Development of an Iot Ecosystem Model for the Task of Semantic Analysis of Internet Posts

Andrii Biloshchytskyi [1], Olha Kravchenko [2], Olena Sipko [2], Rostyslav Lisnevskyi [2], and Dmytro Stvoglaz [3]

[1] *Astana IT University, Turkistan street, Astana 020000, Kazakhstan*
[2] *Taras Shevchenko National University of Kyiv, 60 Volodymyrska Str., Kyiv, 01033, Ukraine*
[3] *Cherkasy State Technological University, 460 Shevchenko bul., Cherkasy, 18000, Ukraine*

Abstract

The work is devoted to the development of an IoT ecosystem model for the semantic analysis of Internet posts. The need to create an IoT ecosystem that will allow offline assessment of the emotional state of Internet posts is substantiated. The work provides a graphic representation of the IoT ecosystem and describes its components. The functional model of the IoT system of semantic analysis of Internet posts is described. The principle of learning a neural network, which is the basis of the operation of the IoT system, is presented. The algorithm of the interlocutor bot, which is a means of receiving Internet posts into the system database, is given. Verification of the work of the IoT system of semantic text analysis was carried out. The result of emotional coloring was obtained with a probability of 94%.

Keywords [1]

Semantic Text Analysis, IoT Ecosystem Architecture, Neural Network, IT Technologies

1. Introduction

The environment (nature, buildings, IT technologies, gadgets, etc.) has a significant impact on a person. In this work, we perceive a person not only as a creature, but as a person. When solving problems, we take into account her personal characteristics that affect her reactions.

The modern world in Ukraine and other countries is undergoing changes. Humanity has reached the threshold value of information perception. The general ecosystem of human existence is not possible without information technologies and the world of the Internet of Things. The work [1] considered the need to use information and digital platforms to create efficient and fast business processes in the ecosystem. The concept of the IoT ecosystem includes three main levels: the level of information collection, the level of data verification and transmission, and the level of data analytics. These are the three main pillars on which the architecture of the IoT ecosystem is built. Another component for the successful operation of the ecosystem is security. Security of data transmission, physical security of the operation of gadgets, application of the ecosystem for security purposes. Our research is devoted to solving the problem of semantic text analysis.The advantages of using the methods of semantic analysis of texts in natural language (Text Mining) for working with textual descriptions of typical attacks and their components contained in the above classification systems are noted. An automated method is proposed for assessing current (i.e., potentially most dangerous) vulnerabilities in industrial control system (ICS) software using semantic analysis methods of descriptions in order to determine a list of threats that are relevant to software vulnerabilities identified at a specific object using security scanners. An example of the proposed techniques application for assessing vulnerabilities of the application software of industrial oil production facility automation subsystem is considered, followed by the formation of a list of relevant threats [2].

Information Technology and Implementation (IT&I-2023), November 20 - November 21, 2023, Kyiv, Ukraine
EMAIL: bao1978@gmail.com (A. Biloshchytskyi); Olha.Kravchenko@knu.ua (O. Kravchenko); sipko.olena@knu.ua (O. Sipko); lisnevskyi.rostyslav@knu.ua (R. Lisnevskyi); grayeyed16@gmail.com (D. Stvoglaz)
ORCID: 0000-0001-9548-1959 (A. Biloshchytskyi); 0000-0002-9669-2579 (O. Kravchenko); 0000-0003-1385-119X (O. Sipko); 0000-0002-9006-6366 (R. Lisnevskyi); 0000-0002-6806-839X (D. Stvoglaz)
© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3624/Short_9.pdf

The task of analyzing the test is relevant at any time, both in peacetime and in military time. The analysis of the text in peacetime is most often aimed at the analysis of the posts of people in social networks in order to identify harm to themselves. In wartime, this task acquires characteristic features of searching for collaborators.

