Towards a Retrospective One-Class Oriented Approach To Parents Detection In Social Media

Alexander Egorov, Timur Sokhin, Nikolay Butakov
ITMO University
St. Petersburg, Russia
al.g.egorov@gmail.com, 245591@niuitmo.ru, alipoov.nb@gmail.com

***Abstract*—Social media is the source of data for different purposes: advertisement, social study, human recruiting. However, usually, we are limited to readily available, structured information: age, gender, education, occupation. We have to work with unstructured data such as texts related to a user if we want to extract more complex, implicit features. We show the case of complex user analysis in social media using textual data. The task we solve is detecting parents on social networks. Our approach works with content that is not generated by a user, but with the content, the user was interested in implicitly - the user liked, or explicitly - the user subscribed to a group, where the content was published. In this paper, we compare classification methods for the task of parents detection on social media. Using mentioned above user's likes and other information it is required to estimate chances if a user has got a child or children already or not. This task is an example of positive-unlabeled learning: data from social networks and media may contain explicit signals about users' parenthood but there is no ground to make a backward conclusion. It can be considered as a case of look-a-like modelling or in other words a one-class classification problem. We propose a retrospective approach that can exploit data from social media to allow building a binary classifier. We compare both these approaches and conclude that the retrospective approach albeit requiring more efforts to be implemented may yield better results. This approach may be useful in similar tasks having look-a-like problem statement.**

I. Introduction

Multiple stockholders are interested in targeting users by their social, economic and other stratification. Particular examples of such stockholders are advertisement, social study and human recruiting. The stratification can be done using social media, which is involved in all fields of human activity. Online social networks are known for their ability to represent users from different sides: there is a massive amount of data about users' behaviour and interests; but the data is noisy, distorted and incomplete. Data complexity is an obstacle between synthetic machine learning achievements and real-world applications.

An example of such a problem that we are faced with is the task of parents detection on social media. It requires knowledge about user's posts, likes and other activities. The social status identification is an example of positive-unlabeled learning: we can be sure that users are parents if they have some explicit features, but if they do not have, they can be either parents or non-parents. Some users indicate that they are parents and they can be used as a positive class. On the other hand, we have users without those explicit features, but we cannot use them as non-parents. Moreover, features, which indicate that the parental status of users are rare, and it is hard to detect them. We analyze two possible solutions:

- building One-Class classifier with a parental kernel;
- binary classification using retrospective data.

The retrospective classifier training assumes that we know exactly when a user changed his/her social status. We can split user activity into two periods and reduce the problem to positive-negative classification. The same approach can be applied to other classification problems or analyses in social media. The main idea is to keep the user's preferences, behaviour and to avoid bias in the dataset. However, we should take into account the possible bias of the social media itself: the size of the community, the target audience. Contributing to this work is the following:

- method for classifying users with an explicit positive class and the absence of a negative class based on retrospective user data;
- the experimental study of the proposed approach.

In Section 2, we analyze related works. In Section 3, we describe how to build a dataset using a retrospective approach, build user representations, and describe the classification methods we use. Section 4 presents experimental studies and basic results. In Section 5, the results are analyzed, and the conclusion about the prospects for using this approach is made.

II. Related works

The task of social or economic status identification in different social media is essential for a great variety of applications.

As it was shown by [1] the number of women who prefer to search for information on social networks is increasing. The knowledge of their statuses makes it possible to provide more appropriate content or for bringing them together in communities. In some works, researchers analyse an income level of people or even food preferences based on their accounts on Twitter [2], [3]. They use specific features for each set of users, label them and then train a simple Gaussian Process [4]. Prediction of political loyalties is also very popular [5], [6].The task of parents detection in social media is not well studied, but some works investigated the dynamic of a user's behaviour changing related to their postpartum status [7], [8]. They show that some mothers demonstrate significant changes expressed in linguistic style, engagement and emotions.

