ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

UDC 004.9

REALIZATION OF THE DECISION-MAKING SUPPORT SYSTEM FOR TWITTER USERS' PUBLICATIONS ANALYSIS

Batiuk T. – Postgraduate student of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

Dosyn D. – Dr. Sc., Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. The paper emphasizes the need for a decision-making system that can analyze users' messages and determine the sentiment to understand how news and events impact people's emotions. Such a system would employ advanced techniques to analyze users' messages, delving into the sentiment expressed within the text. The primary goal is to gain insights into how news and various events reverberate through people's emotions.

Objective. The objective is to create a decision-making system that can analyze and determine the sentiment of user messages, understand the emotional response to news and events, and distribute the data into clusters to gain a broader understanding of users' opinions. This multifaceted objective involves the integration of advanced techniques in natural language processing and machine learning to build a robust decision-making system. The primary goals are sentiment analysis, comprehension of emotional responses to news and events, and data clustering for a holistic view of user opinions.

Method. The use of long-short-term memory neural networks for sentiment analysis and the *k*-means algorithm for data clustering is proposed for processing large volumes of user data. This strategic combination aims to tackle the challenges posed by processing large volumes of user-generated data in a more nuanced and insightful manner.

Results. The study and conceptual design of the decision-making system have been completed and the decision-making system was created. The system incorporates sentiment analysis and data clustering to understand users' opinions and the sentiment value of such opinions dividing them into clusters and visualizing the findings.

Conclusions. The conclusion is that the development of a decision-making system capable of analyzing user sentiment and clustering data can provide valuable insights into users' reactions to news and events in social networks. The proposed use of long-short-term memory neural networks and the *k*-means algorithm is considered suitable for sentiment analysis and data clustering tasks. The importance of studying existing works and systems to understand available algorithms and their applications is emphasized. The article also describes created and implemented a decision-making system and demonstrated the functionality of the system using a sample dataset.

KEYWORDS: natural language processing, convolutional neural network, recurrent neural network, LSTM, *k*-means clustering.

ABBREVIATIONS

NLP is a natural language processing;
CNN is a convolutional neural network;
RNN is a recurrent neural network;
LSTM is a long-short-term memory;
HC is a hierarchical clustering;
KMC is a *k*-means clustering;
DB is a database;
DMS is a decision-making system;
SDS is a sample data set.

NOMENCLATURE

S is a decision-making analysis system;
I is a set of inputs;
O is a set of outputs;
R are the basic rules for processing the flow of input data to the sentiment analysis system;
F is an input data processing parameters;
N is a recurrent neural network;
α is an input validation operator;
β is an input data processing operator;
γ is a the search operator for relevant users after clustering;
P is an improved recurrent model of searching for validated users;
μ is an user authentication operator;
χ is an input data set formation operator;
ω is an operator of list formation and sentiment analysis data;
λ is a validation request resolution operator;
i_1 is a set of authentication data (login, password, set of distributed data);
i_2 is a data storage of publications of a specific social network;
i_3 is a different sets of user data;
i_4 is a specific user request;
o_1 is a clustering requests using the elbow method;
o_2 is a set of updates for the user's profile in the selected social network;

DOI 10.15588/1607-3274-2024-1-16

o_3 is a saving at the request of the DMS user;

r_1 is a set of rules of the data saving algorithm;

r_2 is an operating rules of recurrent neural network;

r_3 is a set of rules of operations of a convolutional neural network;

r_4 is a set of hierarchical clustering rules;

r_5 is a set of k-means clustering rules;

u_1 is a set of levels of data processing;

u_2 is a set of data processing requirements;

u_3 is a set of text validation requirements;

u_4 is a set of multiple levels of creating a linked list of relevant users;

u_5 is a set of LSTM data analysis requirements;

X_{UF} is a the result of an authenticated user.

INTRODUCTION

Creating and implementing an intelligent system for sentiment analysis and clustering publications is a relevant and promising task these days, most communication between people occurs in social networks according to certain situations or circumstances. Each message of a social network user has particular semantics, reflects certain thoughts and analysis of the relevant situation, or is a reaction to a specific event. Modern algorithms and approaches to data analysis make it possible to effectively and relatively analyze large volumes of text data, thanks to which it is possible to determine the average tone of users' reactions to certain events and, as a result, conclude the analyzed content. In this way, it is possible to understand the attitude of different groups of users to a certain type of content, offers, goods, and other market offers, and large brands and corporations actively use these approaches to the analysis of information by determining the tone of the user text and further dividing this messages and, accordingly, the users themselves into clusters for further work with the resulting clusters.

To implement this kind of algorithms, first of all, it is necessary to determine precisely which group of users we want to research, that is, it is essential to have certain keywords that will be used for searching, geolocation, setting tags, etc. This is the starting point of the current research, next it is necessary to load the desired dataset, that is, a pool of user messages stored according to a given predicate. After receiving the file with the saved messages, it is necessary to perform a general check of the file for the correctness of the data and, after that, to digitally structure the data in such a way that it is possible to operate with the data as accurately as possible, that is, they need to be formatted and brought to a single structure. After that, it is necessary to analyze user messages. The general algorithm can be divided into two main sub-algorithms: it is an analysis of the sentiment of user messages, according to a confident approach, where each message has its own sentiment level and rating, and also clustering of messages to divide users and their messages into a number determined during the operation of the intelligent system clusters. It is most convenient to show the obtained results in the form of graphs and charts for visualization and understanding of the general situation in the results of the analysis of user publications. These algorithms are not new and contain vast possibilities for modification and optimization, which will be carried out in further work.

1 PROBLEM STATEMENT

The system of sentiment analysis of user publications S is represented by a tuple simulation model:

$$S = < R, I, F, O, N, \alpha, \beta, \gamma >,$$

where $R = \{r_1, r_2, r_3, r_4\}$, $F = \{u_1, u_2, u_3, u_4, u_5\}$, $I = \{i_1, i_2, i_3, i_4\}$, $O = \{o_1, o_2, o_3\}$.

The main processes of DMS users' publication analysis are "User authentication", "Lemmatization of user data", "Performing sentiment analysis" and "Performing k-means clustering".

The DMS users' publication analysis user authentication process will be described by superposition:

$$X_{UF} = \alpha \circ \beta \circ \mu,$$
$$X_{UF} = \alpha\ (\beta(\mu\ (i_1, i_3, i_4), r_5, u_5), u_4).$$

The process of forming lemmatized user data of DMS users' publication analysis will be described by superposition: $C_{CU} = \alpha \circ \beta \circ \chi$, so

$$X_{UC} = \alpha\ (\beta(\chi(X_{UF}, i_2, i_3, i_4), r_1, u_3), r_2).$$

The process of performing sentiment analysis will be described by superposition:

$$X_{UK} = \omega \circ \gamma \circ \beta \circ \alpha,$$
$$X_{UK} = \beta\ (\gamma(\omega(\alpha(X_{UC}, i_4), i_2), u_5), r_3).$$

The process of performing k-means clustering will be described by superposition:

$$X_{UM} = \lambda \circ \gamma \circ \beta \circ \alpha,$$
$$X_{UM} = \lambda(\gamma(\beta(\alpha(X_{UK}, i_4), i_3), u_3), r_5).$$

2 REVIEW OF THE LITERATURE

Neural networks have become an indispensable part of the work of various companies and corporations. In the article [1], the authors investigate the exact role of deep learning in e-commerce and the principles of working with users using the example of online store networks. The authors investigated the concept of user reviews of certain products of different quality and the impact of current reviews on the sale of products by other users of the system in the future, investigated when precisely and under what conditions users pay the most or least attention to product reviews and, accordingly, how exactly positive feedback is composed whether negative feedback can affect the purchase attractiveness of the product. In the article [2], the authors investigated user comments under videos on the YouTube social network, a dataset with videos about the COVID-19 virus and user comments was selected, as the sentiment of written messages, their frequency, and the activity of writing by users were investigated, accordingly, it was assumed that a part of users with negative comments is bots, due to

DOI 10.15588/1607-3274-2024-1-16

very similar message patterns and almost the same level of negative tone. On the contrary, the authors in the article [3] examined the comments of famous personalities on social networks. They found many comments written at certain selected moments in time and had almost the same positive tone in the range from 0 to 1, which also indicated the unnatural state of these messages.

