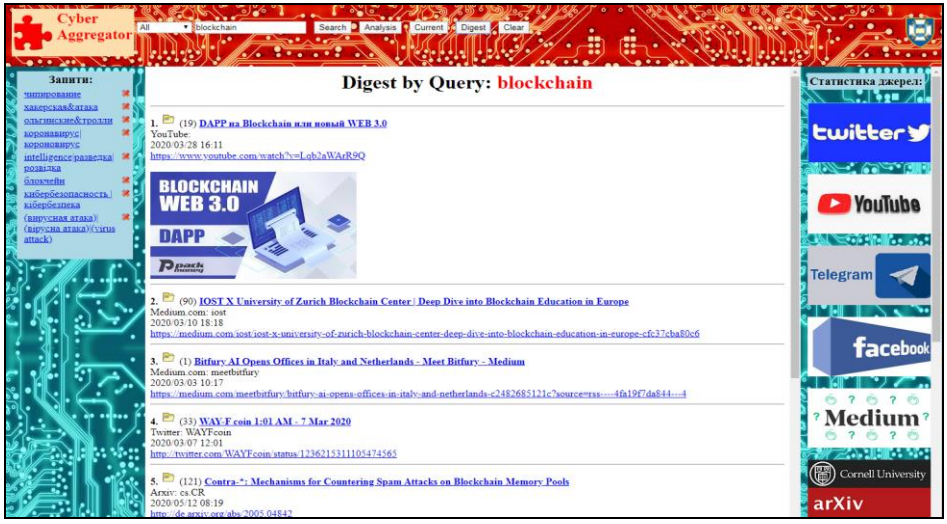


**Figure 6: A fragment of the thematic digest on query «Blockchain».**

$$F(t) = A + B(t_c - t)^m \left[1 + C\cos\left(\omega \log\left(\frac{t_c - t}{T}\right) + \varphi\right)\right].$$

In this model $t_c$ – the critical time (crisis time). The model coefficients $A,\ B,\ \omega,\ \varphi$ are determined using a selection procedure. Using the Sornette model (Forecast key, Fig. 8), you can obtain forecast values for the number of relevant publications based on the monitoring data.

## Conclusions

The paper describes the basic principles of building and using a monitoring and analysis system of social media on cybersecurity, which are based on the concepts of Big Data, Data/Text Mining, Information Extraction, Complex Networks.

The information technologies of creating a system of content monitoring of social networks on certain issues, selection of relevant information from social networks, implementation of search engine for their refinement by users, saving queries as RSS feeds, maintaining personal databases in client applications are suggested and substantiated.

The practical significance of the obtained results is to create a working model of the content monitoring and analysis of social media system, ready for use as a component of decision support systems for information and cyber security.

