DOI: 10.17323/2587-814X.2021.3.35.47

An approach to identifying threats of extracting confidential data from automated control systems based on internet technologies*

Vladimir N. Kuzmin
E-mail: vka@mil.ru

Artem B. Menisov
E-mail: vka@mil.ru

Space Military Academy named after A.F. Mozhaysky
Address: 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia

Abstract

Together with ubiquitous, global digitalization, cybercrime is growing and developing rapidly. The state considers the creation of an environment conducive to information security to be a strategic goal for the development of the information society in Russia. However, the question of how the "state of protection of the individual, society and the state from internal and external information threats" should be achieved in accordance with the "Information Security" and the "Digital Economy of Russia 2024" programs remains open. The aim of this study is to increase the efficiency whereby automated control systems identify confidential data from html-pages to reduce the risk of using this data in the preparatory and initial stages of attacks on the infrastructure of government organizations. The article describes an approach that has been developed to identify confidential data based on the combination of several neural network technologies: a universal sentence encoder and a neural network recurrent architecture of bidirectional long-term short-term memory. The results of an assessment in comparison with modern means of natural language text processing (SpaCy) showed the merits and prospects of the practical application of the methodological approach.

Key words: information security; countering information security threats; confidential data; personal data; machine learning; deep learning; identifying the entities of natural language texts.

Citation: Kuzmin V.N., Menisov A.B. (2021) An approach to identifying threats of extracting confidential data from automated control systems based on internet technologies. *Business Informatics*, vol. 15, no 3, pp. 35–47. DOI: 10.17323/2587-814X.2021.3.35.47

* The article is published with the support of the HSE University Partnership Programme

Introduction

Present-day society is characterized by the increasing role of the information sphere, which is a set of information, information infrastructure, specialists that collect, form, disseminate and use information, as well as a system for regulating the resulting social relations. The information sphere, being a system-forming factor in the life of society, actively influences the state of political, economic, defense, and other components of national security [1]. In turn, the development of information technology has led to a transformation in the understanding of personal space and privacy. Processes that previously took place in the physical (real) world have spilled over into the online environment: e-commerce, search services, social networks, the proliferation of tablets and smartphones, which enable people to be constantly online. As a result, the volumes of confidential information that a person discloses and uploads to the global network, as well as personal data of citizens collected and systematized by various institutions and departments, has increased many times over [2].

Thus, one of the significant threats to national security and the interests of the Russian Federation in the information sphere [1] is the possibility of using confidential information [3–5] at the preparatory and initial stages of organizing and carrying out attacks on the infrastructure of state and commercial organizations [6].

At the same time, models, methods, technologies, and devices for identifying and removing confidential data from open sources are not perfect enough due to the low efficiency of using methods of syntactic data comparison, as well as the lack of global coverage of the open segment of information [7, 8]. In this article, the problematic situation is formulated as the need to ensure effective identification of confidential and personal data from html pages based on the development and implementation of models, methods, and devices for identifying and removing confidential data from open information sources.

1. Background

Improving the efficiency of detecting threats of confidential data leakage can be achieved through the implementation of a number of activities, which include [9]:

- improving the means of monitoring open sources;
- linguistic processing of unstructured data [10];
- identification of references to persons and related confidential data in texts.

In the 1990s, identification of references in texts was first presented as a task of information extraction (named entity recognition, NER) [11], and since then this approach has attracted much attention from researchers. The main purpose of NER is to tag or classify objects (words) in a specific text based on predefined labels or tags (for example, person, location, organization, etc.) [12]. Most of the research relates to the processing of English text, but regardless of the language, three main approaches can be distinguished: based on linguistic rules, based on machine learning algorithms, and hybrid approaches.

The linguistic approach uses rule-based models that are hand-written by linguists. With this approach, a set of rules or patterns is formed in order to distinguish between mentions at a certain place in the text. Several automated systems have been developed [13, 14], in which specialized dictionaries are used, including the names of countries, large cities, organizations, names of people, etc. The main disadvantage of this approach is the need to use a large number of grammar rules in addition to modified usage (style, jargon). Moreover, these systems are practically unsuitable for working with other languages.

Approaches based on machine learning algorithms use a large amount of annotated training data to obtain high-level language knowledge and are classified into two types: supervised and unsupervised. Unsupervised NER models do not require any training data [15] and have the ability to annotate the data themselves. These learning models are not popular in practical use due to the rather low accuracy of identifying entities in the text. On the other hand, supervised learning models require a large amount of high-quality annotated data. Some machine learning algorithms used for NER (Markov models [16–19], maximum entropy method [20–22], decision trees [23–25], support vector machine [26, 27], etc.) have shown reliable results for identifying entities in multilingual texts.

