DOI: https://doi.org/10.15276/hait.07.2024.14
UDC 004.9

Information system for analyzing public sentiment in web platforms based on machine learning

Dmytro I. Uhryn[1)]
ORCID: https://orcid.org/0000-0003-4858-4511; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300
Artem O. Karachevtsev[1)]
ORCID: https://orcid.org/0009-0000-6226-6822; a.karachevtsev@chnu.edu.ua. Scopus Author ID: 36925155800
Yurii Ya. Tomka[1)]
ORCID: https://orcid.org/0000-0002-0495-3090; y.tomka@chnu.edu.ua. Scopus Author ID: 9279702200
Mykyta M. Zakharov[1)]
ORCID: https://orcid.org/0009-0003-5026-3546; jake2000nik@gmail.com
Yuliia L. Troianovska[2)]
ORCID: https://orcid.org/0000-0002-6716-9391; troyanovskaja@gmail.com. Scopus Author ID: 57211747293
[1)] Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubynsky Str. Chernivtsi, 58002, Ukraine
[2)] Odessa Polytechnic National University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ABSTRACT

The systems for studying public sentiment in web platforms are analyzed. Various tools and methods for effectively determining the mood in textual data from web platforms are described, including the formalization of the social graph and the content graph. The process of classifying comments, which includes the systematization and categorization of statements, is investigated. Based on the studied dataset, information on customer reviews and hotel ratings in Europe from the booking.com web platform is selected. Taking into account the requirements of the information system and the results of the analysis, it is determined that in order to obtain better results in determining the emotional connotation of the texts of reviews and messages from users, the most appropriate is the use of machine learning methods, taking into account natural language methods for processing text data. When choosing a text vectorization method for machine learning, the Term Frequency Inverse Document Frequency Vectorizer was chosen as the most effective among the studied methods. The architectural structure of the studied system is proposed, which is aimed at effective interaction between components and modules. The LogisticRegression model is chosen to determine the public mood. An information system has been developed that analyzes public sentiment about objects, uses advanced machine learning technologies to assess the emotional connotation of text comments, and provides users with insights and analysis of the results.

Keywords: Web platform; information system; public mood; propaganda; disinformation; fake; message; text; data mining; artificial intelligence; machine learning

For citation: **Uhryn D. I., Karachevtsev A. O., Tomka Y. Ya., Zakharov M. M., Troianovska Y. L. "Information system for analyzing public sentiment in web platforms based on machine learning".** ***Herald of Advanced Information Technology.*** **2024; Vol. 7 No. 2: 199–212. DOI: https://doi.org/10.15276/hait.07.2024.14**

INTRODUCTION

Due to the rapid development of information technology and the increasing use of web-based platforms for communication and expression of opinions, it is becoming increasingly important to analyze public opinion in the digital space [1, 2], [3]. The media, business and academic community are showing considerable interest in collecting and processing information that reflects the attitudes of citizens towards various aspects of public life.

In this period of digital transformation, web platforms are turning into more than just platforms for interaction and information exchange, but also into a powerful tool for studying the collective consciousness and mood of society.

The growing influence of these networks is giving rise to a new era, where an important task is to develop effective tools for analyzing and determining public sentiment [4, 5]. This is becoming a key task for various fields, including marketing, politics, and social sciences.

Modern web platforms have become an unlimited source of data that reflects public sentiment and opinions. Users express their opinions, impressions and reactions to various events, discuss trends and common interests. In this context, public sentiment analysis plays a key role in understanding the needs, preferences, and reactions of consumers [6, 7].

The increase in the number of users on web platforms leads to the exploitation of information circulating in this virtual space. The need for efficient and automated analysis of this data flow is clearly evident, as it is the key to identifying trends, demand, and public reaction. In this context, the development of an information system for analyzing public sentiment becomes key, given its potential impact on marketing, politics, public opinion research, and many others. It should also be noted that business and marketing companies are actively using various web platforms to make strategic decisions. Therefore, public opinion analysis allows forecasting market trends, evaluating the effectiveness of advertising campaigns, and adapting products to consumer needs [8].

In the field of scientific research, public sentiment analysis is becoming a key tool for studying the social and psychological aspects of human behavior. In this context, the use of advanced information technology methods, such as text classification and sentiment analysis, not only opens up new opportunities for understanding public reactions to events, products, or ideas, but also becomes a key to improving modern approaches to analyzing web platforms.

The relevance of this paper lies in the need to improve the analysis of public sentiment using information technology and machine learning methods. Modern web platforms not only serve as platforms for communication, but also reflect the general response and mood of society. With the ever-increasing flow of information in these networks, it is important to develop tools for accurate and automated data analysis. The amount of information on web platforms exceeds the capabilities of manual processing, so there is a need to develop an information system aimed at efficient and automated analysis of public sentiment. The development of an information system for this purpose is an important step in improving the tools for analyzing data from web platforms.

ANALYSIS OF LITERARY DATA AND PROBLEM DEFINITION

Public opinion analysis systems on web-based platforms can differ in terms of various features and functionality.

Below are a few key features that can distinguish between such systems:

1. Methods of analysis:

1.1. Text analysis: some systems focus on analyzing textual information, taking into account the tone, mood and emotions in texts.

1.2. Visual analysis: Others may use the analysis of visual content, such as photos or videos, to determine sentiment.

2. Data sources:

2.1. Various web platforms: Some systems analyze data from various web platforms such as Twitter, Facebook, Instagram, YouTube, etc.

2.2. Specialization on a specific network: Others may specialize in certain web platforms, taking into account their unique features.

3. The scale of the system:

3.1. Local analysis: Some systems are limited to analyzing data within a specific topic area or local context.

3.2. Global analysis: Others may focus on global analysis of public sentiment, taking into account global trends and events.

4. Use of machine learning technologies:

4.1 Basic methods: some systems may use basic analysis methods such as rules and keywords.

4.2. Advanced methods: Others may implement advanced machine learning methods to improve the accuracy and efficiency of the analysis.

5. Applying:

5.1. Business intelligence: Some systems may be focused on business intelligence applications to determine consumer reactions to products or services.

5.2. Academic research: Others may be used in academic research to study social phenomena and trends.

Each of these features determines the accuracy and effectiveness of the public opinion analysis system in web platforms.

Each of the various web platforms can be viewed as a tuple consisting of two sets: a set of users and a set of content [9].

The social graph of users is a multigraph, which means that the relationship between users can be multidimensional and contain additional structural data:

$$G = \{V, E_1 \dots E_k, p, \delta_1 \dots \delta_k\}, \qquad (1)$$

where V are vectors; $E_1 \dots E_k$ are types of relations in the network; p is actor profile, which is an organizational or group social unit.

Examples of actors include individuals in a group, departments in a corporation, government agencies in a city, or countries in the world. In the case when all actors belong to the same type, for example, people in a group, they form a homogeneous network. $V \rightarrow p$; $\delta_1 \dots \delta_k$ – parameters of the corresponding relationship $E_i \rightarrow \delta_i$.

The network content graph can be represented in a similar way:

$$C = \{T, R_1 \dots R_m, \theta, \gamma_1 \dots \gamma_m\}, \qquad (2)$$

where T is an aggregate of various content in the network; $R_1 \dots R_m$ are interaction between content elements; θ is content parameters $T \to \theta$; $\gamma_1 \dots \gamma_m$ are characteristics of the corresponding connection within the content $R_j \to \gamma_j$.

