Social Media Analysis with Machine Learning

2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) | 978-1-6654-0476-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ElConRus51938.2021.9396713

Adelya M. Khasanova
Higher Engineering School
National Research Nuclear University "MEPhI"
Moscow, Russian Federation
AMKhasanova@mephi.ru

Margarita O. Pasechnik
Department of Computer Systems and Technologies
National Research Nuclear University "MEPhI"
Moscow, Russian Federation
MOPasechnik@mephi.ru

***Abstract*—Social networks have revolutionized the world. We all display personal information on social networks, thereby leaving a digital footprint on the Internet. The analysis of personal information can help companies conduct interviews, as it will give employers a full picture of a person, with a description of his/her personality and social behavior. Among the main problems of the analysis conducted there were the effective working groups formation and the allocation of deviant behavior based on the analysis of information from personal profiles on the social network VKontakte. In the course of this study, the data was collected, pre-processed, analyzed. After that users were organized into groups using machine learning and deep machine learning techniques. The analysis of data from users' social networks was carried out using neural networks and other machine learning methods, the K-means clustering algorithm being used for clustering users by interests.**

Keywords — clustering; k-means; social media; machine learning; text mining

I. Introduction

In the modern world, social networks have become an integral part of everyone's life. We often share sensitive information with other people, tell everyone about our interests, hobbies, so we leave a digital footprint. By analyzing this information, you can get a lot of useful data that can later be used, for example, by employers or teachers. Based on this data, people can be organized into high performance teams where positive atmosphere can help facilitate their work and provide for the effective interaction. It is a great opportunity for employers to identify their future employees during job interviews. In modern realities of the remote format due to the coronavirus, this can simplify the algorithm for selecting eligible employees and reduce the possibility of making a wrong decision. It can also make it possible for schoolteachers and headmasters to assign students to classes in such a way that they could feel comfortable and find like-minded people from young age.

II. Problem Statement

The purpose of this study is to analyze the social network Vkontakte to seek out the best teams and the possibility of determining deviant behavior in the future. To do this, it is necessary to study sources related to the analysis of human behavior on social networks. Then we need to upload all the necessary data of the investigated users' profiles, including the user's groups of interest, content of the publications on his/her personal page, residence, age and other personal information. After that, it is necessary to clear the received data from *junk information* (such as stop words, punctuation, etc.) and only then proceed to the analysis of pure text using neural networks and machine learning.

III. Description of the Research Subject

Behavioral analysis was carried out using the information found on personal user pages under this study. Based on the processing of the digital footprint, users can be divided into specific groups, with similar interests, hobbies, etc. The data from personal pages differs from the ordinary data, since people write there something personal, intimate, something that applies only to them and to their inner worldview. With the help of analysis, we need to understand how people work and communicate with each other, whether they have common topics for conversation, whether they are compatible in character, whether it is convenient for them to see each other in person to discuss business or just personal issues and whether they are able to find a common language in difficult stressful situations. In order to do this, we need to decide what data we need for the analysis, and also combine the list of the data with the capabilities of the library that can be used when downloading data from social networks.

IV. Software Implementation

To achieve the result, the following steps were taken:

A. Uploading data from the VKontakte social network using the vkapi and vk libraries.

B. Preparing data for further analysis (cleaning, forming dataframes, presenting text as a sequence of numbers).

C. Training a neural network to classify the text and generating a sequence of labels for each user (which indicate his interests).

A. Uploading data from a social network using the vkapi and vk libraries.

To begin with, it was necessary to upload data from the social network using the vk and vkapi libraries. These libraries are designed specifically to work with Russian social network Vkontakte, which is used by tens of millions of people throughout Russia and beyond. Almost all data from users' personal pages can be downloaded through these libraries.

For the analysis in this work, the following methods were used and thus the necessary information about users was obtained:

978-1-6654-0476-1/21/$31.00 ©2021 IEEE

TABLE I. FUNCTIONS AND FIELDS TO GET THE REQUIRED DATA

Data	Method	Necessary information
Groups	Groups.Get	id group
	Group.GetById	name, city, country, description, status, activity
User's friends	friends.get	user_id
User's backgroung	users.get	sex, city, country, home_town, education, universities, schools, status, followers_count, relatives, relation, personal, activities, interests, music, movies, tv, books, games
	users.getSubscriptions	name, status