Also, the authors in the article [4] considered several types of user personification models in the Twitter social network using LSTM-neural networks, several critical parameters for each user were taken from the open API and considered as a separate dataset with the subsequent division of users into groups according to their location, description, and profile avatar, and a corresponding neural network was trained to assign users to certain groups. A system was also created [5] that processes author citations in articles and provides an opportunity to analyze the correctness and correctness of the current text and correct it using a recurrent neural network, the main task of which is to process text data and analyze subsequent data based on training with the teacher. Also, in the article [6], the authors considered machine learning based on customer feedback on hotels to understand the situation in this market and further develop a hotel development plan based on positive and negative customer feedback.

3 MATERIALS AND METHODS

The purpose of this work is the implementation of an intelligent system of sentiment analysis and clustering of publications in the Twitter social network. The idea of analyzing the sentiment of user messages or publications in social networks is familiar because several practically implemented systems perform a similar task. At the moment, an important task is the optimization and the most effective use of already existing technologies and the correct selection of models and algorithms to perform a particular task, which may depend both on the size of the input dataset and on the size of individual text tokens [7] within the dataset, or even parameters search of text information for users using a particular system, or only its partial functionality for text research.

The process of analyzing text publications or user comments can be divided into two essential parts: the analysis of the text's sentiment and the clustering of text data. This task is non-trivial and quite complex since many parameters must be considered before creating such an intelligent system for analyzing textual data: the size of the sample, the textual data, and the context to which these data belong. These users also write posts or messages to react to a particular event or set of events happening at a certain time. This all means that it is impossible to create a unique intelligent system that can cover all cases and analyze text data with approximately the same efficiency. One way or another, there are tasks that each specific system can perform efficiently and with maximum accuracy. Still, there are also tasks for which the same set of algorithms cannot function efficiently or accurately. Thus, convolutional neural networks are often used to analyze the text's sentiment, and hierarchical clustering is used for cluster analysis. These algorithms are efficient and time-tested but can only perform analysis on small to medium-sized datasets or data samples. These algorithms can be used without problems for large volumes of data. Still, such research will be inefficient and less accurate, especially if the number of text data units is large and the tokenized object is small [8]. A clear example is the social network Twitter, where one user message can have a maximum of 280 characters to create a post. In order to analyze the sentiment of messages and their clustering as effectively as possible, we will use the LSTM neural network and the *k*-means clustering algorithm, thanks to which we can achieve 10...15% greater efficiency compared to convolutional neural networks and hierarchical clustering. Before describing the functionality of the system and the main algorithms of the long-short-term memory neural network and the *k*-means clustering algorithm, it is worth paying attention to unsolved problems, namely, why convolutional neural networks specifically in the context of user posts and comments on the Twitter social network and conventional hierarchical clustering are inefficient and may show unexpected results in the results of text sentiment analysis and subsequent clustering [9], namely due to nuances in the implementation of these algorithms for working with text and due to the properties of the weights provided by the convolutional neural network, which are important to consider, since the generated model may be invalid if the weights are miscalculated.

A convolutional neural network, or CNN, can be thought of as a set of matrices that make up one large matrix, in which both horizontal and vertical sets of elements, formed in the process of learning a neural network to build a model, can be chosen as vectors for training. In the course of the work, a collection of words is superimposed on each other with the help of vectors since each word is represented by a separate vector of letters that together form a particular image. With the help of the created system predicates, the generated image of vectors representing a set of text data or words is filtered, since the predicate has the same width and length as the value check, it is possible to analyze only part of the vector, therefore, for the efficiency of work, vectors are most often put into temporary matrices for correct and more effective filtering using a given predicate[10]. Since this is a natural language processing task, the filters we employ in the work process have the same width as the length of the investigated text element or a single word. The height, on the other hand, is more static and can usually change its size from 0 to 5 since it is necessary to form certain sequences of values from the received text information, which will later be written into lists of values, accordingly, it is required to limit the number of lists to 5 pieces to be able to carry out effective parallel processing of text information and model training with given limitations of the current neural network [11]. Due to these algorithmic limitations, the main problem of the impossibility of accurate learning and subsequent text analysis using convolutional neural networks arises. The

DOI 10.15588/1607-3274-2024-1-16

text with restrictions will be analyzed, and the result of the analysis of the sentiment of the text will be correct. Still, due to the height restriction of the vector, the neural network can only be trained on posts and messages with polar sentiment values, i.e., –1, 0, or 1 [12], respectively clearly negative, neutral or positive comments without the possibility of their distribution over a certain range, which, on the one hand, is not critical and can be helpful in a general understanding of the sentiment of messages. Still, on the other hand, with the help of such a convolutional neural network, it is impossible to carry out an accurate sentiment analysis. There are also reservations about the speed of operation because such a neural network has low speed and efficiency since the memory algorithm of repeated data is used to save the polar values of the vectors.

If we talk about hierarchical clustering, then it also has certain disadvantages in its use. The biggest drawback is similar to what was described earlier about the convolutional neural network, namely, the inability to perform complex tasks, as well as the problem of low speed and efficiency of such a clustering algorithm. To begin with, it is worth noting that hierarchical clustering is formed based on a tree graph, also known as a dendrogram. Accordingly, when constructing this tree graph, we use an agglomerative approach to work with input data [13]. Having a dendrogram and applying an agglomerative approach, we observe relatively monotonous clusters and simultaneously look for possible connections between them, in the second step, we successively combine clusters into a separate “connection” based on the predicates we need. The main advantage is simplicity and clarity of use, that is, the presence of a convenient tree-like structure, in which clusters can often be seen “by eye” with the correct generation of the model [14]. Unfortunately, the downside of this simplicity is that it can usually only be used for small amounts of data and small datasets. One of the main disadvantages is the complexity of the tree hierarchy, which is clustered in time with an algorithmic complexity of $O(n^2 \log n)$, where n is the number of total data points.

If, on the contrary, we take the *k*-means algorithm, in it, we will use the optimization of a specific objective function, for example, within a certain range of values from k to l, that is, we do not actually have an objective function, and therefore we will have a much more efficient implementation due to the complexity of the algorithm $O(nKm)$, where K is the number of clusters, and m is the number of average values [15]. Also, the problem of hierarchical clustering is a certain staticity, namely the impossibility of undoing the previous steps of the algorithm, that is if we cluster n–1 points. Then it turns out that the connection between the clusters was incorrect, or there were problems creating a tree graph on one of the hierarchy levels. We cannot cancel this step at the program level during execution because either the final dendrogram with distorted values will be obtained, or it will be necessary to stop the program and start the system from the beginning, which is also a big minus compared to the work the *k*-means algorithm, where it is possible to check the correctness of average values using a specified execution condition [16]. Based on the above information, it can be concluded that convolutional neural networks and hierarchical clustering are suboptimal for analyzing the sentiment of user posts and messages in the Twitter social network. Instead, to optimize and speed up the operation of the intelligent system, it is more appropriate choice to use a neural network based on LSTM architecture and the modern *k*-means clustering algorithm.

4 EXPERIMENTS

Speaking about the functionality of the created intelligent system for the analysis of sentiment and clustering of posts in the Twitter social network, first of all, it is worth paying attention to two main algorithms, which will perform the main volume of work in the middle of the system, this is an LSTM neural network, with the help of which an effective and fast analysis of the sentiment of posts and comments by Twitter users and the *k*-means clustering algorithm, which will be used to select the main clusters and distribute the user text according to these clusters. To begin with, it is worth revealing the essence of the neural network that was built.

One of the most popular and effective neural network models focused on processing time series is the Long Short-Term Memory (LTSM) model [17]. It is efficient to use and much more accurate than the previously mentioned convolutional neural network, but at the same time, it is challenging to implement because it is essentially a recurrent neural network whose primary function is to predict a sequence of data from an input dataset and the corresponding problem that must be resolved. For a better understanding of the context of a neural network application, it is worth formulating what is meant by time series analysis for which we are building a neural network. The point is that the input data represent particular points of information that are analyzed in certain time intervals. Thanks to this, it is possible to create and analyze patterns of processes that occur in certain time intervals. Since we have certain related data points [18], we use a Recurrent Neural Network (RNN).

The Long Short-Term Memory neural network, or LSTM, is a separate and special case of recurrent neural networks because this model is even more efficient. After all, it can store a collection of information for extended periods. In the LSTM neural network, two main recurrent problems are overcome: gradients that disappear and gradients that are deformed during the operation of the intelligent system. An LSTM model consists of a memory cell state and three main passes. The memory cell’s state stores the cell’s value for a certain period of time and is a kind of tape that moves and linearly transmits data further along the conveyor with practically no additional deformations. In the LSTM network model, it is possible to add, modify and delete information using the three passes mentioned earlier. Passages help regulate information and are key in the LSTM architecture of a neural network, as thanks to the state of the memory cell,

DOI 10.15588/1607-3274-2024-1-16

the data stream is formed into a particular linear structure and allows for uniform distribution of memory throughout the neural network's operation time in the intelligent system. A neural network is in 3 main states [19]: either data is input or output, or it is forgotten due to distortion or unnecessary information.