The result $A \in V \times T$ is a representation of the relationship between actors and content

$$A = \{L_1 \dots L_n, \varepsilon_1 \dots \varepsilon_n\}, \qquad (3)$$

where $L_1 \dots L_n$ are elements of interaction of network actors in relation to the content; $\varepsilon_1 \dots \varepsilon_n$ are characteristics of the relationship of the relevant elements $L_l \to \varepsilon_l$.

The process of classifying comments on web platforms involves systematizing and dividing statements into different categories or classes according to specific parameters or characteristics. This method is aimed at improving the analysis of extensive interaction on web platforms and allows to effectively solve the problems of filtering unwanted or aggressive statements, identifying thematic and emotional shades, etc.

The classification can be based on various criteria, such as the tone of statements (positive, negative, neutral), topics (sports, politics, entertainment, etc.), the level of aggression, spamming, and features of language analysis (sentiment analysis, keyword detection) [12, 13].

Interaction between network members and content can be analyzed by determining the authorship of posts and comments. Posts can be in the form of text, images, audio or video and are intended to inform network members about a particular event or phenomenon (Fig. 1). Comments represent the attitudes of actors to messages recorded in a certain format.

Classifying comments helps to improve the security and usability of web platforms and allows users to interact more effectively and receive content that meets their interests and standards.

The analysis of web platforms on the Internet is often based on the “egocentric graph”, which is the simplest method. In this graph, the vertices represent the “Ego” and its closest neighbors. Although this approach does not reflect all the characteristics of the network, it can be used to study social roles in the group [14, 15]. Analyzing the network topology can help identify patterns, but for a deeper understanding of the network, you need to take into account characteristics such as centrality, density, average number of paths through an actor, level of structural equivalence, and others.

In some cases, statistical characteristics such as variance, centrality, and histograms of the distribution of node degrees are also important. Patterns of interconnections in a “self-centered network” can reflect the position, social or professional activities of an actor [16]. For example, managers and administrators may share a “bridge” pattern that connects subgroups in an organization.

To describe development teams, the “onion” template may be suitable, where a dense core is surrounded by layers from other structural units. Each social role defines its own structure in the network, which allows it to be categorized by the nature and structure of its connections. This can be useful for identifying potentially useful users or identifying “provocateurs,” spammers, “bots,” and “flooders,” and separating them from other actors.

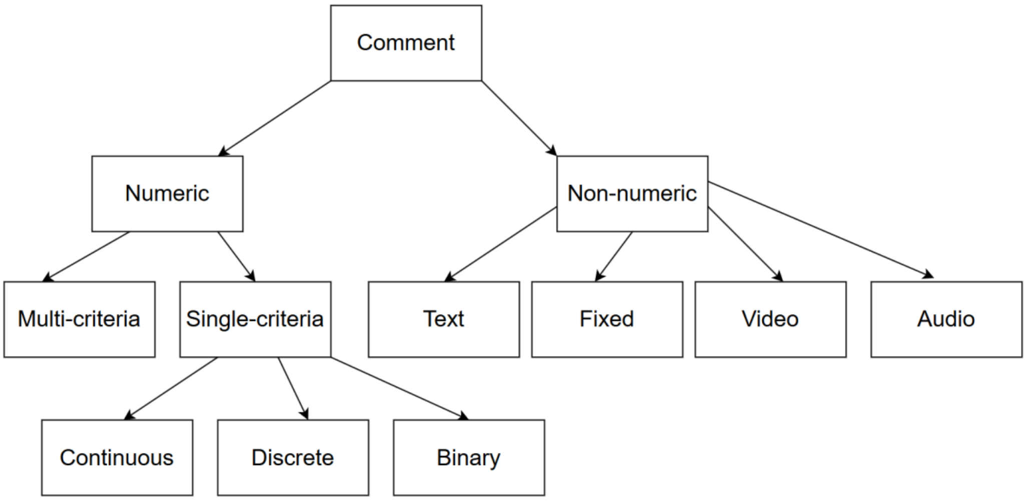


Fig. 1. **Classification of comments in web platforms**
Source: **compiled by the authors**

More complex studies are conducted on the basis of a full-scale network graph. The study of popular online web platforms such as Flickr, YouTube, LiveJournal, Orkut, and Facebook confirms their scale-free nature [15]. This means that the network has a power law distribution of connections $P(k) \sim k^{-\gamma}$ between actors. In other words, most actors have a similar number of connections, and the share of actors with an excessively large number of connections is rare. Thus, an online network can be represented as a structure with hierarchically connected nodes, similar to computer networks [16, 17[, [18].

Scale-free networks are highly stable against random element failures. At the same time, they are sensitive to targeted attacks – the exclusion of elements with the largest number of connections can lead to a loss of component connectivity [19, 20]. To split a web platform into small components, it is enough to remove about 1-10 % of the elements with the highest vertex degree. These properties are important in the analysis of web platforms, as the development of the theory of resilience and vulnerability is directly related to the analysis of information flow. Communication between nodes is assumed to be mainly through concentrators - actors with the largest number of connections. Controlling hubs is almost identical to controlling the entire network, while controlling a large number of individual actors is usually not justified.

To provide reliable estimates of the characteristics of web platforms, a full-scale graph must be taken into account. The methods for obtaining appropriate data samples are relatively simple, and the process is easy to parallelize. However, the samples themselves may not justify the material and computational costs required to obtain them. Therefore, for many researchers, obtaining, let alone analyzing, a complete network may seem impossible.

To adequately assess the characteristics of the network, it is sufficient to obtain a representative sample that has the same characteristics as the full network. Such a network can be created by randomly selecting a subset of actors. To ensure the uniformity of the samples, there are various methods, among which the most famous are [21, 22], [23]: walkover in width; random walkover; reweighted random walkover; random walkover of Metropolis-Hastings.

In general, the width and random walk methods can provide the required amount of data, but the sample obtained by these methods is far from uniform and has biased characteristics. In particular, Metropolis-Hastings and re-weighted random walks allow you to get indicators that are close to a uniform sample with unbiased characteristics. To find statistically unbiased network characteristics, it is recommended to consider a subgraph of about 3-10 thousand actors.

To analyze and display data collected from the web platform, it is necessary to use specialized software such as UCINET, Pajek, Gephi, ORA, NetMiner, StOCNET, MultiNet, GUESS, as well as NodeXL applications and libraries igraph (R, Python, C), libSNA, NetworkX (Python), SNA (R), SNAP (Gauss), SNAP (C++) [15]. To process and store user profile data, it is structured by extracting features and attributes from the account web page. It is also important to identify a list of keywords and normalize them using approaches such as the bag of words model and stemming [15, 24],[25].

In the bag-of-words model, the text is represented as an unordered set of words, without regard to grammar or word order. To obtain information from the user's profile, it is necessary to perform stemming to find the basis of each word. One of the most popular and effective stemming algorithms is Porter's stemmer, which includes prefix and suffix extraction.

However, this method has its drawbacks:

1) in Ukrainian and Russian, where vowels and consonants alternate, the word is often truncated to a too short base;

2) the method is sensitive to spelling errors;

3) the set of word-forming parts is different for different languages and their word-forming rules [26, 27], [28].

An alternative to stemming is fuzzy word retrieval, which is based on using the Levenshtein metric to find words in a dictionary. This metric is defined by the minimum number of corrections required to transform a word into a word from the dictionary. The advantages of this method are error tolerance, language independence, and the ability to compare word similarity in steps. The conceptual approach to this information technology for analyzing web platforms is discussed in [9, 10], [15].