Deep learning is a subset of machine learning, which is a combination of multiple layers of representation-based data processing at multiple levels of abstraction. Deep learning methods have recently been widely used due to their outstanding performance compared to other methods for solving various problems, including natural language processing.

There are two main architectures that are widely used to extract textual representation at the character or word level [28]:

- models based on convolutional neural networks (CNN) [29];
- models based on recurrent neural networks (RNN) [30].

More modern neural network architectures based on various combinations of convolutional and recurrent networks show the best results in solving many problems [31–34]. These models show significant versatility because they can be applied to multiple languages with unified network architecture.

Analysis of research in this area has shown the advantages of deep learning algorithms for solving the problem of identifying confidential data from html pages.

2. Methods

This study proposes a methodological approach that includes the use of two different neural network technologies:

- neural network recurrent architecture of bidirectional long short-term memory (BLSTM), which has shown an improvement in the quality of identifying entities for solving other problems [35, 36];
- a universal sentence encoder that allows scaling the approach for texts in different languages [37].

Combining a bidirectional long short-term memory (BLSTM) neural network and sentence encoder involves the following steps (*Figure 1*):

1. Cleaning the html page from markup and other service information. Highlighting the text part located on the html page.
2. Text preprocessing.
3. Presentation of text features the transformation of primary features into a vector representation.
4. Presentation of BLSTM-based sentences: obtaining a high-level representation of features (semantics) from the features of stage 3.
5. Construction of a common vector of features: combining the features of the sentence of the lexical and semantic levels from stages 3 and 4 in order to form the final vector of features.
6. Classification: detection of confidential data.

Let's take a closer look at each of the stages of identifying confidential entities on html pages.

Stage 1. Cleaning up the html page. At this stage, it is necessary to solve the problems of extracting the actual text from the content of the html page and cleaning the text from special characters. The process begins by building the document object model (DOM) of the html page by parsing tags. The DOM provides a representation of the structure of an html page. The text nodes of the DOM are retrieved for further use: the text is filtered and cleaned,