Accordingly, the work of the neural network is implemented using the previously mentioned three passes. The first is the "forgetting" pass, which is responsible for removing information that, due to distortion or use, is no longer needed to analyze the text's sentiment so that it can be removed and make room for the following information at the input. Thanks to this, the model becomes more efficient at each neural network step. This pass has two main inputs: the hidden state of the previous memory cell and the current input at this step. These data are multiplied with previously created weight matrices, after which a certain displacement factor is added. Next, a sigmoid function is applied, which produces a result from 0 to 1, thanks to which the neural network "knows" which information can be "forgotten" and which data can be passed on. If we have a value of 0, the information about this state of the memory cell can be deleted, if the value is close to 1, then all the information about the cell must be saved and passed on. The vector output of the sigmoid function, obtained as a result of working out part of the neural network model, must be multiplied by the state of the memory cells that were not deleted during the work, and the result passed on [20], as a result of which the first pass is formed, which performs the vital function of removing all unnecessary for analysis of the sentiment of the text of the memory cells. Figure 1 depicts the main aspects of the LSTM neural network architecture.

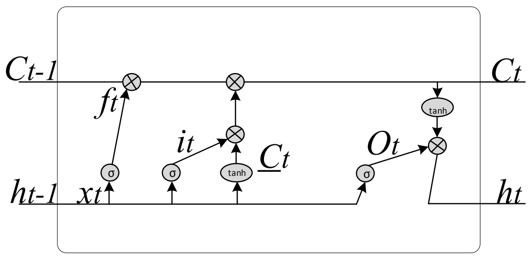

Figure 1 – Architecture of Long Short-Term Memory neural network

The second is the "entrance" passage. This pass is used to add information to the state of the memory cell. Initially, the allocated values to be added to the cell are adjusted using a sigmoid function, the inputs are still the hidden state of the previous memory cell and the current input of the algorithm step. During operation, a vector is created that contains all or almost all possible values that must be added to the state of the memory cell. This happens using the tanh function, so the tangent outputs values from –1 to 1. Next, you need to determine [21] the value of the sigmoid function, a regulatory function. The available values of the regulatory function must be multiplied by the previously created vector of values; all vital information for the intelligent system to analyze the sentiment of a certain text is added to the state of the memory cell using the data addition operation, thanks to which you can be sure that only valid and necessary information, which was previously filtered and checked, was added through the "input" pass during network operation.

Last is the "output" pass, which is used to use the current information that is available at a given point in time and display the most relevant results. First, a separate vector is created after applying the tanh function to the state of the memory cell, where the output value ranges from –1 to 1. The same sigmoid function again acts as a regulatory function, that is, it is used to regulate precisely those values that need to be output from the vector by using the two previously mentioned inputs, namely the hidden state of the previous memory cell and the current input of information at this step. The value of the sigmoid function must be multiplied by a vector, and the obtained result of the operation is used as the output value. Also, the neural network sends the result to the hidden state of the following memory cell, which, accordingly, is a modern solution, thanks to which the LSTM neural network [22] is the most efficient and convenient in predicting sequences and performs exceedingly well in the task of analyzing the sentiment of posts on the Twitter social network.

It is also worth clarifying the importance of the sigmoid activation function during the "forgetting" pass operation because instead of distributing the values between –1 and 1, the input values are distributed between 0 and 1. This helps to update the changed data in time or "forget" the distorted ones. that is, those that will no longer be of any use in analyzing the text's sentiment. The division is made precisely between 0 and 1 through mathematical multiplication since any number we multiply by 0 will result in 0, and vice versa, any number multiplied by one will remain unchanged. Thanks to this, the sigmoid function helps to effectively determine which data should be "forgotten" or deleted, which should be updated to the current pitch value, and which should be kept and passed on to the next "input" pass. So, Figure 2 shows the step-by-step algorithm of the Long Short-Term Memory neural network.

In addition to implementing the LSTM neural network, an important part of the intelligent system for sentiment analysis and clustering of posts in the Twitter social network is the correct implementation of clustering using the *k*-means algorithm. As already described earlier, the usual hierarchical clustering, although it is pretty popular and has certain clearly expressed advantages, is still not suitable for our task, namely for processing the exact values of the text's sentiment and, accordingly, the selection of the necessary clusters. To begin with, it is worth clarifying that clustering is the process of dividing a volume of data of a certain size into several clearly defined groups similar in structure in such a way that the values of data points in one group are more similar to the importance of other data points in the same group than other points data that were assigned to other groups. It is

DOI 10.15588/1607-3274-2024-1-16

also worth noting that clustering is a learning algorithm without a teacher due to its work features.

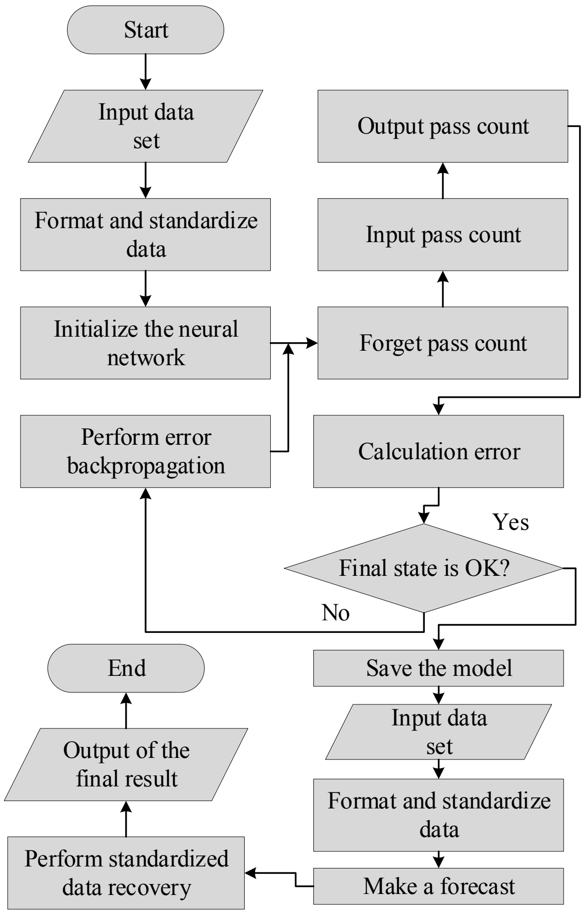

Figure 2 – LSTM neural network algorithm

k-means clustering is an algorithm that does not need to mark the input data in the learning process, unlike learning algorithms with a teacher. *k*-means divides objects into clusters in such a way that all objects within a cluster are similar to each other and dissimilar to objects in different clusters. The Latin letter k stands for a number that represents the number of clusters to be created. Also, the essence of the algorithm is to find the best or optimal value of the clusters to be used for better work with sentiment analysis. The *k*-means clustering procedure itself is a relatively simple and straightforward mathematical problem. To begin with, a general notation should be defined. For example, we have from C_1 to C_k sets that have indices of observations in each cluster, these sets satisfy two main properties: firstly, each statement belongs to at least one of the k existing clusters, and secondly, the clusters cannot overlap each other, that is, one set of observations can belong to only one unique cluster. The basis of *k*-means clustering is the idea that the best clustering is the one where the variation of values within the cluster is minimal, that is, the variation within a particular cluster C_k is a measure of the magnitude of the cluster observation, where one cluster differs from another, this is the problem that is solved *k*-means clustering within an intelligent system.

It is worth noting the importance of the parameter k, which determines the number of clusters, a certain approximate value can often determine it simply by estimating the size of the dataset and the data it contains. For our task of analyzing the text's sentiment, the "elbow" method was chosen, allowing you to more accurately determine the number of clusters required for work by running the *k*-means algorithm with a different set of clusters to determine the optimal value empirically. The "elbow" method involves finding a certain metric to estimate how good the clustering result is for different k values by finding the "elbow" point to separate all unnecessary further values and choose the optimal one up to the specified point. A sharp drop in values in the corresponding graph means that the clustering value is being optimized, but there is a point where the sharp drop in values stops falling and stabilizes. This is the same "elbow" point. That is, all values after the "elbow" point should be discarded, and only those where there is a decline in the average sum of squares of the observation values within each cluster, where the fall is the largest – this point will most likely represent the optimal number of clusters.

