Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2018"

Moscow, May 30—June 2, 2018

MACHINE LEARNING CLASSIFICATION OF USER INTERESTS ACROSS LANGUAGES AND SOCIAL NETWORKS

Mikhalkova E. V. (e.v.mikhalkova@utmn.ru),
Ganzherli N. V. (n.v.ganzherli@utmn.ru),
Karyakin Y. E. (y.e.karyakin@utmn.ru),
Grigoryev D. A. (Grigd2013@gmail.com)

Tyumen State University (University of Tyumen), Tyumen, Russia

Being a matter of cognition, user interests should be apt to classification independent of the language of users, social network and the essence of interest itself. To prove it, we built a collection of English and Russian Twitter and Vkontakte community pages manually classified according to the interests of their followers. First, we created a model of Major Interests (MaIs) with the help of expert analysis and then classified the mentioned set of pages using machine learning algorithms (SVM, Neural Network, Naive Bayes, Logistic Regression, Decision Trees, k-Nearest Neighbors) trying different optimization techniques. We take three interest domains that are typical of both English and Russian-speaking communities: football, rock music, vegetarianism. The results of classification show a greater correlation between Russian-Twitter and English-Twitter pages. The Logistic Regression with Bernoulli bag-of-words model proves to be the most effective classification algorithm.

Keywords: interest discovery, social networks, natural language processing, optimization

КЛАССИФИКАЦИЯ ИНТЕРЕСОВ ПОЛЬЗОВАТЕЛЕЙ ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ В РАЗНЫХ ЯЗЫКАХ И СОЦИАЛЬНЫХ СЕТЯХ

Михалькова Е. В. (e.v.mikhalkova@utmn.ru),
Ганжерли Н. В. (n.v.ganzherli@utmn.ru),
Карякин Ю. Е. (y.e.karyakin@utmn.ru),
Григорьев Д. А. (Grigd2013@gmail.com)
Тюменский государственный университет, Тюмень, Россия

1. Introduction

Social networks provide people with an opportunity to form social clusters that share interests not only sporadically but on a regular basis (circles of fans of different music, books, kinds of sports, etc.). Every circle communicates these interests creating lots of linguistic data to attract new followers and support interests of the existing ones. Researchers often use these data in content-based user models to classify interests of particular users. As a rule, such models are tested on a corpus of one language downloaded from one social network. However, being a matter of cognition, user interests should be independent of the language in which they are expressed and the network where users communicate them, when we try to process them with different algorithms.

To see if the performance of machine learning algorithms is the same for two different languages and two networks, we tested them in three internationally popular interest domains: football, rock music, vegetarianism. For the present research, we collected three datasets from two different networks: the English (I) and Russian (II) corpora from Twitter and the Russian corpus (III) from Vkontakte.[1] Then, we tried to classify the corpora according to the interests of users with such machine learning instruments as SVM, Neural Network, Naive Bayes etc.[2]

One of the most problematic issues in this classification was to find three classes that have enough data in the two networks and two languages. For example, there are lots of pages devoted to football in Vkontakte and Twitter, in Russian and English. However, there appeared to be few Russian vegetarian pages in Twitter, the same was the case for Russian historical reenactors who are only widely present in Vkontakte.

[1] The dataset can be found at https://github.com/evrog/TSAAP.

[2] Within the scope of this research, we focus on the classification of pages to learn about the language of social groups in social networks. We apply machine learning and statistical analysis to observe linguistic data rather than to partition a social network into **clusters** of groups and individuals with similar interests. The latter is a practical task that currently has no universally acknowledged solution.

The problem of small scale datasets[3] is nowadays addressed quite often irrespective of the branch of science [Huda et al. 2017] in various fields like dialectology, archeology, biology [Plekhanova et al. 2018; Steyerberg et al. 2001] and, closer to our research, recommender systems [Li et al. 2018]. It follows that if we are striving to build a system of interest classification for a social network, we should, first of all, focus on the inherent properties that show in the small but prominent samples and can serve as a standard in large-scale research.

