Smirnov *EPJ Data Science* ( 2020) 9:27
https://doi.org/10.1140/epjds/s13688-020-00245-8

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

# Estimating educational outcomes from students' short texts on social media

Ivan Smirnov[1*]

*Correspondence:
ibsmirnov@hse.ru
[1]Institute of Education, National Research University Higher School of Economics, Moscow, Russia

**Abstract**

Digital traces have become an essential source of data in social sciences because they provide new insights into human behavior and allow studies to be conducted on a larger scale. One particular area of interest is the estimation of various users' characteristics from their texts on social media. Although it has been established that basic categorical attributes could be effectively predicted from social media posts, the extent to which it applies to more complex continuous characteristics is less understood. In this research, we used data from a nationally representative panel of students to predict their educational outcomes measured by standardized tests from short texts on a popular Russian social networking site VK. We combined unsupervised learning of word embeddings on a large corpus of VK posts with a simple, supervised model trained on individual posts. The resulting model was able to distinguish between posts written by high- and low-performing students with an accuracy of 94%. We then applied the model to reproduce the ranking of 914 high schools from 3 cities and of the 100 largest universities in Russia. We also showed that the same model could predict academic performance from tweets as well as from VK posts. Finally, we explored predictors of high and low academic performance to obtain insights into the factors associated with different educational outcomes.

**Keywords:** Academic performance; Prediction; Social media; Text; Transferability

## 1 Introduction

In the past decade, digital trace data has become an integral part of social science research [1, 2]. One of the advantages of such data is that it can be used to predict otherwise unknown characteristics of people, allowing researchers to conduct studies on a large scale. For example, on an aggregate level, mobile phone metadata were used to predict wealth [3], images of street scenes were used to predict voting preferences [4], and texts of books were used to predict subjective well-being [5]. On an individual level, socio-demographic characteristics such as gender, ethnicity, age, and income were predicted from profile images [6], tweets [7, 8], and Facebook posts [9].

The work in this domain is typically focused on basic, often categorical, demographic variables. In such cases, the ground truth data required for training and validating the predictive model could be relatively easily obtained. In contrast, studying more complex human characteristics would require linking extensive survey data with digital traces on an

individual level—a task that is known to be rather challenging [10]. Examples of such work include predicting personality (see [11] for review) and mental health status from social media activity (see [12] for review). These complex individual-level characteristics were predicted from Facebook likes [13], posts on Facebook [14] or Twitter [15], and Instagram images [16].

A common limitation of this line of work is a reliance on voluntary response samples: participants are typically recruited via purpose-built applications or crowdsourcing platforms such as Amazon Mechanical Turk. This approach might not be problematic if the model could be externally validated. For basic demographic variables, the external validation is typically achieved by comparing model prediction with existing data on an aggregated level (e.g., census data [4], governmental data [3], Eurobarometer survey [5]). However, for more complex characteristics, such data are rarely available, and in most cases, the validation is limited to cross-validation on the existing dataset. As a result, the predictive power of the models on out-of-sample cases is not well known.

In our study, we built a model to predict the educational outcomes of students from their posts on social media. Educational outcomes are of particular interest because they are complex, continuous characteristics that—in contrast to basic demographic variables—can be reliably measured only by extensive standardized tests. Because it is also closely related to many important life outcomes [17–21], the ability to predict it on a large scale might be particularly valuable both for researchers and policy-makers.

Despite there being a large body of research related to the prediction of academic performance (see [22, 23] for review), such approaches typically (though not always [24]) rely on internal data from the educational organization (i.e., information about library logins [25], class attendance [26, 27], or data from learning management systems [28]); thus, they are limited to one educational institution. As a result, the extent to which it is possible to predict educational outcomes on a population level is not well understood.

In this study, we used data from a nationally representative panel study entitled "Trajectories in Education and Careers" (TrEC) [29] that tracks 4400 students who participated in the Programme for International Student Assessment (PISA) [30]. In addition to survey data, this dataset contains information about public posts on a popular Russian social networking site—VK for those participants who agreed to share their VK data ($N = 3483$). We combined unsupervised learning of word embeddings on a large corpus of VK posts (1.9B tokens) with a simple supervised model trained on individual posts to predict PISA scores from texts. We then tested the predictive performance of the model in different contexts. In particular, we used the model to predict the rankings of schools and universities based on the public posts of their students on VK. We further tested the generalizability of the model by applying it to users' tweets rather than VK posts. Finally, we used the model to explore the differential language use of participants.