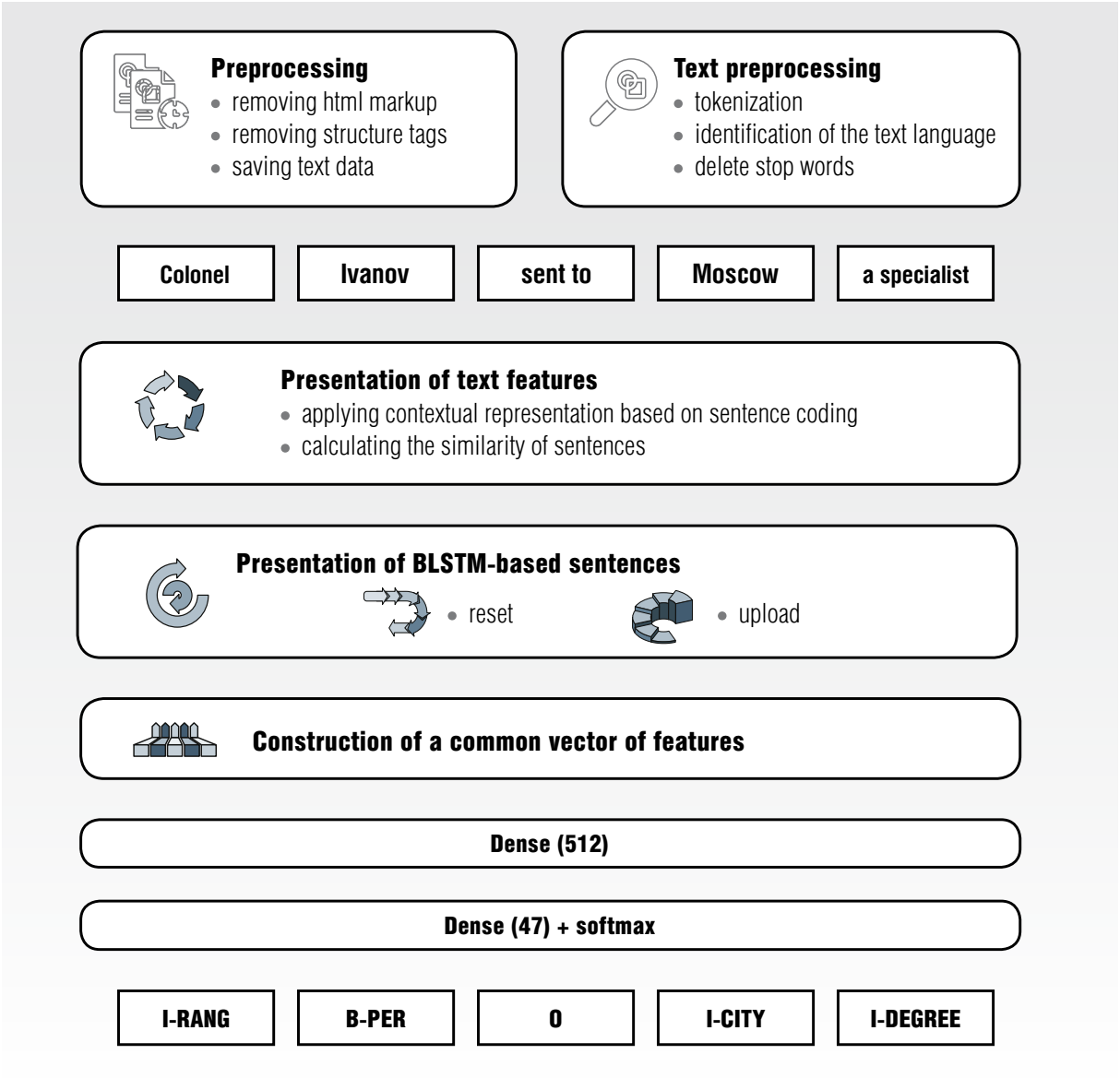

Fig. 1. Scheme of the methodological approach developed for extracting confidential data

excluding scripts and html structure symbols such as navigation lists, style tags, tables, and miscellaneous frames. This step also removes punctuation marks and special characters.

Stage 2. Text preprocessing. At this point, the resulting text segments are tokenized to obtain individual tokens (words). Then stop words are removed – common words in different languages that have the least semantic meaning. First, the language of the text is identified, and then the list of stop words for that language is determined. This does not contradict the goal of providing a language-independent method, as the language finder and stop word list for different languages are available in open-source software libraries.

In complex natural languages (such as Russian), the same word can take different forms (cases), and all word forms that differ in prepositions and endings can be included in the frequency analysis dictionary. Because of this, the size of the vocabulary can greatly increase and,

accordingly, the size of the training data set, which can cause a decrease in system performance and a deterioration in the generalizing abilities of the classifier (overfitting). To solve this problem, additional text preprocessing measures are used: lemmatization and stemming.

To solve the problem of identifying confidential data, these additional measures were specifically omitted for the following reasons:

- stemming (and especially lemmatization) requires dictionaries of the languages in which the text is written;
- these measures search normal-form dictionaries for all words in the text, which can significantly degrade performance.
- confidential data often contains information in the form of numbers and special characters (for example, numbers of certificates and other documents).

Stage 3. Presentation of text features. The presentation of the features of the text is one of the main stages of its linguistic processing. To solve the problem of identifying confidential data, it is proposed to use a pre-trained algorithm for representing sentences (or sequences of words) called the Universal Sentence Encoder (USE) [37]. USE is a sentence coding algorithm released by Google in 2018 that aims to provide a presentation at the sentence level, not at the word or character level.

The USE algorithm was implemented using two approaches:

- application of contextual representation based on the coding of sentences;
- evaluating the similarity of proposals.

The USE algorithm was first used for English [37], and then it was implemented for multilingual texts [38].

Stage 4. Presentation of BLSTM-based sentences. This step uses a bidirectional recurrent block as an extended version of a recurrent neural network. The lightweight block structure using two gates (reset and upload) improves the efficiency of solving the gradient disappearance problem compared to the long short-term memory (LSTM) architecture [39], which consists of three gates (input, output, and forgetting).

Stage 5. Construction of a common vector of features. At this stage, the features of the sentence of the lexical and semantic levels are combined in order to preserve all possible distinctive features of the text. The union is done in the concatenation layer.

Stage 6. Classification. In the classification step, an initially tightly coupled layer of 512 neurons is used to process the distinctive characteristics of the text from the feature concatenation layer. Next, a layer with a softmax activation element is used to calculate the probability that the given word belongs to one of 57 classes. The dataset is annotated using the IOB labeling format [40] and includes 28 confidential information classes (B- and I-tags) and one additional class (O-tag). Therefore, the model allows classification into 57 classes.

3. Results

This section describes the experiment and the parameters that were used to train and validate the proposed approach.

3.1. Dataset

The dataset used for research was provided by the Kaspersky Innovation Hub[1] as part of the 2020 Digital SuperHero (Fintech & Security) Russian hackathon[2]. The data-

[1] https://www.kaspersky.ru/ihub/
[2] https://www.dshkazan.ru/finsec/

set consists of 28 classes of confidential data (*Table 1*). This table also presents the distribution of entries by classes of sensitive data that were used to train and validate the models proposed in the approach. In total, the dataset consists of almost 900 html pages, of which 833 html pages were used as the training dataset and 70 html pages were used for model validation.

