Identifying Life Satisfaction of Social Network Users

Evelina Bronnikova de Menezes[1]

[1] Patrice Lumumba Peoples' Friendship University of Russia, Moscow, Russia

evelina.menezes@gmail.com

Abstract. Well-being is an integral aspect of the research aimed at supporting and improving the quality of life. Life satisfaction is not only one of its contributing factors, but also one of the most popular ways to assess the quality of an individual's life. The field of psychology has been developing for many years, collecting and analyzing people's behavior. Now, more than ever before, we have the means to further expand this research. The Internet age is capable of providing an unimaginable amount of data that can further improve the understanding of human nature and artificial intelligence provides the necessary means to analyze it. In this paper we apply various methods of artificial intelligence, such as machine learning and topic modelling, to identify the psychological characteristics of VKontakte social network users using the Satisfaction with Life Scale. The results of the experiment show that a better accuracy is achieved by applying binary classification methods to our data and that socially active users tend to display a higher level of satisfaction with life.

Keywords: Satisfaction with life, Social Networks, Machine Learning, Topic Modelling.

1 Introduction

The use of artificial intelligence to study online users is becoming more popular every year. The developed methods are often used for commercial purposes, to study the interests and critics of the audience in order to improve marketing strategies, for example. However, it is also noticeable the growth in the use of this technology to analyze more nuanced aspects such as psychological disorders and subjective well-being (SWB). Satisfaction with life (SWL) is a subjective assessment of a person's quality of life, depending on a group of internal and external factors [1-5]. It is also considered one of the components of SWB, often used as a measurement of the latter.

In this paper, we apply different artificial intelligence methods to identify the psychological characteristics of users in the Vkontakte social network using the Satisfaction with Life Scale (SWLS), as well as evaluate the resulting predictions.

In Section 2, related works are reviewed, in Section 3 we present our dataset of Vkontakte users, in Section 4 we describe our methods and in the last sections, we present and discuss the results of the experiments.

2 Related Work

Over the years multiple approaches were taken to study the different aspects of the human psyche, varying from machine learning methods to neural networks. The standard first step is to collect user data which can be divided into general data – text messages, likes on social media, etc. and psychological data – for example, test results from different screening tests. The analysis of user texts is usually conducted using language patterns that predict SWB. These patterns can be formed by an open or closed dictionary approach. The task of predicting the results of screening tests, especially SWLS, is often solved by regression algorithms, while classification methods are usually used to create additional input features. Among other algorithms, Random Forest regression (RFR) and Elastic Net regression stand out in this area as the more popular and efficient.

The RFR algorithm is chosen because of its interpretability, nonlinear assumptions, efficiency and accuracy. RFR uses a set of decision trees, the results of which are combined into one final result. Its principle is to build several binary decision trees by bootstrapping samples randomly selected from the training dataset. Each tree is trained on a randomly selected sample of training data, and predictions are made by a majority vote of these trees. Thus, RFR is able to minimize errors due to bias and variance. In turn, the Elastic Net regression method combines Ridge regression and Lasso regression, combining their penalty coefficients.

An example of the presented approach is in Collins et al. [5] where the following features were considered: static ego (Big Five[1], age, number of friends, likes, etc.), temporal ego (data calculated using LIWC, reflecting the user's sentiment at the time of writing the text), link features (SWL scores for couples and friends); and RFR was used to predict the SWL of the users.

In a study by Chen et al. [6], the dataset of which included SWL results and Facebook status updates, Elastic Net regression was used to select informative features among the topics of feelings generated by LIWC and LDA for a random forest model. Texts after preprocessing (removal of stop words and links, conversion of emoticons into words, etc.) were subjected to the calculation of polarity - evaluation of positivity (+1) or negativity (-1) of the post according to a predefined list of words valence. The result is the sum of the valence words divided by the number of status updates for each user. The texts were also subjected to topic clustering using LIWC, followed by feature reduction for extraction.