Because we make use of data from a nationally representative sample, our results could be generalized on a population level, albeit for one age cohort. Our study also benefits from the fact that academic performance measured by a standardized test is publicly available for both high school and university levels in Russia, allowing for external validation of the predictive model.

The main aim of this study was to understand to what extent educational outcomes can be predicted from posts on social media. Note that the prediction here is understood as identification of patterns in data (i.e., correlations between academic performance and

text of posts, rather than future forecasting). Although the estimates provided by such models cannot be used to predict future outcomes, they are nevertheless valuable. For instance, Russian students take standardized tests only once per cohort, making it extremely difficult to estimate any value that is added to students' learning by an educational organization [31]. A comparison of predicted scores for the same users at different points in time might help to overcome this issue and shed light on the factors contributing to students' progress. The ability to estimate educational outcomes on a large scale could also help to uncover previously unknown factors that are associated with low or high academic performance.

## 2 Data & methods

### 2.1 TrEC data

We used data from the Russian Longitudinal Panel Study of Educational and Occupational Trajectories (TrEC) [29]. The study tracks 4400 students from 42 Russian regions who took the PISA test in 2012 [30]. We used PISA reading scores as a measure of students' academic performance. PISA defines reading literacy as "understanding, using, reflecting on and engaging with written texts in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" and considers it a foundation for achievement in other subject areas within the educational system and also a prerequisite for successful participation in most areas of adult life [32]. PISA scores are scaled so that the OECD average is 500 and the standard deviation is 100, while every 40 score points roughly correspond to the equivalent of one year of formal schooling [30].

In 2018, publicly available information from the social networking site VK was collected for 3483 TrEC participants who provided informed consent for the use of this data for research purposes. Note that while the initial sample was representative of the 9th-grade high school Russian students in 2012, the social network data is not necessarily representative. There were no publicly available posts for 498 users. The median number of public posts for remaining users was 35. We removed posts that contain URLs from our data set to account for potentially automated postings, and we also excluded re-posts and posts with no text. This resulted in the final data set of 130,575 posts from 2468 users.

### 2.2 Continuous-vocabulary approach

There are two general strategies for identifying correlations between user characteristics and their behavior as manifest through language: a closed vocabulary analysis that relies on a priori word category human judgments and an open vocabulary analysis that extracts patterns from data and that does not rely on any a priori word categories [9]. We adopted the latter approach; however, in contrast to previous studies [9, 33, 34], we used continuous rather than discrete word representations. For that purpose, we trained the fastText model [35] on the VK corpus (1.9B tokens, vocabulary size is 2.5M) to obtain vector representations of Russian words (the model is available at [36]). We used a simple tokenizer function that defines a word (token) as any uninterrupted sequence of symbols from Cyrillic or Latin alphabet, or a hyphen. We also substitute all numbers by a single ⟨num⟩ token and all URLs by a single ⟨url⟩ token.

We represented each post as a 300-dimensional vector by averaging over vector representations of all its constituent words. These post representations were used to train a linear regression model to predict the PISA scores of the posts' authors (the model is

available at [36]). By construction, the predicted text score is equal to the average predicted score of its constituent words.

The advantage of this *continuous-vocabulary approach* approach in comparison to discrete methods is that it allows the incorporation of rich knowledge about the language structure learned from the training of unsupervised word embeddings. It enables, for instance, the computation of meaningful scores even for words that are not present in the training dataset. This property is particularly valuable for small datasets that are typical for studies that combine survey data with digital traces. As we demonstrate, this approach outperforms common alternatives such as TF-IDF models (see Results).

At the same time, the continuous-vocabulary strategy is simpler than state-of-the-art ANN methods [37] and, as a result, allows straightforward interpretation of the predictions and exploration of the differential language use by users, as we demonstrate in the Results section.

### 2.3 External validation: high schools and universities data

VK provides an application programming interface (API) that enables the downloading of information systematically from the site. In particular, downloading user profiles from particular educational institutions and within selected age ranges is possible. For each user, obtaining a list of their public posts is also possible. According to the VK Terms of Service: "Publishing any content on his/her own personal page, including personal information, the User understands and accepts that this information may be available to other Internet users, taking into account the architecture and functionality of the Site." The VK team confirmed that their public API could be used for research purposes.

