Social Media Content Analysis with Machine Learning Tools*

Dmitriy V. Anashkin[1[0000-0002-0750-1217]], Kostyantyn A. Malyshenko[1[0000-0002-3453-2836]]

[1]V.I. Vernadsky Crimean Federal University, Simferopol, Russian Federation,
docofecon@mail.ru

Abstract. The article is devoted to the development of a computer program that will allow researching the content of social networks using computer analysis methods to predict the information response of users to a new publication and visualize a user graph. The aim of the study is the implementation of an algorithm that allows to improves the methods of automatic collection of information from a social network. It is proposed to improve the methods of collecting information in combination with computer learning tools, in particular, using neural networks. The work consists of formulating and solving the problem of using classical machine learning methods to predict the sentiment of a user reaction to new news in the form of quantitative markers. The work on the algorithm was divided into two stages. At the first stage of the study, a software solution was implemented that builds a graph of a social network user using the internal means of the computer mathematics system "Wolfram Mathematica". The process of modeling the local user network of individual communities is described. Further work carried out is based on the interaction of two network development environments, namely "VKAPI" and "Wolfram Cloud". The second stage of the research is writing a special software module for collecting publications from the news feed of the VKontakte social network at a given request. The final stage of the second stage is checking and evaluating the performance of the trained model on new data.

Keywords: Machine Learning, Social Media, Computer Mathematics, Wolfram Mathematica, Social Graph, API, Recommendation System.

1 Introduction

In the modern world, a colossal amount of information is generated every minute, be it data from satellites, statistics of cellular operators, or news reports. Thanks to the development of network communication protocols, humanity has the opportunity to quickly exchange a large amount of information and content, which in the later stages provoked the emergence of social interaction networks.

Social networks continue to confidently penetrate more and more distant areas of the everyday life of any person. Today there are no longer just group networks with a user interface, but full-scale commercial projects that accumulate the experience of a wide range of narrow areas. These areas include social engineering, marketing, PR (advertising), online education. The modern social network is an interdisciplinary field. It brings together mathematicians, IT specialists, economists, and sociologists. It is worth noting that the results obtained by specialists from different disciplines quickly find their application in the social network. This is possible due to the close relationship between its components.

Now, not a single commercially successful project is complete without an additional platform to attract a potential audience. Therefore, with the active integration of social media into global processes, the need for a thorough analysis of endlessly flowing content is increasing. It is of great value for research, as it allows you to draw useful conclusions for making important decisions on related issues.

There are many areas of social media analysis. T.V. Batura [12] identifies four main branches: structural, resource, normative and dynamic. Each of these areas requires close attention to one or another component of the global structure of social networks.

The very process of analyzing social networks involves the use of specific methods and tools for working with a variety of heterogeneous data. This problem has become the impetus for the development of a new branch of science, which is directly involved in the description of connections of different densities and intensities that have arisen in the course of social interaction and communication.

The science of social network analysis (SNA) is characterized by the following defining properties:

1. Embodying the idea of the importance of social connections.
2. Collecting data reflecting social connections.
3. Use of graphic images to visualize the results.
4. Use of mathematical, statistical, and computational models.

Modern research requires a powerful and promising environment for programming, processing, and visualization of information objects. In such an environment, tools are required to send queries to the social network to obtain accessible and structured analysis results.

The computer mathematics system “Wolfram Mathematica” fully complies with these requirements. It is known as the most powerful computing system in the world, along with well-known competing software packages such as Matlab, Maple, and SciLab. Mathematica differs from its predecessors and competitors primarily in that it serves as a single platform for immediate deployment and is an effective solution for conducting research and computing work and interactive demonstration of scientific results of publishing quality. The environment "Wolfram Mathematica" allows the user to solve time-consuming tasks and understand the most complex concepts, using a large arsenal of ready-made tools and packages for the development and implementation of new functions.

In addition, the flexibility of the system and the ease of mastering the syntax of the programming language makes it possible to use Wolfram technologies for educational purposes. We cannot ignore the presence in the system of special means for interacting with the API of most well-known social networks. For example, "Wolfram" offers the user ready-made personal analytics for Facebook and Instagram in the form of his social graph and diagrams reflecting user statistics (activity, friends, check-ins, etc.).