The clustering algorithm is essentially a method of dividing the observation indices in each cluster so that the goal of the last equation of selecting *k* clusters is minimized. The problem is quite tricky because there are k^n ways to partition n observations into clusters, but there is an algorithm that can be used to find a local optimum for the *k*-means optimization problem. The *k*-means clustering algorithm itself consists of two global steps. To begin with, it is necessary to choose a random value from 1 to k for each of the data observations that were separately selected from the dataset, they represent certain initial values for the clusters. This procedure should be repeated until the values of the clusters stop changing. For each of the k clusters, it is necessary to calculate the cluster's centroid. The resulting centroid of the *k*-th cluster is a separate vector of the average values of the observation indices in the *k*-th cluster. Each cluster needs to be assigned an observation identifier whose centroid is the closest, it is necessary to make sure that all clusters are stable and do not contain uncertainties, as this can prevent the correct distribution of observation indices relative to clusters. The *k*-means clustering algorithm is shown in Figure 3.

Having outlined the logic and functionality of the LSTM neural network and the *k*-means clustering algorithm, it is clear how sentiment analysis and clustering of posts and messages of users of the Twitter social network will take place within the system. In addition, it is necessary to understand the logic of the intelligent system as a whole. Neural network and clustering algorithms perform only a certain function in the system. The system itself consists of many processes, which is not surprising since we interact with a genuine

DOI 10.15588/1607-3274-2024-1-16

user and push back from the parameters set by the user at a certain time when this user works with the system.

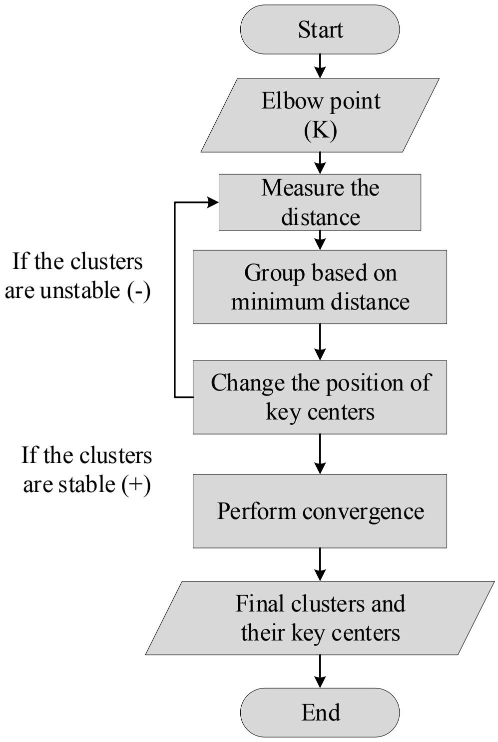

Figure 3 – The *k*-means clustering algorithm

The use case diagram is the most convenient diagram for displaying the general concept of user interaction, the system, and other elements of the overall structure because it mainly consists of actors and their use cases. In this intelligent system of sentiment analysis and clustering of textual data, there is a single actor, the "User", who is a separate entity in relation to the system and is depicted separately from other usage options, representing the main aspects of the system's operation. The diagram of use cases itself is shown in Figure 4.

In the diagram, the intelligent system is depicted inside a rectangle, it consists of several prominent use cases responsible for saving tweets, that is, posts and comments of users of the social network Twitter, and there are also use cases that are responsible for formatting text messages, data to the desired form, performing LSTM analysis of the sentiment of text data, performing cluster analysis, outputting all the necessary received results in the format that is most suitable according to the received data and ending the system. In addition to the usual use cases that perform the main functions of the intelligent system, the diagram shows inclusion and extension options that help to explain in more detail the essence of the operation of the prominent use cases. For example, the use case that describes saving tweets contains inclusion options that describe setting parameters in a particular format and saving the formed dataset in a .csv file. There is also an extension option that illustrates the process of selecting the main search keywords. The use case that describes the formatting of text data only has extension options that describe removing redundant characters from text, standardizing existing data, and removing text processors that were loaded by default. Also, another use case that is responsible for LSTM sentiment analysis of text data only has options for expanding the possibilities, such as performing the calculation of all current network layers and memory cell state and three main passes, performing sentiment prediction of text information, i.e., posts and comments of a Twitter user and final saving of the created model. Then there is the option of using cluster analysis, including options for expansion and inclusion, such as finding the "elbow" point, determining the final clusters and their key centers, and implementing convergence. It is also worth noting the last two options of use, namely the results' output and the intelligent system's termination. The output of the results has only options to include, namely

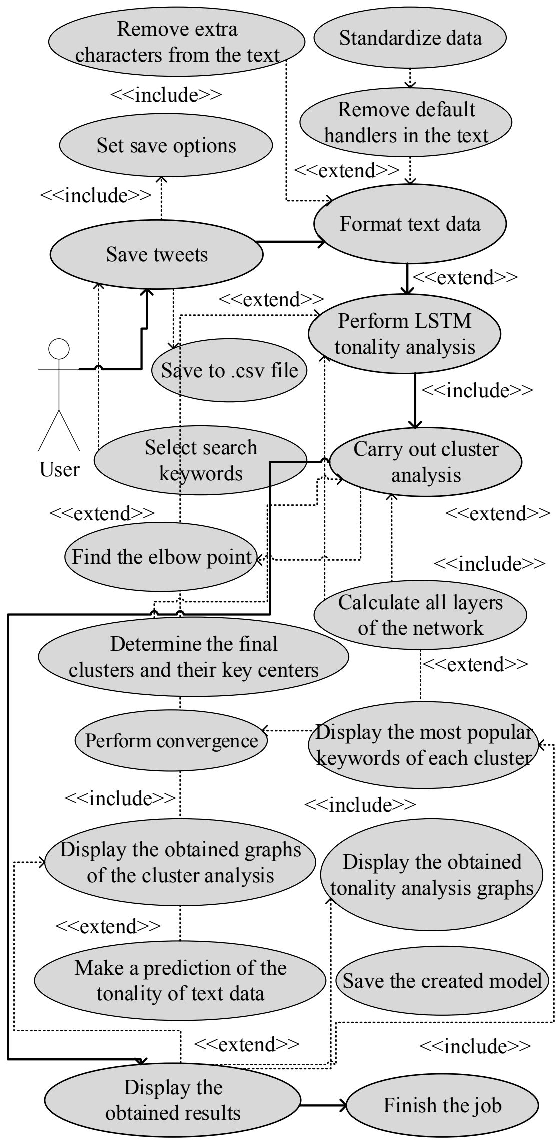

Figure 4 – Diagram of options for using the intelligent system of sentiment analysis and clustering of publications

DOI 10.15588/1607-3274-2024-1-16

the output of the tone of user posts and comments and the output of the cluster analysis in the form of textual information and using charts for a more detailed explanation of the obtained data according to the search keywords.

Having described the general structure of the intelligent system with the help of a diagram of use cases, it is necessary to more precisely describe the design of the created system and its functionality, for this the activity diagram is ideal, with its help, you can highlight the entities of the system with the help of tracks, the conditions for performing calculations and calculations, clearly show the state of operation and its interaction with other states within the initial and final state of the intelligent system of sentiment analysis and clustering of posts in the Twitter social network. It is also convenient to display all existing branching of flows during work and their results. To begin with, it is worth highlighting two main entities, namely "User" and "Server", on the activity diagram they are presented in the form of separate tracks, accordingly, it is possible to describe their context not only in space but also in time, that is, to understand what processes in the intelligent system will occur at a certain point in time and how they will interact with each other. The first is the initial state, which describes the functional start of the system, then the states of actions belonging to the user of the system interact sequentially with each other, i.e. entering save parameters, setting keywords and other actions, or states of actions that are completely encapsulated within the logic of a separate instance class of task execution. The constructed activity diagram is shown in Figure 5.

Each state of action can be considered independently of others. Also, the "Server" entity in the work process checks the condition for the admissibility of saving the created model or the need to implement the process of reverse propagation of the error. In addition to conditions, the "Server" entity contains parallelization of action states, where at the same time, during the execution of two different predicates, the change of crucial centers can be divided into both the measurement of distances between clusters and the implementation of convergence with already existing clusters as a result of the calculation of the centroid of the cluster containing divided indices observation in an intelligent system. So, after describing the functionality of the system using object-oriented diagrams, for a better understanding of the context of the system's work, it is necessary to describe its main elements and relationships using a functional diagram, the data flow diagram is the most convenient for performing this task, as it additionally shows the system, as a set of processes that interact with each other during the entire life cycle of the system. The data flow diagram of the intelligent system is shown in Figure 6.