3.2. Metrics

Several metrics were used to assess the effectiveness of the proposed approach. First, precision was measured, which is a known metric for evaluating any machine or deep learning model and characterizes the proportion of correctly classified objects out of their total. The precision is calculated as follows:

$$Pr = \frac{TP}{TP + FP}, \quad (1)$$

where TP – number of true positives;

FP – number of type I-errors (false positives).

The recall metric was then evaluated, which is the number of correctly identified features from the total number of features in the dataset. This metric is calculated as follows:

$$R = \frac{TP}{TP + FN}, \quad (2)$$

where TP – number of true positives;

FN– number of type II-errors (false negative).

Finally, the $F1$-measure is calculated based on the precision and recall values:

$$F1 = 2\frac{Pr \cdot R}{Pr + R}. \quad (3)$$

where Pr – value of precision;

R – value of recall.

Table 1.

Used dataset classes

No	Name	Feature	Frequency
1	PASSPORT	Passport data	1
2	DRIVER_LIC	Driver's license data	1
3	CAR_START	Starting to use the car	8
4	CAR	State registration number of the car	28
5	EDU_START	Start date of education	35
6	DEATHDATE	Date of death	83
7	AGE	Age	120
8	EDU_END	Date of graduation	145
9	BIRTHDATE	Date of Birth	146
10	ZIPCODE	Postcode	218
11	FACULTY	Faculty	219
12	HOBBY	Hobby	315
13	START	Start time of education or work	338
14	EMAIL	E–mail	354
15	END	End of work	355
16	STREET	Street	364
17	EDU	Education	389
18	TEL	Telephone	409
19	DEGREE	Rank, position	459
20	STATE	Subdivision	912
21	NICKNAME	Social network profile (login)	1220
22	COUNTRY	Country	1254
23	INDUSTRY	Direction of activity	1597
24	CITY	City	1824
25	GENDER	Sex	1960
26	FUNC	Job responsibilities	2517
27	ORG	Name of the organization	3667
28	PER	Full name (or part thereof)	7682

3.3. Quality of revealing confidential data

Currently, there are many software solutions on the market for open-source information analysis and artificial intelligence technologies (for example, Amazon, ABBYY, IBM Watson, MS Azure, and Palantir). However, for comparison with the approach we developed, the spaCy[3] framework was adopted, which is explained by the following economic and technical reasons:

- support for over 60 languages;
- the ready-to-industrial application system for teaching linguistic models;
- the presence of extensible components for recognizing named entities, tagging parts of speech, parsing dependencies, segmenting sentences, classifying text, lemmatization, morphological analysis, linking entities, etc;
- support for custom models on PyTorch, TensorFlow, and other neural network frameworks;
- the presence of built-in visualizers for syntax and NER;
- relatively simple integration of the model into automated systems.

A comparison of the results presented at the hackathon[4] and obtained using the spaCy framework showed that the approach developed shows an increase in the quality of detecting confidential data by 21% for all classes of confidential data (*Table 2*). The approach ensured the achievement of an average $F1 = 0.55$, an average precision $Pr = 0.57$, and an average recall $R = 0.67$.

4. Discussion

To demonstrate the practical importance of the proposed methodological approach for identifying confidential data from the internet, the results obtained were compared with other compositions of neural network technologies (*Table 3*).

As can be seen from *Table 3*, the BLSTM model used in the approach we developed is superior to the LSTM model in all the specified classes. It can be noted that the smallest improvement was achieved in the EDU_END class, while the highest was in the PER class. This can be explained by the distribution of data across classes and their sizes. Returning to *Table 1*, you can see that the PER class has the largest volume compared to other classes (7682 named entities), while the EDU_END class has the smallest volume (145 named entities).

However, for such classes of confidential data as ORG and COUNTRY, lower quality is observed, which is due to the following reasons:

- he objective advantage of identifying these classes by the spaCy framework, since the possibility of identifying the ORG class is present in all spaCy linguistic models for different languages, trained on a large amount of data; it is also possible to identify the COUNTRY class using additional spaCy classes (GPE and LOC);
- the intersection of entities of different classes in the training set: ORG, STATE, EDU, and FACULTY.
- The reasons for improving the quality of detection of confidential data include:
- the capabilities of the USE algorithm in terms of processing the semantic presentations of sentences and phrases. Using only the USE algorithm as a classifier for the same dataset, as shown in *Table 3* (USE), achieved an average $F1 = 0.42$. Although the overall results of the USE algorithm for classification are lower than those achieved with the approach we developoed, it can be seen that

[3] https://www.spaCy.com
[4] https://www.dshkazan.ru/finsec/

Table 2.