We created a list of high schools in Saint Petersburg ($N$ = 601), Samara ($N$ = 214), and Tomsk ($N$ = 99) and then accessed the IDs of users who indicated on VK that they graduated from one of these schools. We removed profiles with no friends from the same school, profiles that already belong to the TrEC data set, and users who indicated several schools in their profiles. The public posts of the remaining users were downloaded and our model was applied to them to get a prediction of the users' academic performance. We then estimated the educational outcomes of a school by averaging the predictions of its students' performance. Overall, 1,064,371 posts from 38,833 users were used at this stage of analysis. The same procedure was performed to obtain a prediction of academic performance for students from the 100 largest universities in Russia ($N_{\text{users}}$ = 115,804; $N_{\text{posts}}$ = 6,508,770). Moscow State University was excluded from the analysis as it is known to be a default choice for bots and fake profiles. Even the application of the aforementioned filtering method does not allow reliable data to be obtained for this university, i.e. there is still an order of magnitude more user profiles than the real number of students for a given cohort.

We used data from the web portal "Schools of Saint Petersburg" [38] to obtain the average performance of schools' graduates in the Unified State Examination (USE). This is a mandatory state examination that all school graduates should pass in Russia. The USE scores for Samara were provided by the web portal "Zeus" [39]. The USE scores of the Tomsk schools along with the data on university enrollees [40] were collected by the Higher School of Economics.

This information was used to check if the scores predicted from social media data correspond to the ranking of schools and universities based on their USE results. Note that

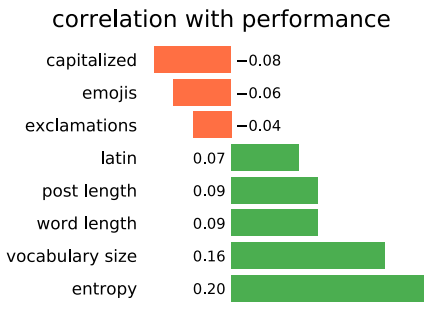


**Figure 1** Pearson correlation between common text features and academic performance. The use of capitalized words, emojis and exclamations (average number per post normalized by the post length in tokens) is negatively correlated with performance. The use of Latin characters, average post and word length, vocabulary size and entropy of users' texts are positively correlated with academic performance

individual USE and PISA scores are not perfectly correlated [41], however it might be assumed that both test measure the same underlying academic ability. It means that our model was tested not only on a different set of users but that the measure of academic performance was also different from the training settings.

### 2.4 Twitter data

VK allows users to indicate links to other social media accounts, including Twitter, in their profiles. Only a small proportion of users provide links to their Twitter accounts. In our sample information, about 665 Twitter accounts were available for the Saint-Petersburg data set (less for other cities) and 2836 Twitter accounts were available for the university data set. This allowed the analysis to be performed only for the university data set. The latest tweets of these 2836 users were downloaded via Twitter's API. Note that unlike tweets, VK posts are not limited to 140 or 280 characters. However, most VK posts are short texts (85.2% of the posts are less than 140 characters and 92.6% of the posts are less than 280 characters in our sample).

## 3 Results

### 3.1 Prediction

We first explored the predictive power of common text features with respect to academic performance. We found a small negative effect for the use of capitalized words ($P = 2 \times 10^{-3}$), emojis ($P = 7 \times 10^{-3}$), and exclamations ($P = 0.05$), as seen in Fig. 1. The use of Latin characters ($P = 5 \times 10^{-3}$), average post length ($P = 2 \times 10^{-4}$), word length ($P = 4 \times 10^{-4}$), and vocabulary size ($P < 10^{-10}$) are positively correlated with academic performance. The strongest correlation was found for the information entropy of users' texts (Pearson's $r = 0.20$, $P < 10^{-15}$).

We used a TF-IDF model to obtain a base-line prediction of academic performance from the users' posts. We selected the 1000 most common unigrams and bigrams from our corpus, excluding stop words, for the Russian language. We then applied a TF-IDF transformation to represent posts as 1000-dimensional vectors and then trained a linear regression model on individual posts to predict the academic performance of their authors. The correlation between predicted and real scores is $r = 0.285$. Here, and for the following models, we report results on the user level obtained using leave-one-out cross-validation, i.e. scores for posts of a certain user were obtained from the model trained on posts of all other users. We obtained significantly better results with a model that used word-embeddings (see Methods). We also find that embeddings trained on the VK corpus outperform models trained on the Wikipedia and Common Crawl corpora (Table 1).