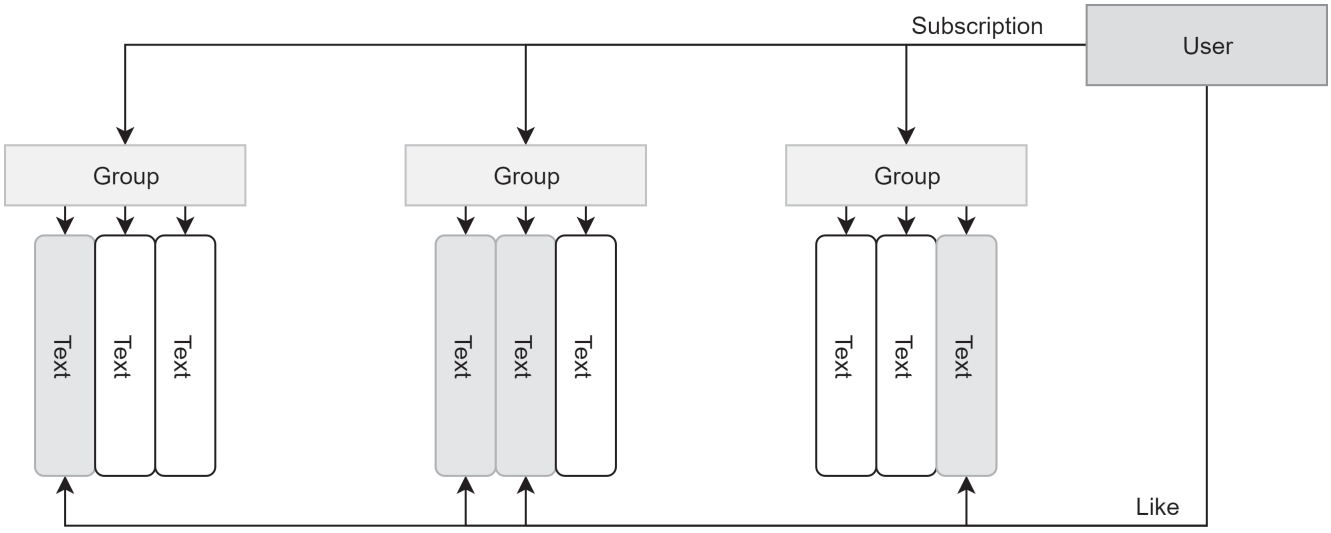

Fig. 1. User-text possible relations

A large number of works are devoted to Predict Personality Traits on social networks [9], [10], [11]. In [10], the authors classify Facebook users using the five-factor model of personality. To train the model, they use the user profile data, as well as information about the total number of likes, posts, etc. The disadvantages of this approach include that the topics of liked texts and posts are not taken into account. The authors [12] classify users based on text content generated by the users themselves. Authors use Latent Dirichlet Allocation (LDA) as a topic model. In [9] pre-extracted linguistic features from LWIC, SPLICE are used and SNA to infer personality traits.

In [13] authors attempted to identify Starbucks-lovers. They labelled users with a clear indication of love for Starbucks in their profiles and sampled random profiles as a negative class. In its core, they solved the positive-unlabelled problem [14]. We can see, that straightforward approach, when we use only profiles data, gives us a low recall, while textual data allows us to achieve high precision and recall at the same time. However, their approach with negative sampling is not applicable in the case of parents detection - the ratio of Starbucks-lovers to other people is much less than the ratio of parents to non-parents.

As mentioned approaches to user classification based on explicitly set of features have one drawback - positive and negative classes may be too different and can be easily separated. We want to build the dataset, which will allow the classifier to detect only parental features of the user without any other biases.

III. The approach

A user's social media profile is a partial reflection of a real human. Some users do not provide complete information about themselves. Therefore, not all users can be found by the completed profile fields. In the profile, there is a field about the presence of children. Parents can be found through it, but users cannot be considered as non-parents if they left this field empty.