2. Interest discovery by means of NLP

There exists a variety of content-based models of user interests. These models make use of keywords, interests enlisted in profiles, tags attached to posts etc. Such data serve as the classification basis in works of [Bonhard 2006; Firan, 2007; Dugan 2007; Li 2008, Sen 2009, Guy 2010][4]. However, they are often very unreliable and hard to formalize. Interest discovery has now become a separate branch of user modelling.

In regard to social networks, NLP provides several approaches to interest discovery. [Piao 2011] view interests as terms and named entities extracted from a collection of user tweets. In works of [Mccallum 2005; Ramage 2010; Ahmed 2011], interests are viewed as topics distributed across users' tweets. The authors apply variations of Latent Dirichlet Allocation suggested by [Blei 2003] as the main method of topic analysis to scale user messages down to a particular topic. [Wang 2014] describe the User Message Model that is designed particularly for microblogs to reduce data sparseness and topic diversity.

Interests can be represented as concepts in an ontology. The latter often includes named entities. [Bakalov 2009] suggest a hybrid user model that makes use of ontologies to specify user interests. Interests are either extracted as keywords from the content of visited pages or can be manually specified by a user. [Al-Kouz 2012] describe another approach where the system creates a semantic graph of interests based on the "entities" mentioned in tweets. Entities are words denoting real-world phenomena that have an encyclopedic description. For reference, the authors used the currently deprecated knowledge base Freebase[5]. At the same time, a recent study of [Piao 2016] demonstrates that "concept-based representations of user interests using a KB" add efficiency to the model, but then there is no need to add "rich semantic information from a KB to extend the interests of users."

[3] As well as class imbalance [Sitompul et al. 2018].

[4] In recommender systems, tags and keywords in profiles define a scope of users that share similar interests. According to [Guy 2009], this process is called *collaborative* filtering. [Pazzani 1999] suggests *demographic* filtering that infers types of users with a common interest based on their age, gender, education etc. mentioned in profiles. With the rise of the social network analysis, many researchers, for example [Groh 2007], attempt to objectivize communities with the help of social graphs (*social* filtering). A more detailed account of these approaches is given by [Burke 2002].

[5] http://www.freebase.com. Before the widespread use of knowledge bases, linguists often used WordNet [Stefani 1999]. More recent approaches like [Shen 2013] use DBpedia.

3. Modelling social nature of interests

It appears that interest discovery in social networks is a two-sided problem. First, regarding the number of published posts and comments, although in social networks linguistic content is abundant, it is often very hard to structure. Second, user interests themselves are an arcane matter: some researchers view them as topics, tags, keywords, etc. We will call the interest that attracts users to a page, the Major Interest (MaI). In the present research, we will attempt to classify a number of community pages based on three MaIs: football, rock music, vegetarianism.

3.1. Community pages

In our research, we will focus on community pages, e.g. accounts of public value that represent institutions, authorities, famous people, leaders of social groups, events, etc. They exist in all networks known to us (Twitter[6], Vkontakte[7], Facebook[8], LiveJournal[9] etc.). Many researchers already use data from such pages together with a user's individual page content but view them as complementary material. Usually, but not necessarily, such accounts have many followers (typically, more than 1,000).

Concerning the content downloaded for analysis, from Vkontakte, we obtained posts, comments to posts, and comments from the so-called "board". As for Twitter, the only content available there is tweets.

3.2. Data survey

Observations show that for an expert it is quite easy to bind a community page to one certain MaI based on user comments and tweets and to find other pages with a similar MaI (the same kind of sports, music style, etc.). Many pages even provide links to other recommended pages. However, on the same page, users can mention a variety of different interest domains especially if they are related hyponymically (a style of music and its substyles), antonymically (a football team vs. its opponent in a championship), pragmatically (a football team and a stadium where it trains). Therefore, to define the basis of classification, i.e. MaIs that are not just microtopics and the pages that are devoted to these MaIs, we conducted an expert-based survey.