**Table 1** Predictive power of the models measured as Pearson correlation between real and predicted outcomes. The results were computed using leave-one-out cross-validation

| | Correlation coefficient (LOOCV) |
|---|---|
| TF-IDF | 0.284 |
| fastText (Wiki) | 0.335 |
| fastText (CC) | 0.359 |
| fastText (VK) | 0.420 |

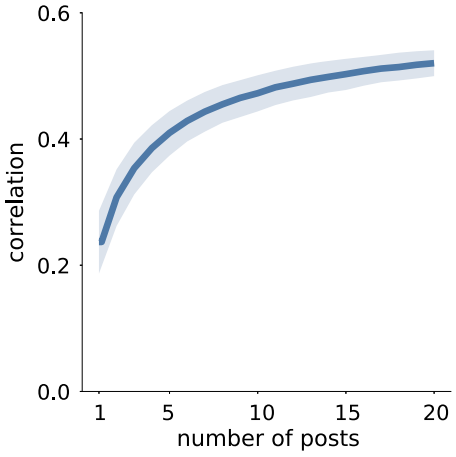


**Figure 2** The predictive power of the model depending on the number of posts used per user. First, users with at least 20 posts were selected. Then, for each user, $N$ of their posts were selected to predict their academic performance ($N = 1, \ldots, 20$). The shaded region corresponds to the bootstrapped 90% confidence interval

**Table 2** Model performance in discrimination between different proficiency levels measured as the area under the ROC curve

| | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Level 6 |
|---|---|---|---|---|---|---|
| Bellow level 1 | 0.573 | 0.665 | 0.774 | 0.870 | 0.951 | 0.987 |
| Level 1 | | 0.604 | 0.723 | 0.833 | 0.930 | 0.971 |
| Level 2 | | | 0.621 | 0.748 | 0.864 | 0.919 |
| Level 3 | | | | 0.649 | 0.797 | 0.889 |
| Level 4 | | | | | 0.666 | 0.770 |
| Level 5 | | | | | | 0.571 |

The predictive power of a model depends on the number of posts available for each user (see Fig. 2). If only one post is available per user, the predictive power is rather low ($r = 0.237$). However, it increases with the number of posts available, reaching $r = 0.541$ for 20 posts per user.

In addition to predicting raw scores, we analyze how well the model could distinguish between low- and high-performing users. To help interpret what student scores mean in substantive terms, the PISA scale is divided into six proficiency levels [30]. In Table 2, we report the ability of our model to distinguish between students of different proficiency levels. The performance of the model is measured as the area under the ROC curve (AUC). According to the OECD, Level 2 is a baseline proficiency that is required to participate fully in modern society [30]. Students who do not meet this baseline are considered as low-performing. High-performing students are those who achieve proficiency Level 5 or higher. The ability of our model to distinguish between low- and high-performing students, i.e. between Levels 0 & 1 and Levels 5 & 6 is 93.7%.

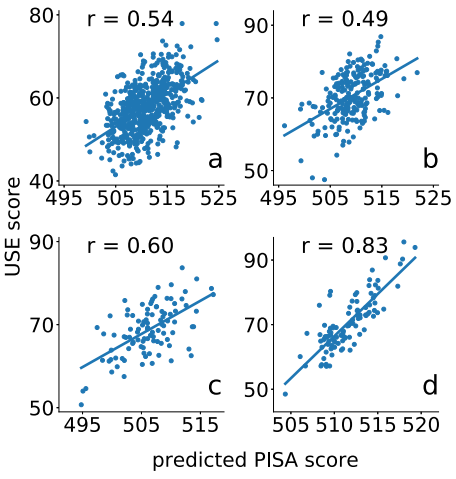


**Figure 3** Correlation between the predicted and real performance of schools and universities. Pearson's correlation coefficients between predicted school scores and the USE scores of their graduates were computed for Saint-Petersburg (**a**), Samara (**b**) and Tomsk (**c**). The correlation between predicted university scores and the USE scores of their enrollees was also computed for the 100 largest Russian universities (**d**)