On web platforms, you can conduct various surveys and questionnaires with subsequent analysis using sociometric methods. The information received from users from the web platform contains various types of uncertainties due to various “non-factors”. To take into account the uncertainties of the initial information for analyzing the web platform, it is proposed to use fuzzy logic in [12, 13], [29, 30].

THE PURPOSE AND OBJECTIVES OF THE RESEARCH

The main goal of this study is to develop an information system for analyzing public sentiment in web platforms using data mining methods.

Main objectives of the research:

1. Selection and application of machine learning methods for text analysis: evaluation of different machine learning methods, such as Logistic Regression, to determine their effectiveness in identifying public sentiment from text data on web platforms.

2. Development and optimization of text processing algorithms: creation of text processing algorithms that will help determine the emotional connotation of comments and messages from web platforms using natural language methods.

3. Development of an information system for analyzing public sentiment on web platforms: creation of a software system that can analyze texts, comments and reviews on a web platform to identify emotional tone and public sentiment.

4. Validation and testing of the developed information system: validation and testing of the developed system on real data to confirm its effectiveness and accuracy in determining public sentiment on web platforms.

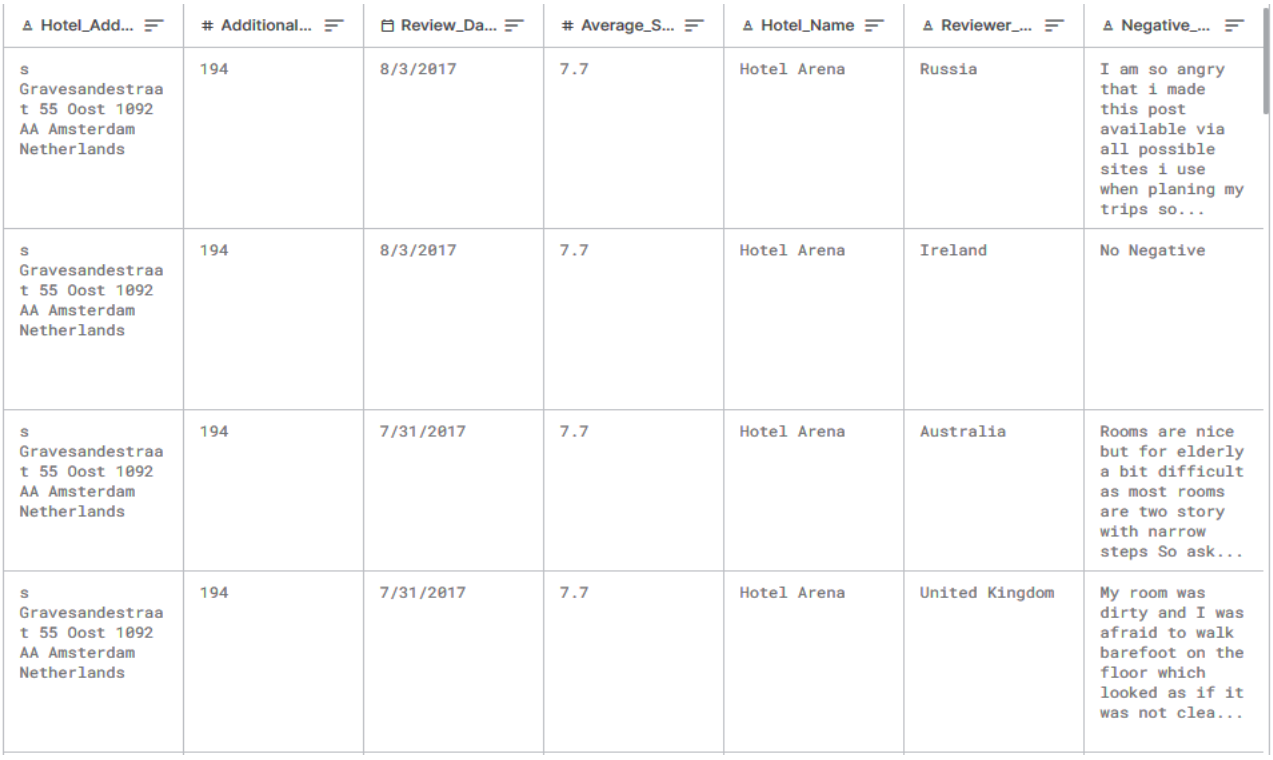
MATERIALS AND RESEARCH METHODS

The study analyzes various tools and methods for effectively determining sentiment in text data from web platforms. The main goal of this study is to select the optimal model for further use in a service for analyzing public sentiment.

For the successful functioning of a machine learning model, it is important to clearly define the characteristics of the input data. In our case, the input data is represented by texts from the web platform (Fig. 2), such as reviews, comments, and messages. In addition to textual data, we consider the possibility of using additional parameters, such as time of publication, number of likes, reposts, etc. This allows for a more comprehensive analysis of public sentiment and provides more accurate results.

The dataset under study contains information on 515,000 customer reviews and ratings of 1493 hotels across Europe, collected from the Booking.com platform. The csv file includes 17 fields, including the key ones: date of the comment, hotel name, positive and negative reviews, and the text of the comment.

Taking into account the requirements of the information system and the results of the analysis, it was determined that in order to obtain better results in determining emotional connotation of texts



Hotel_Add...	# Additional...	Review_Da...	# Average_S...	Hotel_Name	Reviewer_...	Negative_...
s Gravesandestraat 55 Oost 1092 AA Amsterdam Netherlands	194	8/3/2017	7.7	Hotel Arena	Russia	I am so angry that i made this post available via all possible sites i use when planing my trips so...
s Gravesandestraat 55 Oost 1092 AA Amsterdam Netherlands	194	8/3/2017	7.7	Hotel Arena	Ireland	No Negative
s Gravesandestraat 55 Oost 1092 AA Amsterdam Netherlands	194	7/31/2017	7.7	Hotel Arena	Australia	Rooms are nice but for elderly a bit difficult as most rooms are two story with narrow steps So ask...
s Gravesandestraat 55 Oost 1092 AA Amsterdam Netherlands	194	7/31/2017	7.7	Hotel Arena	United Kingdom	My room was dirty and I was afraid to walk barefoot on the floor which looked as if it was not clea...

Fig. 2. **Samples of raw data from the collection of “515K Hotel Reviews Data in Europe”**
***Source:* compiled by the authors**

of reviews and messages from users, the most appropriate approach is to use machine learning methods. In particular, natural language processing (NLP) methods of sentiment analysis have been chosen to process textual data and determine semantic coloration [9, 10], [12].

In choosing the optimal approach to text data processing, two natural language methods are taken into account to reduce them to their basic form:

1. Stemming, this involves discarding affixes in order to preserve the main root of the word. For example, “Running” can be simplified to “run”.

2. Lemmatization, a more complex process that takes into account grammatical rules and reduces words to their normal form (lemmas). For example, “Better” can be reduced to “good”.

After a detailed analysis of both methods, it was decided to use lemmatization for processing natural language words. This choice is justified by the fact that lemmatization meets the requirements of the task, where accuracy and grammatical correctness are important, which perfectly corresponds to to the needs of the product. On the other hand, stemming is more often used to simplify words to their basic root when accuracy is not the main priority.