The data is downloaded in json format. An example of a file for the uploaded user data is shown in Fig. 1. Dataframes were formed to work with a neural network for further analysis using machine learning based on json files, (Fig 2).


```
{
  "response": [{
    "first_name": "Lindsey",
    "id": 210700286,
    "last_name": "Stirling",
    "can_access_closed": true,
    "is_closed": false,
    "sex": 1,
    "verified": 1,
    "nickname": "",
    "bdate": "21.9.1986",
    "city": {
      "id": 5331,
      "title": "Los Angeles"
    },
    "country": {
      "id": 9,
      "title": "США"
    },
    "has_photo": 1,
    "has_mobile": 1,
    "interests": "Family, Friends, Dancing, Music",
    "books": "Harry Potter series, The Book of Mormon",
    "tv": "The Ellen Degeneres Show",
    "games": "Village Idiot, Hearts, Who What When Where Why Ho
    e board games/card games)",
    "movies": "Anne of Green Gables, That's Dancing, Freedom Wr
    "Movies:" http://www.youtube.com/user/lindseystomp http://w
    be.com/user/lindseytime",
```


Fig. 1. Example of json file

1	it's time to get up
2	
3	50 favorite children's books
4	kinesis ru a project that is definitely worth
	...
15195	Nastya happy birthday stay bright beuti ...
15196	Happy Birthday
15197	Happy birthday Nastya all the best to yo...
15198	Happy Birthday
15199	Nastya is happy birthday of all the bright...

Fig. 2. Example of dataframe file

When working with these vk and vkapi libraries, the following restrictions within the VKontakte social network were considered:

TABLE II. CONSTRAINTS AND THEIR SOLUTIONS

Constraint	Solution
Private profiles	These users have been removed from the list of researched
Limited number of identical requests per day	A function was written that, when this limitation appeared, the program was put into standby mode for the next day
Limited frequency of identical requests	Methods were added that, when this limitation appeared, program execution was stopped for the required time, and then the data unloading was restored

B. Preparing data for further analysis.

For the further work with the text that is in the dataframe fields obtained in the previous step, you need to clear it from:

- punctuation characters;
- Russian and English stop words;
- tab characters;
- digits.

Also, all the data in the dataframe was checked for NaN values. When uploading data about groups and users and converting them to the .csv format, all fields containing empty strings became NaN fields in the dataframe. The dataframe was analyzed for the presence of NaN fields, and almost 20% of the data had to be deleted from the file, as they could have degraded the quality of the model. (Fig. 3).

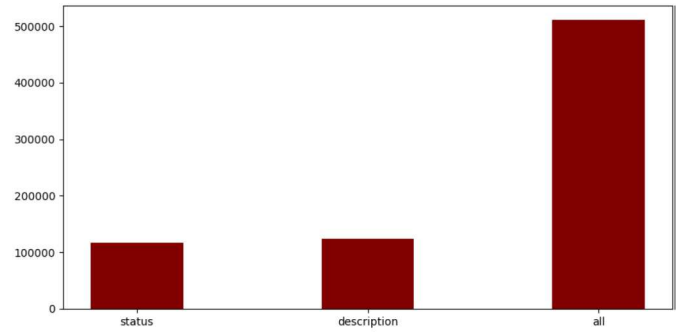

Fig. 3. Number of NaN data in dataframe

Also, in the uploaded text, all words were combined into logically coherent sentences, which means they had one root, but different:

- declensions;
- cases;
- numbers.

To carry out tokenization, it was necessary to lemmatize all words in the dataframe. Lemmatization was carried out using the nltk library package of the WordNetLemmatizer () instance and the lemmatize () function. After conducting lemmatization, incoherent sentences were obtained, however, words in different forms were reduced to the same initial form, which improves the learning of the neural network.

C. Neural network training

Data on the groups of the users under study was used to train the neural network, based on the GRU recurrent controlled neuron. Group names, statuses and descriptions were features, and the activity field was used as a target variable.

A recurrent neural network was built on the obtained data. It was applied to user data to assign targets to a field that stores the texts of all posts on a given user's page. First, the text was prepared for building the model using the Text Preprocessing Tokenizer utility class to vectorize textual data, turning each element of the text either into a sequence of integers (where each integer is the index of the token in the dictionary), or into a vector in which the value of each token can be binary, or can be presented based on “a Bag of Words” method. We indexed all the words in the dataframe (Fig. 4). Due to the fact that the main audience of the Vkontakte social network is Russian-speaking, the analyzed data were also downloaded in Russian, but for convenience they were translated into English.

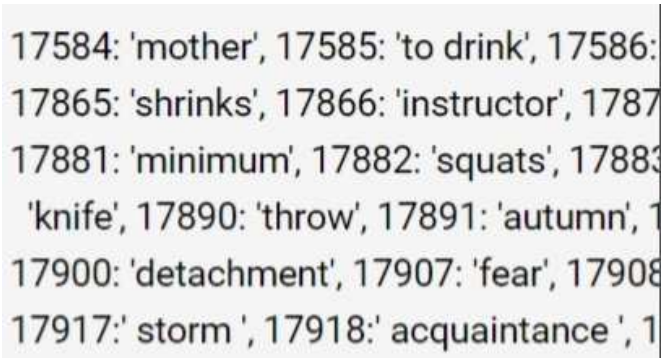

Fig. 4. Indexing words in a dataframe using tokenization

The model was built based on tokenized text, the target variable was the group subject (activity field). 400,000 records were used to train neural networks.