Usually, users do not set this profile field, and we should use other activities. Mostly, global social media assumes that texts and images are the main content generated by their community. In this work, we focus on the relation between a user and texts. The user connects to the post that he liked, and to the group in which it was published.

A. Dataset

The dataset building is not trivial in our case. The full pipeline is provided in Fig. 2. For data collection, both we use the following constraints: we collect users with explicit *only one child* status on their pages; we collect users with children under 15 years of age.

In this research, we detect only parents with one child in the family. The task of detecting parents with two or more children is more difficult, because in the same year parents can like posts related to children's education and posts related to pregnancy and childbirth, which will eventually lead to classification errors. The information about children's ages is required to be able to collect activities of these users from the past and perform the analysis. Our way to collect non-parents is retrospectively. For time-dependent data (likes, posts, comments), we can look into the past of parents and use them as non-parents - Fig. 2. The problem here is to avoid the bias of data in terms of size and topics: social media grows all the time, new people join. Details about this part are provided below.

B. Retrospective

Using the approach represented in Fig. 2, we collected a set of users, which are parents at the current moment. These users also represent non-parents in their past. We choose a year in the past when part of users were parents and the other - non-parents and collect only their activities in this year. The splitting allows us to build a representative dataset with reliable classes in the same period. Because all non-parents are obtained from current parents, the closer the selected year is to the current year, the fewer non-parents will remain according