The already mentioned SWLS is a screening questionnaire consisting of 5 statements evaluated by a seven-point system. It was designed for mass surveys of respondents about the level of subjective satisfaction with their lives. SWLS was proposed in 1985 by Diener et al., the scale was adapted and tested into Russian by Leontiev and Osin in 2003 [7-10] and allows us to calculate the SWL score on a 0-35 scale.

As mentioned above, topic modeling methods are most often used to extract features, as for example in the above-mentioned study by Chen et al. or in Gkotsis et al. [11], where topic modeling was used to classify data from the social network Reddit that was related to different types of mental disorders. Thus achieving more than 80% accuracy

[1] five broad dimensions of human personality

in binary classification. There are other successful examples of using this method to study the mental state of users in social networks such as Twitter and Facebook [12].

Nevertheless, even without applying topic modeling to predictive models, this method can be a very convenient resource for studying a dataset. This method allows, for example, to observe the prospect of a field of study through the consideration of mentions of that field in documents of various periods [13].

3 Dataset

The volunteers of this study were required to take a SWL test and provide the results as well as general information about themselves such as age (12-79), gender and text messages (2007-2019). Thus was collected the information of 1908 users, from which:

- 1340 completed the SWL test;
- 1659 provided their general data;
- 249 completed the test, but did not provide general data;
- 568 submitted general data, but not the test scores;
- 1091 completed the SWL test and provided general data.

As the scope of our interest is the investigation of SWL score and profile of the user, we cleaned the dataset from users that did not provide this score and the required general data. Afterwards we removed all non-Cyrillic words, punctuation and stop words from users' text messages, then tokenized and lemmatized the text using the nltk library and pymorphy2 analyzer [14]. This process left some users without textual messages, which led to another cleanup. The data before any changes is referred as *initial data*, the data after the first cleaning as *cleaned data*, and the data after the text processing as *pre-classification data* (see Table 1).

Table 1. Dataset statistics at various stages of preparation

Observed data	Initial data	Cleaned data	Pre-classification data
Number of users	1908	1091	1069
Males	427 (31.87%)	336 (30.8%)	320 (29.93%)
Females	913 (68.13%)	755 (69.20%)	749 (70.07%)
Age	24.73 ± 6.83	25.08 ± 6.91	25.10 ± 6.86
SWLS score	15.81 ± 6.71	16.06 ± 6.77	16.07 ± 6.76
Total number of posts	112573	81706	81675

Note: The numbers are presented as a mean value ± standard deviation.

The next step was the further analysis of the dataset in search of possible useful features. As such, the feature "post frequency" (number of text messages per user) was created. It is worth noting that the more significant correlations were observed between message frequency and SWL score (r=0,088), as well as message frequency and age (r=0,238).

In preparation for the classification task were created two copies of the pre-classification dataset. One for binary and the other for ternary classification. For binary classification, SWL scores were converted to 0 (dissatisfied), if 0 ≤ SWL score ≤ 17 and 1 (satisfied), if 18 ≤ SWL score ≤ 35. As for ternary classification, the scores were classified as one of the following:

- 0 (dissatisfied), if score 0-14;
- 1 (satisfied), if score is 15-25;
- 2 (very satisfied), if score is 26-35.

If we look at the percentage in the resulting set according to the ternary classification: 47.61% are not satisfied, 42.1% are satisfied and 10.3% are very satisfied. And in the binary set, 61% are not satisfied, while 38.96% are satisfied.

It was also of interest to see if gender has any influence in the results.

Table 2. Comparison of satisfaction between men and women

	Male	**Female**
Mean SWLS score	16.16	16.02
Binary classification: dissatisfied, satisfied (%)	61.25, 38.75	60.88, 39.12
Ternary classification: dissatisfied, satisfied, extremely satisfied (%)	49.38, 39.69, 10.94	46.86, 43.12, 10.01

As we can see in Table 2, the parameters are very close in value, but women appear to be slightly more satisfied than men in the binary and ternary sets. This observation may be explained by the fact that women evaluate situations differently from men. Because of this, a woman's score may be higher than a man's, even under the same circumstances [15].