When choosing a text vectorization method for further use in machine learning, two approaches were considered: Word Embedding and TF-IDF Vectorizer.

After analysis and testing, the following results of model performance were obtained (Table 1).

Table 1. **Comparative results of text vectorization methods**

Text vectorisation method	Precision	Recall	F1 score	support
Word Embedding	0.956098	0.859649	0.905312	None
TF-IDF Vectorizer	0.985294	0.881579	0.930556	None

Source: **compiled by the authors**

Even though the Word Embeddings method uses a more modern and sophisticated approach. It seemed that it could lead to better results. However, in fact, TF-IDF Vectorizer proved to be more effective.

The text data was converted into numeric vectors using TF-IDF. This vectorization step is key because it converts text into a set of numerical values that can be used to train models.

The advantage of TF-IDF Vectorizer is explained by several factors:

1. Consideration of all words: TF-IDF Vectorizer takes into account all available words in the text, which allows you to extract more information from longer documents. Even the extra 7 % of documents that contain more than 20 words represent a significant amount of information.

2. Less potential data overload: The use of Word Embeddings can lead to data overload due to the larger vocabulary. However, given that only a small percentage of documents are longer than 20 words, this overload may not be optimal in terms of model accuracy.

3. Less noisy signal: Word Embeddings may contain noisier signal compared to TF-IDF Vectorizer. Taking into account more hidden information can lead to the creation of false patterns in the model, which can negatively affect its accuracy.

The structure of the studied system is proposed by the architecture, which is aimed at ensuring effective interaction between various components and modules of the system. Below is a detailed overview of each element of the information system architecture diagram (Fig. 3).

The diagram shows the complex structure of the information system, which includes various components for efficient processing and display of data in the context of public opinion analysis.

Let's consider the main elements and their functionality:

1. Web server is the central element that interacts with all other components. It processes requests from client browsers, interacts with PostgreSQL and Sentiment Analyzer to process and send data.

2. PostgreSQL is a database component that stores and interacts with object categories. It provides the web server with access to the necessary data through the Return View.

3. Sentiment analyzer (Logistic Regression) – used to analyze public sentiment based on text comments. The results of the analysis are transmitted to the web server for display to the user.

4. Client browser maps - the client browser interacts with the Google Map API, sends requests to the server and displays changes on the map according to the received data.

5. Client browser sentiment analysis interaction – user interaction with the interface for entering text questions, sending them to the server for analysis and receiving mood results.

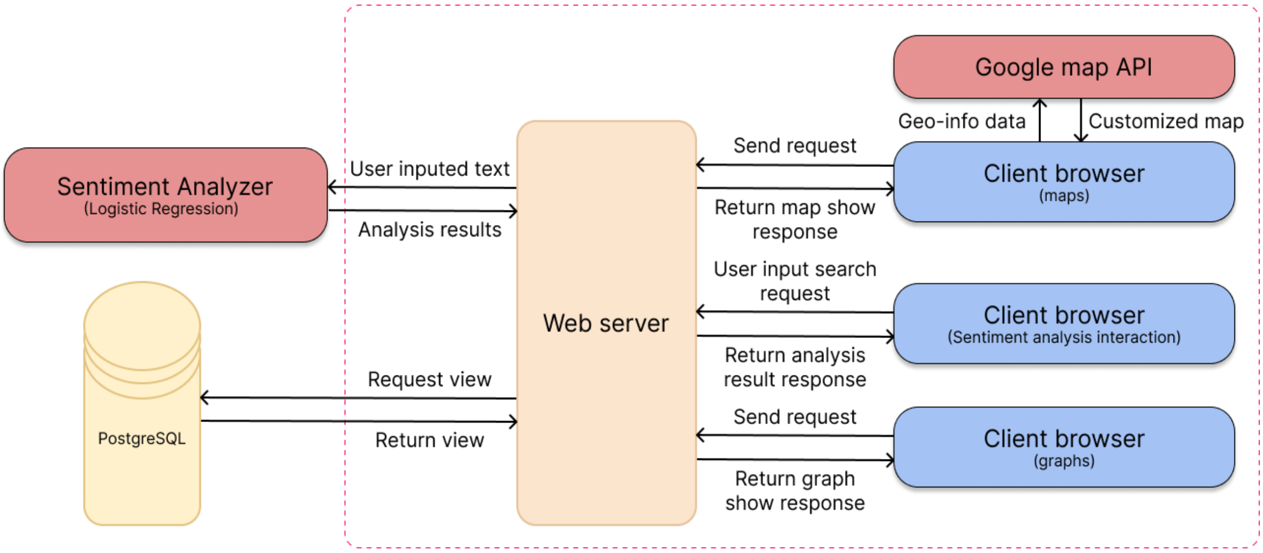


Fig. 3. **Architecture of information systems**
Source: **compiled by the authors**

6. Client browser graphs – the client browser receives graphs and sentiment charts from the server to display to the user.

7. Google Map API – interacts with the client browser to obtain geographic data and create modified cartography. Transmits and receives geographic data to build customized maps.

RESEARCH RESULTS

Accuracy is one of the criteria for evaluating classification models. Informally, it represents the percentage of predictions that the model identified correctly.

Formally, accuracy is defined as follows:

$$Pr = \frac{N_{CF}}{N_{AF}},$$

where Pr – accuracy; N_{CF} – the number of correct predictions; N_{AF} – the number of all predictions.

In the case of binary classification, the accuracy can be calculated using the criterion of relative positive and negative ratings

$$k = \frac{Res_{sip} + Res_{sin}}{Res_{sip} + Res_{sin} + Res_{iip} + Res_{iin}},$$

where Res_{sip} is correctly identified positive results, Res_{sin} is correctly identified negative results, Res_{iip} is incorrectly identified positive results, Res_{iin} is incorrectly identified negative results.

The models for determining public sentiment are selected and their accuracy is evaluated. Among the selected classifiers are DecisionTreeClassifier, RandomForestClassifier, SVC, LogisticRegression, KNeighborsClassifier, BernoulliNB. After using cross-validation for each model, the results on the accuracy and errors of their functioning were obtained (Table 2).

Table 2. **Conclusions of the results on the accuracy and error of the models**

Model	Accuracy, %	Error rate, %
DecisionTreeClassifier	97.2903	2.7097
RandomForestClassifier	97.0301	2.9699
SVC	95.8301	4.1699
LogisticRegression	98.0301	1.9699
KNeighborsClassifier	98.0301	1.9699
BernoulliNB	94.7999	5.2001

Source: **compiled by the authors**

The model error is defined as the percentage of incorrectly classified examples out of the total number of test data. Among the different models, Logistic Regression and KNeighborsClassifier show the lowest error.

An assessment of the accuracy and error of the models leads to the following conclusions:

1. DecisionTreeClassifier showed high accuracy, but may be prone to overfitting.

2. The RandomForestClassifier showed good accuracy, but may require significant training resources.

3. SVC model has lower accuracy compared to other models.

4. LogisticRegression is characterized by high accuracy and training speed, making it an attractive option.

5. KNeighborsClassifier demonstrates accuracy on par with other models, but can be costly in terms of computing resources.

6. BernoulliNB has slightly lower accuracy compared to other models.

Taking into account the results and requirements of the information system for determining public opinion, the LogisticRegression model was chosen.