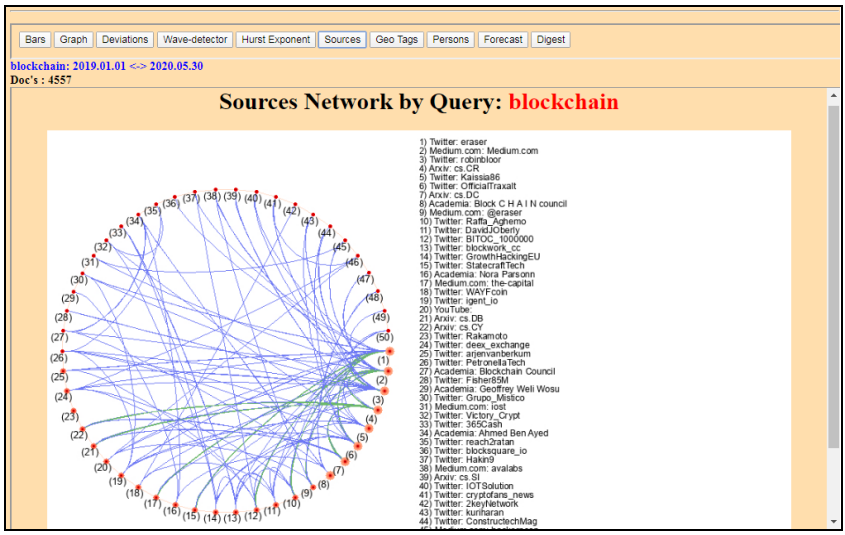


**Figure 7: The network of interconnected information sources on which messages are published on query «Blockchain».**

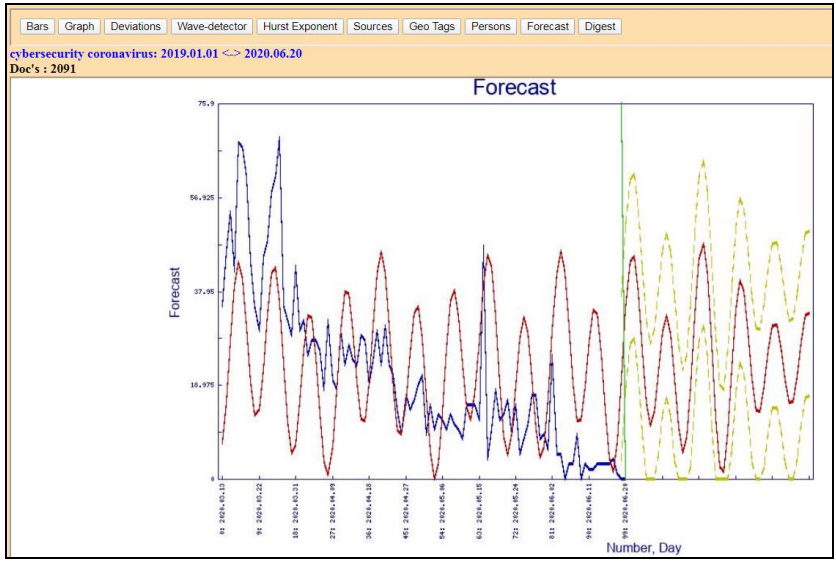


**Figure 8: Forecast line according to the Sornette algorithm for the time series on query «cybersecurity coronavirus».**

The work is introduced into the educational process of the Institute of Special Communications and Information Protection of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute."

By the level of readiness, the system can be estimated as TRL (Technology readiness levels) 4 or 5 (Technology validated in lab or Technology validated in relevant information environment).

The study of the system and the principles of its construction are included in the educational process as a demonstration of a stack of modern technologies for working with big data. In the future, it is planned to introduce the system into the state authorities of Ukraine.

The main area of further research is the automatic generation of domain models, semantic maps, and scenarios of activities in the cybersecurity field.

## Acknowledgements

This research was supported by CyRADARS project (SPS G5286 "Cyber Rapid Analysis for Defense Awareness of Real-time Situation") in the frame of the NATO Science for Peace and Security program.

## References

1 Dmytro Lande, Igor Subach, and Yu. Boyarinova, *Fundamentals of the Theory and Practice of Data Mining in the Field of Cyber Security: A Textbook* (Kyiv: Institute of Special Communications and Information Protection of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," 2018), http://dwl.kiev.ua/art/oiad/.

2 Danah Boyd and Kate Crawford, "Critical questions for Big Data," *Journal Information, Communication & Society* 15, no. 5 (2012): 662-679, https://doi.org/10.1080/1369118X.2012.678878.

3 Robert Layton and Paul Watters, *Automating open source intelligence: algorithms for OSINT* (Rockland, MA: Syngress Media, 2016), www.bookdepository.com/Automating-Open-Source-Intelligence-Robert-Layton/9780128029169.

4 Babak Akhgar, P. Saskia Bayerl, and Fraser Sampson, *Open Source Intelligence Investigation: From Strategy to Implementation* (Springer International Publishing AG, 2016), https://doi.org/10.1007/978-3-319-47671-1.

5 Uffe Kock Wiil, ed., *Counterterrorism and Open Source Intelligence* (Wien: Springer-Verlag, 2011), https://doi.org/10.1007/978-3-7091-0388-3.

6 Edward J. Appel, *Cybervetting: Internet Searches for Vetting, Investigations, and Open-Source Intelligence* (Taylor & Francis Group, 2014), https://doi.org/10.1201/b17651.

7 John Foreman, *Data Smart: Using Data Science to Transform Information into Insight* (Wiley, 2013), http://dx.doi.org/10.4258/hir.2014.20.3.243.

8 Nathan Marz, James Warren, *Big Data: Principles and Best Practices of Scalable Realtime Data Systems* (Manning, 2012), https://dl.acm.org/doi/book/10.5555/2717065.

[9] Davy Cielen, Arno Meysman, and Mohamed Ali, *Introducing Data Science. Big Data, Machine Learning, and More, Using Python Tools* (Manning Publications Co., 2016), https://dl.acm.org/doi/book/10.5555/3051941.

[10] Krish Krishnan, *Data Warehousing in the Age of Big Data* (Elsevier, 2013), https://doi.org/10.1016/C2012-0-02737-8.

[11] David Easley and Jon Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, 2010), https://doi.org/10.1017/CBO9780511761942.

[12] Giancarlo Ragozini and Maria Prosperina Vitale, *Challenges in Social Network Research: Methods and Applications: Lecture Notes in Social Network* (Springer, 2020), https://doi.org/10.1007/978-3-030-31463-7.