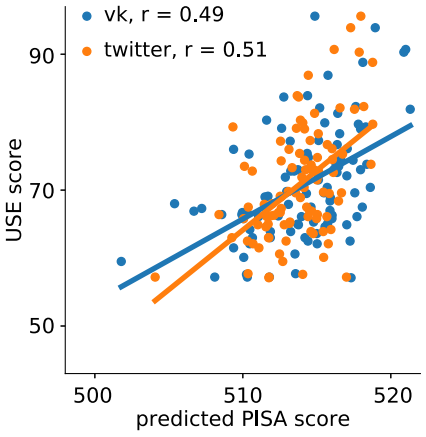


**Figure 4** Comparison of predictions based on VK and Twitter data. While estimates from Twitter and VK vary for individual universities the overall performance of the model is similar for both cases. Note that the performance of the model is rather low due to the limited number of users per university for whom both VK and Twitter data is available

### 3.2 Transfer

Figure 3 shows the correlation between the predicted performance of schools (a)–(c) and universities (d) and the USE scores of their graduates or enrollees. In all four cases, we find a relatively strong signal despite the fact that the VK sample might not be representative and that academic performance was measured differently than in training settings, being available only in aggregated form and from secondary sources.

Intriguingly, we find that the substitution of VK posts by tweets doesn't substantially alter the resulting performance (see Fig. 4)). For fair comparison, we use VK data only for those users for whom Twitter data was also available. This is why the performance of the model is substantially lower than in Fig. 3(d), where all available VK data was used.

### 3.3 Differential language use

We explored the resulting model by selecting 400 words with the highest and lowest scores that appear at least 5 times in the training corpus. A t-SNE representation of the embeddings produced by our model helped [42] to identify several thematic clusters (Fig. 5). High performing clusters include English words (above, saying, yours, must), words related to literature (Dandelion, Bradbury, Fahrenheit, Orwell, Huxley, Faulkner, Nabokov, Brodsky, Camus, Mann, Shelley, Shakespeare), words related to reading (read, reread, published,

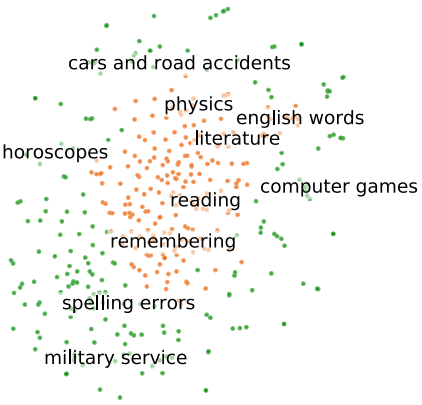


**Figure 5** t-SNE representation of the words with the highest and lowest scores from the training data set. High performing clusters (orange) include English words and words related to literature, physics, or thinking processes. Low performing clusters (green) include spelling errors and words related to horoscopes, military service, or cars and road accidents

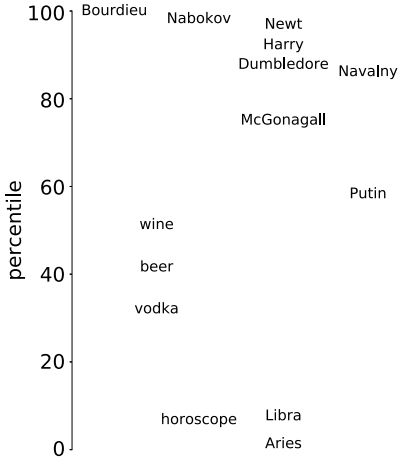


**Figure 6** Ranking of selected words by their predicted score (translated from Russian). The application of the model to words from different domains confirms its face validity

book, volume), words related to physics (universe, hole, string, theory, quantum, Einstein, Newton, Hawking), and words related to thinking processes including various synonyms of "thinking" and "remembering."

Low performing clusters include common spelling errors and typos, names of popular computer games, words related to military service (army, to serve, military oath), horoscopes (Aries, Sagittarius), and cars and road accidents (traffic collision, General Administration for Traffic Safety, wheels, tuning).

The use of continuous word representations allows one to compute scores even for words that were not present in the training data set. We computed the scores for all 2.5M words in our vector model and made it available for further exploration [36]. This could be used for exploratory analysis to get insights into differential use of language with respect to academic performance and could be applied to various domains, from literature to politics or food (Fig. 6).