The main advantages of this choice include:

1. High accuracy: the model demonstrates one of the highest accuracies compared to other classifiers.

2. Efficiency: LogisticRegression is characterized by high training and prediction speed, which is a key aspect for the service.

3. Reliability: the model works efficiently when processing a limited amount of data, which makes it practical to use in real-world conditions.

Therefore, given the above, we can choose LogisticRegression for a service for recognizing public sentiment.

SentimentScope is a developed information system designed to analyze public sentiment about various objects, such as hotels, products, personalities (media or political), brands, and other popular topics. The system uses advanced machine learning technologies to assess the emotional color of text comments and provides users with insights and analysis of the results.

User-flow of the SentimentScope information system:

1. The user opens the SentimentScope information system in his browser and logs in. After successful login, the user gets access to the main functionality of the application. On the main page, after authorization, the user selects the section of interest (for example, hotels, products, personalities, brands, etc.). The main page with the section selection is shown in Fig. 4.

2. The user can familiarize himself with the most popular objects in the selected section (Fig. 5) and analyze their sentiment. If the user has a specific object in mind, they can use the search field to find it.

3. After selecting a specific object, the user is taken to a page with a detailed analysis (Fig. 6). Here, the user receives information on public sentiment, statistics, and data visualization in the form of graphs.

DISCUSSION OF THE RESULTS

The need to analyze public sentiment in the digital space is becoming increasingly important, as information technology and web platforms are becoming key components of modern society. Web platforms are turning into unlimited sources of data that reflect the collective consciousness and mood of the public. The development of an information system for effective analysis of this data is

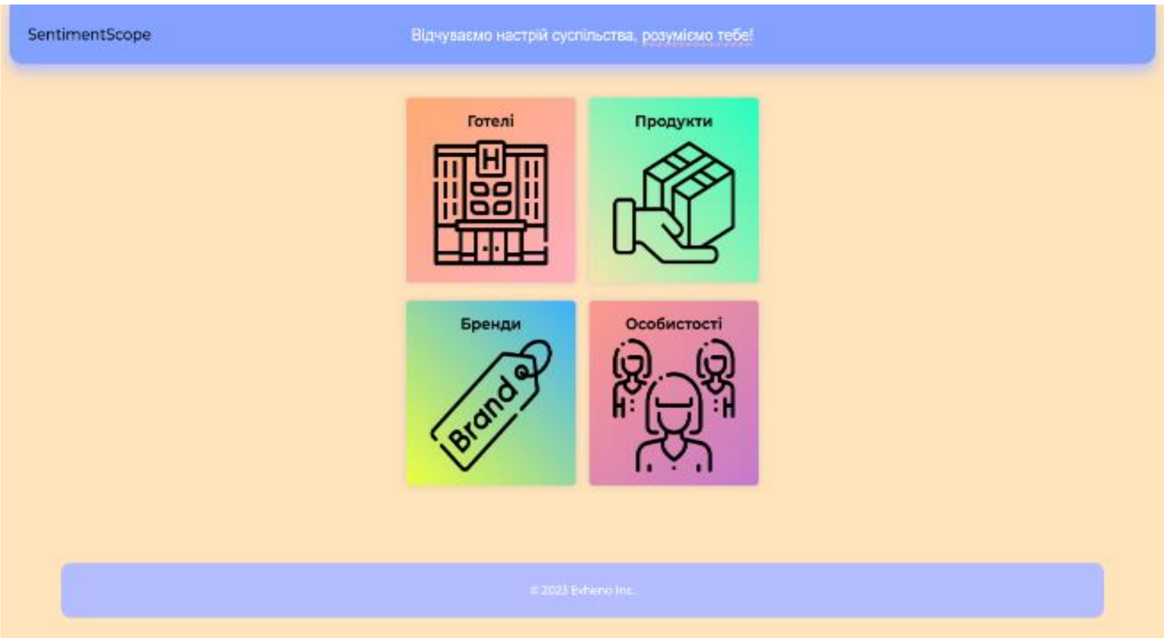
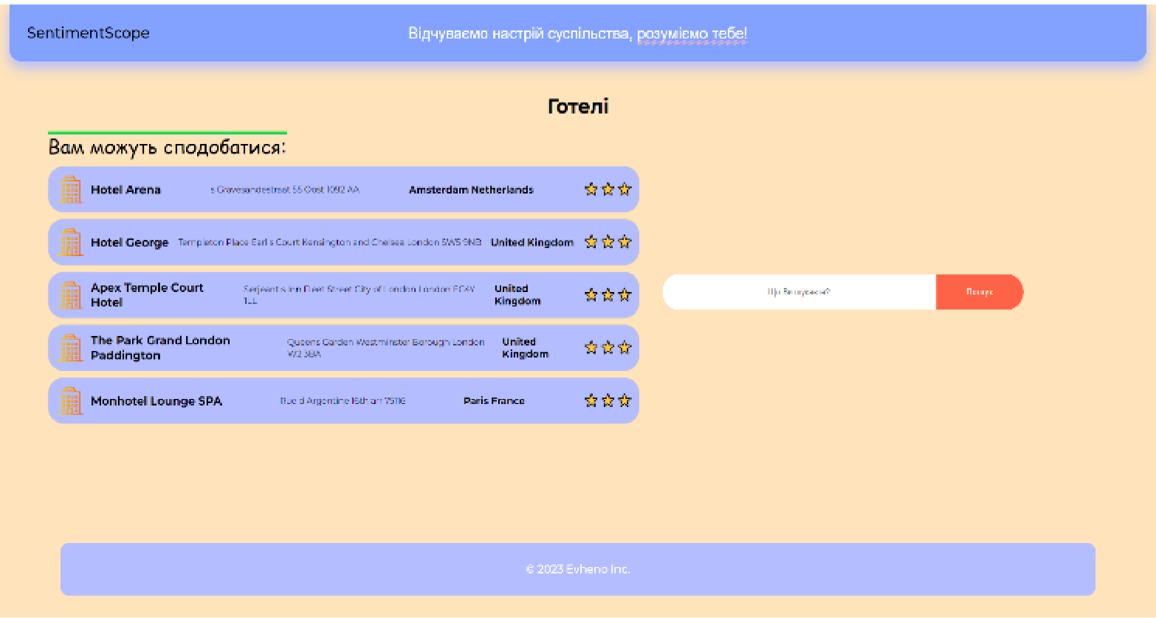


Fig. 4. **Main page of the information system**
Source: **compiled by the authors**