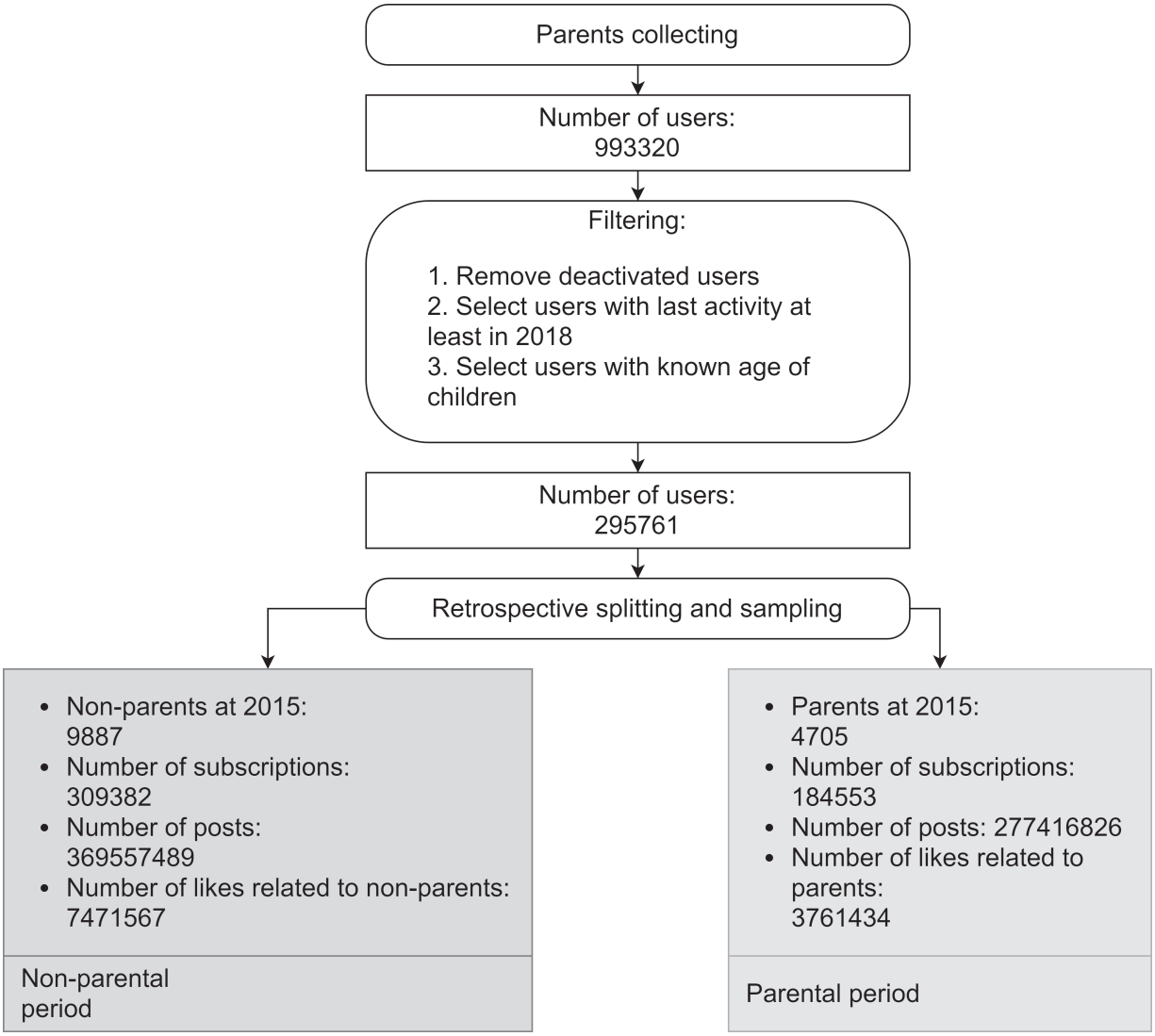

Fig. 2. The pipeline of the dataset building with retrospective user splitting

to the selected split. In our case, the best choice was to go back two years to get a training dataset, based on which it will be possible to classify current users of the social network based on actual data. But because we want to demonstrate the effectiveness of this approach, we took 2015 for the training dataset, and 2017 for the test one. Training and test datasets consist of different users. We did not collect users data before 2015, because there are not enough users with high activity and required conditions. The data after 2017 was not used due to few users reporting a baby after 2017 (presumptive non-parents). The validation dataset shows how well the classifier would work if it were possible to find a reliable sample of parents and non-parents for the current year.

For these users, we collected data from the Russian social network VKontakte for the year of 2015 (2017): comprising of texts that they liked in that period. For collecting data, we used a data crawler for social media, which was described in [15]. We do not use posts written by users themself due to a large number of reposts and picture-only publications.

As it was shown in [1], the difference in a user activity becomes significant over a three-month period. So we are interested in users who became parents before 2014 (2017) and users who became parents after 2016 (2018). Our dataset building pipeline is the following:

- select parents with the known ages of their children - explicit parents;
- select users with a year of childbirth before 2014 and after 2016 (before 2016 and after 2018 accordingly for the test);
- calculate sex and ages proportions on all of the parents;
- split them into two groups of age per sex: 0-24 and 24-40.

C. User representation

We approach the representation of users as a natural language processing task and solve the problem of interpreted

texts vectorization using topic modelling (TM). There are three levels of vectorization.

Post-level vectorization (set of vectorized "liked" texts) - each text, which is related to a user ("liked" by him), is converted into a vector using TM. TM inferring gives us a probability for each topic. The probabilities are combined into a vector. Vectors are combined into a set.

Groups-level vectorization (set of vectorized descriptions of the subscriptions in which a user set "likes") - we select subscriptions groups with at least one user activity ("like") and convert descriptions of these groups into vectors using TM.

Interest-level vectorization includes (vectorized combined "liked" texts; vectorized combined descriptions of subscriptions in which a user set "likes"):

- set of posts, which were "liked" by a user, combined into a single text corpus. This corpus is converted into the vector using TM,
- set of descriptions of subscriptions with at least one user activity combined into a single text corpus. This corpus is converted into the vector using TM.

The last two levels of vectorization are useful if we want to reduce the influence of an unbalanced number of likes of different users. Also, it allows reducing the amount of data to process and combine. It becomes vital in the case of thousands of posts for each user.