The diagram consists of 7 consecutive data flows, where each flows into the next and transmits certain information within the request. Each thread fulfills all the requirements for transactions, that is, the intelligent system transmits data without the risk of its loss: if the transaction does not take place or information is lost or distorted in the process, the transaction will simply stop the current thread and return control to the previous thread, which generally ended without distorting existing user information. Also, under the number 1 and the name R on the diagram are data stores, in our case, it is a text file with the extension .csv for saving the dataset and a server for dynamically saving data between requests. Based on the completed description of the main essence of the system being created, its functionality, the LSTM neural network implementation, and the *k*-means clustering algorithm, an experiment was conducted to create a complete dataset of publications and comments on the social network Twitter, according to the keywords specified by the user, and an analysis of sentiment was carried out individual messages and corresponding division into final clusters.

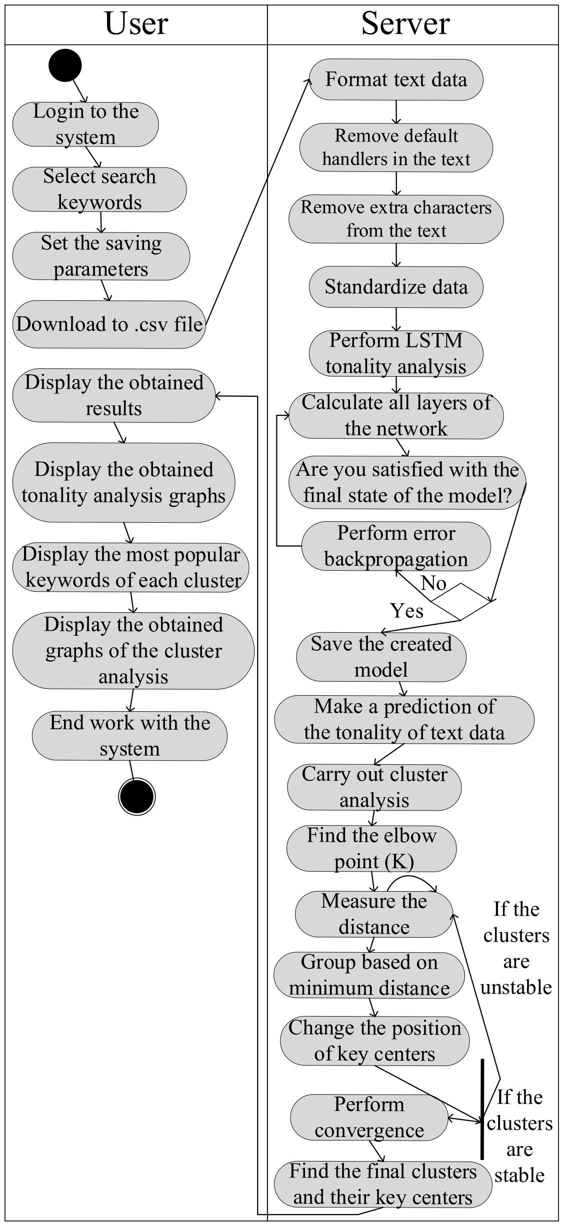

Figure 5 – Activity diagram of the intellectual system of sentiment analysis and publication clustering

DOI 10.15588/1607-3274-2024-1-16

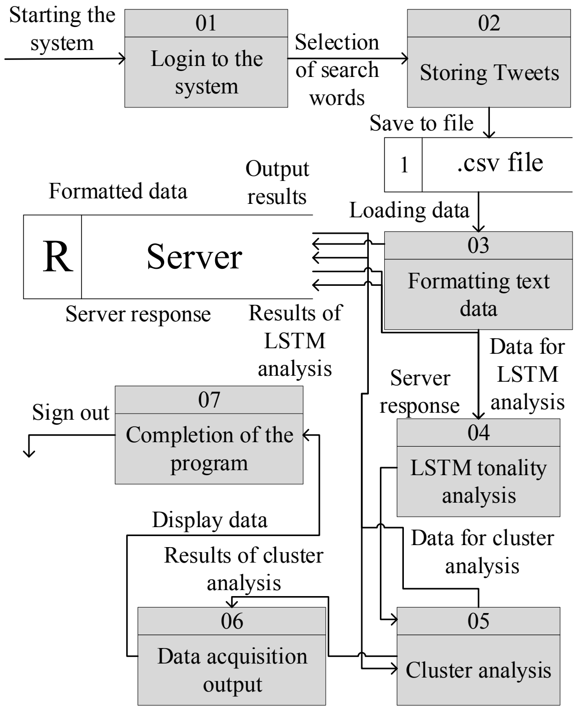

Figure 6 – Diagram of data flows of the intelligent system of sentiment analysis and clustering of publications

5 RESULTS

An experiment was conducted on applying the implemented intelligent system of sentiment analysis and clustering of posts in the social network Twitter, in which the work of the created LSTM neural network and the *k*-means clustering algorithm was tested. First, you need to find data to analyze. For maximum relevance, it was decided to make a separate dataset and download all user comments and posts on the Twitter social network over the past few months. Before downloading data, it was necessary to register a Twitter developer account and obtain four keys: consumer key, consumer secret key, access token, and access token secret. All of them are required to use the official Twitter API to save posts.

Since Russia attacked Ukraine in February 2022, it became an event that caused a corresponding reaction, both in Ukraine and abroad. People on social media, including Twitter, have been actively discussing the war and its related topics since February. Therefore, it was decided to analyze the tone of publications and comments of Ukrainian users regarding the war. The search was carried out using the geo-tag "Ukraine" and tweets were searched for the past six months, starting from July 2022, for posts and comments containing the mandatory keywords "Ukraine" and "war", as well as the optional words "missile", "offensive" and "invasion". Given the API limit for downloading post instances, it was decided to download 2,000 tweets matching the specified keywords. To form the dataset, only specific parameters of downloaded tweets are needed, accordingly, those that are not used must be discarded by looping through all downloaded tweets and saving the user's nickname, profile description, number of tweets, number of subscriptions, and followers, and most importantly, the text of the publication or comment All data is initially stored in the data dictionary and converted to the required format using the DataFrame function for data processing.

Considering the peculiarity of the created intelligent system, it was decided that the most convenient format for working with text data of this type using a neural network and a clustering algorithm will be the .csv format since it is universal and almost every programming language or data processing tool has functionality for working with files of this type. So, a dataset was created, which was processed during the training of the LSTM model of the neural network, and data was distributed into clusters. Before that, it became necessary to process all text data, that is, to bring them to one form for the most efficient processing.

Since we have a ready-made dataset, we can perform sentiment analysis, but before that, we need to format and standardize all the text so it can be conveniently analyzed. All hashtags were removed from the text, as they distorted the post's content or comment. All words have been reduced to lowercase, all URL links have been removed, and all special characters not needed to determine the text's tone have been drawn. In the end, all unnecessary spaces and single characters in the text were removed, and all system characters that were added utilizing the Twitter social network were removed, after which an LSTM analysis of the sentiment of the messages was performed, an example of the results of which is shown in Figure 7.

	text	Sentiment Score	\ Overall Sentiment
0	rt mtracey this woman literally works forus go...	0.000000	Neutral
1	rt smelyansky_igor 6000 branches of ukrposhta ...	0.000000	Neutral
2	snekotron fine with me we all knew they were c...	0.208333	Positive
3	rt euromaidanpress vilnius lithuania protest i...	0.000000	Neutral
4	rt andrew__roth putin is losing the war facing...	0.000000	Neutral

Figure 7 – An example of the results of determining the sentiment of the text by means of LSTM analysis

In the course of the work, the value of not only the general sentiment of words but also the polarity of the emotion was determined quite precisely, that is, the user can express a certain emotion with different strength and intensity according to the situation and the written publication or comment, accordingly, we determined the sentiment of the text instance of the data, as well as the emotional range, is from –1.0 – the most negative text, to 1.0 – the most positive, everything in between – the text with a certain sentiment and scope, there is also a value of 0, which means an average neutral text. The analysis of the text was carried out according to the previously described algorithm of memorizing certain moments in time from the test sample on which the neural network was trained. To begin with, we determine the language in which the text message was written, for convenience, only tweets in English were selected, after which two functions, spellcheck and correct, are performed, where we check the correctness of the written words using basic algorithms for working with text. Next, according to the algorithm, we perform word systematization, that is, using the definitions method, we obtain a list of possible meanings for the word according to the trained model and

DOI 10.15588/1607-3274-2024-1-16

choose the closest value. A neural network containing two levels of 100 elements, i.e., 50 elements per level, was generated. We set the value of spatial descent, take 500 epochs for the correct processing of 2000 messages, 0.5 is the value of spatial descent, and, accordingly, we set a similar value of 0.5 for gradient descent. Added a density value of 1 and a sigmoid activation function for the LSTM algorithm and correct handling of memory cell states. As a result, our recurrent model consists of an embedded layer, an LSTM layer, and a density layer, in which the sigmoid function is responsible for the native activation process. The training is carried out with a training batch size of 20 elements and a distribution factor of 0.2. The next step was to extract the level of emotionality of the sentiment of the text for a more comprehensive analysis of the text sample, but first, a general experiment was carried out on a test sample of 2000 words.