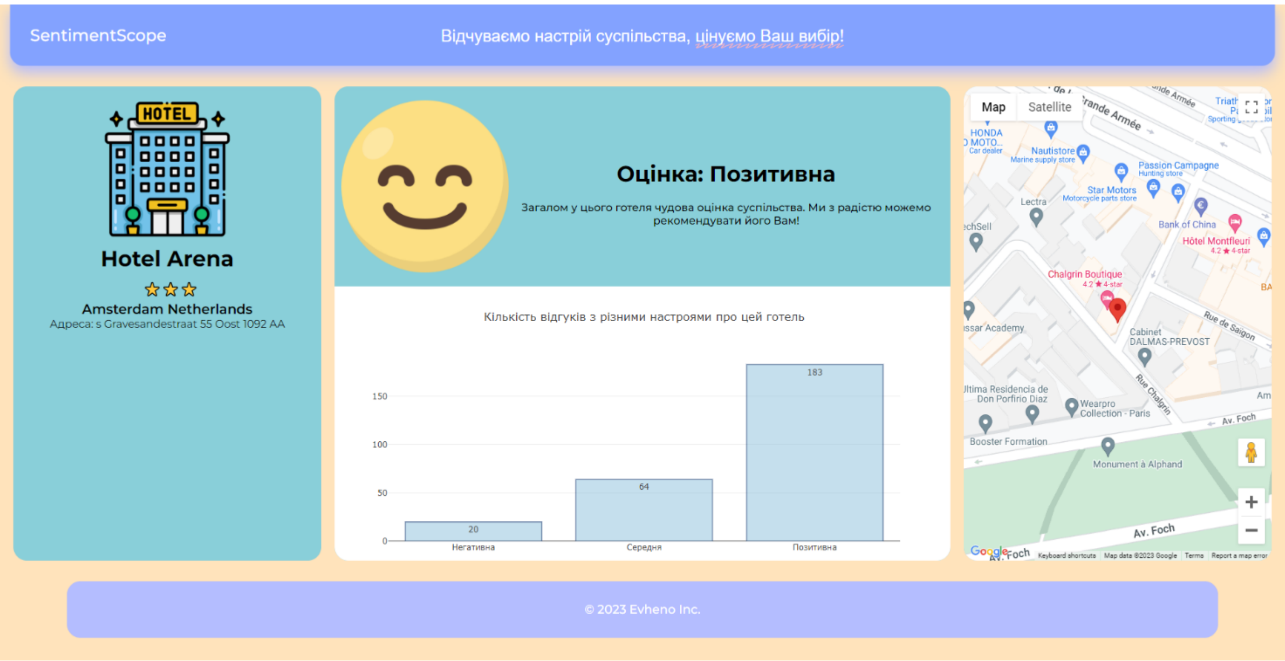


Готелі

Вам можуть сподобатися:

Hotel Arena	s Gravesandestraat 55 Oost 1092 AA	Amsterdam Netherlands	☆☆☆
Hotel George	Templeton Place Earl s Court Kensington and Chelsea London SW5 9NB	United Kingdom	☆☆☆
Apex Temple Court Hotel	Serjeant s Inn Fleet Street City of London London EC4Y 1LL	United Kingdom	☆☆☆
The Park Grand London Paddington	Queens Garden Westminster Borough London W2 3BA	United Kingdom	☆☆☆
Monhotel Lounge SPA	Rue d Argentine 16th arr 75116	Paris France	☆☆☆

Fig. 5. **Page "Hotels" of the information system**
Source: **compiled by the authors**



Hotel Arena
☆☆☆
Amsterdam Netherlands
Адреса: s Gravesandestraat 55 Oost 1092 AA

Оцінка: Позитивна

Загалом у цього готеля чудова оцінка суспільства. Ми з радістю можемо рекомендувати його Вам!

Кількість відгуків з різними настроями про цей готель

Fig. 6. **Sentiment analysis results page for selected hotel**
Source: **compiled by the authors**

becoming strategically important, especially in the context of influencing the areas of marketing, politics and public opinion research. The use of advanced information technology methods in analyzing web platforms not only opens up new opportunities for understanding public reaction, but also identifies areas for improving current approaches. The growing number of users on web platforms leads to the need for automated analysis of large amounts of data circulating in the virtual space. Public sentiment analytics is becoming key for business, marketing, and research, allowing predicting trends and adapting strategies to the needs of consumers. In a world where the amount of information exceeds the capabilities of manual analysis, the development of innovative information technologies and services for public sentiment analysis is becoming an important step in improving tools for working with data from web platforms. Such work has great potential for further development and use in various industries, contributing to the development of scientific research, political analysts, marketers and other areas of society.

CONCLUSIONS

This research is aimed at determining the importance of conducting public opinion analysis in the digital space due to the rapid development of information technology and the expanding use of web platforms, which is aimed at developing an information system for analyzing public opinion.

The systems for analyzing public sentiment in the web platforms were analyzed and compared by various characteristics and functionality. The analysis of various tools and methods aimed at effectively determining the mood in text data from web platforms was carried out. The social graph and the network content graph were formalized. The relationship between actors and content is modeled. The process of classifying comments on web platforms is described, which includes the systematization and distribution of statements into different categories or classes according to specific parameters or characteristics. The author analyzes specialized software for processing data collected from a web platform.

The data set under study is based on information on customer reviews and hotel ratings across Europe collected from the Booking.com web platform.

Taking into account the requirements of the web service and the results of the analysis, it is determined that the use of machine learning methods is most appropriate for obtaining better results in determining the emotional connotation of texts of reviews and messages from users. When choosing the optimal approach to text data processing, two natural language methods are taken into account to reduce them to their basic form: stemming and lemmatization. After a detailed analysis of both methods, it was decided to use lemmatization to process natural language words. This choice is justified by the fact that lemmatization meets the requirements of the task, where accuracy and grammatical correctness are important, which perfectly matches the needs of the product. On the other hand, stemming is more often used to simplify words to their basic root when accuracy is not the main priority.

In choosing a text vectorization method for further use in machine learning, two approaches were considered: Word Embedding and TF-IDF Vectorizer. Both methods showed high F1 accuracy: 90.5% for Word Embeddings and 93.1% for TF-IDF Vectorizer. However, TF-IDF Vectorizer actually proved to be more efficient at converting text data into numeric vectors. This stage of vectorization is key, as it converts text into a set of numerical values for further use in model training.

The structure of the studied system is proposed by an architecture aimed at ensuring effective interaction between various components and modules of the system.

Models for determining public sentiment were selected and their accuracy was evaluated. Among the selected classifiers, after using cross-validation for each model, the results on the accuracy and errors of their functioning are obtained. Taking into account the results and requirements for an information system for determining public opinion in web platforms, the LogisticRegression model was chosen.

An information system has been developed to analyze public sentiment about various objects, such as hotels, products, personalities (media or political), brands, and other relevant topics. Using advanced machine learning technologies to evaluate the emotional tone of text comments, the system provides users with insights and analysis of the results.

REFERENCES

1. Ulichev O., Meleshko Ye., Sawicki D. & Smailova S. “Computer modelling of dissemination of informational influences in social networks with different strategies of information distributors”. *Proc. SPIE*. 2019; 111761T, https://www.scopus.com/authid/detail.uri?authorId=57212037312.
DOI: https://doi.org/10.1117/12.2536480.

2. Ugryn, D. I., Ushenko, Y. O., Dovgun, A. Y., & Kalancha A. D. “Intelligent system for identifying user trust rating”. *Optoelectronic Information-Power Technologies*. 2023; 46 (2): 150–158,

https://www.scopus.com/authid/detail.uri?authorId=57163746300. DOI: https://doi.org/10.31649/1681-7893-2023-46-2-150-158.

3. Hamborg, F. & Donnay K. "NewsMTSC: A Dataset for Multi-Target-dependent Sentiment Classification in Political News Articles". *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. 2021. p. 1663–1675, https://www.scopus.com/authid/detail.uri?authorId=57195416222. DOI: https://doi.org/10.18653/v1/2021.eacl-main.142.

4. Liang, B., Su, H., Gui, L., Cambria, E. & Xu, R. "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks". *Knowledge-Based Systems*. 2022; 235: 107643, https://www.scopus.com/authid/detail.uri?authorId=57201841373.
DOI: https://doi.org/10.1016/j.knosys.2021.107643.