First, we downloaded comments from 4,000 random Vkontakte community pages that contained from 22 to 100,523 words. Next, we asked a sociologist and a marketing specialist to find several active communities with common interests, i.e. such community pages where people actively interact about something they share an interest for. The result set included four communities whose MaI is one of the following: 1. rock music, 2. historical reenactment, 3. football, 4. vegetarianism. All these MaIs are international and represented by pages in Russian as well as in English.

[6] https://twitter.com/

[7] https://vk.com/. One of the most popular Russian social networks.

[8] https://www.facebook.com/

[9] https://www.livejournal.com/

We chose sample discussions from Vkontakte pages where people talk about things related to the MaIs. For control, a sample with several disparate objects of interest was chosen.

10 experts (certified and employed linguists, sociologists, marketing specialists) gave their opinion on what community manifests itself in every sample. We instructed experts to define if authors in the sample dialogue *are* a community and, if yes, explain why they think so. Thus, the expert answers were formulated freely without the aim of interest attribution. Some of them preferred to just name the community ("vegans", "rockers"); some stated the object of interest ("vegetarianism", "rock music"). If these keywords were mentioned, we assigned 1 point to the answer (a True Positive answer); if no or some other keywords were mentioned ("music addicts" instead of "rockers"), we assigned 0 points. The answers were put in a ranking table (cf.: Table 1 in Appendix). In general, agreement between the experts can be considered reliable, as Krippendorff's $\alpha=0.82$ (>0.8). To see which samples relate to the most unanimous decision, we calculated percentage of True Positive answers in every column (percent agreement).

Determining adherence of the authors of comments to communities of football fans, vegetarians, and historical reenactors, the raters showed perfect agreement. Fans of rock music were not as easy to define (only 50% of raters recognized them). The control group also provided a highly reliable result[10] that allows us to state that the raters were not apt to see communities in any text we offer them.

3.3. Feature set

After the data survey, we searched Vkontakte and Twitter for pages that attract fans of 1. rock music, 2. historical reenactment, 3. football, 4. vegetarianism. The search showed that historical reenactment has no Russian accounts in Twitter. Hence, we had to exclude it from the further research. For each class in the three corpora (I. English-Twitter, II. Russian-Twitter, III. Russian-Vkontakte), we managed to find a different number of pages from which we downloaded tweets and comments.

Normalization. We parse Twitter pages with our tweet preprocessing software[11]. It has a special treatment of mentions (they start with "@", e.g. "@WhoopiGoldberg" becomes a two word group "whoopi goldberg") and hashtags (e.g. "ElectrikBLOOM" becomes "electric bloom"). In Vkontakte, we remove URLs, attachments and emoji. All texts are converted to lowercase, symbols and punctuation marks are removed.

Lemmatization. The sets are processed as in Normalization (see above), but before the change of case we lemmatize English texts with NLTK Lemmatizer [Bird 2009] and Russian texts with Pymystem3[12].

The properties of the sets are reflected in Table 1.

[10] We assigned 1 point for this sample if the expert directly expressed doubt in describing the community, e.g. wrote "Don't know", "I doubt this is a community at all", or left the field blank.

[11] "Preprocessing tweet" at https://github.com/evrog/PunFields. Its full description can be found in [Mikhalkova 2018].

[12] https://github.com/nlpub/pymystem3

Table 1. Feature set. F—football, R—rock music, V—vegetarianism, T—Twitter, Vk—Vkontakte, En—English, Ru—Russian. Total No. is given as follows: *tokens* first, then *types* (no duplicates). denotes the mean of the scores. In Twitter, the maximum No. of comments downloaded from a page is 1,000; in Vkontakte, the maximum number of wall posts and comments to posts available for download is 100.

