An approach to the training dataset formation for assessing the sentiment degree of social network posts using machine learning

Andrey Konstantinov
Ulyanovsk State Technical University
Ulyanovsk, Russia
adwaises@mail.ru

***Abstract*—This article describes an approach to the formation of a training set for assessing the emotional coloring of social network posts. The dataset is formed in an automated mode. The input values of the algorithm are 2.5 million posts of a social network, the output values is a training set neural network. The algorithm for the formation of the training set is based on selection using copyright symbols for expressing emotions and key phrases. The quality of the training set is checked during the training of the multilayer perceptron by the set obtained and experiments. The accuracy of determining the emotional coloring of posts of a social network by a neural network is about 67%.**

Keywords—data analysis, sentiment analysis, natural language processing, social network

I. Introduction

The study of social networks is becoming increasingly important every year due to the growing need to ensure public safety and monitor public sentiment. An analysis of posts can help assess changes in the mood of many users and find application in political and social studies, including consumer research.

Currently, neural networks are used to solve various problems in the field of intelligent data processing. The deployment of a neural network is carried out in two stages.

- Choice of neural network architecture.
- Creation of a training dataset [1].

The training dataset preparation phase takes a lot of time. In many cases, the expert analyzes and generates a training dataset in manual mode and spends a lot of time.

The purpose of this work is to develop an experimental model of a software system for determining the emotional coloring of posts on a social network based on copyright symbols for expressing emotions.

The main tasks are presented below.

- Analysis of the subject area, which includes the determination of the source data for the formation of the training dataset and classes of emotional coloring of posts;
- A review of existing solutions and studies that were proposed by Russian researchers;
- Development of a methodology for the formation of a training dataset, which is based on the methods of linguistic analysis of text information;
- Software implementation;
- Conducting experiments that show the effectiveness of determining the emotional coloring of a post with a trained neural network.

II. Analogs

Currently, the works of Russian researchers offer various methods for the formation of training datasets.

The first method of generating training datasets is described in [2]. The essence of the method is to minimize the training dataset. The training dataset should fully describe the behavior of the model. To minimize the amount of experimental data when training a neural network, it is proposed to synthesize the missing training pairs from a previously constructed mathematical model.

In [3], several methods for the formation of a training dataset are described. The first way is the software generation. The essence of the method is to vary as many parameters as possible during the sampling process.

The second way is sampling. The essence of the method is to set the distribution in the space of objects. This method is used to examine not all data, but only meaningful parts.

The next method is the natural modification of the base object. The training set is obtained by modifying the parameters.

A fourth example is fetching from a database of objects. The bottom line is to group objects into groups. Moreover, the objects of a certain group will be closer to each other, and further from different groups.

In [4], a research prototype of a text tonality analyzer is described, which implements a step-by-step process of text processing. At the first stage, the text is divided into separate sentences, and sentences into separate words. At the second stage, a morphological analysis of each word, lemmatization and determination of parts of speech are performed. The listed stages of the sentence analysis are necessary for the exact matching of the words found to the tonal dictionary. Tonal dictionaries are used for Russian-language text with a volume of about 35,000 words. In the dictionary, each word corresponds to a tonal score. This indicator is a set of five values. Each value determines the degree to which a word belongs to one of the classes: extremely negative, negative, neutral, positive, extremely positive.

Also in the course of work, software systems and modules that perform sentiment analysis of texts were considered. The SentiFinder module [5] defines three types of tonality of Russian-language texts: positive, negative and neutral. Tonality is defined relative to a given tonality object within a single sentence or throughout a document. The average accuracy for the three types of tonality is about 87%.

There are some thesauruses specifically marked out taking into account the emotional component. Such dictionaries are necessary for computer programs in the analysis of the tonality of the text. WordNet-Affect is a semantic thesaurus in which concepts are associated with emotions and are represented using words with an emotional component [6]. WordNet-Affect also uses additional emotional labels to separate synsets according to their emotional valency. To do this, four additional emotional labels are defined: positive, negative, ambiguous, and neutral.

SentiWordNet is a lexical-semantic thesaurus, the first version of which was developed in 2006 [7]. This system is the result of the process of automatic annotation of a set of synonyms by its degree of positivity, negativity and objectivity. Using SentiWordNet provides more than a 20% increase in accuracy compared to the first version [8].

SenticNet is another semantic thesaurus for working with sets of emotional concepts [9]. SenticNet is used to design intelligent applications for analyzing the emotional component of the text. The main purpose of SenticNet is to simplify the process of machine recognition of conceptual and emotional information that is transmitted using natural language. The main difference between the considered thesauruses is that SentiWordNet and WordNet-Affect provide the linking of words and emotional concepts at the syntactic level and do not allow to reveal the semantic component.