Comparison of model results to identify confidential data

No	Class ID	Precision		Recall		*F1*	
		Proposed approach	spaCy	Proposed approach	spaCy	Proposed approach	spaCy
1	PER	0.9082	0.35	0.87076	0.98	0.88908	0.51579
2	ORG	0.1932	0.44	0.93319	0.75	0.32008	0.55462
3	FUNC	0.1872	0.15	0.51531	0.79	0.27466	0.25213
4	CITY	0.584	0.23	0.58657	0.83	0.58528	0.36019
5	NICKNAME	0.7695	0.16	0.94219	0.88	0.84714	0.27077
6	COUNTRY	0.8086	0.45	0.51655	0.92	0.63038	0.60438
7	GENDER	0.9379	0.18	0.29631	0.06	0.45034	0.09
8	INDUSTRY	0.5283	0.3	0.2601	0.08	0.34858	0.12632
9	STATE	0.1797	0.17	0.72339	0.83	0.28783	0.2822
10	EMAIL	0.8901	0.34	0.41612	0.19	0.56712	0.24377
11	STREET	0.7692	0.25	0.69175	0.8	0.72844	0.38095
12	TEL	0.8774	0.5	0.5833	0.76	0.70075	0.60318
13	EDU	0.171	0.16	0.80821	0.8	0.28224	0.26667
14	ZIPCODE	0.6034	0.41	0.77374	0.85	0.67807	0.55318
15	DEGREE	0.6585	0.35	0.93206	0.1	0.77175	0.15556
16	START	0.3947	0.27	0.43456	0.84	0.41368	0.40865
17	EDU_END	0.1366	0.45	0.55193	0.3	0.21896	0.36
18	END	0.0298	0.38	0.86437	0.92	0.05754	0.53785
19	AGE	0.9287	0.18	0.83912	0.76	0.88164	0.29106
20	HOBBY	0.1011	0.3	0.5287	0.21	0.16975	0.24706
21	BIRTHDATE	0.8076	0.17	0.92066	0.5	0.86043	0.25373
22	FACULTY	0.9578	0.34	0.97121	0.45	0.96448	0.38734
23	CAR	0.4014	0.25	0.99413	0.85	0.57185	0.38636
24	DEATHDATE	0.6176	0.5	0.73078	0.96	0.66942	0.65753
25	EDU_START	0.9323	0.16	0.78325	0.02	0.85131	0.03556
26	CAR_START	0.3057	0.15	0.146	0.2	0.19761	0.17143
27	PASSPORT	0.8492	0.23	0.72283	0.53	0.78096	0.32079
28	DRIVER_LIC	0.5959	0.16	0.44169	0.54	0.50733	0.24686
Mean value:		**0.57588**	**0.285**	**0.67067**	**0.59643**	**0.55381**	**0.34157**

Table 3.

Comparison of the results of the work of different compositions of the approach

No	Class ID	Precision				Recall				*F1*			
		Proposed approach	LSTM	w/o USE	USE	Proposed approach	LSTM	w/o USE	USE	Proposed approac	LSTM	w/o USE	USE
1	PER	0.9082	0.788899	0.7554	0.2513	0.87076	0.3161	0.9247	0.4166	0.88908	0.45136	0.83153	0.31349
2	ORG	0.1932	0.910661	0.7594	0.3073	0.93319	0.0655	0.6615	0.4548	0.32008	0.12222	0.70711	0.36673
3	FUNC	0.1872	0.564068	0.2792	0.8071	0.51531	0.2536	0.0691	0.5931	0.27466	0.34992	0.11074	0.68375
4	CITY	0.584	0.979289	0.1985	0.5707	0.58657	0.4723	0.3151	0.8761	0.58528	0.63722	0.24358	0.69113
5	NICKNAME	0.7695	0.958619	0.6594	0.9251	0.94219	0.9439	0.3993	0.0077	0.84714	0.9512	0.4974	0.01521
6	COUNTRY	0.8086	0.682749	0.9073	0.8715	0.51655	0.1534	0.7579	0.4965	0.63038	0.25058	0.82588	0.63264
7	GENDER	0.9379	0.761573	0.3181	0.0151	0.29631	0.7039	0.2468	0.4464	0.45034	0.73159	0.27794	0.02926
8	INDUSTRY	0.5283	0.155593	0.2821	0.1805	0.2601	0.2655	0.8068	0.0553	0.34858	0.19621	0.41799	0.08467
9	STATE	0.1797	0.109872	0.0281	0.2822	0.72339	0.2258	0.8331	0.873	0.28783	0.14783	0.05438	0.42655
10	EMAIL	0.8901	0.712207	0.4924	0.5611	0.41612	0.3041	0.3899	0.8125	0.56712	0.42625	0.4352	0.66379
11	STREET	0.7692	0.154987	0.7198	0.9594	0.69175	0.9779	0.9269	0.7743	0.72844	0.26757	0.8103	0.85698
12	TEL	0.8774	0.036389	0.7911	0.4411	0.5833	0.1756	0.5942	0.4934	0.70075	0.06028	0.67864	0.46582
13	EDU	0.171	0.26002	0.0648	0.5232	0.80821	0.1515	0.7556	0.2679	0.28224	0.19149	0.11942	0.35438
14	ZIPCODE	0.6034	0.203505	0.9409	0.2833	0.77374	0.1067	0.5478	0.7517	0.67807	0.13998	0.69245	0.41156