		No. of pages	Total No. of words	Total No. of lemmes	No. of words per comment, tweet: $\bar{x}$, mode	No. of comments per page, $\bar{x}$
Vk Ru	F	39	738684, 91486	664972, 76657	18.61, 1	992
	R	109	1212731, 136866	1166159, 87589	24.91, 1	438
	V	127	759066, 103800	717531, 62372	58.16, 6	101
T Ru	F	33	334457, 38115	330653, 19130	10.64, 12	924
	R	37	312911, 53721	305206, 31538	10.26, 14	802
	V	32	192643, 45042	188852, 26000	11.66, 14	500
T En	F	97	1366312, 33321	1726604, 29766	14.36, 15	971
	R	96	960542, 47507	1328049, 40503	11.69, 9	846
	V	100	1189804, 51769	1616783, 43110	12.41, 8	949

As mentioned above, in their study of user interests, researchers mainly appeal to keywords, topics, named entities etc. However, when we asked experts from the Data survey to analyze what makes them think that a page attracts a certain social group, they also pointed at terminology and special meaning of common words, derivation (i.e. words with the same stems: **veg**an, **veg**etarian, **veg**etarianism, etc.) and unique vocabulary (e.g. **hoolie** for football fans). We tend to think that interests cannot be bound to a certain topic or a set of semantically related topics (e.g. football-sports). They are rather like umbrella terms to a combination that singles out an interest from similar ones. For example, the combination of "game-field-ball" differentiates football from hockey ("game-ice-stick"). To make sure that the machine learning classifier learns enough about these differences, we need a sufficient set of words. In the present study, we experiment with 1,000 most frequent items (word forms and lemmes).

4. Community pages classification

We used several machine learning algorithms to classify community pages that represent one of the MaIs: 1. football, 2. rock music, 3. vegetarianism.

4.1. Interclass classification

Cross-validation. The sets of pages being of different size, we split some of the longer texts into smaller ones to create a collection of 200 texts of different length (the total of 1800 texts). We randomly split each set into 5 equal parts to apply 5-folds cross-validation. Further, we analyze the average of F1-scores.

Classification algorithms that we chose for the survey are often met in NLP tasks like spam detection, sentiment analysis and the like: Support Vector Machine, Neural Network, Naive Bayes, Logistic Regression, Decision Trees, *k*-Nearest Neighbors. Their structure and implementation in the Python library Scikit-learn [Pedregosa et al. 2011] are described in the documentation of the library.

Optimization parameters. We used the following optimization strategies to compare the performance of classifiers. In SVM, we experimented with four kernel functions: linear, polynomial, Radial Basis Function, sigmoid. In Naive Bayes, we separately tried three well-known algorithms based on Bayes' theorem: Bernoulli, Multinomial, and Gaussian.

The Scikit-Learn implementation of Neural Networks uses a Multi-layer Perceptron algorithm. Unlike Logistic Regression, it learns **non-linear** dependencies with the help of hidden layers. We tested the default model with 1 hidden layer of 100 neurons. We also experimented with two solver functions: "*lbfgs* is an optimizer in the family of quasi-Newton methods" and *adam* is "a stochastic gradient-based optimizer" [Pedregosa et al. 2011].

For all the classifiers, we tested three data models: Bernoulli—absence or presence of a word denoted by 0 and 1 correspondingly; Frequency distribution—presence of a word denoted by its frequency in the training vocabulary denoted by a whole number in the interval [0;+∞); Normalized frequency—presence of a word denoted by normalized frequency in the training vocabulary in the interval [0;1]. The lemmatized texts are analyzed separately from the normalized texts.

We also do not exclude stop-words for the following reason. As we deal with social groups, their **use of some stop-words is significant**. First, in Frequency models, such stop-words as "I", "we" and "they" have different frequency. This frequency can show differences in groups' values. For example, some groups can be more focused on collectiveness and use "we" more often than "I"; some can be more competitive using "they", "their", etc. In Bernoulli models, stop-words are not so significant as they are likely to be present in nearly all the texts. However, in cases of short texts, they can be quite important (for example, if there are only words like "we", "us", "our", the group can belong to a more "collective" type).