The model continues to run within 500 epochs, but instead of one block of memory, we have three different ones, called functional cells. Each cell has a certain state that can change according to the learning process and transfer the model's state. Accordingly, there is a hidden state that is unique to each cell and is not accessible from the others, and a distributed layer that is common to all three cells and, accordingly, can be changed by choosing the optimal value, which is carried out by selecting the maximum value at each of the steps of learning the subjective models. We will carry out not only a general analysis of sentiment but also an analysis of emotional, subjective sentiment in the range from –1 to 1. In this way, it is possible to understand how strongly the Twitter social network users expressed positive or negative emotions within the limits of the publication or comment, which is a reaction to military actions. Such an uneven distribution can be explained by the fact that people react to good news during the war much more emotionally and impulsively, rejoicing in something positive, as can be seen on the graph, some of the comments and publications reach 0.8 emotionality value and, on the contrary, negative comments written by people after several months of war much less emotional. A graph of the subjectivity of sentiment is shown in Figure 8. From this graph we can see the highest density of positive and negative emotions in the text instances has the same value of 0.5.

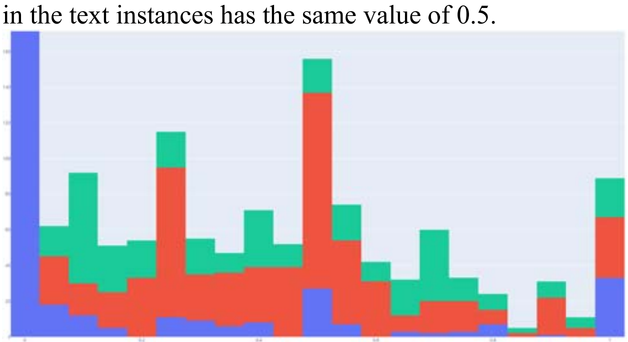

Figure 8 – Subjectivity of sentiment

So, after analyzing the sentiment of user posts and comments, we saw how many positive, negative, and neutral tweets were written during a specific period using certain thematic keywords and the distribution between their objectivity. Next, clustering was carried out by the *k*-means method to see the distribution of text messages of different tonalities by clusters. The clustering process consists of 3 main steps: initialization of the model, adjustment of model fit parameters, and forecast for further processing and division into clusters. To begin with, n was set – the number of clusters, which can be from 3 to 5, according to the number of text elements, and random state – the state of the clustering process, which is a random number necessary for the initialization of the data model.

After determining the number of clusters, it is time to distribute the text instances by these clusters, since there are only 3 of them, it is convenient to divide them into positive, neutral, and negative. To do this, the groupby.mean function was used, the essence of which is to determine the average value for each cluster. Next, standardization is carried out, for which it is necessary to subtract the cluster value from the sample mean and divide it by the mean deviation, after which the mean value is removed, and the unit variance is scaled. The distribution of user text data by clusters is depicted in Figures 9 and 10 using different graphs. Therefore, the PCA plot reflects the general structure of the data, and the TSNE plot best demonstrates the relationship between neighbors. For a more convenient display of data, 40 text instances were displayed on the PCA graph instead of all possible objects, and 400 text instances were displayed on the TSNE.

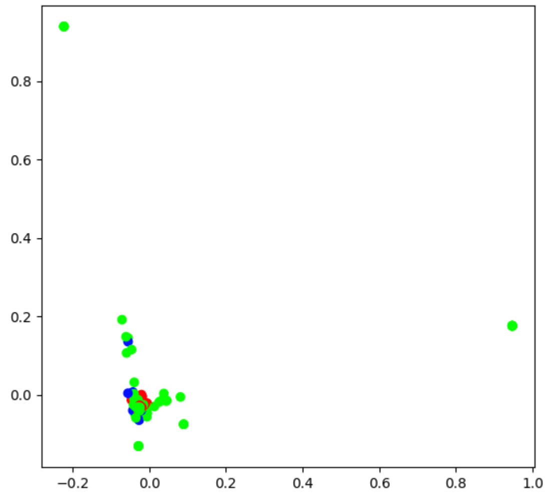

Figure 9 – Distribution graph of PCA type clusters

In the end, the most popular keywords for each cluster were extracted, from which certain conclusions can be drawn or trends can be observed, for this, the average value was calculated for all dimensions of the created model, which were grouped in each cluster. The next step was sorting the arrays of average values of each cluster in descending order and selecting the first ten elements,

DOI 10.15588/1607-3274-2024-1-16

shown in Figure 11, where zero cluster contain negative user posts and comments, the first cluster – positive, and the second – neutral.

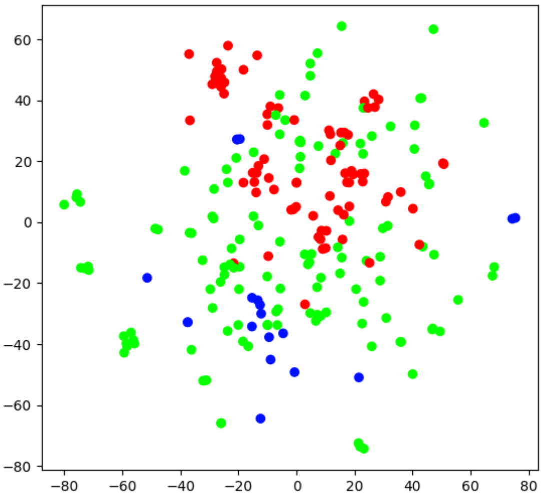

Figure 10 – Distribution graph of TSNE-type clusters

```
Cluster 0
one,country,nato,stop,people,putin,rt,russia,war,ukraine

Cluster 1
apmassaro3,attack,supporter,russia,civilian,today,russian,military,ukraine,rt

Cluster 2
monday,ukraine,oct,10,defense,pic,russian,air,rt,missile
```

Figure 11 – The most popular keywords

6 DISCUSSION

After experimenting, it can be concluded that the used combination of algorithms, namely the LSTM sentiment analysis neural network and the *k*-means clustering algorithm, works effectively and allow more accurate analysis of datasets than their counterparts, especially when the datasets reach large sizes and it is necessary to accurately analyze the sentiment of the text together with the level of emotionality and effectively distribute the number of clusters required for further research. Figure 12 shows a comparison of the work of combinations of the CNN neural network and the hierarchical clustering algorithm, which have the complexity of the $O(n^2)$ algorithm, and the LSTM neural network and the *k*-means clustering algorithm created in the intelligent system, which have the complexity of the $O(n\log n)$ algorithm, respectively, the implemented algorithms in an intelligent system, they work at least 10...15% faster and more efficiently than the previously described convolutional neural network and hierarchical clustering, and with large volumes of data they can work up to 20% better.