No	Class ID	Precision				Recall				*F1*			
		Proposed approach	LSTM	w/o USE	USE	Proposed approach	LSTM	w/o USE	USE	Proposed approac	LSTM	w/o USE	USE
15	DEGREE	0.6585	0.299845	0.3623	0.9814	0.93206	0.8859	0.242	0.0327	0.77175	0.44805	0.2902	0.06328
16	START	0.3947	0.436453	0.8491	0.716	0.43456	0.7234	0.3869	0.9749	0.41368	0.54442	0.53161	0.82562
17	EDU_END	0.1366	0.806458	0.8001	0.8182	0.55193	0.3229	0.2334	0.4665	0.21896	0.46118	0.3614	0.59419
18	END	0.0298	0.858825	0.7621	0.4577	0.86437	0.9984	0.3264	0.242	0.05754	0.92338	0.45707	0.31662
19	AGE	0.9287	0.239638	0.4984	0.4532	0.83912	0.3478	0.8059	0.8474	0.88164	0.28376	0.61586	0.59055
20	HOBBY	0.1011	0.116531	0.9497	0.9496	0.5287	0.5741	0.0152	0.4049	0.16975	0.19373	0.02987	0.56771
21	BIRTHDATE	0.8076	0.133296	0.9161	0.9204	0.92066	0.6129	0.15	0.3458	0.86043	0.21897	0.25785	0.50276
22	FACULTY	0.9578	0.995736	0.0912	0.8146	0.97121	0.0517	0.6599	0.3144	0.96448	0.09824	0.16023	0.4537
23	CAR	0.4014	0.378413	0.2613	0.821	0.99413	0.3192	0.6812	0.0832	0.57185	0.34627	0.37774	0.15108
24	DEATHDATE	0.6176	0.634855	0.9917	0.3356	0.73078	0.3214	0.8013	0.3979	0.66942	0.42675	0.88636	0.3641
25	EDU_START	0.9323	0.5145	0.4761	0.5342	0.78325	0.1459	0.0167	0.0504	0.85131	0.22736	0.03235	0.09205
26	CAR_START	0.3057	0.4658	0.8526	0.8066	0.146	0.9981	0.5294	0.7014	0.19761	0.63518	0.6532	0.75031
27	PASSPORT	0.8492	0.689894	0.3281	0.5283	0.72283	0.6247	0.3017	0.9222	0.78096	0.6557	0.31436	0.67176
28	DRIVER_LIC	0.5959	0.390814	0.511194	0.012905	0.44169	0.233	0.0786	0.6786	0.50733	0.29193	0.13618	0.02533
Mean value:		**0.57588**	**0.5085532**	**0.56595**	**0.57602**	**0.67067**	**0.43841**	**0.48062**	**0.4922**	**0.55381**	**0.38138**	**0.42167**	**0.42732**

the use of the USE algorithm only for classification of some other classes (STREET) is higher than that of the proposed approach ($F1 = 0.81$ for USE, versus 0.73 for the proposed approach). Therefore, further research should do more work to improve the classification of individual datasets;

- the proposed sequence of stages and architecture of the neural network. In particular, by accumulating the USE and BLSTM algorithm, an improvement of the average $F1$-measure by 15% was achieved.

Conclusion

The huge volume of unstructured data on the internet disseminated daily creates a need for the development of effective methods for searching and extracting information. Extracting confidential data is a complex classification task for natural language texts, which becomes even more difficult when applied to html pages due to their special properties and complex structure. This article presents a new deep learning approach for identifying confidential data that has proven to be effective over others.