D. Texts vectorization

is performed using an ARTM (Additive Regularization of Topic Models) [16] approach, which builds multi-objective models by adding the weighted sums of regularizers to the optimization criterion. This approach is more flexible than LDA-based methods and makes it possible to determine background topics or less frequent topics [17]. The main advantage of the topic modelling is the possibility to interpret vectors in a straightforward way. Example of topics is provided in Table I.

We use regularization to detect not only main topics but also background topics about parents and children, which consist of frequently used terms and which would be ignored by LDA. Some creators of groups on social media pointed out the categories to which the groups belong. We use the BigARTM library [16] to create the topic model. For the training dataset, we collected 15000 descriptions of groups with following tags: 'Products for children', 'Parents and children', 'Pregnancy, childbirth', 'Kindergarten', 'Baby food', 'Baby clothes and shoes'. For the summary of the used parameters is provided in Table II. There are 100 main topics and 10 background topics, so the size of the vectors is 110.

TABLE I. TOPICS EXAMPLES

Type and number	Terms
Main 32	kindergarten, mentor
Main 41	mother, family, parent, love
Main 66	kid, toy, newborn
Main 73	school, learning, study
Back 2	child, children's, home, game

TABLE II. TOPIC MODEL TRAINING PARAMETERS

Parameter	Value
Main topics number	100
Background topics number	10
Epochs for the training	29
Smooth-Sparse Phi Regularizer	0.5
Smooth-Sparse Theta Regularizer	0.5
Decorrelator	1e-6

E. User classification

Since there are no known works on this task, we need some baseline. As it mentioned in the Introduction, it is impossible to collect a representative set of non-parents with the same activities, social statuses, etc. All parents have common interests related to the upbringing of children, and thus their Interest-level vectors can be located close to each other in the feature space, so we apply the one-class classification. One-Class Supports Vector Machine [18] requires only "positive" class, and the further inference is applied in terms of the distance to the kernel.

The retrospective approach allows working in a binary domain. We use a multilayer perceptron which is a simple but an effective classification method. To use this method, we merge texts into one. In addition, we use a long short-term memory (LSTM) [19] neural network, which allows each text to be used as a separate feature. Despite the fact that the vectors of user's texts are independent of each other, LSTM can detect the hidden relations between them.

The OneClassSvm implementation was taken from the scikit-learn library. Parameter nu=0.5, gamma=0.1, kernel= "rbf". The lstm implementation was taken from the keras library. CuDNNLSTM was used with 100 units and on top of it a dense layer with two outputs. The MLP implementation was taken from the keras library. We used a network of four dense layers with the number of units 100, 200, 100 and 2, respectively. For both neural networks, we use the same optimization: Adam [20] with learning rate 1e-4, 30 epoch of training. The other optimization techniques, as well as the other neural network architectures, including those with a large number of layers, did not give better results.

We state the following experiments using liked posts and descriptions of subscriptions with at least one "like" from user:

- One-class classification with united texts of liked posts / descriptions.
- Binary classification with united texts of liked posts / descriptions.
- LSTM classification with a set of liked posts texts / descriptions.

IV. EXPERIMENTAL STUDY

A. Classification with liked posts and texts subscriptions

For each classifier, we calculate precision and recall metrics. The users from the training datasets are split into 5 folds, 4 folds are used for the training and one fold for validation. Additionally, we test the algorithm on users with likes in 2017. Training was used five times for each algorithm, and the metrics were averaged. The results of training classifiers are presented in Table III.