5. Birjali, M., Kasri, M. & Beni-Hssane, A. "A comprehensive survey on sentiment analysis: Approaches, challenges and trends". *Knowledge-Based Systems.* 2021. 226: 107134, https://www.scopus.com/authid/detail.uri?authorId=57192214153.
DOI: https://doi.org/10.1016/j.knosys.2021.107134.

6. Shanti Pragnya, S. "VADER (Valence Aware Dictionary and sentiment Reasoner) Sentiment Analysis". 2022. Available from: https://medium.com/mlearning-ai/vader-valence-aware-dictionary-and-sentiment-reasoner-sentiment-analysis-28251536698. – [Accessed: March, 2023].

7. Prokipchuk, O. & Vysotska, V.. "Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning". *Radio Electronics, Computer Science, Control*. 2023; 2: 103–116, https://www.scopus.com/authid/detail.uri?authorId=57226532359. DOI: https://doi.org/10.15588/1607-3274-2023-2-11.

8. Ahmed, S. & Kumar, A.. "Classification of censored tweets in chinese language using XLNet". *Proceedings of Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics*. 2021; 136–139, https://www.scopus.com/authid/detail.uri?authorId=57385522100. DOI: https://doi.org/10.18653/v1/2021.nlp4if-1.21.

9. Vysotska, V., Mazepa, S., Chyrun, L., Brodyak, O., Shakleina, I. & Schuchmann, V. "NLP tool for extracting relevant information from criminal reports or fakes/propaganda content". *Computer Sciences and Information Technologies: 17th International Conference*. 2021; 93–98, https://www.scopus.com/authid/detail.uri?authorId=24484045400. DOI: https://doi.org/10.1109/CSIT56902.2022.10000563.

10. Prokipchuk, O., Vysotska, V., Pukach, P., Lytvyn, V., Uhryn, D., Ushenko, Y. & Hu, Z. "Intelligent analysis of Ukrainian-language tweets for public opinion research based on NLP methods and machine learning technology". *International Journal of Modern Education and Computer Science (IJMECS)*. 2023; 15 (3): 70–93, https://www.scopus.com/authid/detail.uri?authorId=57226532359.
DOI: https://doi.org/10.5815/ijmecs.2023.03.06.

11. Uhryn, D. I., Halochkin, O. V., Khostiuk, A. V. & Ushenko, O. G. "Complex protection of information in operating systems". *Optoelectronic Information and Energy Technologies*. 2022; 2 (44): 44–48, https://www.scopus.com/authid/detail.uri?authorId=57163746300. DOI: https://doi.org/10.31649/1681-7893-2022-44-2-44-48.

12. Perdoor, S. "Fake news detection with LSTM and NLP". – Available from: https://www.kaggle.com/code/superrajdoor/fake-news-detection-with-lstm-and-nlp-prorew1. – [Accessed: March, 2024].

13. Al-Oraiqat, A. M., Ulichev, O. S., Meleshko, Y. V., AlRawashdeh, H. S., Smirnov, O. O. & Polishchuk, L. I. "Modeling strategies for information influence dissemination in social networks". *J Ambient Intell Human Comput.* 2022: 13: 2463–2477, https://www.scopus.com/authid/detail.uri?authorId=57200504918. DOI: https://doi.org/10.1007/s12652-021-03364-w.

14. Chen, Y. C. "A novel algorithm for mining opinion leaders in social networks". *World Wide Web*. 2019; 22: 1279–1295. DOI: https://doi.org/10.1007/s11280-018-0586-x.

15. Russell, M. A. "Mining the social web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub and More". *O'Reilly Media, Inc*. 2019.

16. Tomka, Yu. Ya., Talakh, M. V., Dvorzhak, V. V. & Ushenko, O. G. “Practical aspects of forming training/test samples for convolutional neural networks”. *Optical-Electronic Information and Energy Technologies*. 2022; 1 (43): 24–25, https://www.scopus.com/authid/detail.uri?authorId=9279702200. DOI: https://doi.org/10.31649/1681-7893-2022-43-1-24-35.

17. Talakh, M., V. Holub, S. V. Luchsheva, P. O. & Turkin, I. B. “Intelligent monitoring of air temperature by the DATA of satellites and meteorological stations”. *International Journal of Computing*. 2022: 21 (1): 120–127, https://www.scopus.com/authid/detail.uri?authorId=57211567133. DOI: https://doi.org/10.47839/ijc.21.1.2525

18. George, B., Omer, O. J., Choudhury, Z. & Subramoney, A. V. “A unified programmable edge ma trix processor for deep neural networks and matrix algebra”. *ACM Transactions on Embedded Computing Systems.* 2022; 21 (5): 63:1–63:30, https://www.scopus.com/authid/detail.uri?authorId=57209139155. DOI: https://doi.org/10.1145/3524453.

19. Pinzon-Arenas, J. O., Kong, Y., Chon, K. H. & Posada-Quintero, H. F. “Design and evaluation of deep learning models for continuous acute pain detection based on phasic electrodermal activity”, *IEEE Journal of Biomedical and Health Informatics*. 2023: 27 (9): 4250–4260, https://www.scopus.com/authid/detail.uri?authorId=57201298021.
DOI: https://doi.org/10.1109/JBHI.2023.3291955.

20. Uhryn, D., Ushenko, Y., Kovalchuk, M. & Bilobrytskyi, D. “Modelling a system for intelligent forecasting of trading on Stock Exchanges”. *Security of infocommunication systems and internet of things (SISIOT).* 2023; 1 (2): 02002, https://www.scopus.com/authid/detail.uri?authorId=57163746300. DOI: https://doi.org/10.31861/sisiot2023.2.02002.

21. Lytvyn, V., Lozynska, O., Uhryn, D., Vovk, M., Ushenko, Y. & Hu, Z. “Information technologies for decision support in industry-specific geographic information systems based on swarm intelligence”, *Modern Education and Computer Science*. 2023; 2: 62–72, https://www.scopus.com/authid/ detail.uri?authorId=56446930100. DOI: https://doi.org/10.5815/ijmecs.2023.02.06.

22. Nápoles, G ,Van Houdt, G. &Mosquera, C. “A review on the long short-term memory model”. *Artif Intell Rev* .2020: 53: 5929–5955, https://www.scopus.com/authid/detail.uri?authorId=37861926800. DOI: https://doi.org/10.1007/s10462-020-09838-1.

23. Nápoles, G., Griffioen, N., Khoshrou, S. & Güven, Ç. “Feature Importance for Clustering”. *Lecture Notes in Computer Science*. 2024; 14469, https://www.scopus.com/authid/detail.uri?authorId=37861926800. DOI: https://doi.org/10.1007/978-3-031-49018-7_3.

24. Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W. & Müller, K. -R. “From Clustering to Cluster Explanations via Neural Networks”. *IEEE Transactions on Neural Networks and Learning Systems*. 2024: 35 (2): 1926–1940, https://www.scopus.com/authid/detail.uri?authorId=57214096720. DOI: https://doi.org/10.1109/TNNLS.2022.3185901.

25. Marra, F., Gragnaniello, D., Cozzolino, D. & Verdoliva, L. “Detection of GAN-Generated Fake Images over Social Networks”. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*”. 2018. p. 384-389, https://www.scopus.com/authid/detail.uri?authorId=56256597400. DOI: https://doi.org/10.1109/MIPR.2018.00084.