In case of some classifiers, Scikit-learn offers more instruments for optimization (penalty parameter C and types of loss function in SVM, activation function in neural networks, etc.). Some of them might have been overseen in the experiment, as their scope creates a huge field for testing, or can subject the classifier to over-tuning. However, with our data openly published, we hope they will be further studied in other projects.

4.2. Results of experiment

Table 2 (Appendix) demonstrates average results of F1-score after a 5-fold cross validation. First of all, it shows that lemmatization slightly increases the performance (by about 3%): the sum of -scores of the lemmatized texts is 262.752 versus 254.186 of the normalized texts.

Second, the Bernoulli model is the most effective one by mode: it has 18 scores of 1.0 when the two other models have only 4 such scores together, and by mean: 0.845 against 0.753 for plain and 0.795 for normalized frequencies.

Third, the best performing algorithm is Logistic Regression with Bernoulli model. The sum of its -scores equals 17.71. The second best score (17.664) belongs to the Neural Network (*lbfgs*) with Bernoulli model which hints at the lack of necessity to complicate a Logistic Regression classifier with a non-linear model. The third place belongs to the Multinomial Bayes with plain frequencies (17.5).

SVM models, even the linear one, were not as successful compared to the Logistic Regression. From this, we can assume that the word combinations that help to differentiate between two classes are more different in their core and have blended, noisy margins.

Concerning normalization of word frequencies, it appears to improve performance of such algorithms as SVM with RBF and sigmoid kernels. Without it, SVM 'Sigmoid' shows the lowest results in the ranking table. However, it can also decrease the result. Surprisingly, it derated the average result of Multinomial Naive Bayes from 0.972 to 0.51.

4.3. Statistical analysis

We will now try to analyze differences in classification of the three datasets according to the MaI, the language of user communication and the network where the texts were posted. For the analysis we will use the -scores from Table 2, Appendix. First, we will normalize the Table excluding classifiers that gave lower results in the either of the two sets: normalized, lemmatized.

For every MaI, the total sum of $\overline{F1}$-scores and sum dependent on the language and network is shown in Table 2. We use sums for the first four columns (Total, Vk Ru, T Ru, T En) as every number in each of these columns characterizes sets of the same size. E.g. there are 24 F-scores in "Football, Vk Ru" and "Rock, Vk Ru": 12 F-scores of Bernoulli models and 12 of Frequency models. Hence, there is no need to average them. However, the other four columns characterize sets of different sizes. For example, the "Football—Vkontakte" set of F-scores has 24 items, whereas "Football—Twitter" has 48 items (2 $\overline{F1}$-scores for every English and Russian text corpus of Football fans). Likewise, "Football—Ru" has 48 scores, and "Football—En" has 24 scores. Therefore, we average data in the last four columns. The size of the set of scores in Total is 24×3=72; in Vk Ru, T Ru, T En, it is 24 each.

Table 2. Comparison of MaI $\overline{F1}$-scores. F—football, R—rock music, V—vegetarianism, T—Twitter, Vk—Vkontakte, En—English, Ru—Russian. denotes the mean of the scores

MaI	Total	Vk Ru	T Ru	T En	Vk, $\bar{x}$	T, $\bar{x}$	Ru, $\bar{x}$	En, $\bar{x}$
				Normalized texts				
F	33.976	10.240	11.826	11.910	0.853	0.989	0.919	0.993
R	33.138	10.064	11.334	11.740	0.839	0.961	0.892	0.978
V	32.906	9.8080	11.302	10.796	0.817	0.962	0.880	0.983
				Lemmatized texts				
F	34.282	10.430	11.932	11.920	0.869	0.994	0.932	0.993
R	34.776	10.398	11.624	11.754	0.867	0.974	0.918	0.980
V	33.708	10.272	11.622	11.814	0.856	0.977	0.912	0.985