Looking at the statistics for all available classes, we can conclude that there are about 500 classes in the data. But most of them are represented by a small number of copies. To balance the dataset, 100 classes were selected, which occurred in the dataframe more than 1000 times.

The model was first trained on raw data, where there were extra characters and fields of the NaN type. The proportion of correct answers of the constructed model on the training set is 70%, on the test set - 60%. (Fig. 5).

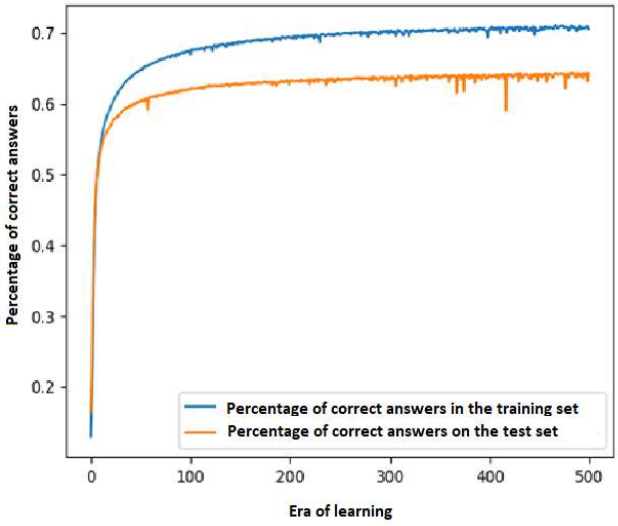

Fig. 5. The Result of training model on on raw data

The proportion of the correct answers on the cleaned data was increased both on training values and on test values up to 80% and 70%, respectively (Fig. 6).

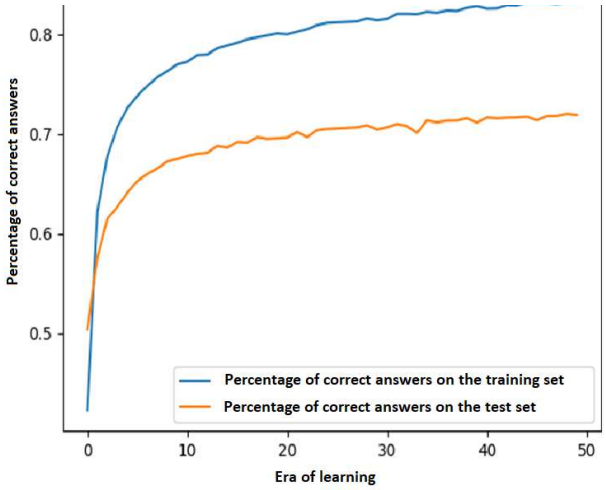

Fig. 6. The result of training the model on cleaned data

D. Clustering based on assigned user interest labels

After processing the data and assigning certain target variables to all user fields, the dataframe was arranged in the following form:

TABLE III. .DATAFRAME FOR CLUSTERING

id	groups' targets	posts' targets	sex	city	...
id 1	target1, target2, target3	target1, target2, target3	woman	Moscow	...

The next stage is to carry out the clustering itself on the prepared dataframe. The k-means algorithm was used, that allowed to determine effective commands from the data received from the VKontakte social network.

V. CONCLUSION

The purpose of this project was to analyze the social network Vkontakte in order to determine the best teams and to identify possible further deviant behavior. It showed how to prepare data for the necessary analysis and then do the clustering. In the future, the possibility of such a division into effective teams can be implemented in practice, for example, in various business companies or schools.

REFERENCES

[1] Sozykin A. V. Review of methods of teaching deep neural networks // Bulletin of the South Ural State University. Series: Computational Mathematics and Informatics. - 2017. - T. 6. - No. 3.

[2] Jones K. S. What is the role of NLP in text retrieval? //Natural language information retrieval. – Springer, Dordrecht, 1999. – C. 1-24.

[3] Korenius T. et al. Stemming and lemmatization in the clustering of finnish text documents //Proceedings of the thirteenth ACM international conference on Information and knowledge management. – 2004. – C. 625-633.

[4] Ramadhani A. M., Goo H. S. Twitter sentiment analysis using deep learning methods //2017 7th International annual engineering seminar (InAES). – IEEE, 2017. – C. 1-4.

[5] Alsayat A., El-Sayed H. Social media analysis using optimized K-Means clustering //2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA). – IEEE, 2016. – C. 61-66.

[6] Korgutlova N. A. et al. Formation of academic groups and project teams based on collecting data about students // Electronic libraries. - 2018. - T. 21. - No. 3-4. - S. 193-208.