4 Methods

4.1 Features

To create additional features, we used the polarity method [6]. We determined the polarity of each word or expression by using the free polarity dictionaries RuSentiLex (2017) [16] and KartaSlovSent [17]. Their contents were combined and the score was converted to +1, 0 or -1 accordingly. As result we got positive, negative and overall polarity.

4.2 Topic modeling

In order to find and study possible common trends or interests of users from different satisfaction categories we created topics. For this we used LDA models from the gensim library, which is easy to use and allows the visualization of topics as bubbles.

The tests were carried out on 3 groups from the dataset formed according to the ternary classification criteria:

1. general (texts of all users);
2. positive (texts of users satisfied with life);
3. negative (texts of users who are not satisfied with life).

At the first stage of topic modeling, only 2 topics were set. They can be described as "Public life" (topic 1) - mentions of politics and public events; and "Personal life" (topic 2) - congratulations and more mentions of positive terms like "love" and "good" (see Table 3).

Table 3. The ratio of words used in each topic, depending on the group

Group	Topic 1	Topic 2
General	61.3%	38.7%
Positive	92%	8%
Negative	50.3%	49.7%

When 10 topics were set, it was noticed that the negative group had fewer significant topics than the positive one. While for the positive group we could distinguish 5 topics with a more or less uniform amount of use, the negative group had only 3, where the first was much more common than the second and third. This distribution suggests that users from the positive group are much more involved in public life than users from the negative group.

In the second stage of these trials we compared the following topic models: LDA, LDAMallet, TF-IDF and HDP. LDA models are based on variational Bayes inference and are considered optimal because of their speed, but not the most accurate. LDAMallet models use Gibbs sampling and are more accurate than LDA. TF-IDF models use a TF-IDF corpus instead of the matrix of terms as input data for the LDA model. HDP models can be considered as an extension of the LDA model, as they also use Dirichlet placements, but unlike the LDA do not require specifying the number of topics. The selection of the optimal number of topics was based on the calculation of the coherence value (Coherence Model) for models with 2 to 40 topics. The general group served as input data. The last highest result before the first decrease in value was selected as the best for each model. For all models except HDP (150 topics) the optimal number of topics was 14 with the coherence value being LDA=0.3048, LDA(TF-IDF)=0.3591, LDAMallet=0.3874 and HDP=0.4926. Despite the HDP model performing better, it should be kept in mind that an increase in the number of topics entails an increase in the number of overlapping topics, which complicates the process of describing and differentiating each.

Further analysis showed that 324 out of 1069 documents had less than 100 words. In other words, 30.3% of the documents. Observations also showed that among the most frequent words in the dataset there were those that did not carry a significant semantic meaning or because of their frequency, negatively affected the creation of topics (being present in almost all topics). After the removal of some of those words, the previous experiment was repeated once more (see Table 4).

Table 4. Results of different models

Models	**LDA**	**LDA (TF-IDF)**	**Mallet LDA**	**HDP**
Number of topics	8	8	8	150
Coherence value	0.4107	0.5127	0.408	0.5145

The frequency of studied topics with a threshold exceeding 0.3 showed that in general, the most popular topics in VKontakte messages are congratulations.

At last, 4 simple topics inspired by those observed in the previous steps were manually created: “life” – general quotidian terms, “citizen” – words related to politics and public events, “gratitude” – wishes and felicitations, “vacation” – words related to free time such as games and movies.

4.3 Models

Separate models were built for classification and regression. From the initial list of algorithms, one or several that showed a higher result compared to the others were selected. After that, the chosen models were tested with different sets of features and the result was compared once more. The different features were grouped as follows: general info (age, gender, post frequency), text (users’ textual messages), SWL score, polarity (positive, negative and overall), topics.