The purpose of the research is to write and implement an executable program in the computer mathematical system "Wolfram Mathematica", which allows solving the problem of using machine learning and computer analysis methods in the compilation of predictive models trained on open information sources of social networks.

2 Building and Analyzing a Social Graph

2.1 Subject Area Overview

Since social media can be of different structures and sizes, there is no one way to represent them across all applications. However, the network model can be defined as some object - a substitute for the original object, retaining the fundamental properties of the network, but adapted for certain tasks. In this case, the original object is a social network. The application of the models is because any social network with its internal interconnections has a complex branching structure. As a rule, a network is a closed set containing at least n≥2 elements that are connected by a certain type of connection. Mathematically, the concept of a network is described using graph theory and its applications. Graph (G) is a mathematical object uniquely determined by three non-empty sets (V, E, I), where V is the set of vertices of the graph, E is the set of graph edges, I is the set of incidence relations.

Since social media can be of different structures and sizes, there is no one way to represent them across all applications. However, the network model can be defined as some object - a substitute for the original object, retaining the fundamental properties of the network, but adapted for certain tasks. In this case, the original object is a social network. The application of the models is because any social network with its internal interconnections has a complex branching structure.

As a rule, a network is a closed set containing at least n≥2 elements that are connected by a certain type of connection. Mathematically, the concept of a network is described using graph theory and its applications. Graph (G) is a mathematical object uniquely determined by three non-empty sets (V, E, I), where V is the set of vertices of the graph, E is the set of graph edges, I is the set of incidence relations.

$$G = \begin{Bmatrix} V \subset X \leftrightarrow \mathbb{N} \\ E \subset V \times V \\ I\colon \forall t \in E \to I(t) = \{x, y\} \end{Bmatrix}$$

In modern computer sociology, the concept of a node in a graph is replaced by the concept of an actor (user), respectively, the edges are considered as connections between these same actors. Note that actors are separate individual, corporate, or collective social units. Thus, a new concept is brought up for consideration a social graph.

A social graph is a kind of graph, the nodes of which are represented by social objects (user profiles with various attributes, communities, media content), and the edges are represented by social connections between them. This type of network, such as an implicit social graph, has become widespread in practice. This is a graph that does not indicate the explicit social connections of the actor, but it can be deduced by considering the interactions between the user and the social unit (group, friend).

The foundations of social graph analysis were laid back in 1994 in the book Social Network Analysis: Methods and Applications by S. Wasserman and K. Faust [10]. The principles outlined in this book are used in network analytics to this day.

Let's present a basic list of graph network models that can be found in the literature today:

– *Real-world networks*. One of the main characteristics inherent in this type of network is the fact that connections between nodes depend on a certain probability. In real networks, such as social networks, this means that the likelihood of communication between two nodes is higher if there is any relationship between them (friendly, professional). However, social networks differ from other types of networks, mainly because they are easier to divide into communities. Consequently, this characteristic affects the distribution of degrees, usually with a positive correlation of degrees and a high level of clustering.

– *Random networks*. From a theoretical point of view, random networks, best known as random graphs, can be considered the intersection of graph theory and probability theory. Any random graph can be described as a probability distribution or as a random process that will be used to create connections between any pair of nodes in the graph. Therefore, to generate a random graph, it is only necessary to randomly add (following a predetermined probability distribution) consecutive edges between any pair of nodes. Random graphs are commonly used to compare the structure and properties of any graph.

– *Small world networks*. Many real-world networks have two main properties: the first is related to the distance between two nodes (which is usually small), while the second is related to the transitivity or clustering factor of these networks (which is usually relatively high). However, when real networks, such as those associated with social networks, are analyzed, it is quite common to find one property that does not appear in random graphs. This property is related to the degree of transitivity, which in these real networks has a higher value than in random graphs. This property appears due to the social behavior of the network (usually the user's friends are likely to be friends too), which means that the transitivity on the graph or the value of the clustering coefficient will have more importance on social networks.

– *Scale-free networks*. This type of network is usually defined based on its power distribution. In these networks, the degree of distribution of nodes will follow the strength of the law (at least asymptotically). That is, the fraction P (k) of nodes in the network that have k connections to other nodes goes to large values of k as $P(k) \sim k^{-\gamma}$, where γ is a parameter whose value is usually in the range $2 < \gamma < 3$, although sometimes it can go beyond these boundaries.