[13] Mehmet Kaya, Jalal Kawash, Suheil Khoury, and Min-Yuh Day, *Social Network Based Big Data Analysis and Applications* (Springer International Publishing, 2018), https://doi.org/10.1007/978-3-319-78196-9.

[14] Mehmet Kaya, Özcan Erdoğan, and Jon Rokne, *From Social Data Mining and Analysis to Prediction and Community Detection* (Springer International Publishing, 2017), https://doi.org/10.1007/978-3-319-51367-6.

[15] Katharina Zweig, *Network Analysis Literacy: A Practical Approach to the Analysis of Networks* (Wien: Springer-Verlag, 2016), https://doi.org/10.1007/978-3-7091-0741-6.

[16] Matthew Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram* (O'Reilly Media, 2019), http://dx.doi.org/10.1080/15536548.2015.1046287.

[17] Matthew Russell, *21 Recipes for Mining Twitter* (O'Reilly Media, 2011), http://shop.oreilly.com/product/0636920018261.do.

[18] Army Techniques Publication No. 2-22.9 (FMI 2-22.9), *Headquarters Department of the Army,* ATP 2-22.9 (Washington, DC, 10 July 2012), https://fas.org/irp/doddir/army/atp2-22-9.pdf.

[19] Dmytro Lande and Ellina Shnurko-Tabakova, "OSINT as a part of cyber defense system," *Theoretical and Applied Cybersecurity* 1, no. 1 (2019): 103-108, https://doi.org/10.20535/tacs.2664-29132019.1.169091.

[20] Dmytro Lande, "Information Streams Analysis in the Global Computer Networks," *Visnyk NAS of Ukraine* 3 (2017): 46-54, https://doi.org/10.15407/visn2017.03.045.

[21] Aleksandr Dodonov, Dmitry Lande, Vitaliy Tsyganok, Oleh Andriichuk, Sergii Kadenko, and Anastasia Graivoronskaya, *Information Operations Recognition: From Nonlinear Analysis to Decision-Making* (Lambert Academic Publishing, 2019), https://www.morebooks.shop/store/gb/book/information-operations-recognition/isbn/978-620-0-27697-1.

[22] P.A. Kisel'ov and Dmitry Lande, "Development of Software for Analysis and Forecasting of Information Operations," *Proceedings of the scientific-practical conference of cadets (students), graduate students, doctoral students and young scientists "Topical issues of special information and telecommunications systems,"* Kyiv: Institute of Special Communications and Information Protection of National

Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute,” 2019, pp. 180-181, http://dwl.kiev.ua/art/dz19-3/.

[23] Oleksandr Dodonov, Dmytro Lande, Nesterenko, O., and Boris Berezin, “Approach to forecasting the effectiveness of public administration using OSINT technologies: Information technology and security,” *Proceedings of the XIX International Scientific and Practical Conference ITS-2019, (*Kyiv: Engineering, 2019): 230-233, http://dwl.kiev.ua/art/itb2019-5/.

[24] Dmytro Lande, Igor Subach, and Artyom Sobolyev, Computer program of social networks content monitoring on cybersecurity "CyberAggregator" (Cyber Aggregator), Ukraine, Certificate of registration of copyright to the work N 91831 from 31 July 2019 (2019), http://dwl.kiev.ua/art/AS/as91831/.

[25] Dmytro Lande, N. Kalyan, and O. Matiishin, “Social Media Aggregation System on Cybersecurity,” *Proceedings of the XVII All-Ukrainian scientific-practical conference of students, graduate students and young scientists “Theoretical and applied problems of physics, mathematics and computer science,”* 25 - 26 April 2019, Kyiv, Ukraine, pp. 10-11, http://dwl.kiev.ua/art/ipt191.

[26] Didier Sornette, *Why Stock Markets Crash: Critical Events in Complex Financial Systems* (Princeton University Press, 2004), https://doi.org/10.23943/princeton/9780691175959.001.0001.

[27] Didier Sornette and Susanne von der Becke, “Financial Market and Systemic Risks,” in *Market Risk and Financial Markets Modeling* (Berlin, Heidelberg: Springer-Verlag, 2012), https://doi.org/10.1007/978-3-642-27931-7.

## About the Authors

Professor Dmytro **Lande**, Dr. of Sci., is Head of Department in the Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, and senior researcher in the Institute of Special Communications and Information Protection of National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute,” Kyiv, Ukraine.

Assoc. professor Igor **Subach**, Dr. of Sci., is head of department in the Institute of Special Communications and Information Protection of National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute,” Kyiv, Ukraine.

Prof. Alexander **Puchkov**, PhD., is Director of the Institute of Special Communications and Information Protection of National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute,” Kyiv, Ukraine.