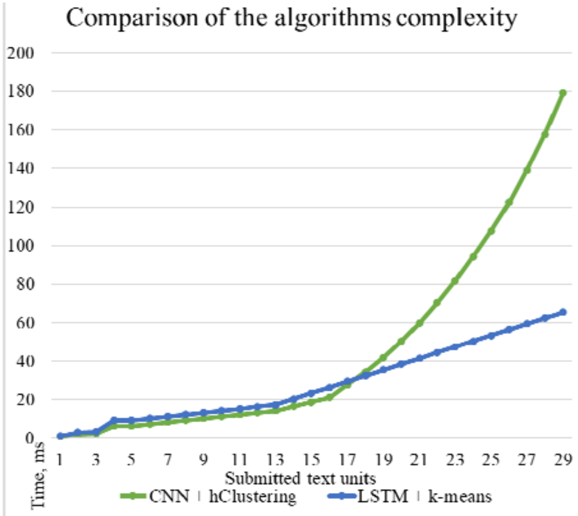

Figure 12 – Comparison of the complexity of algorithms

CONCLUSIONS

During the work, an intelligent system of sentiment analysis and clustering of posts in the Twitter social network was implemented, with the help of which the user can enter certain parameters and keywords that will be used to download posts and comments from the Twitter social network for analyzing the sentiment of messages, their emotional evaluation and division into clusters. To begin with, the importance of creating such an intelligent system was described, and the analysis of recent research and literature sources to understand what has already been implemented and what still needs to be implemented and, accordingly, repeat the advantages and avoid the disadvantages. The application of neural networks of various directions and clustering algorithms was mainly considered. The purpose of the work was described, and it was explained why implementing a modern LSTM neural network, and the *k*-means clustering algorithm is an urgent task to increase the system's efficiency and improve the accuracy of the final results. The functionality of the system was also described, namely the essence of the work of its main algorithms and the general implementation of the main functional components, and presented with the help of diagrams that depict both the block diagrams of the algorithms and the options for using and interacting with all the created data flows of the intelligent system as a whole.

In the course of the work, an actual experiment was carried out using the created intelligent system, during which keywords and search parameters for publications and comments in the Twitter social network were set, and a set of tweets was downloaded, which was saved and formed into a dataset. After that, formatting, lemmatization, and standardization of text instances were carried out for the correct operation of the algorithms. Concerning this text, a Long Short-Term Memory neural network was trained and used, with the help of which both a general sentiment analysis was carried out with the division of the text into negative, neutral, and positive, and an analysis of the emotional value of the text, i.e., the distribution from –1.0, i.e., the maximum possible negative to 1.0 – the maximum possible positive, and the value and subjectivity of sentiment are reflected. Similarly, a cluster analysis of the text was carried out using the *k*-means clustering algorithm, a model was created and trained, and the optimal number of clusters for the downloaded dataset was selected using the

DOI 10.15588/1607-3274-2024-1-16

"elbow" method. After dividing the text data into clusters, the formed clusters were graphically displayed using PCA and TSNE graphs. The average value was calculated for the current dimensions of the created model, and the ten most popular words in each cluster were displayed, the image of the text elements of each cluster was demonstrated using a three-dimensional graph, and the difference in efficiency and complexity between the CNN neural network and hierarchical clustering and the LSTM neural network was shown and *k*-means clustering algorithm implemented in the intelligent system.

ACKNOWLEDGEMENTS

The study was conducted within the framework of joint scientific research of the Department of Information Systems and Networks of the "Lviv Polytechnic" National University on the topic "Research, development and implementation of intelligent distributed information technologies and systems based on the resources of databases, data warehouses, data spaces and knowledge in order to accelerate the processes of formation modern information space using neural networks and data clustering algorithms". Scientific research was also carried out as part of the initiative research topic of the Department of Information Systems and Networks of the "Lviv Polytechnic" National University on the topic "Development of intelligent distributed systems based on the ontological dynamic system analysis approach for the purpose of integrating information resources as a data distribution services".

REFERENCES

1. Alhakiem H. R. Setiawan E. B. Aspect-Bas1ed Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi),* 2022,Vol. 6, No. 5, pp. 840–846. DOI: 10.29207/resti.v6i5.4429
2. Huang Xiaoran et al. Machine Learning Modeling of Vitality Characteristics in Historical Preservation Zones with Multi-Source Data, *Buildings*, 2022, Vol. 12, No. 11, P. 1978. DOI: 10.3390/buildings12111978
3. Lampropoulos G., Keramopoulos E. Virtual Reality in Education: A Comparative Social Media Data and Sentiment Analysis Study, *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, 2022, Vol. 10, No. 03, p. 19–32. DOI: 10.3991/ijes.v10i03.34057
4. Wang Y., Chen Zh., Fu Ch. Synergy Masks of Domain Attribute Model DaBERT: Emotional Tracking on Time-Varying Virtual Space Communication, *Sensors,* 2022, Vol. 22, No. 21, P. 8450. DOI: 10.3390/s22218450
5. Karyukin Vl. et al. On the development of an information system for monitoring user opinion and its role for the public, *Journal of Big Data*, 2022, Vol. 9, No. 1. DOI: 10.1186/s40537-022-00660-w
6. Albahli Saleh et al. A Machine Learning Method for Prediction of Stock Market Using Real-Time Twitter Data, *Electronics,* 2022, Vol. 11, No. 20, P. 3414. DOI: 10.3390/electronics11203414
7. Li Qizhi et al. A New Sentiment-Enhanced Word Embedding Method for Sentiment Analysis, *Applied Sciences,* 2022, Vol. 12, No. 20, P. 10236. DOI: 10.3390/app122010236
8. Abulhaija Sabreen et al. Mobile Applications Rating Performance: A Survey, *International Journal of Interactive Mobile Technologies (iJIM)*, 2022, Vol. 16, No. 19, pp. 133–146. DOI: 10.3991/ijim.v16i19.32051
9. Kamath A. N., Shenoy S. S., Kumar N. S. An overview of investor sentiment: Identifying themes, trends, and future direction through bibliometric analysis, *Investment Management and Financial Innovations,* 2022, Vol. 19, No. 3, pp. 229–242. DOI: 10.21511/imfi.19(3).2022.19
10. Zhang M. et al. Intelligent Vehicle Sales Prediction Based on Online Public Opinion and Online Search, *Sustainability,* 2022, Vol. 14, No. 16, P. 10344. DOI: 10.3390/su141610344
11. Pawełoszek I. Towards a Smart City – The Study of Car-Sharing Services in Poland, *Energies,* 2022, Vol. 15, No. 22, P. 8459. DOI: 10.3390/en15228459
12. Li Chih-yuan et al. Public Health Policy Monitoring through Public Perceptions: A Case of COVID-19 Tweet Analysis, *Information,* 2022, Vol. 13, No. 11, P. 543. DOI: 10.3390/info13110543
13. Corti L. et al. Social media analysis of Twitter tweets related to ASD in 2019–2020, with particular attention to COVID-19: topic modelling and sentiment analysis, J*ournal of Big Data,* 2022, Vol. 9, No. 1. DOI: 10.1186/s40537-022-00666-4
14. Liu H. Online review analysis on various networks' consumer feedback using deep learning, *IET Networks,* 2022, Vol. 11, pp. 234–244. DOI: 10.1049/ntw2.12045
15. Wang Yili et al. Sentiment Analysis of Twitter Data, *Applied Sciences,* 2022, Vol. 12, No. 22, P. 11775. DOI: 10.3390/app122211775
16. Kim D., Kim Y.-J., Jeong Y.-S. Graph Convolutional Networks with POS Gate for Aspect-Based Sentiment Analysis, *Applied Sciences*, 2022, Vol. 12, No. 19, P. 10134. DOI: 10.3390/app121910134
17. Opiła J. On Employing of Extended Characteristic Surface Model for Forecasting of Demand in Tourism, *Interdisciplinary Description of Complex Systems*, 2022, Vol. 20, No. 5, pp. 621–639. DOI: 10.7906/indecs.20.5.8
18. Bagate R. A., Suguna R. Sarcasm detection of tweets without #sarcasm: data science approach, *Indonesian Journal of Electrical Engineering and Computer Science,* 2021, Vol. 23, No. 2, P. 993. DOI: 10.11591/ijeecs.v23.i2.pp993-1001
19. Asgari T. et al. Identifying key success factors for startups With sentiment analysis using text data mining, *International Journal of Engineering Business Management,* 2022, Vol. 14, pp. 435–453. DOI: 10.1177/18479790221131612
20. Kowalska-Styczeń A. et al. Game problem of assigning staff to project implementation, *Decision Making: Applications in Management and Engineering,* 2023, Vol. 6, No. 2, pp. 691–721. DOI: 10.31181/dmame622023713
21. Lytvyn V. et al. Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology, *Mathematics,* 2023,Vol. 11, No. 4, P. 904. DOI: 10.3390/math11040904
22. Prokipchuk O. et al. 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,* 2023, Vol. 15, No. 3, pp. 70–93. DOI: 10.5815/ijmecs.2023.03.06

Accepted 04.01.2024.
Received 22.02.2024.