Considered scientific works describe only general recommendations for the formation of the training dataset but do not provide methods or algorithms that would allow the formation of a high-quality training dataset for sentiment analysis in an automated mode. The accumulated knowledge in the study of research can be used in the performance of this work.

III. Models and algorithms

The most popular method for creating a training dataset is the selection by keywords and phrases. When using this method, dictionaries of copyright symbols of expression of emotions and dictionaries of key phrases are used.

Dictionaries of copyright symbols of expression of emotions were compiled by an expert. Each dictionary is compiled for a specific emotion and contains several copyright symbols for expressing emotions. Dictionaries of key phrases were found on the Internet and supplemented by analyzing posts on the social network.

At the first stage, posts are selected based on dictionaries of copyright symbols for expressing emotions. As input information, 2.5 million posts from the database are taken. If a post contains an author’s symbol for expressing emotions, then it belongs to a specific class and is added to the corresponding list.

In the second stage, posts are selected based on dictionaries of key phrases. The input information is the lists that were received at the previous stage. At this stage, the lemmatization of each post word is performed. Then the post is checked for the content of each word from the dictionary. If the post contains a phrase, then it belongs to a specific class of emotional coloring. At the output, the data is written into text files, each of which contains a training dataset of a particular class of emotional coloring.

A neural network only works with vectors, so texts must be represented in vector form. To represent the training dataset in the form of vectors, the word2vec algorithm was used [10]. Initially, a list of all the words in the posts is compiled. Previously, all words were reduced to the initial form using lemmatization. Then, vectors are created whose size is equal to the size of the list of all words. After the vector is set to 1 if the word occurs in the post, otherwise 0 if not.

A multilayer perceptron with three layers was used as a neural network. The number of neurons in the first layer is equal to the size of the list of all dictionary words. The number of neurons in the second layer is equal to the size of the first divided by 50. The size of the second layer was selected by conducting many experiments. For a dictionary of 2000 words, the size of the second layer will be 400 neurons. The number of neurons of the third layer is equal to three since we need to determine seven emotions.

After training the neural network, a test set is input. Each post of the set is also transformed into a vector based on the dictionary that was obtained during the training of the neural network.

A. Formal Description of the System

Formally, the process of selecting posts can be represented by a flowchart in Figure 1. The flowchart describes the process of selecting posts for the formation of a training dataset. Each stage of the selection contains the processes of selecting posts for each specific emotion.

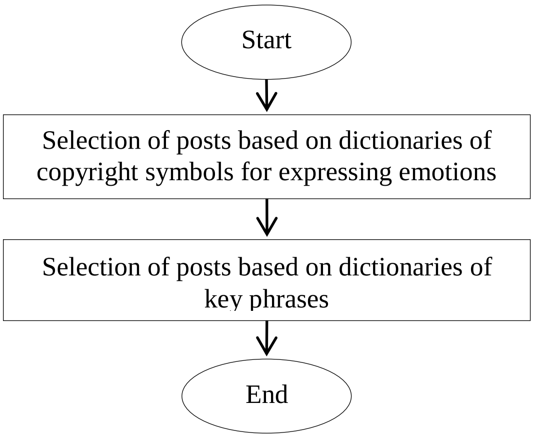
Start
Selection of posts based on dictionaries of copyright symbols for expressing emotions
Selection of posts based on dictionaries of key phrases
End

Fig. 1. Post selection process.

At the first stage, posts are selected based on dictionaries of copyright symbols of expression of emotions for each class of emotional coloring of the text. In the second stage, posts are selected based on dictionaries of key phrases. In the third stage, posts are selected whose length is less than the specified length. A length restriction was introduced because training the neural network in large posts reduces the accuracy of recognition of the emotional coloring of the text [11].

Formally, a lot of dictionaries by which posts are selected can be represented by the formula (1)

$$D = \{D^E, D^W\} \quad (1)$$

where D^E is a set of dictionaries with copyright symbols for expressing emotions, D^W - many dictionaries with keywords and phrases.

In turn, many dictionaries with copyright symbols for expressing emotions can be represented by the formula (2)

$$D^E = \{D^E_{joy} , D^E_{sad} , D^E_{surp} , D^E_{anger} , D^E_{disg} , D^E_{cont} , D^E_{fear}\} \quad (2)$$

where D^E_{joy} – dictionary with emotion «joy», D^E_{sad} – dictionary with emotion «sad», D^E_{surp} – dictionary with emotion «surprise», D^E_{anger} – dictionary with emotion «anger», D^E_{disg} – dictionary with emotion «disgust», D^E_{cont} – dictionary with emotion «contempt», D^E_{fear} – dictionary with emotion «fear».