The main goal of developing a new approach is to provide more detailed results for practical applications in the field of natural language processing and information security. The approach we developed uses a neural network technology of bidirectional long short-term memory in combination with a multilingual universal sentence encoder. ■

References

1. *The Information Security Doctrine of the Russian Federation*. Approved by Decree of the President of the Russian Federation No 646 of 5 December 2016. Available at: http://pravo.gov.ru/proxy/ips/?docbody=&prevDoc=102161033&backlink=1&&nd=102417017 (accessed 01 February 2021) (in Russian).
2. Shaydullina V.K. (2019) Big data and personal data protection: the main theoretical and practical issues of legal regulation. *Society: Politics, Economics, Law*, no 1, pp. 51–55 (in Russian). DOI: 10.24158/pep.2019.1.8.
3. Federal Law of the Russian Federation No 152-FZ of 27 July 2006 "*About personal data*." Available at: http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261 (accessed 01 February 2021) (in Russian).
4. Decree of the President of the Russian Federation No 188 of 06 March 1997 "*On approval of the list of confidential information*." Available at: http://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102046005 (accessed 01 February 2021) (in Russian).
5. Federal Law of the Russian Federation No 79-FZ of 27 July 2004 "*About the state civil service of the Russian Federation*." Available at: http://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102088054 (accessed 01 February 2021) (in Russian).
6. Information message "*On the development of the methodological document of the FSTEC of Russia* "*Methodology for determining information security threats in information systems*" No 240/22/1534 of 9 April 2020. Available at: https://fstec.ru/normotvorcheskaya/informatsionnye-i-analiticheskie-materialy/2071-informatsionnoe-soobshchenie-fstek-rossii-ot-9-aprelya-2020-g-n-240-22-1534 (accessed 01 February 2021) (in Russian).
7. Los V.P., Nikulchev E.V., Pushkin P.Yu., Rusakov A.M. (2020) Information-analytical system of monitoring implementation of legislation requirements by operators of personal information. *Information Security Problems. Computer Systems*, no 3, pp. 16–23 (in Russian).
8. Kozin I.S. (2018) Providing personal data protection in an information system based on user behavior analytics. *Information and Control Systems,* no 3, pp. 69–78 (in Russian). DOI: 10.15217/issn1684-8853.2018.3.69.
9. Ivichev V.A., Ignatova T.V. (2013) Technologies of identification and depersonalization of Data. *ECO*, vol. 43, no 2, pp. 168–179 (in Russian).
10. Platonov A.A., Psaryov A.A (2019) Approach to identifying program defects based on mutually complementing semantics. *Proceedings of the Mozhaisky Military Space Academy*, no 666, pp. 148–157 (in Russian).

11. Grishman R. (1994) Whither written language evaluation? Proceedings of the *Workshop on Human Language Technology. Plainsboro, New Jersey, 8–11 March 1994*, pp. 120–125.
12. Minkov E., Wang R.C., Cohen W. (2005) Extracting personal names from email: Applying named entity recognition to informal text. Proceedings of the *Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, 6–8 October 2005*, pp. 443–450.
13. Grishman R. (1995) The NYU System for MUC-6 or Where's the Syntax? Proceedings of the *Sixth Message Understanding Conference (MUC-6). Columbia, Maryland, 6–8 November 1995*, pp. 167–175.
14. Wakao T., Gaizauskas R., Wilks Y. (1996) Evaluation of an algorithm for the recognition and classification of proper names. Proceedings of the *16th International Conference on Computational Linguistics (COLING 1996). Copenhagen, 5–9 August 1996*, vol. 1, pp. 418–423.
15. Alfred R., Leong L.C., On C.K., Anthony P. (2014) Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, vol. 4, no 3, pp. 300–306. DOI: 10.7763/IJMLC.2014.V4.428.
16. Salleh M.S., Asmai S.A., Basiron H., Ahmad S. (2018) Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis. *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no 2–7, pp. 121–126.
17. Zhou G.D., Su J. (2002) Named entity recognition using an HMM-based chunk tagger. Proceedings of the *40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA, 6–12 July 2002*, pp. 473–480.
18. Morwal S., Jahan N., Chopra D. (2012) Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing*, vol. 1, no 4, pp. 15–23. DOI: 10.5121/ijnlc.2012.1402.
19. Morwal S., Jahan N. (2013) Named entity recognition using hidden Markov model (HMM): an experimental result on Hindi, Urdu and Marathi languages. *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no 4, pp. 671–675.
20. Borthwick A. (1999) *A maximum entropy approach to named entity recognition* (PhD Thesis). New York: New York University.
21. Chieu H.L., Ng H.T. (2002) Named entity recognition: a maximum entropy approach using global information. Proceedings of the *19th International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, 24 August – 1 September 2002*, vol. 1, pp. 1–7. DOI: 10.3115/1072228.1072253.
22. Chieu H.L., Ng H.T. (2003) Named entity recognition with a maximum entropy approach. Proceedings of the *Seventh Conference on Natural language learning at HLT-NAACL 2003. Edmonton Canada, 31 May 2003*, pp. 160–163.
23. Speck R., Ngomo A.C.N. (2014) Ensemble learning for named entity recognition. Proceedings of the *13th International Semantic Web Conference (ISWC 2014). Riva del Garda, Italy, 19–23 October 2014*, pp. 519–534.
24. Paliouras G., Karkaletsis V., Petasis G., Spyropoulos C.D. (2000) Learning decision trees for named-entity recognition and classification. Proceedings of the *ECAI Workshop on Machine Learning for Information Extraction (ECAI 2000). Berlin, 21 August 2000*, pp. 1–6.
25. Szarvas G., Farkas R., Kocsor A. (2006) A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. Proceedings of the *9th International Conference on Discovery Science (DS 2006). Barcelona, Spain, 7–10 October 2006*, pp. 267–278. DOI: 10.1007/11893318_27.
26. Mansouri A., Affendey L. S., Mamat A. (2008) Named entity recognition approaches. *International Journal of Computer Science and Network Security*, vol. 8, no 2, pp. 339–344.
27. Ekbal A., Bandyopadhyay S. (2010) Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical, Computer, and Systems Engineering*, vol. 4, no 2, pp. 155–170.
28. Li J., Sun A., Han J., Li C. (2020) A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* (Early access). DOI: 10.1109/TKDE.2020.2981314.
29. Ma X., Hovy E. (2016) End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv:1603.01354v5.*