To analyze the significance of differences between sets, we used Mann-Whitney test. It supports the following hypotheses:

1. Lemmatized and non-lemmatized (normalized) sets may come from different distributions (i.e. differences in their results are statistically significant): statistic=5296.5, pvalue=0.243 (>0.05), *two-sided*.
2. Differences in the Twitter-Russian and Twitter-English sets are insignificant: statistic=3407.0, pvalue=0.001 (<0.05), *two-sided*. However, the Vkontakte-Russian set underscores significantly compared to the Twitter-English (statistic=161.0, pvalue=1.0, *greater*) and Twitter-Russian (statistic=703.0, pvalue=0.99, *greater*) sets.
3. Vegetarianism and Rock Music are very likely to score less than Football: statistic=1671.0, pvalue=0.99, *greater,* and statistic=1612.5, pvalue=0.99, *greater* correspondingly.

Also, there appears to be no correlation between the experts' difficulty to classify the rock music fans (see 3.2 Data survey) and the ML classification which was successful enough.

In general, the Vkontakte set seems to actually provide a lower performance than Twitter. We suppose that the difference is caused by more noise which can be due to the normalization software. In Twitter, we have a processor for hashtags and mentions that turns them into clear word forms. In Vkontakte, we simply remove all kinds of attachments.

As for the languages, if we do not take into account the mentioned processor, there seems to be no significant differences between the Russian and English languages. For the both languages, lemmatization is a useful tool.

5. Conclusion

Summing up, with due normalization, languages do not influence the ML classification of interests. However, the social network can be an important factor. Which social network features decrease the performance on Vkontakte sets requires more research. Also, there can be differences due to the interest itself (in our case, vegetarianism and rock music were significantly less supple in classification than football).

Concerning the classifiers, we have assumed, on the grounds that the Logistic Regression has the best score, that interest classification is more focused on the core of a set of features rather than the margins. We also faced the efficiency of the Bernoulli model. I.e. word frequencies are not as important in classification as the absence or presence of characteristic features.

If we consider this experiment in terms of a practical application to classify all social network pages according to "user interests", the data in our research is, of course, much more structured. In a real network, it will be hard to get as many expert-classified pages as there are user interests. However, the findings of this research can be helpful in developing practical tools of their discovery.

References

1. *Ahmed A., Low Y., Aly M., Josifovski V., Smola A. J.* (2011), Scalable distributed inference of dynamic user interests for behavioral targeting, In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 114–122.
2. *Alabandi G. A.* (2017), Combining Deep Learning with Traditional Machine Learning to Improve Classification Accuracy on Small Datasets, M.S. thesis submitted to the Graduate Council of Texas State University.
3. *Al-Kouz A., Albayrak S.* (2012), An interests discovery approach in social networks based on semantically enriched graphs, In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, IEEE, pp 1272–1277.
4. *Bakalov F., König-Ries B., Nauerz A., Welsch M.* (2009), A hybrid approach to identifying user interests in web portals, In: IICS, pp 123–134.
5. *Bird S., Loper E., Klein E.* (2009), Natural Language Processing with Python, O'Reilly Media Inc.
6. *Blei D. M., Ng A. Y., Jordan M. I.* (2003), Latent dirichlet allocation, Journal of machine Learning research, 3 Jan, 993–1022.
7. *Bonhard P., Sasse M. A.* (2006), 'Knowing me, knowing you'—using profiles and social networking to improve recommender systems, BT Technology Journal, 24(3), pp. 84–98.
8. *Burke R.* (2002), Hybrid recommender systems: Survey and experiments, User modeling and user-adapted interaction, 12(4), pp. 331–370.
9. *Dugan C., Muller M., Millen D. R., Geyer W., Brownholtz B., Moore M.* (2007), The dogear game: a social bookmark recommender system, In: Proceedings of the 2007 international ACM conference on Supporting group work, ACM, pp. 387–390.
10. *Firan C. S., Nejdl W., Paiu R.* (2007), The benefit of using tag-based profiles, In: Web Conference, LA-WEB 2007, Latin American, IEEE, pp. 32–41.
11. *Groh G., Ehmig C.* (2007), Recommendations in taste related domains: collaborative filtering vs. social filtering, In: Proceedings of the 2007 international ACM conference on Supporting group work, ACM, pp. 127–136.
12. *Guy I., Zwerdling N., Carmel D., Ronen I., Uziel E., Yogev S., Ofek-Koifman S.* (2009), Personalized recommendation of social software items based on social relations, In: Proceedings of the third ACM conference on Recommender systems, ACM, pp. 53–60.
13. *Guy I., Zwerdling N., Ronen I., Carmel D., Uziel E.* (2010), Social media recommendation based on people and tags, In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 194–201.
14. *Huda R. K., Banka H.* (2017), Efficient feature selection and classification algorithm based on PSO and rough sets, Neural Computing and Applications, 1–17.
15. *Li R., Ye X., Zhou H., Zha H.* (2018), Learning to Recommend via Inverse Optimal Matching, arXiv preprint arXiv:1802.03644.
16. *Li X., Guo L., Zhao Y. E.* (2008), Tag-based social interest discovery, In: Proceedings of the 17th international conference on World Wide Web, ACM, pp. 675–684.