Today, there are several methods of text analysis. In manuscript [3] proposed two innovative unsupervised approaches for the combined modeling and interrelated accomplishment of the two tasks. Both approaches rely on respective Bayesian generative models of topics, contents and clusters in textual corpora. Such models treat topics and clusters as linked latent factors in document wording. In particular, under the generative model of the second approach, textual documents are characterized by topic distributions, that are allowed to vary around the topic distributions of their membership clusters. Within the devised models, algorithms are designed to implement Rao-Blackwellized Gibbs sampling together with parameter estimation. These are derived mathematically for carrying out topic modeling with document clustering in a simultaneous and interrelated manner. A comparative empirical evaluation demonstrates the effectiveness of the presented approaches, over different families of state-of-the-art competitors, in clustering real-world benchmark text collections and, also, uncovering their underlying semantics. Besides, a case study is developed as an insightful qualitative analysis of results on real-world text corpora.

The text often contains images in the form of text that need to be analyzed. It is not only the comments on the image that are important, but also the posts on the image themselves. [4] presents Parallel Deep Fusion Generative Adversarial Networks (PDF-GAN), which use two new components to mitigate inconsistent semantics and bridge the semantic gap between text and images.

Experiments evaluating the generalization of scientific articles and news reports show that the proposed framework achieves comparable or higher correlations with human judgments than popular evaluation models. This study also supports the role of a network of semantic connections in the representation and analysis of texts [5]. The study [6] evaluated the empirical impact of certain syntactic and semantic features on the accuracy of note-taking. An improved feature set that jointly uses syntactic and semantic features is proposed. No one will deny the fact that the use of neural networks to create text analysis software products is relevant. A neural network (artificial neural network) is a mathematical model that mimics the structure and functioning of biological neural networks in order to solve a variety of tasks, such as classification, regression, prediction, and generation [7].

The analysis of short texts and texts with a large set of words has significant differences. The authors of the article [8] propose a new Set-CNN approach. It is aimed at using text convolution neural network based on semantic expansion for short text classification.

The analysis of short texts of comments and reviews under the product in the online store allows to analyze the intentions of customers regarding the purchase of a series of products. Analysis of posts in social networks allows obtaining data on the psychological state and mood of the population. Using this analysis for peaceful purposes brings positive results for the development of society and technology.

Italian online news is analyzed using semantic network analysis and text analysis. For this, they use Semantic Brand Score. In the process of measuring the importance of energy community media, ideas for implementation and public awareness of energy communities are identified [9].

With the increasing complexity of the building process, it is difficult for project stakeholders to retrieve large and multi-disciplinary building information models (BIMs). A natural language interface (NLI) is beneficial for users to query BIM models using natural language. However, parsing natural language queries (NLQs) is challenging due to ambiguous name descriptions and intricate relationships between entities. To address these issues, this study proposes a graph neural network (GNN)-based semantic parsing method that automatically maps NLQs into executable queries. Firstly, ambiguous mentions are collectively linked to referent ontological entities via a GNN-based entity linking model. Secondly, the logical forms of NLQs are interpreted through a GNN-based relation extraction model, which predicts links between mentioned entities in a heterogeneous graph fusing ontology and NLQ texts. The experiment based on 786 queries shows its outstanding performance. Moreover, a real-world case verifies the practicability of the proposed method for BIM model retrieval [10].

In manuscript [11] proposed a new systematic methodology to address these issues and identify technology opportunities using a hierarchical semantic network and dual link prediction. The proposed methodology consists of three modules: 1) constructing the hierarchical semantic network based on

SAO structures extracted from patents; 2) identifying technology opportunities in this semantic network through probabilistic-based link prediction; and 3) evaluating these opportunities via similarity-based link prediction. The viability and usefulness of the proposed methodology is proved by empirical analysis of the exploitation technology in the coal seam gas (CSG) industry. The results show that the hierarchical semantic network, including semantic and co-word relationships, can improve prediction accuracy.

So, after considering various directions and methods of applying semantic text analysis, we came to the opinion that this topic is relevant. Since the text we are analyzing is written by a person, the question arose to investigate the components of the ecosystem and build its architecture. Develop an IoT ecosystem model for semantic analysis of Internet posts. To check in practice the operation of the system for semantic analysis of Internet posts.

2. Analysis of the constituent components of the IoT ecosystem architecture for the task of semantic analysis of Internet posts

The Internet of Things combines real things into virtual systems capable of solving completely different tasks. The key idea is to connect all connectable objects together, connect them to a network to collect data and make decisions based on it [12].