УДК 004.9

РЕАЛІЗАЦІЯ СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ АНАЛІЗУ ПУБЛІКАЦІЙ КОРИСТУВАЧІВ TWITTER

Батюк Т. М. – аспірант кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

Досин Д. Г. – д-р техн. наук, старший науковий співробітник, професор кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

DOI 10.15588/1607-3274-2024-1-16

АНОТАЦІЯ

Актуальність. У статті наголошується на необхідності створення системи прийняття рішень, яка може аналізувати повідомлення користувачів і визначати настрої, щоб зрозуміти, як новини та події впливають на емоції людей. Така система використовуватиме передові методи для аналізу повідомлень користувачів, заглиблюючись у почуття, виражені в тексті. Основна мета – отримати уявлення про те, як новини та різноманітні події відбиваються на емоціях людей.

Метою дослідження є створення системи прийняття рішень, яка зможе аналізувати та визначати настрої повідомлень користувачів, розуміти емоційну реакцію на новини та події та розподіляти дані в кластери, щоб отримати ширше розуміння думок користувачів. Ця багатогранна мета передбачає інтеграцію передових методів обробки природної мови та машинного навчання для створення надійної системи прийняття рішень. Основними цілями є аналіз настроїв, розуміння емоційних реакцій на новини та події та кластеризація даних для цілісного уявлення про думки користувачів.

Метод. Для обробки великих обсягів даних користувача пропонується використання нейронних мереж довгострокової пам'яті для аналізу настрою та алгоритму *k*-середніх для кластеризації даних. Ця стратегічна комбінація спрямована на вирішення проблем, пов'язаних із обробкою великих обсягів даних, створених користувачами, у більш глибокий та цілеспрямований спосіб.

Результати. Виконано дослідження та концептуальне проектування системи прийняття рішень та створено систему прийняття рішень. Система включає аналіз настроїв і кластеризацію даних для розуміння думок користувачів і цінності настроїв таких думок, розділяючи їх на кластери та візуалізуючи результати.

Висновки. Розробка системи прийняття рішень, здатної аналізувати настрої користувачів і кластеризувати дані, може надати цінну інформацію про реакцію користувачів на новини та події в соціальних мережах. Запропоноване використання нейронних мереж довгострокової пам'яті та алгоритму *k*-середніх вважається придатним для аналізу настроїв і завдань кластеризації даних. Підкреслюється важливість вивчення існуючих робіт і систем для розуміння доступних алгоритмів і їх застосування. У статті також описано створену та впроваджену систему прийняття рішень та продемонстровано функціональність системи на прикладі набору даних.

КЛЮЧОВІ СЛОВА: обробка природної мови, згорткова нейронна мережа, рекурентна нейронна мережа, LSTM, кластеризація *k*-середніх.

ЛІТЕРАТУРА

1. Alhakiem H. R. Aspect-Bas1ed Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion / Hanif Reangga Alhakiem, Erwin Budi Setiawan // Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). – 2022. – Vol. 6, No. 5. – P. 840–846. DOI: 10.29207/resti.v6i5.4429
2. Machine Learning Modeling of Vitality Characteristics in Historical Preservation Zones with Multi-Source Data / [Xiaoran Huang et al.] // Buildings. – 2022. – Vol. 12, No. 11. – P. 1978. DOI: 10.3390/buildings12111978
3. Lampropoulos G. Virtual Reality in Education: A Comparative Social Media Data and Sentiment Analysis Study / Georgios Lampropoulos, Euclid Keramopoulos // International Journal of Recent Contributions from Engineering, Science & IT (iJES). – 2022. – Vol. 10, No. 03. – P. 19–32. DOI: 10.3991/ijes.v10i03.34057
4. Wang Y. Synergy Masks of Domain Attribute Model DaBERT: Emotional Tracking on Time-Varying Virtual Space Communication / Ye Wang, Zhenghan Chen, Changzeng Fu // Sensors. – 2022. – Vol. 22, No. 21. – P. 8450. DOI: 10.3390/s22218450
5. On the development of an information system for monitoring user opinion and its role for the public / [Vladislav Karyukin et al.] // Journal of Big Data. – 2022. – Vol. 9, No. 1. DOI: 10.1186/s40537-022-00660-w
6. A Machine Learning Method for Prediction of Stock Market Using Real-Time Twitter Data / [Saleh Albahli et al.] // Electronics. – 2022. – Vol. 11, No. 20. – P. 3414. DOI: 10.3390/electronics11203414
7. A New Sentiment-Enhanced Word Embedding Method for Sentiment Analysis / [Qizhi Li et al.] // Applied Sciences. – 2022. – Vol. 12, No. 20. – P. 10236. DOI: 10.3390/app122010236
8. Mobile Applications Rating Performance: A Survey / [Sabreen Abulhaija et al.] // International Journal of Interactive Mobile Technologies (iJIM). – 2022. – Vol. 16, No. 19. – P. 133–146. DOI: 10.3991/ijim.v16i19.32051
9. N Kamath A. An overview of investor sentiment: Identifying themes, trends, and future direction through bibliometric analysis / Aditi N Kamath, Sandeep S. Shenoy, Subrahmanya Kumar N. // Investment Management and Financial Innovations. – 2022. – Vol. 19, No. 3. – P. 229–242. DOI: 10.21511/imfi.19(3).2022.19
10. Intelligent Vehicle Sales Prediction Based on Online Public Opinion and Online Search / [Mingyang Zhang et al.] // Sustainability. – 2022. – Vol. 14, No. 16. – P. 10344. DOI: 10.3390/su141610344
11. Pawełoszek I. Towards a Smart City–The Study of Car-Sharing Services in Poland / Ilona Pawełoszek // Energies. – 2022. – Vol. 15, No. 22. – P. 8459. DOI: 10.3390/en15228459
12. Public Health Policy Monitoring through Public Perceptions: A Case of COVID-19 Tweet Analysis / [Chih-yuan Li et al.] // Information. – 2022. – Vol. 13, No. 11. – P. 543. DOI: 10.3390/info13110543
13. Social media analysis of Twitter tweets related to ASD in 2019–2020, with particular attention to COVID-19: topic modelling and sentiment analysis / [Luca Corti et al.] // Journal of Big Data. – 2022. – Vol. 9, No. 1. DOI: 10.1186/s40537-022-00666-4
14. Liu H. Online review analysis on various networks' consumer feedback using deep learning / Huajin Liu // IET Networks. – 2022. – Vol. 11. – P. 234–244. DOI: 10.1049/ntw2.12045
15. Sentiment Analysis of Twitter Data / [Yili Wang et al.] // Applied Sciences. – 2022. – Vol. 12, No. 22. – P. 11775. DOI: 10.3390/app122211775
16. Kim D. Graph Convolutional Networks with POS Gate for Aspect-Based Sentiment Analysis / Dahye Kim, YoungJin Kim, Young-Seob Jeong // Applied Sciences. – 2022. – Vol. 12, No. 19. – P. 10134. DOI: 10.3390/app121910134
17. Opiła J. On Employing of Extended Characteristic Surface Model for Forecasting of Demand in Tourism / Janusz Opiła // Interdisciplinary Description of Complex Systems. – 2022. – Vol. 20, No. 5. – P. 621–639. DOI: 10.7906/indecs.20.5.8
18. Bagate R. A. Sarcasm detection of tweets without #sarcasm: data science approach / Rupali Amit Bagate, R. Suguna // Indonesian Journal of Electrical Engineering and Computer Science. – 2021. – Vol. 23, No. 2. – P. 993. DOI: 10.11591/ijeecs.v23.i2.pp993-1001
19. Identifying key success factors for startups With sentiment analysis using text data mining / [Tina Asgari et al.] // International Journal of Engineering Business Management. – 2022. – Vol. 14. – P. 435–453. DOI: 10.1177/18479790221131612
20. Game problem of assigning staff to project implementation / [Agnieszka Kowalska-Styczeń et al.] // Decision Making: Applications in Management and Engineering. – 2023. – Vol. 6, No. 2. – P. 691–721. DOI: 10.31181/dmame622023713
21. Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology / Vasyl Lytvyn [et al.] // Mathematics. – 2023. – Vol. 11, No. 4. – P. 904. DOI: 10.3390/math11040904
22. Intelligent Analysis of Ukrainian-language Tweets for Public Opinion Research based on NLP Methods and Machine Learning Technology / [Oleh Prokipchuk et al.] // International Journal of Modern Education and Computer Science. – 2023. – Vol. 15, No. 3. – P. 70–93. DOI: 10.5815/ijmecs.2023.03.06

DOI 10.15588/1607-3274-2024-1-16