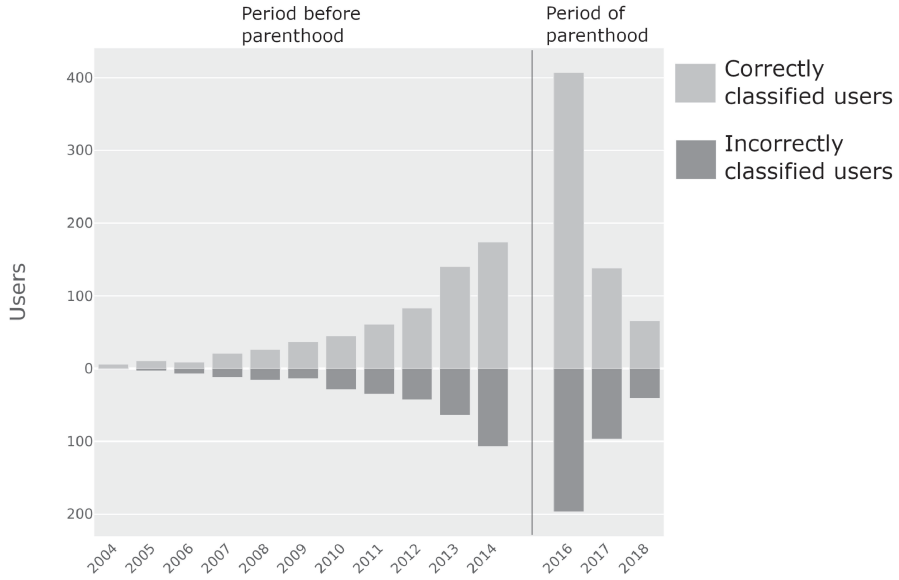

Fig. 3. The numbers of correctly and incorrectly classified users depend on child's birth years

Test results demonstrate decreasing metrics for all selected classification approach, especially in terms of recall. The LSTM-based classification shows the highest stability in the results, which can be explained by levelling the problem of averaged vectors.

We assumed that the sizes of texts of liked posts could be small, and the topic model could not detect suitable topics. So instead of likes, the descriptions of subscriptions in which a user liked posts in 2015 and 2017 are used.

The same three classifiers were used for subscriptions: One-Class SVM, MLP, LSTM. The validation scheme is the same — results of classification based on subscriptions, in which users liked in 2015 presented in Table IV. The trend towards a drop in efficiency for test data remains; however, in general, results demonstrate greater accuracy of this approach. In this case, LSTM is superior to other classification methods both for validation data in 2015 and test data in 2017.

B. Discussion

According to the results given above, we can conclude that events related to the parental status are rare: combining texts of posts/description leads to decrease in the classification efficiency, while the entire set of individual events gives both stable and high precision/recall (LSTM classifier). Moreover, our analysis of parent-related groups of kindergartens, schools and maternity hospitals demonstrates that subscription to these groups is not related to the parental status of a user. Furthermore, the analysis of user's activity related to non-parental status - for example, posts in dating groups, where people publish summaries about themselves, including fields such as "has no child", demonstrates that it is not a category definition criterion.

In addition to that, we had a hypothesis that the results of classification depending on the age of the child: while the child is small, parents are more interested in posts related to the upbringing and health of children, children's products, etc. and more often like in such groups, but over time the number

TABLE III. USER CLASSIFICATION BASED ON LIKES FOR A GIVEN PERIOD

Model	Precision	Recall	F1-score
Validation split (2015)			
	Non-parent		
One-Class SVM	0.48	0.46	0.47
MLP	**0.61**	0.62	0.61
LSTM	0.61	**0.63**	**0.62**
	Parent		
One-Class SVM	0.48	0.50	0.49
MLP	0.61	**0.61**	**0.61**
LSTM	**0.62**	0.60	0.61
Test split (2017)			
	Non-parent		
One-Class SVM	0.52	**0.69**	**0.59**
MLP	**0.54**	0.59	0.57
LSTM	0.49	0.49	0.49
	Parent		
One-Class SVM	0.48	0.31	0.37
MLP	0.50	0.45	0.48
LSTM	**0.50**	**0.60**	**0.54**