The dataset was divided into training, test and validation (80/10/10) and vectorized using TF-IDF, which calculates the significance of a word based on the its frequency of occurrence in one document and in the entire corpus. The models were built using a pipeline consisting of a transformer (ColumnTransformer[2]) and different models as estimators. After that, the model with the best result was selected.

At the intermediate stages, the best results in classification were shown by support vector machine and random forest models. In the regression task, ridge regression, elastic net and random forest were the best performing.

5 Results

The distribution of the created topics in section 4.2 showed that for binary classification, the biggest difference in the frequency of use falls on the topics "citizen" and

[2] https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html

"gratitude". While in the ternary classification, the topic "vacation" is added to this list (see Tables 5 and 6).

The results obtained support the theory that more satisfied users have a greater interest in social activities, since the biggest difference between opposite groups is reflected precisely on the topic of public interests - "citizen". The reason for this may be a sociopsychological factor, in particular extraversion – a personality trait that affects a person's need to search for social interactions, which are important to humans as social animals. The explanation for this result may also be related to the concept of satisfaction, as the fulfillment of public expectations or duties. A closer connection with public life may also be reflected in the “gratitude” topic, since it is assumed that a social person has a wider circle of friends and acquaintances. The significance of the topic "vacation" may hint at the ability of a satisfied individual to rest more often and thus maintain a balance between work and rest.

Table 5. Binary classification users who used the created topics, %

Topic	**Users**	
	Dissatisfied	Satisfied
Life	78.68	83.45
Citizen	40.79	49.64
Gratitude	66.72	75.54
Vacation	72.54	79.38

Table 6. Ternary classification users who used the created topics, %

Topic	**Users**		
	Dissatisfied	Satisfied	Very satisfied
Life	78.99	82	86.36
Citizen	38.11	49.11	52.73
Gratitude	65.22	73.78	78.18
Vacation	71.9	76.22	86.36

As for the models created in section 4.3, results showed that for the ternary classification the features text, general info and polarity combined had the most positive effect, unlike binary classification, where the feature general info practically does not affect the final result. This was probably caused by the obvious imbalance of the third category ("very satisfied") in the ternary classification compared to the other 2. The topics created in 4.2 showed a slight improvement in classification and prediction results for the worst performing models, but not for the best models (see Table 7).

Table 7. Comparison of the best results obtained with and without topics as features

Features	Type	Model	Results			
			MAE	Accuracy	RMSE	R-2
Text + Polarity + Topics	BC	RFC	0.3627	0.6373	0.6022	-0.6
Text + Polarity	BC	**RFC**	**0.316**	**0.6839**	**0.5622**	**-0.3945**
Text + General info + Polarity + Topics	TC	SVC	0.4663	0.5751	0.7411	-0.397
Text + General info + Polarity	TC	**RFC**	**0.4611**	**0.5855**	**0.7446**	**-0.4102**
Text + General info + Polarity + Topics	R	ElasticNet	5.5631	-	6.5701	-0.0217
Text + General info + Polarity	R	**ElasticNet**	**5.5308**	**-**	**6.6493**	**-0.0465**

Note: BC – Binary classification, TC – Ternary classification, R – Regression.

Nevertheless, it is noticeable that the results for ternary classification and regression using topics and without using them are not so different as in the case of binary classification. Perhaps this indicates that topic modeling has a greater influence on the result of these former approaches. Therefore, it can be assumed that in case of improving the topic modeling, such a mixed approach can lead to a better result than the one we got now.

6 Conclusion

In this study, we considered multiple approaches to the identification of the psychologic characteristics of VKontakte users, mainly the prediction of SWL score and detection of influential factors. For prediction, we used regression and classification models. For the behavior analysis were created several topic models.