As practice shows, visualization is one of the advantages of using graphs for analyzing social networks. There are three methods for visualizing such graphs: the "planarization" method, the "orientation" method, and the "directed forces" method. The essence of the "planarization" method is that if a graph is not planar (that is, it has intersections of edges in its two-dimensional projection), it is artificially made by it. To do this, artificial "pseudo" -verts are placed at the intersection points of the edges, and then one of the methods for constructing a planar graph can be applied to the graph. The "orientation" method is based on a similar principle: an undirected graph is artificially transformed into a directed one, after which one of the directed graph visualization methods can be applied to the graph.

Thus, the choice of a method for visualizing a social graph will significantly affect its metrics (density, division into clusters) and informative value.

2.2 Integration of Wolfram Mathematica and Social Network API

To get started, it is necessary to integrate the computer mathematical system and the program interface of the social network. The authors decided to use the API of the Russian-language social network "VKontakte" for the work. VKAPI has some advantages, thanks to which it is possible to carry out complex work with open data of a social network while saving time on various intermediate operations (for example, installing and configuring additional software). So, connecting and using the VKontakte API is not difficult for any registered user of this social network. In addition, the API query format is relatively simple and scalable to provide detailed data useful for research.

The first stage of the VKontakte API integration is the creation of a web application that processes the received requests to send them to the server. After filling out the form, the user is prompted to select the type of application to create. For the full and convenient use of all the capabilities of the VKAPI program interface, the type of the created application "Standalone" was chosen. Next, you need to get a unique client access key, which will later be used to generate a request to the social network interface. The access token is provided to the user through the OAuth2 proprietary authorization protocol. To do this, a link is generated in the address bar of the browser used, containing information about the newly created application and a request to grant the necessary rights to work with the server's functionality. The rights "wall", "friends" and "offline" were obtained for the conducted research.

The responses returned by the server to the client request are in JSON (JavaScript Object Notation) format. It is a cross-platform textual data exchange format that has a nested structure and uses the key: value construction as the main syntax. This format is especially convenient for presenting lists and compiling data samples based on predefined features.

After receiving a unique key (token), you can proceed to the second stage of integration of the platforms used. To do this, a token variable is declared in the Wolfram Mathematica notebook (in the format of a text string) and equated to the value of the received client access key (access_token). We implement a function that allows you to work with a repository of methods stored on the server (Fig. 1).

```
VK[Method_, Params_] :=
  Import["https://api.vk.com/method/" <> Method <> "?" <> Params <>
  |импорт                                  |метод
    "&access_token=" <> token <> "&v=5.21", "JSON"];
```

Fig. 1. Wolfram Mathematica code that makes a request to the VKontakte social network API.

Wolfram Mathematica has a set of tools that allow you to make POST / GET requests and process JSON files, formatting the return values of these files into rules of the form Rule [{1} → {2}]. Characters that appear as pattern names on the left are treated as a local rule. Any character or expression on the right side of the rule is considered as its condition. Mathematica allows you to combine data types within the same list in a simple enumeration. This turns out to be extremely convenient when processing data received from a social network.

3 Social Graph Programming

Obtaining information about a user of a social network, his profile data is carried out by similar means. Let's say you need to get a photo of the user, his name, and surname. For this, the users.get method will be used with the user.id option passed in. Next, we list the fields whose values we want to get using this method. We need fields "photo", "first_name", "last_name". To represent the values returned in JSON format as a list, let's write a function that clears the response from the "key" parameter (Fig. 2).

```
VKInfo[ID_, OptionsPattern[{Photo → False}]] :=
           |шаблон опций          |ложь
  Module[{}, If[OptionValue[Photo] == True,
  |программн··· |··· |значение опции      |истина
    {#[[2]] <> " " <> #[[3]], Import[#[[1]]]} & /@
                              |импорт
     ({"photo", "first_name", "last_name"} /.
       VK["users.get", "user_ids=" <> ToString@ID <> "&fields=photo"][[1, 2]]),
                                      |преобразовать в строку
    #[[1]] <> " " <> #[[2]] & /@
     ({"first_name", "last_name"} /. VK["users.get", "user_ids=" <> ToString@ID][[1, 2]])]];
                                                                  |преобразовать в строку
```

Fig. 2. A function that collects additional information about a user using his online identifier.