The "IoT Ecosystem": includes all the components that enable businesses, governments and users to connect their IoT devices, including control panels, panels tools, networks, gateways, analytics, data storage and security [12, 13]. The architecture of the IoT ecosystem depends on the problem to be solved. But in general, the main components of the IoT architecture can be listed.

The components of the IoT architecture are the data acquisition area, the data transmission network, and the data analysis unit (Figure 1).

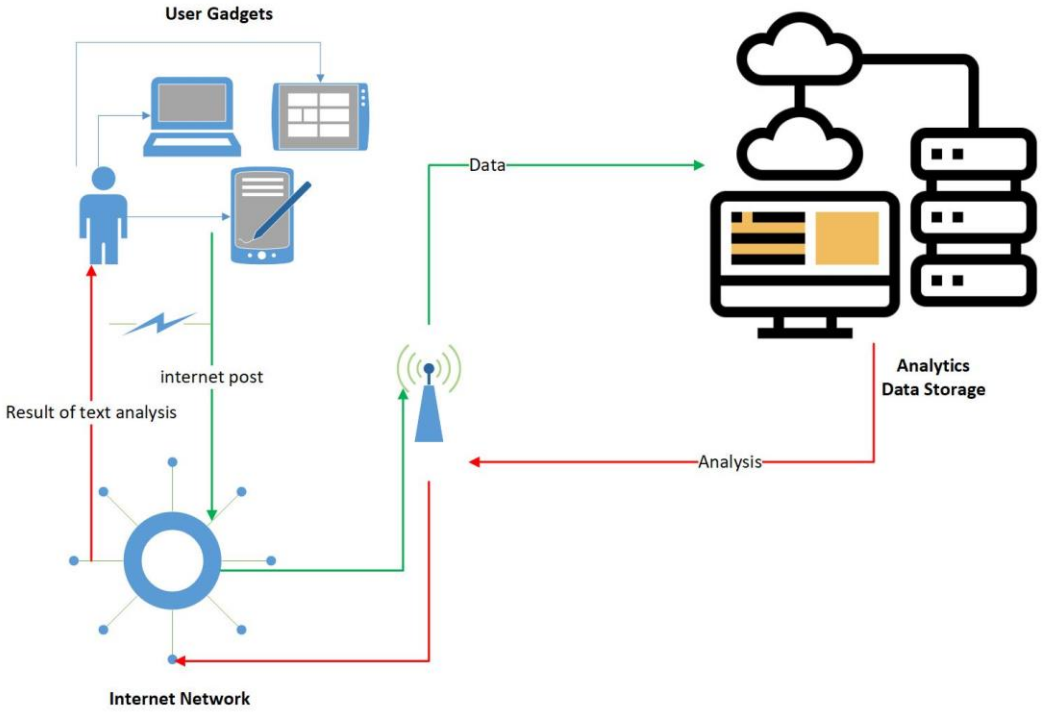

Figure 1: IoT ecosystem architecture for semantic analysis of Internet posts

The data collection area includes the user's gadgets and the user himself. The user (upper left corner of Figure 1) is an integral part of the ecosystem as the initiator of Internet posts [14]. The results of the analysis of Internet posts are sent to the user of the gadget from which the request was made, as a result of the analysis of the Internet post.

The data analysis unit (upper right corner of Figure 1) consists of a database that stores both input and output data, a semantic analysis system, and hardware.

The data transmission network consists of a wired and wireless part of data transmission. The user makes an Internet post using his own gadget. The post is sent to the application database with the help of the Internet. Then, if necessary, the text is analyzed by the semantic text analysis system and sent to the requesting user's gadget. The protection of information is ensured by coding of the data transmission network, provided by the provider and the support system of the cloud environments of social networks and the storage environment of the semantic data analysis system.

The hardware of the process of data transmission, reception and storage is not the purpose of this study. The prototype of the system for semantic analysis of Internet posts is placed on a stationary PC and searches for data with the help of a conversational bot [15]. The interlocutor bot is written in Ruby on Rails. The task of the interlocutor bot is to receive data from social networks. The bot was tested on posts, comments and posts by Facebook users.