The investigation showed that binary classification using textual messages and polarity scores performed better, than the others. While the results of topic modelling indicate that users who are more satisfied with life tend to pay more attention to the public and are presumably more socially active than the dissatisfied.

In future work, we plan to explore the possibility of using neural network approaches and language models to improve the quality of determining the level of life satisfaction among Russian-speaking users of the VKontakte social network.

References

1. Hall, A.: Life satisfaction, concept of. In: AC Michalos, ed. Encyclopedia of Quality of Life and Well-Being Research, pp. 3599 – 3601. Springer Netherlands: Dordrecht (2014). https://doi.org/10.1007/978-94-007-0753-5_1649
2. Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L.: Subjective Well-Being: Three Decades of Progress. Psychological Bulletin 125(2), 276-302 (1999).
3. Hayes, N., Joseph, S.: Big Five correlates of three measures of subjective well-being. In: Personality and Individual Differences 34(4), pp. 723-727 (2003).
4. Andreenkova, N. V.: Comparative analysis of life satisfaction and its determining factors. In: Monitoring obschestvennogo mneniya, 5(99), pp. 189-215 (2010).
5. Collins, S., Sun, Y., Kosinski, M., Stillwell, D., Markuzon, N.: Are You Satisfied with Life? Predicting Satisfaction with Life from Facebook. In: Social Computing, Behavioral-Cultural Modeling, and Prediction. 8th International Conference, pp. 24-33. SBP 2015. https://doi.org/10.1007/978-3-319-16268-3_3
6. Chen, L., Gong, T., Kosinski, M., Stillwell, D., Davidson, R.L.: Building a profile of subjective well-being for social media users. PLoS ONE 12(11): e0187278 (2017). https://doi.org/10.1371/journal.pone.0187278
7. Psylab.info - psychodiagnostics encyclopedias. Life satisfaction scale, https://psylab.info/Шкала_удовлетворённости_жизнью, last accessed: 2022/04/28.
8. Psylab.info – psychodiagnostics encyclopedias. Life satisfaction scale, https://psylab.info/Шкала_удовлетворённости_жизнью/Бланк, last accessed: 2022/04/28.
9. Osin, E.N., Leontiev, D.A.: Approbation of Russian-language versions of two scales of express assessment of subjective well-being. In: Materials of the III All-Russian Sociological Congress. Moscow: Institute of Sociology of the Russian Academy of Sciences, Russian Society of Sociologists (2008).
10. Luhmann, M., Lucas, R.E., Eid, M., Diener, E.: The Prospective Effect of Life Satisfaction on Life Events. In: Social Psychological and Personality Science 4, pp. 39-45 (2013).
11. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J.P., et al.: Characterisation of mental health conditions in social media using Informed Deep Learning. Sci Rep 7, 45141 (2017). https://doi.org/10.1038/srep45141
12. Tadesse, M., Lin, H., Xu, B., Yang, L.: Detection of depression-related posts in reddit social media forum. In: IEEE Access, vol. 7, pp. 44883-44893 (2019).
13. Liu, S., Zhang, R., Kishimoto, K.: Analysis and prospect of clinical psychology based on topic models: hot research topics and scientific trends in the latest decades, In: Psychology, Health & Medicine 26(4), pp. 395-407 (2021).
14. Morphological analyzer pymorphy2, https://pymorphy2.readthedocs.io/en/stable/, last accessed 2023/04/17
15. Montgomery M.: Reversing the gender gap in happiness. In: Journal of Economic Behavior & Organization 196, 65-78 (2022).
16. Kotelnikov, E., Peskisheva, T., Kotelnikova, A., Razova, E.: A comparative study of publicly available Russian sentiment lexicons. In: Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings 7, pp. 139-151 Springer International Publishing (2018).
17. Kulagin, D.I.: Publicly available sentiment dictionary for the Russian language KartaSlovSent. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”[Komp’yuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”] (20), pp. 1106-1119 (2021).