30. Li P.-H., Dong R.-P., Wang Y.-S., Chou J.-C., Ma W.-Y. (2017) Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. Proceedings of the *2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 7–11 September 2017*, pp. 2664–2669.

31. Fu J., Liu P., Zhang Q. (2020) Rethinking generalization of neural models: A named entity recognition case study. Proceedings of the *34th AAAI Conference on Artificial Intelligence. New York, USA, 7–12 February 2020*, vol. 34, no 05, pp. 7732–7739. DOI: 10.1609/aaai.v34i05.6276.

32. Al-Smadi M., Al-Zboon S., Jararweh Y., Juola P. (2020) Transfer learning for Arabic named entity recognition with deep neural networks. *IEEE Access*, vol. 8, pp. 37736–37745. DOI: 10.1109/ACCESS.2020.2973319.

33. Lin B.Y., Lee D.-H., Shen M., Moreno R., Huang X., Shiralkar P., Ren X. (2020) Triggerner: Learning with entity triggers as explanations for named entity recognition *arXiv:2004.07493v4.*

34. Cho M., Ha J., Park C., Park S. (2020) Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics*, vol. 103, article ID 103381. DOI: 10.1016/j.jbi.2020.103381.

35. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. (2016) Neural architectures for named entity recognition. *arXiv:1603.01360v3.*

36. Huang Z., Xu W., Yu K. (2015) Bidirectional LSTM-CRF models for sequence tagging. *arXiv:1508.01991v1.*

37. Cer D., Yang Y., Kong S.-Y., Hua N., Limtiaco N., John R.S., Constant N., Guajardo-Cespedes M., Yuan S., Tar C., Sung Y.-H., Strope B., Kurzweil R. (2018) Universal sentence encoder. *arXiv:1803.11175v2.*

38. Yang Y., Cer D., Ahmad A., Guo M., Law J., Constant N., Abrego G.H., Yuan S., Tar C., Sung Y.-H., Strope B., Kurzweil R. (2019) Multilingual universal sentence encoder for semantic retrieval. *arXiv:1907.04307v1.*

39. Greff K., Srivastava R.K., Koutník J., Steunebrink B.R., Schmidhuber J. (2016) LSTM: A search space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no 10, pp. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.

40. Song Y., Kim E., Lee G.G., Yi B.-K. (2004) POSBIOTM-NER in the shared task of BioNLP/NLP-BA2004. Proceedings of the *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). Geneva, Switzerland, 28–29* August *2004*, pp. 103–106.

About the authors

Vladimir N. Kuzmin

Dr. Sci. (Mil.), Professor;

Leading Researcher, Military Institute (Science and Researching), Space Military Academy named after A.F. Mozhaysky, 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;

E-mail: vka@mil.ru

ORCID: 0000-0002-6411-4336

Artem B. Menisov

Cand. Sci. (Tech.);

Doctoral Student, Space Military Academy named after A.F. Mozhaysky, 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;

E-mail: vka@mil.ru

ORCID: 0000-0002-9955-2694