The program module contained in the body of the function accepts as input the conditions specified in the form of boolean variables (true, false). By default, the Photo variable is set to false, which means that when requesting data from the API, the value of this field will not be returned. Further, the "VKInfo" option will be used when "drawing" a social graph to replace its vertices with photos from the profiles of the corresponding friends.

The main characterizing motive for finding a user in a social network is his interaction with friends or social groups of different sizes. Therefore, to reflect a visual picture of the user's social circle, it is necessary to get a list of his friends. In VKAPI, this function is called "friends.get". The response of the parser (query tool) is a list of user identifiers that are in a network "friendship" with the element being examined.

Thus, based on the above remarks, let us single out the general algorithm for constructing a social graph:

1. Formation of the study group.
2. Interaction with the provider of the social network (data parsing).
3. Getting a list of the user's friends.
4. Formation of an egocentric graph.
5. Finding common friends of the researched user and users in the circle of interaction.
6. Building a communication graph.
7. Combining the communication graph and the egocentric graph.
8. Clearing the graph from side elements (blocked, deleted users, users who have denied access to the profile).
9. The choice of laying the network. Selecting the graph clicks and dividing it into clusters.

Research into egocentric (personified) networks is limited to the study of the social connections of one individual. This approach provides representative examples of the social environment of individual elements and is compatible with statistical generalization methods for large networks. Thus, the ability to calculate the most likely environment for a particular social unit increases.

4 Research of a Social Network Using Artificial Intelligence Methods

In recent years of active development of science, many new research opportunities have appeared. Increasingly, the use of the latest developments and algorithms of machine vision is observed when performing various tasks. Such a direction of computer science as artificial intelligence has managed to exert a strong influence on scientific research and has become a prerequisite for the emergence of a new tool for specialists in the field of data analysis [5]. But applying machine learning requires a persistent source of data. Currently, the popularity of social media is growing and it is they that can act as a regularly updated data bank for scientists and analysts.

Data that represents information about a user on social networks can be schematically divided into 4 target types: profile, behavior, text content, and visit monitoring data. To extract the features of the model, it is enough to assess the relative influence of the feature, reliability, and generalization of the potential for classification.

Some social media data is easy to sort, classify, and structure (these include forms, historical data), but the bulk of the content is unstructured and requires more careful analysis. Also, when collecting data, you can face the following problems: limited access or blocking for automatic collection of information, data privacy - the user has the right to set privacy settings, which makes many profiles attributes inaccessible from the outside, poor data structure - some social networks either do not provide an API or install restrictions on working with it, which makes it impossible to use it by third-party programs.

The data required for machine representation is obtained in several stages. First, a corpus of texts for research is collected, which are pre-cleared of special characters to avoid errors in the operation of the machine algorithm.

In the social network "VKontakte" for the study, thematic publications were selected that correspond to the user's request. For the sample, publications were used with the keyword "Eurovision Song Contest". For this, the API method "newsfeed.search" was used, which searches for publications in the database. The data was received in JSON format using an algorithm that works up to $n \leq 200$ publications, where 200 is the maximum value returned by the function, as stated in the official documentation, and was set with the keys "text" (publication text), "id" (record identifier), "Owner_id" (identifier of the user or group that published the news). Then the results were split into two fixed tables. The first table consisted of the texts of records in Russian and was a corpus. The second nx2 table mapped the publication to a pair of "id" and "owner_id" values.

Further work with the corpus assumed its literal translation into English. Further, each element of the corpus was divided into a list consisting of all the words found in the text.

The paired values "like" and "comment" was obtained after sending the second table to the API system and processed by the methods "likes.getList" and "wall.getComments", respectively.

The authors of this work, through the internal functions of "Wolfram Mathematica", implemented a multilayer neural network (multilayer perceptron), which predicts user reactions based on transformations of vectors of fixed dimensions.

MLP (Multi-Layer Perceptron) network consisted of three functional layers:

«Embedding Layer» – embedding layer. It stores the words of all words in the dictionary. A dictionary is a matrix consisting of words and their vector representations, while words that are close in semantic meaning have a close location in the R^n-dimensional space. The dimension R^n is a tunable parameter, and the larger n, the higher the stability of the dictionary and the better the model.