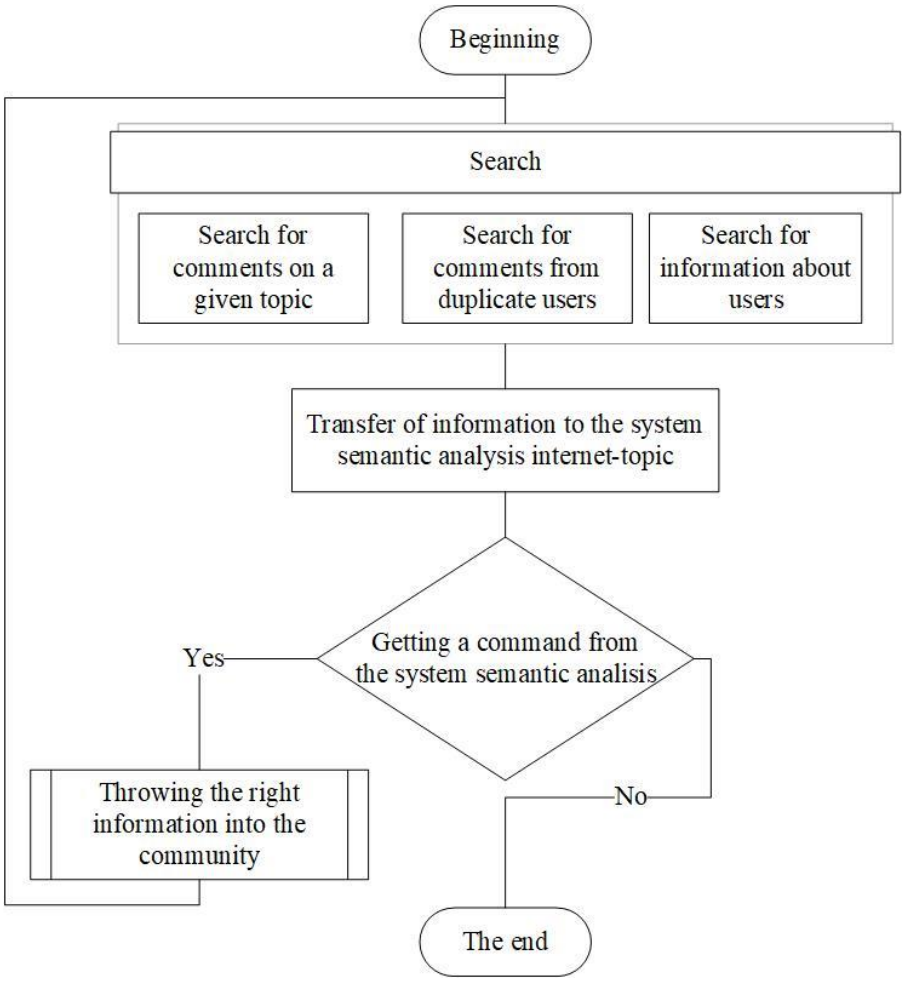

Figure 2: The algorithm of the bot companion for system semantic analysis internet-topic

Figure 2 describes the algorithm of the interlocutor bot, which provides data for the system of semantic analysis of Internet posts.

Algorithm of the interlocutor bot:

Step 1 Settings before starting work

Step 2 Search for posts and information about contributors from open sources

Step 3 Submission of the text of Internet posts to the system of semantic analysis of Internet posts

Step 4 Execution of Step 1 - Step 3 before executing the system command or changing the settings.

The system of semantic analysis of Internet posts is built on the basis of a neural network. It reflects the emotional coloring of an Internet post. The works [16,17] describe the tasks of text analysis by means of a neural network. Figure 3 shows their classification. In the process of classification, the text is compared with already existing ones. Characteristic words and constructions are checked. For the

emotional study of the text, it is necessary to take into account not only the constructions of the words in the sentence, but also the connections between the words.