TABLE IV. USER CLASSIFICATION BASED ON SUBSCRIPTIONS FOR A GIVEN PERIOD

Model	Precision	Recall	F1-score
Validation split (2015)			
	Non-parent		
One-Class SVM	0.46	0.43	0.44
MLP	0.61	**0.64**	0.63
LSTM	**0.65**	0.63	**0.64**
	Parent		
One-Class SVM	0.47	0.50	0.48
MLP	0.63	0.59	0.61
LSTM	**0.64**	**0.66**	**0.65**
Test split (2017)			
	Non-parent		
One-Class SVM	0.46	0.46	0.46
MLP	0.51	**0.55**	**0.53**
LSTM	**0.53**	0.52	0.52
	Parent		
One-Class SVM	0.46	0.45	0.46
MLP	0.52	0.47	0.49
LSTM	**0.53**	**0.54**	**0.53**

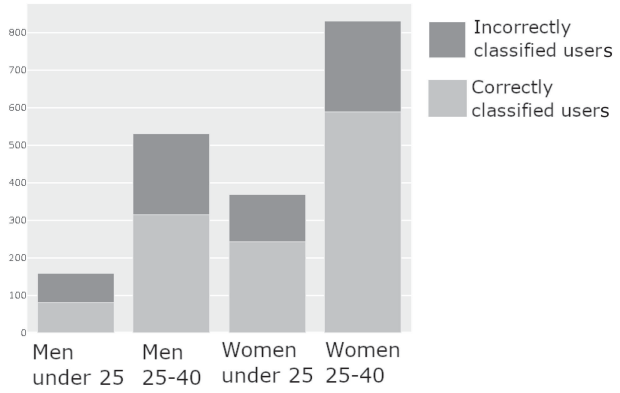

Fig. 4. The distribution of correctly and incorrectly classified users depending on their ages and genders.

of likes decreases. That could lead to more a higher error rate for parents with older children and a lower error rate for a newborn. But in reality, there is no explicit dependency found - Fig. 3. Also we found no statistically significant changes in general user activity on social networks over the three-year periods before and after birth.

Additionally, we built the distribution of correctly and incorrectly classified users based on their gender and age - Fig. 4. The graph shows that in the group of women from 25 to 40, the percentage of correctly classified people is slightly higher than in other groups. Apparently, among users of this group, the percentage of likes of "parent" posts is higher than among other groups, so the error rate among them is slightly less.

We did not classify liked post into parent and non-parent, preferring to use all the posts at once. Because when filtering texts, some useful information may be lost. In particular, we expect that with the birth of a child, activities in other groups may change, for example, spending less time in humorous groups and more devoted to the family.

V. Conclusion

In this work, we solve the problem of user classification on parent and non-parents categories. The proposed approach to the positive-unlabelled problem allows us to reduce that to the binary domain and also keep individual hidden user's features: fields of interests and behaviour on social media. The difficulty of the data is a lack of samples with the class opposite to the target. Using simple text representation, we built a set of classification models and analyzed their efficiency. The best result we achieved using LSTM model, which takes into account every single feature and can detect even rare events. Our baseline, One-Class SVM, is not effective, and it allows us to conclude that using only one - positive - class in a given task is inefficient.

The proposed retrospective approach to user classification on social categories allows us to be sure that we did not catch the difference in some other feature mentioned above. While the absolute values of the efficiency metric are not very high, we suppose that this is the influence of the chosen method of presenting text data, which can be replaced by state-of-the-art neural language models such as BERT. We are going to continue working in the field of determining the social statuses of users of social networks that are not amenable to the classical split of the samples into positive and negative cases. The provided results demonstrate the potential of this approach, even with such simple instruments as topic modelling and classical classifiers.