A ready-made model stored on the Wolfram Neural Net Repository under the name GloVe 50 Dimensional Word Vectors Trained on Wikipedia and Gigaword 5 Data was used like a dictionary. This model was released in 2014 by the Computer Science Department of Stanford University and was trained using an original method called Global Vectors (GloVe). It encodes 400,000 tokens as unique vectors in a 50-dimensional space, with all tokens outside the dictionary being encoded as a null vector.

«Recurrent Layer» – a layer with directed links between elements. The output of the neuron in this layer can be fed back to the input. This structure allows you to create a

structural semblance of "memory" and sequentially process incoming data for their subsequent storage. The recurrent layer allowed us to process the sequence of vectors included in the proposal to build a vector for the entire publication, which helped to form a dataset for training a linear layer with R^2 - directional vector at the output and R^{50} at the input.

«Sequence Last Layer» – the last layer of the network. Used to convert a sequence of token vectors into a sentence vector based on the semantic arrangement of words in the dictionary. The advantage of this layer is that if a word from the text is absent in the dictionary, then this is compensated by the transformation.

The last step in building a neural network model was adding a Linear Layer, trained on vectors of submitted publications and pairs of values from the table (like a comment). In this case, the output value was a vector of dimension R ^ 2, which describes the number of user tags (like and comment, respectively) under the newly published post.

The parameters of the neural network implemented in the Wolfram Mathematica language are shown in Fig. 3. A detailed scheme of the algorithm that predicts user reaction to new news is shown in Fig. 4.

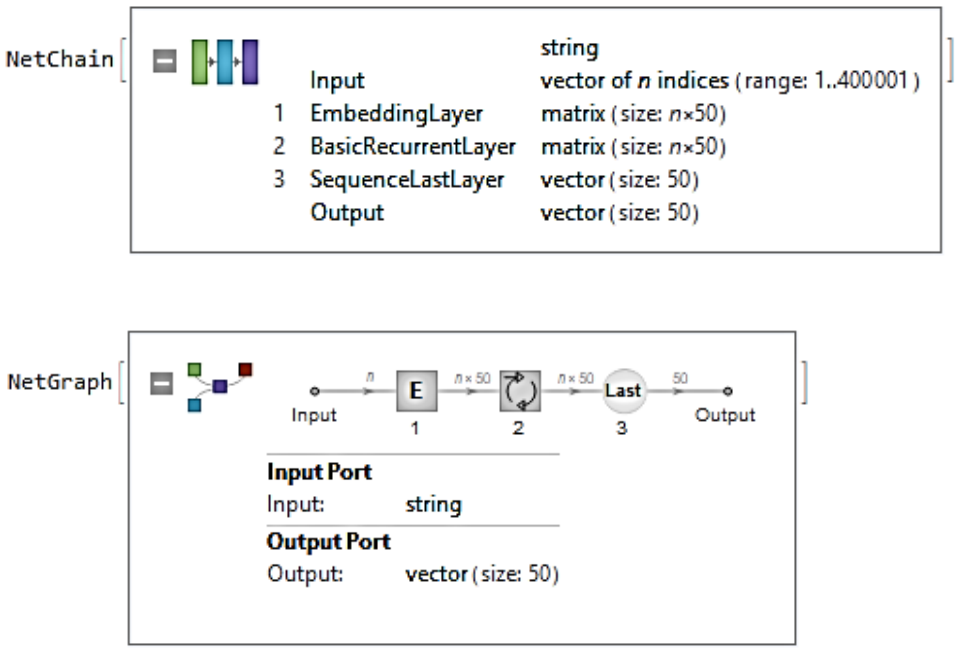

Fig. 3. Parameters of a neural network that converts text into a 50-dimensional vector.

The objective function of the neural network, implemented by the authors, was optimized using the "Adam" (Adaptive moment estimation) method. It combines the idea of motion accumulation and weaker renewal of weights for typical signs.