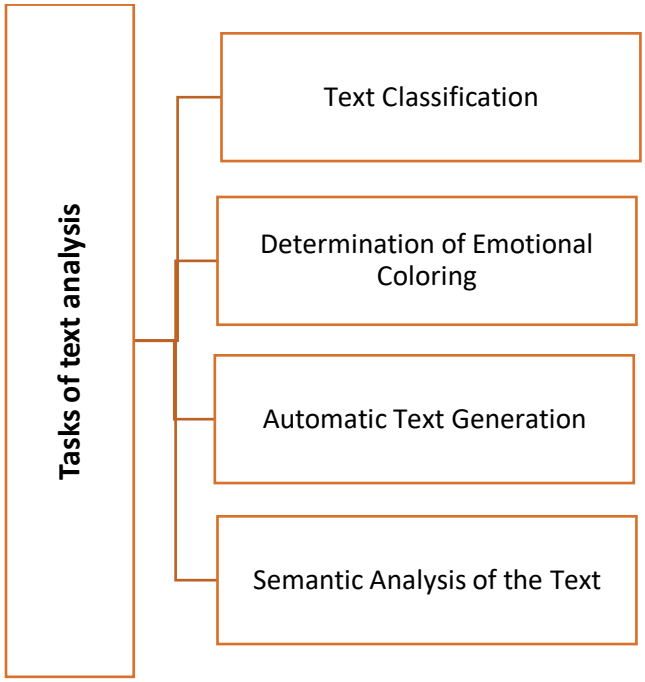

Figure 3: Tasks of text analysis

Semantic analysis requires analyzing grammatical errors, counting the number of words and relating them to groups. Integral speech recognition systems based on neural networks are investigated in the article [18]. The author conducted an experiment with the LSHTM neural network using an encoder-decoder model, which is based on the attention-based models. The result of the experiment showed a Character Error Rate (CER) equal to 8.01% and a Word Error Rate (WER) equal to 17.91%. This result proves the possibility of getting a good ASR model without the use of the Language Model (LM) [18].

The system for semantic analysis of Internet posts uses a neural network on a device with a graphics core. The system includes a self-learning unit for further processing of Internet posts with the help of a teacher.

3. IoT model of the system of semantic analysis of Internet posts

At the first stage of working with the Internet post, we turn the text into a combination of zeros and ones. The neural network works with the digital description of the text. Let's turn the text into dense vectors. Each word is a separate token. The neural network learns with the teacher. Therefore, we compile a test and training data set. The training data set contains answers for semantic analysis of Internet posts for emotional coloring. Figure 4 shows part of the neural network.

At the entrance, we matter x_t and return the value h_t. Feedback allows information to be transmitted step by step to each node in the network. The neural network used in the system is recurrent. Connects previously obtained results with future results. One of the characteristics of a recurrent neural network is to define an error function for each training step. We will use binary cross-entropy (1) when solving multitask classification.

$$H(p,q) = H(p) + D_{KL}(p||q) \tag{1}$$

when $H(p)$- entropy p;

$D_{KL}(p||q)$- Kullback-Leibler divergence from q to p.

The results of testing the neural network, which is the basis of the method of semantic analysis of Internet posts for the emotional coloring of the text, are shown in Figure 5.

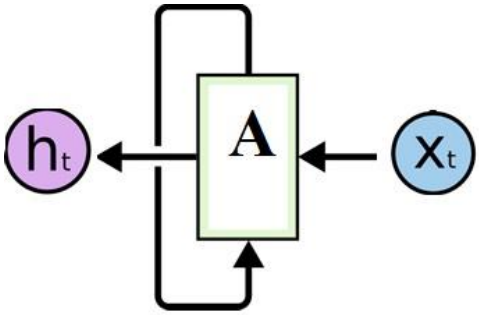

Figure 4: Scheme of a part of a neural network for system semantic analysis internet-topic