17. *McCallum A., Corrada-Emmanuel A., Wang X.* (2005), Topic and role discovery in social networks, Computer Science Department Faculty Publication Series.
18. *Mikhalkova E., Karyakin Y., Voronov A., Grigoriev D., Leoznov A.* (2018), PunFields at SemEval-2018 Task 3: Detecting Irony by Tools of Humor Analysis, In: Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018).
19. *Pazzani M. J.* (1999), A framework for collaborative, content-based and demographic filtering, Artificial intelligence review, 13 (5–6), pp. 393–408.
20. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.* (2011), Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, pp. 2825–2830.
21. *Piao G., Breslin J. G.* (2016), Interest representation, enrichment, dynamics, and propagation: A study of the synergetic effect of different user modeling dimensions for personalized recommendations on Twitter, Springer International Publishing, Cham, pp. 496–510.
22. *Piao S., Whittle J.* (2011), A feasibility study on extracting twitter users' interests using nlp tools for serendipitous connections, In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE Third International Conference on, IEEE, pp. 910–915.
23. *Plekhanova E., Nuzhdin S. V., Utkin L. V., Samsonova M. G.* (2018), Prediction of deleterious mutations in coding regions of mammals with Transfer learning, Evolutionary Applications.
24. *Ramage D., Dumais S. T., Liebling D. J.* (2010), Characterizing microblogs with topic models, ICWSM 10: 1–1.
25. *Sen S., Vig J., Riedl J.* (2009), Tagommenders: connecting users to items through tags, In: Proceedings of the 18th international conference on World wide web, ACM, pp. 671–680.
26. *Shen W., Wang J., Luo P., Wang M.* (2013), Linking named entities in tweets with knowledge base via user interest modeling, In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 68–76.
27. *Sitompul O. S., Nababan E. B.* (2018), Optimization model of K-Means clustering using artificial neural networks to handle class imbalance problem, In IOP Conference Series: Materials Science and Engineering, Vol. 288, No. 1, p. 012075, IOP Publishing.
28. *Stefani A., Strapparava C.* (1999), Exploiting nlp techniques to build user model for web sites: the use of wordnet in siteif project, In: Proc. 2nd Workshop on Adaptive Systems and Comparison of Interest Classifying Model User Modeling on the WWW.
29. *Steyerberg E. W., Eijkemans M. J., Harrell F. E. Jr, Habbema J. D.F.* (2001), Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets, Medical Decision Making, 21(1), 45–56.
30. *Wang Q., Xu J., Li H.* (2014), User message model: A new approach to scalable user modeling on microblog, In: Asia Information Retrieval Symposium, Springer, pp. 209–220.