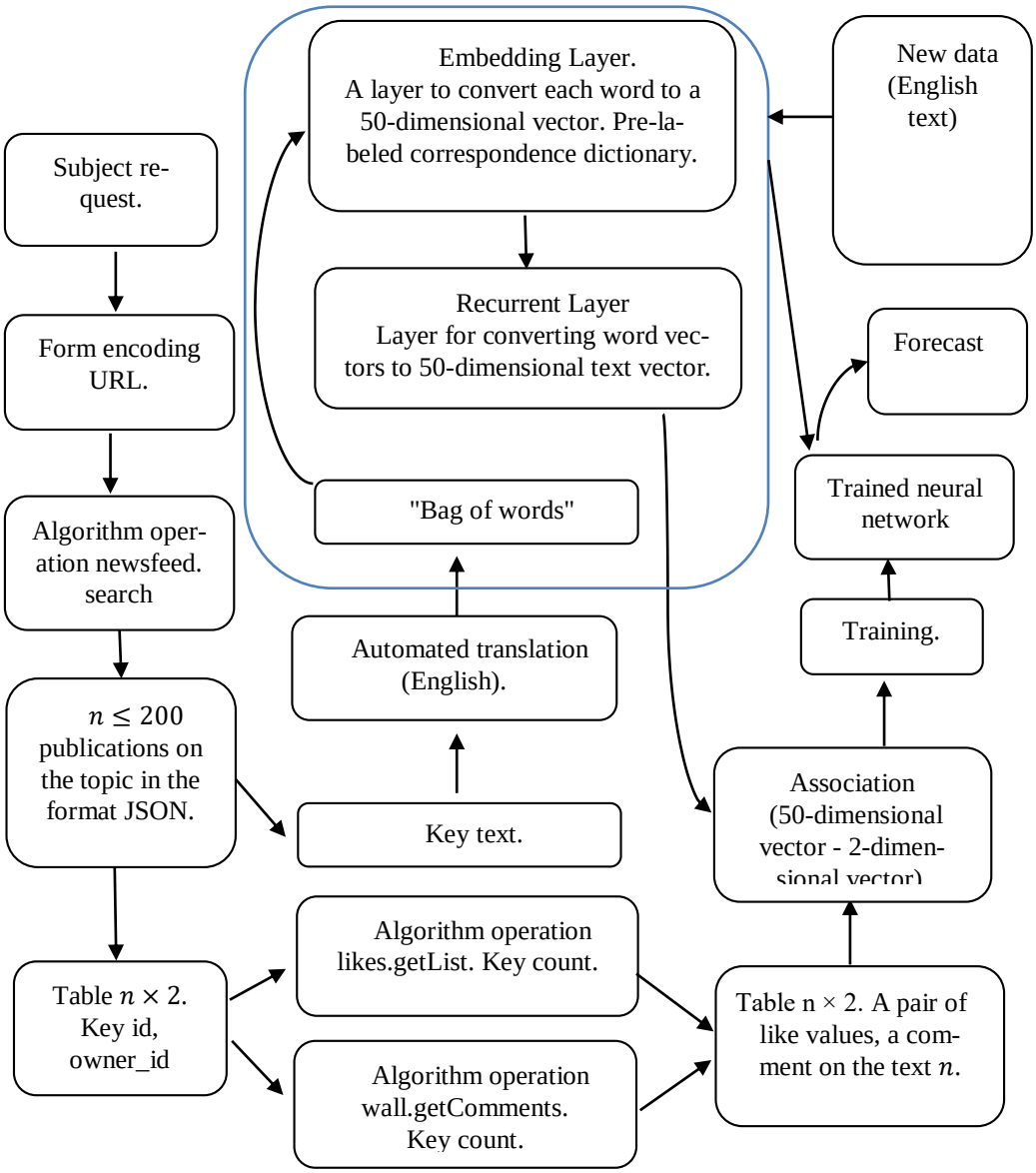

Fig. 4. Diagram illustrating a neural network prediction algorithm for social network content.

The presented algorithm was chosen due to its clear advantages: it is simple to implement, computationally efficient, and not memory-intensive. In addition, it is well suited for tasks that involve working with large amounts of data. In practice, this algorithm is currently used by default and has clear advantages over "SMSProp".

Thus, the above-described neural network construction method can be used to predict user marks under a newly published news item. To predict the number of "I like" marks and the number of comments, the text of publications on a given topic is vectorized, after which it is converted into a vector of dimension (R50) R^{50}. Next, a training sample is compiled, consisting of text vectors and the corresponding publication evaluation vectors of dimension R^2. New data, on which it is required to make a forecast, is submitted in the form of vectorized text, after which a prediction can be obtained on the output layer in the form of a pair of values "like" and "comment".