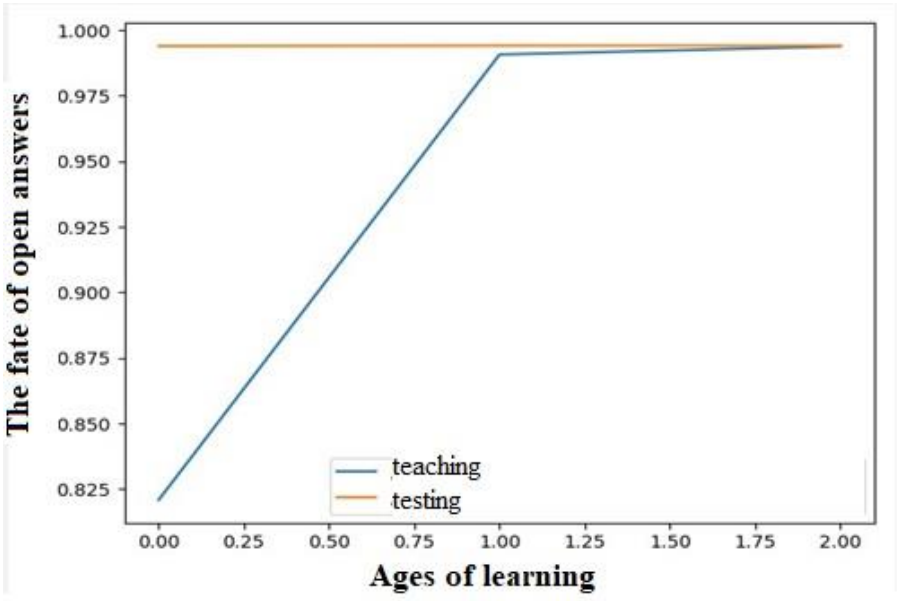

Figure 5: Results of neural network testing

At the training stage, the neural network learned to classify Internet posts with a probability of more than 97% in 1 epoch (Figure 5). At the test stage, the neural network reached the maximum classification value. Carrying out step-by-step testing, we have a score in the range of numbers from 0 to 1 for each class. Figure 7 shows the functional model of the IoT system for the semantic analysis of Internet posts. Common functions include: system update; information about the system; system retraining; text analysis; introduction of new educational data.

Analysis of the use of IoT system for the semantic analysis of Internet posts Let's download the training sample in .csv format. Table 1 shows an example of a training sample with the following characteristics: ID, text, and emotional coloring of the text.

```
[[7, 4882, 3050, 6, 245, 333]]
[[  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   7 4882 3050   6 245 333]]
Result your text: I wasnt vandalising you little shit
[[0.76732504 0.08257934 0.45622438 0.0281907  0.37333727 0.07307157 0.02634966]]
```

Figure 6: Results of step-by-step analysis of the text

The emotional coloring of the text presented in Table 1 is represented by 1-presence of emotional coloring, 0- absence of emotional coloring in the text. In the process of learning a neural network under the guidance of a teacher, the following parameters were obtained to obtain an optimal solution:

— the number of epochs — 2 and 5;

— number of iterations for early termination — 250;

— network error calculation — Binary crossentropy;

— learning speed coefficient — 0.05
— optimizer — Adam.

5 epochs are selected for system training. As early as the third epoch, we get stable values from the neural network (Figure 8).

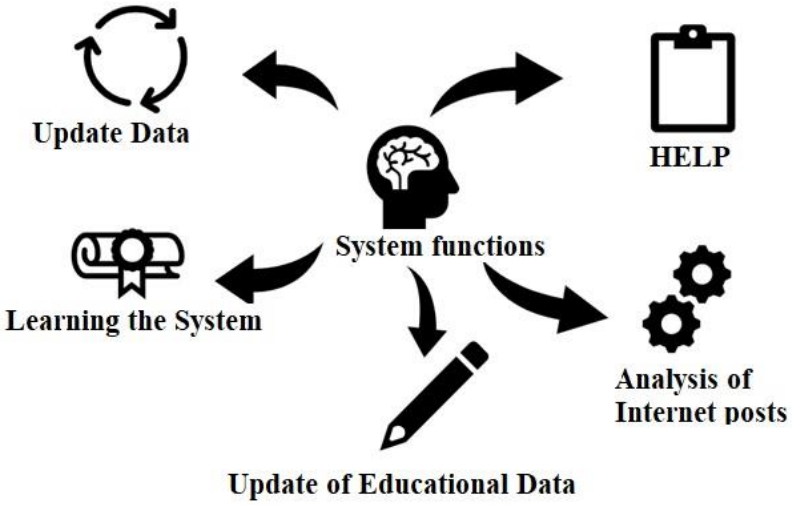

Figure 7: Functional model of the IoT system for the semantic analysis of Internet posts

Table 1

An example of a training sample IoT system for the semantic analysis of Internet posts

ID	Text	Emotional coloring of the text
00000001	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 22:01, March 12, 2023 (UTC)	0,0,1,0,0,0
00000002	OCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1,1,1,0,1,0
00000003	Be a man and lets discuss it-maybe over the phone?	0,0,0,0,0,0
00000004	Bye! Don't look, come or think of comming back! Tosser	1,0,0,0,0,0