In turn, many dictionaries with keywords can be represented by the formula (3)

$$D^W =\{D^W_{joy} , D^W_{sad} , D^W_{surp} , D^W_{anger} , D^W_{disg} , D^W_{cont} , D^W_{fear}\} \quad (3)$$

where D^W_{joy} – dictionary with emotion « joy», D^W_{sad} – dictionary with emotion « sad», D^W_{surp} – dictionary with emotion «surprise», D^W_{anger} – dictionary with emotion «anger», D^W_{disg} – dictionary with emotion «disgust», D^W_{cont} – dictionary with emotion «contempt», D^W_{fear} – dictionary with emotion «fear».

Each process of selecting posts for a specific emotion is associated with a dictionary with the author's symbols for expressing emotions of DE and a dictionary of DW key phrases.

The process of testing the training dataset can be represented by a flowchart in Figure 2.

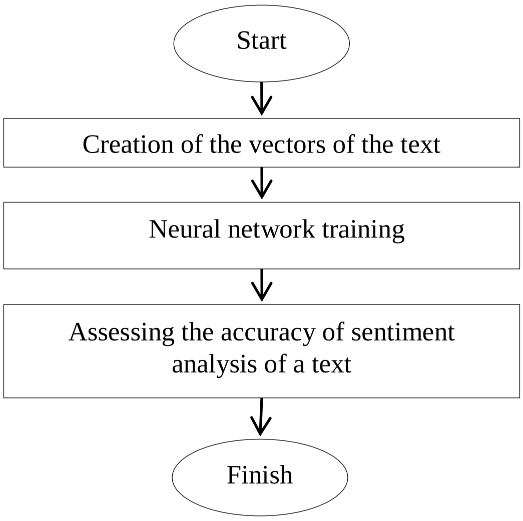
Start
Creation of the vectors of the text
Neural network training
Assessing the accuracy of sentiment analysis of a text
Finish

Fig. 2. Learning set validation process.

At the first stage, a set of vectors is formed using the word2vec algorithm. Next is the training of the neural network. And then an assessment of the accuracy of determining the emotional coloring of the text using a test set.

IV. Software implementation

To evaluate the effectiveness of the developed approach to the formation of the training dataset, a software system was implemented.

The system reads data from the database, dictionaries with copyright symbols of expression of emotions and keywords for each emotion, lemmatization, the formation of a training dataset and training the neural network.

First, dictionaries are read with copyright symbols for expressing emotions, and then posts are selected. After that, dictionaries with key phrases are read, then the posts are lemmatized and the key phrases are selected. Then the selected posts are saved in text files. After the formation of the training dataset, training and testing of the accuracy of determining the emotional coloring of posts by the neural network takes place.

When building the software system, the following libraries were used.

Lucene Russian Morphology is a library of morphological analysis [12]. This library performs a morphological analysis of the word. The library allows you to perform lemmatization of the source word in Russian and get information about part of speech. Lucene uses vocabulary base morphology with some heuristics for unknown words and supports homonyms.

Encog Machine Learning Framework is a machine learning library [13]. The library supports various learning algorithms. The main advantage of the library is the neural network algorithms. The library contains classes for creating a wide range of networks and supports classes for normalizing and processing data for these neural networks. Multithreading is used to provide optimal learning performance on multicore machines.

PostgreSQL JDBC Driver is a library that provides access to the PostgreSQL database [14]. The library provides a connection to the database and interaction with it. As parameters, the library accepts the database address and port, login, and password for the connection. Further, the library receives SQL queries to the database input and returns the data.

V. Experiments

We will evaluate the quality of the generated training dataset as the accuracy of determining the emotional coloring of the text by a neural network.

For the experiments, the following parameters were chosen: a different number of posts in the training set and two methods of text processing - stemming and lemmatization. The accuracy of the system was measured at test posts, each of which belongs to one category.

The quality of the training dataset will be defined as the number of correct conclusions divided by the number of test posts. The experimental results are shown in Table 1.

TABLE I. Stemming and lemmatization experiments

Count posts	Stemming	Lemmatization
20	4/7	6/7
50	6/7	7/7
100	4/7	7/7
200	4/7	7/7
300	5/7	7/7

The experiments performed show that the training dataset, formed with the method of lemmatization, is obtained better than with the method of stemming. Table 1 shows that the accuracy of the recognition of posts by a neural network is much higher when a training dataset is formed using the lemmatization method. The experimental results are also presented in the form of a graph in Figure 3.

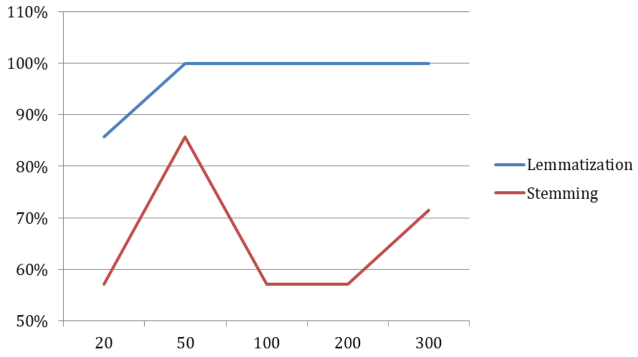

Fig. 3. Stemming and lemmatization.