26. Lindemann, B., Müller, T., Vietz, H., Jazdi, N. & Weyrich, M. “A survey on long short-term memory networks for time series prediction”, *Procedia CIRP.* 2021; 99: 650–655, https://www.scopus.com/authid/detail.uri?authorId=57201983372.
DOI: https://doi.org/10.1016/j.procir.2021.03.088.

27. Zoican, S., Zoican, R. & Galatchi, D., “Terrestrial Traffic Forecasting using Graph-based Neural Networks”. *2023 16th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, Nis: Serbia. 2023. p. 143–146, https://www.scopus.com/authid/detail.uri? authorId=6507127464. DOI: https://doi.org/10.1109/TELSIKS57806.2023.10315720.

28. Daniel Davis & Chia-Chu Chiang. “Natural Language Processing for Detecting Undefined Values in Specifications”. 17th Annual System of Systems Engineering Conference (SOSE), Rochester, NY, USA, 2022. p. 191-196, https://www.scopus.com/authid/detail.uri?authorId=58429978200.

DOI: https://doi.org/10.1109/SOSE55472.2022.9812647.

29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. "Going deeper with convolutions". *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA: USA 2015. p. 1–9, https://www.scopus.com/authid/detail.uri?authorId=56122593000. DOI: https://doi.org/10.1109/CVPR.2015.7298594.

30. Kovalchuk, M. L., Lucin, P., Gorsky, M. P & Soltys, I. V. "Design and creation of an information system for analytical data processing". *Optical-electronic information and energy technologies*. 2022; 2: 26–31, https://www.scopus.com/authid/detail.uri?authorId=36158034500. DOI: https://doi.org/10.31649/1681-7893-2022-44-2-26-31.

Conflicts of Interest: the authors declare no conflict of interest

Received 12.03.2024
Received after revision 07.05.2024
Accepted 15.05.2024

DOI: https://doi.org/10.15276/hait.07.2024.14
UDC 004.9

Інформаційна система аналізу громадського настрою у веб-платформах на основі машинного навчання

Угрин Дмитро Ілліч[1)]
ORCID: https://orcid.org/0000-0003-4858-4511; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300
Карачевцев Артем Олегович[1)]
ORCID: https://orcid.org/0009-0000-6226-6822; a.karachevtsev@chnu.edu.ua. Scopus Author ID: 36925155800
Томка Юрій Ярославович[1)]
ORCID: https://orcid.org/0000-0002-0495-3090; y.tomka@chnu.edu.ua. Scopus Author ID: 9279702200
Захаров Микита Миколайович[1)]
ORCID: https://orcid.org/0009-0003-5026-3546; jake2000nik@gmail.com
Трояновська Юлія Людвигівна[2)]
ORCID: https://orcid.org/0000-0002-6716-9391; troyanovskaja@gmail.com. Scopus Author ID: 57211747293
[1)] Чернівецький національний університет ім. Ю. Федьковича, Коцюбинського, 2, 58002. Чернівці, Україна
[2)] Національний університет «Одеська політехніка», проспект Шевченка, 1. Одеса, Україна

АНОТАЦІЯ

Проведено аналіз систем для вивчення громадського настрою у веб-платформах. Описано різні засоби та методи для ефективного визначення настрою у текстових даних з веб-платформ, включаючи формалізацію соціального графу та графу контенту. Досліджено процес класифікації коментарів, що включає систематизацію та розподіл висловлювань на категорії. На основі дослідженого набору даних відібрана інформація про відгуки від клієнтів та оцінки готелів у Європі з веб-платформи booking.com. З урахуванням вимог інформаційної системи та результатів аналізу визначено, що для отримання кращих результатів у визначенні емоційного відтінку текстів відгуків та повідомлень від користувачів найбільш відповідним є застосування методів машинного навчання, враховуючи методи природної мови для обробки текстових даних. У виборі методу векторизації тексту для машинного навчання обрано Term Frequency Inverse Document Frequency Vectorizer як більш ефективного серед досліджених методів. Запропонована архітектурна структура досліджуваної системи, що спрямована на ефективну взаємодію між компонентами та модулями. Обрано модель LogisticRegression для визначення громадського настрою. Розроблена інформаційна система, що аналізує громадський настрій щодо об'єктів, використовує передові технології машинного навчання для оцінки емоційного відтінку текстових коментарів і забезпечує користувачам інсайти та аналіз результатів.

Ключові слова: веб-платформа; інформаційна система; громадський настрій; пропаганда; дезінформація; фейк; повідомлення; текст; інтелектуальний аналіз даних; штучний інтелект; машинне навчання

ABOUT THE AUTHORS

Dmytro I. Uhryn - Doctor of Engineering Sciences, Associate professor, Computer Science Department.
Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubynsky Str. Chernivtsi, 58002, Ukraine
ORCID: https://orcid.org/0000-0003-4858-4511; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300

***Research field*:** Information technologies for decision support; swarm intelligence systems; branch geoinformation systems

Угрин Дмитро Ілліч - доктор технічних наук, доцент кафедри Комп'ютерних наук. Чернівецький національний університет ім. Ю. Федьковича. Коцюбинського, 2. Чернівці, 58002, Україна

Artem O. Karachevtsev - PhD (Physical and Mathematical Sciences), Assistant Professor, Computer Science Department. Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubynsky Str. Chernivtsi, 58002, Ukraine.
ORCID: https://orcid.org/0009-0000-6226-6822; a.karachevtsev@chnu.edu.ua. Scopus Author ID: 36925155800

***Research field*:** Computer science; software engineering; web development; biomedical optics

Карачевцев Артем Олегович - кандидат фізико-математичних наук, асистент кафедри Комп'ютерних наук. Чернівецький національний університет ім. Ю. Федьковича, Коцюбинського, 2. Чернівці, 58002, Україна.

Yurii Ya. Tomka - PhD (Physical and Mathematical Sciences), Associate Professor, Computer Science Department.
Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubynsky Str. Chernivtsi, 58002, Ukraine.
Scopus Author ID: 9279702200;
ORCID: https://orcid.org/0000-0002-0495-3090; y.tomka@chnu.edu.ua.

Research field: Software engineering; computer science; digital image processing; deep learning; laser polarimetry

Томка Юрій Ярославович - кандидат фізико-математичних наук, доцент кафедри комп'ютерних наук, Чернівецький національний університет ім. Ю. Федьковича, Коцюбинського, 2. Чернівці, 58002, Україна.

Zakharov Mykyta Mykolaiovych - PhD student, Computer Science Department. Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubynsky Str. Chernivtsi, 58002, Ukraine.
ORCID: https://orcid.org/0009-0003-5026-3546; jake2000nik@gmail.com

***Research field*:** Computer Science; machine learning

Захаров Микита Миколайович - аспірант кафедри комп'ютерних наук, Чернівецький національний університет ім. Ю. Федьковича, Коцюбинського, 2. Чернівці, 58002, Україна.

Yuliia L. Troianovska - Senior Lecturer, Information System Department. Odessa Polytechnic National University. 1, Shevchenko Ave. Odessa, Ukraine
ORCID: https://orcid.org/0000-0002-6716-9391; troyanovskaja@opu.ua. Scopus Author ID: 57211747293

***Research field*:** Geographic information systems; machine learning; neural networks

Трояновська Юлія Людвигівна - старший викладач кафедри Інформаційних систем. Національний університет «Одеська політехніка», проспект Шевченка, 1. Одеса, 65044, Україна