This approach can be used to detect events in a person's life that are reflected in changes in their interests and their long-term manifestation (for example, the birth of a child, getting higher education, military service, etc.). The birth of children is just such an event. While children are young, parents are interested in caring for them, their health, education, and choice of toys. Then the child grows up, many send their children to kindergartens, then to schools. Parents have interests related to the child's education. Over the years, parents take care of their children. And it can be reflected in their activities on social media. The approach described in this article uses a set of liked texts over a period of time, the more texts in this array that can be used to judge that a given event has occurred, the higher the likelihood that the classifier will be able to detect this event. If the event that has happened is insignificant in a person's life, it will not affect his interests in the long term, which means that the approach could be ineffective.

Acknowledgement

The Russian Science Foundation, Agreement #17-71-30029 with co-financing of Bank Saint Petersburg.

References

[1] C. Zhu, R. Zeng, W. Zhang, R. Evans and R. He, "Pregnancy-related information seeking and sharing in the social media era among expectant mothers: Qualitative study," *J Med Internet Res*, vol. 21, p. e13694, Dec 2019.

[2] V. Lampos, N. Aletras, J. K. Geyti, B. Zou and I. J. Cox, "Inferring the socioeconomic status of social media users based on behaviour and language," in *ECIR*, 2016.

[3] S. Abbar, Y. Mejova and I. Weber, "You tweet what you eat: Studying food consumption through twitter," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, (New York, NY, USA), p. 3197–3206, Association for Computing Machinery, 2015.

[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[5] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins and L. Ungar, "Beyond binary labels: Political ideology prediction of twitter users," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 729–740, Association for Computational Linguistics, July 2017.

[6] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from twitter data," *Social Science Computer Review*, vol. 31, no. 6, pp. 649–679, 2013.

[7] M. D. Choudhury, S. Counts and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France. CHI 2013.*, April 2013.

[8] M. De Choudhury, S. Counts and E. Horvitz, "Major life changes and behavioral markers in social media: Case of childbirth," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, (New York, NY, USA), p. 1431–1442, Association for Computing Machinery, 2013.

[9] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality predictions based on user behavior on the facebook social media platform," *IEEE Access*, vol. 6, pp. 61959–61969, 2018.

[10] A. M. R. Alireza Souri, Shafigheh Hosseinpour, "Personality classification based on profiles of social networks' users and the five-factor model of personality," *Human-centric Computing and Information Sciences*, vol. 8, Aug 2018.

[11] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong and J. Sun, "Deep learning-based personality recognition from text posts of online social networks," *Applied Intelligence*, vol. 48, pp. 4232–4246, Nov 2018.

[12] Y. J. Yezheng Liu, Jiajia Wang, "Pt-lda: A latent variable model to predict personality traits of social network users," *Neurocomputing*, vol. 210, p. 155–163, Oct 2016.

[13] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: User classification in twitter," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, (New York, NY, USA), p. 430–438, Association for Computing Machinery, 2011.

[14] K. Sechidis, B. Calvo and G. Brown, "Statistical hypothesis testing in positive unlabelled data," in *Machine Learning and Knowledge Discovery in Databases* (T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, eds.), (Berlin, Heidelberg), pp. 66–81, Springer Berlin Heidelberg, 2014.

[15] N. Butakov, M. Petrov, K. Mukhina, D. Nasonov and S. Kovalchuk, "Unified domain-specific language for collecting and processing data of social media," *Journal of Intelligent Information Systems*, vol. 51, 05 2018.

[16] K. Vorontsov, O. Frei, M. Apishev, P. Romov and M. Dudarenko, "Bigartm: Open source library for regularized multimodal topic modeling of large collections," pp. 370–381, 2015.

[17] T. Sokhin and N. Butakov, "Semi-automatic sentiment analysis based on topic modeling," vol. 136, pp. 284–292, 2018. 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece.

[18] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," *The Knowledge Engineering Review*, vol. 29, no.3, pp. 345–374, 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 1735–1780, Nov. 1997.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.