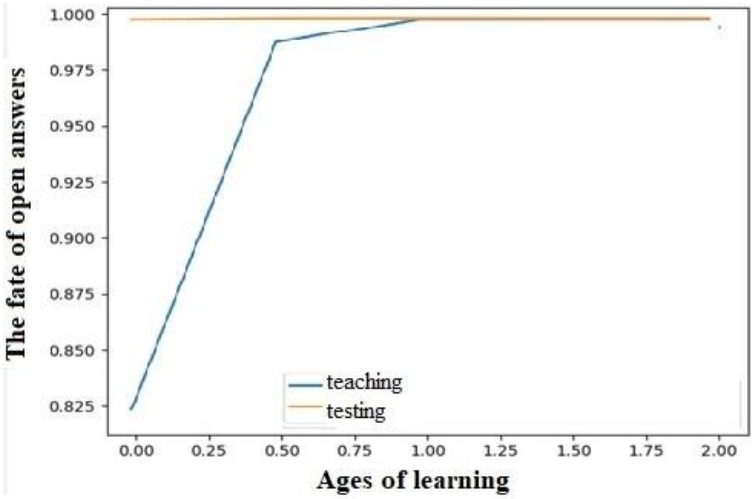

Figure 8: Results of neural network testing

After training, we load the text into the system for analysis. The results of the system are shown in Figure 9.

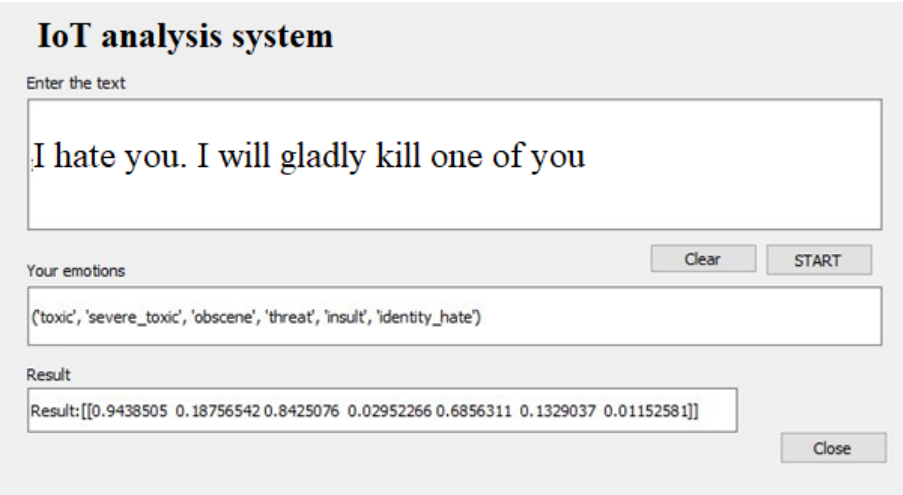

Figure 9: System performance results IoT system for the semantic analysis of Internet posts

So, the text is 94% toxic, 18% moderately toxic, and 84% contains obscene language.

4. Summary

The result of this study is a description of the functional model of the IoT ecosystem for the semantic analysis of Internet posts. The work provides the structure of the IoT ecosystem and describes its components. Data verification was carried out on the example of a previously trained neural network. The network determined with 94% accuracy the emotional color of a short text, which was obtained with the help of a conversational bot from the Facebook social network.

5. References

[1] Kyzymenko K. S., Kravchenko M. O. Development of an innovative ecosystem model for defense enterprises of Ukraine No. 19 (2021): Economic Bulletin of NTUU "Kyiv Polytechnical Institute". DOI: https://doi.org/10.20535/2307-5651.19.2021 (In Ukranian)

[2] V. Vasilyev, A. Vulfin, V. Gvozdev, A. Nikonov, Semantic Text Analysis Technology Application in Assessing Current Threats and Software Vulnerabilities, IFAC-PapersOnLine,Volume 54, Issue 13,2021,Pages 599-604,ISSN 2405-8963, https://doi.org/10.1016/j.ifacol.2021.10.515

[3] Gianni Costa, Riccardo Ortale, Hierarchical Bayesian text modeling for the unsupervised joint analysis of latent topics and semantic clusters, International Journal of Approximate Reasoning, Volume 147, 2022, Pages 23-39, ISSN 0888-613X, https://doi.org/10.1016/j.ijar.2022.05.002.