5 Discussion

The study of network interactions of users, as well as the information that is transmitted between them, is an important task of modern information technology.

The rapid development of means of semantic processing of natural language has become the impetus for the development of a new wave of technological innovations that are successfully implemented in human everyday life. Voice assistants, automatic text recognition systems are now indispensable attributes of many digital systems and devices. The dynamics of the development of natural language processing systems have been repeatedly considered in scientific works. The authors of publications [6, 7] describe the dynamic properties of such systems and emphasize the importance of the consistent application of algorithms to achieve the stated result.

Computer analysis allows scattered data to be adapted for global use. This opportunity has become feasible thanks to the development of mathematical and programmable systems that present visual data in a visual form. The authors of the work set their particular task in the development of an algorithm for visualizing the graph of a social network user. Special literature was studied [1, 8], which made it possible to actualize the main issues of the problem being solved and come close to its direct implementation.

The authors of this article in two stages obtained two software models that allow working with the API of the social network. A similar approach to the collection and use of information from social networks has already been considered by the authors of works [1, 2]. However, the fundamental difference between this study and the previous ones lies in the fact that an integrated approach to the implementation of software solutions has been produced. Also, software was used to carry out the work, the use of which required the preparation of a complex algorithmic code.

6 Conclusions

Based on the open information of the social network "Vkontakte", as well as the available search algorithms and information objects in it, programmable solutions were implemented that allow operating the information received at a high level.

The result of the first stage of the research was a software solution that builds a graph of a user of a social network using the internal means of the computer mathematics

system "Wolfram Mathematica". The second stage of the study made it possible to implement a software module for collecting publications from the news feed of the VKontakte social network at a given request.

In the future, it is planned to refine and improve the presented models for their possible real implementation. Also, the development vector on the topic under consideration is planned to present a software package that, based on feed-forward neural networks, will automatically generate news summaries based on information collected from the social network using the methods presented in the article.

References

1. Camacho, D., et al.: The Four Dimensions of Social Network Analysis: An Overview of Research Methods, Applications, and Software Tools. 2020, 73.
2. Khattak, A., et al.: Fine-Grained Sentiment Analysis for Measuring Customer Satisfaction Using an Extended Set of Fuzzy Linguistic Hedges. International Journal of Computational Intelligence Systems 13(1). 2020, 744-756
3. Artificial Intelligence (AI) in Social Media Market By Technology (Machine Learning & Deep Learning, Natural Language Processing), By Component (Services & Solutions), By Organization Size, By Applications, By Industry Vertical, And Segment Forecasts To 2027. Information and Communication Technology. 2020, 285. URL: https://www.reportsanddata.com/report-detail/artificial-intelligence-ai-in-social-media-market
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 5. 2017. PP. 135-146
5. Müller, A.: An introduction to machine learning with Python. A guide for data scientists. Andreas Müller, Sara Guido. M: Alpha Book. 2018. PP. 480.
6. Sheng, Y., Subhash, K. A.: Survey of Prediction Using Social Media. Department of Computer Science, Oklahoma State University. Stillwater, Oklahoma, U.S.A. URL:
7. Popova, E.P., Leonenko, V.N.: Predicting the Reaction of Users on Social Networks Using Machine Learning Methods. Scientific and technical bulletin of information technologies, mechanics, and optics. 2020.Vol. 1. No 1. PP. 118-124.
8. Danilov, D.E.: Analysis of Clustering of Social Graphs Using Wolfram Mathematica. Mathematics and its applications in modern science and practice. 2014. PP. 257-261.
9. Freeman, L.C.: Centrality in social networks: Conceptual classification. Social Networks, 1979. PP. 215-239.
10. Wasserman S., Faust K. Social network analysis: methods and applications. Cambridge, New York, Cambridge University Press. 1994, 825.
11. Tyakunov, A.S., Slavsky, V.V.: Using Wolfram Mathematica Tools to Assess the Effectiveness of the Vkontakte Social Network Community. Works of the seminar on geometry and mathematical modeling: a collection of articles. Issue 2. / editor-in-chief E.D. Rodionov. Barnaul: Publishing house of Alt. un-ta, 2016. PP.94-99.
12. Batura, T.V.: Methods of Analysis of Computer Social Networks. NSU Bulletin. Series: Information technology. 2012.T.10. No. 4. PP. 13-28.