Additionally, 1,400 posts were submitted to the neural network. 200 posts from each class. The experimental results are presented in Table 2.

TABLE II. EXPERIMENT RESULTS

Emotion	Total	+	-
Joy	200	148	52
Sad	200	154	46
Anger	200	110	90
Surprise	200	126	74
Fear	200	101	99
Disgust	200	151	49
Contempt	200	121	79
Sum:	1400	936	464
Percent:		0.,669	0.331

Experiments show that the neural network correctly recognizes emotion with an accuracy of 67%. Best of all, a neural network determines joy, sadness and disgust with an accuracy of about 75%. The results of the experiment are also presented in the form of a graph in Figure 5.

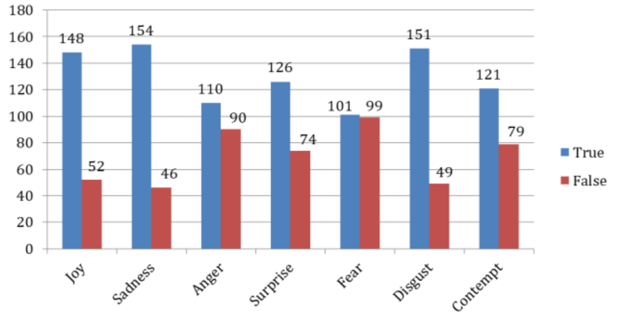

Fig. 4. Experiment results.

VI. Conclusion

As a result of the robots, an expert system was developed to determine the emotional coloring of social network posts. The training dataset is created in an automated mode using dictionaries of copyright symbols for expressing emotions and dictionaries of key phrases. The neural network correctly determines the class of emotional coloring of the post with an accuracy of 67%. The neural network recognizes emotions of joy, sadness and disgust with an accuracy of 75%.

In the future, it is planned to improve the training dataset generation algorithm. Compiled dictionaries will be expanded and updated. To test the set, neural networks of various architectures, for example, deep learning, will be used.

Acknowledgment

This work was supported by the Russian Federal Property Fund. Projects No. 18-47-730035 and 18-47-732007.

References

[1] Yu.V. Vizilter, V.S. Gorbatsevich and S.Y. Zheltov, “Structure-functional analysis and synthesis of deep convolutional neural networks,” Computer Optics, vol. 43, no. 5, pp. 886-900, 2019. DOI: 10.18287/2412-6179-2019-43-5-886-900.

[2] D.A. Grishelenok and A. A. Kovel, “Using the results of mathematical planning of an experiment in the formation of a training dataset of a neural network: article,” Krasnoyarsk: SibSAU, 2010.

[3] I.L. Kaftannikov and A.V. Parasich, “Problems of forming a training dataset in machine learning problems,” Bulletin of SUSU. Series Computer technology, control, electronics, vol. 16, no. 3, pp. 15-24, 2016.

[4] R.V. Posevkin and I.A. Immortal, “The use of sentiment analysis of texts to assess public opinion,” Scientific and Technical Journal of Information Technologies, Mechanics, and Optics, vol. 15, no. 1, pp. 169-171, 2015.

[5] SentiFinder module [Online]. URL: eurekaengine.ru.

[6] Thesaurus WordNet [Online]. URL: http://wndomains.fbk.eu/wnaffect.html.

[7] V. Moshkin, N. Yarushkina and I. Andreev, “The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet,” IEEE 12th International Conference on Developments in eSystems Engineering (DeSE), Kazan, Russia, pp. 576-580, 2019. DOI: 10.1109/DeSE.2019.00110.

[8] Thesaurus SentiWordNet [Online]. URL: http://sentiwordnet.isti.cnr.it.

[9] SenticNet Thesaurus [Online]. URL: https://sentic.net.

[10] Word2Vec Algorithm [Online]. URL: https://neurohive.io/ru/.

[11] I.A. Rycarev, D.V. Kirsh and A.V. Kupriyanov, “Clustering of media content from social networks using BigData technology,”Computer Optics, vol. 42, no. 5, pp. 921-927, 2018. DOI: 10.18287/2412-6179-2018-42-5-921-927.

[12] Library of morphological processing Russian Morphology: Russian [Online]. URL: https://github.com/AKuznetsov/russianmorphology.

[13] Neural network library Encog Machine Learning Framework [Online]. URL: https://www.heatonresearch.com/encog/.

[14] PostgreSQL JDBC Driver Database Access Library [Online]. URL: https://jdbc.postgresql.org /.