[4] Zhaorui Tan, Xi Yang, Zihan Ye, Qiufeng Wang, Yuyao Yan, Anh Nguyen, Kaizhu Huang, Semantic Similarity Distance: Towards better text-image consistency metric in text-to-image generation, Pattern Recognition, Volume 144, 2023, 109883,ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2023.109883.

[5] Mengyun Cao, Hai Zhuge, Automatic evaluation of summary on fidelity, conciseness and coherence for text summarization based on semantic link network, Expert Systems with Applications, Volume 206, 2022, 117777, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.117777.

[6] Begum Mutlu, Ebru A. Sezer, Enhanced sentence representation for extractive text summarization: Investigating the syntactic and semantic features and their contribution to sentence scoring, Expert Systems with Applications, Volume 227, 2023, 120302, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.120302.

[7] Neural network. https://termin.in.ua/neyromerezha/
[8] Yajian Zhou, Jiale Li, Junhui Chi, Wei Tang, Yuqi Zheng, Set-CNN: A text convolutional neural network based on semantic extension for short text classification, Knowledge-Based Systems, Volume 257, 2022, 109948, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2022.109948.
[9] C. Piselli, A. Fronzetti Colladon, L. Segneri, A.L. Pisello, Evaluating and improving social awareness of energy communities through semantic network analysis of online news, Renewable and Sustainable Energy Reviews, Volume 167, 2022, 112792, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2022.112792.
[10] Mengtian Yin, Llewellyn Tang, Chris Webster, Jinyang Li, Haotian Li, Zhuoquan Wu, Reynold C.K. Cheng, Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks, Automation in Construction, Volume 152, 2023, 104902, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2023.104902.
[11] Zhenfeng Liu, Jian Feng, Lorna Uden, Technology opportunity analysis using hierarchical semantic networks and dual link prediction, Technovation, Volume 128, 2023, 102872, ISSN 0166-4972, https://doi.org/10.1016/j.technovation.2023.102872.
[12] Internet of things technologies. Training manual [Electronic resource]: training. manual for study specialty 126 "Information systems and technologies", specialization "Information maintenance of robotic systems" / B. Yu. Zhurakovskyi, I.O. Zeniv; KPI named after Igor Sikorsky – Electronic text data (1 file: 12.5 MB). – Kyiv: KPI named after Igor Sikorskyi, 2021. – 271 p.
[13] Tripathy B. Internet of Things (IoT): TeChnologies, AppliCations, Challenges and Solutions (англ.) / B. Tripathy, J. Anuradha. – Florida: CRC Press, 2017. – 334 c.
[14] Steven T. Smith, Edward K. Kao, Danelle C. Shah, Olga Simek, Donald B. Rubin. Influence Estimation on Social Media Networks Using Causal Inference. IEEE Statistical Signal Processing Workshop (SSP), 2018. Available at: https://ieeexplore.ieee.org/document/8450823
[15] Kravchenko O.V., Umanets I.S., Gevko O.O. Using Conversator Bots for Opinion Formation Web Community Participant// Intelligent decision-making systems and problems of computational intelligence: materials of the international of science conf., p. Iron Port, May 21-25, 2019 - Kherson: Vyshemirsky V.S. Publishing House, 2019. - PP.93-94 ISBN 978-617-7783-02-1 (electronic edition)
[16]Judicial expertise in civil proceedings // Linguistic expertise. URL:https://stud.com.ua/6784/pravo/lingvistichna_ekspertiza.
[17]Azhnyuk L.V. // Linguistic expertiseas a legaltool. URL: http://ekmair.ukma.edu.ua/bitstream/handle/123456789/11606/Azhniuk_Linhvistychna_ekspertyza.pdf
[18] Bekarystankyzy, A., & Mamyrbayev, O. (2023). END-TO-END SPEECH RECOGNITION SYSTEMS FOR AGGLUTINATIVE LANGUAGES. Scientific Journal of Astana IT University, 13(13), 86–92. https://doi.org/10.37943/13IMII7575