## 4 Discussion

In traditional cross-validation settings, one data set is randomly split into two parts: the model is trained on one part, and then its performance is assessed using the other part. In our case, we use two different data sources (VK and Twitter) and different measures of academic performance (PISA scores and USE scores). Some additional limitations were

that the PISA scores were collected in 2012, while most of the posts were written much later, and that the USE scores were available only at the aggregated level. Despite these limitations, we found a relatively strong signal in the data. For instance, we were able to explain 69% of the variation in universities' scores using information about VK posts of users from these universities. While the result for tweets was significantly lower ($R^2$ = 0.26), this was probably at least partly due to the smaller sample size. Also, note that the prediction of individual scores substantially depends on the number of available posts. If only one post per user is used for prediction, then 6% of the variation in their academic performance could be explained by our model. This number rises to 29% if 20 posts are used.

The ability to predict the ranking of educational organizations might seem trivial given that direct ranking information is readily available. However, USE scores are measured only once per cohort, making it extremely hard to estimate any added value provided by an educational organization. A comparison of predicted scores for the same users at different points in time might shed light on factors contributing to the students' progress.

We demonstrate how domain-specific unsupervised learning of word embeddings allows predictive models to be trained using relatively small labeled data sets. One reason is that even words that are rare in the training or testing data sets could be valuable for prediction. For instance, even if the word "Newt" never occurs in the training data set, the model could assign a higher score to posts containing it. This would happen if the model learns from the training data set that words from the Harry Potter universe are indicative of high performing students and learns from the unsupervised training that "Newt" belongs to this category, i.e. this word is close to other Harry Potter related words in the vector space. This might make the use of continuous word representation preferable to common approaches relying on counting word frequencies. As our approach does not depend on a particular language, source of texts, or target variable (i.e. academic performance could be substituted by income or depression), it could be applied to a wide variety of settings.

Our results also suggest that models trained on text data could be successfully transferred from one data source to another. While this certainly might be useful in some applications, it also means there is a greater risk to users' privacy. If users of platform A do not disclose an attribute X on it, then there is no data to train a model to predict X from digital traces on platform A. However, if X is disclosed on platform B, and both platforms collect short texts from users, then it becomes possible to predict X from digital traces on A given access to data from B. In recent years, face recognition technology has raised particular privacy concerns because of its potential omnipresence and the inability of people to hide from it. In a similar way, digital traces in the form of short texts are ubiquitous, and our results suggest that they allow, if not to identify a person, then at least to predict potentially sensitive private attributes.

**Acknowledgements**

Author thanks the Open Data University Research Consortium [43] for providing data on public posts of VK users. Author thanks Ilyuhin B.V., vice rector for informatization and education quality assessment, TOIPKRO, and the Centre of General and Extracurricular Education, HSE University, for providing data on USE scores of Tomsk schools. Author thanks Yulia Torgasheva, head of Zeus web-portal for providing data on USE scores of Samara schools. The TrEC project is supported by the Basic Research Program of the National Research University Higher School of Economics.

**Funding**

This work was supported by a grant from the Russian Science Foundation (project N019-18-00271).

**Availability of data and materials**
The models and the word rankings are available in the Open Science Framework repository http://doi.org/10.17605/OSF.IO/9PBKR. TrEC data cannot be publicly shared but is available to interested researchers upon request https://trec.hse.ru/en/data.

**Competing interests**
The authors declare that they have no competing interests.

**Author's contributions**
Author read and approved the final manuscript.

## Publisher's Note



Received: 25 June 2020 Accepted: 24 August 2020 Published online: 01 September 2020

**References**

1. Golder SA, Macy MW (2014) Digital footprints: opportunities and challenges for online social research. Annu Rev Sociol 40:129–152
2. Lazer D, Radford J (2017) Data ex machina: introduction to big data. Annu Rev Sociol 43:19–39
3. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264):1073–1076
4. Gebru T, Krause J, Wang Y, Chen D, Deng J, Aiden EL, Fei-Fei L (2017) Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proc Natl Acad Sci 114(50):13108–13113
5. Hills TT, Proto E, Sgroi D, Seresinhe CI (2019) Historical analysis of national subjective wellbeing using millions of digitized books. Nat Hum Behav: 1–5
6. An J, Weber I (2016) # greysanatomy vs.# yankees: demographics and hashtag use on Twitter. In: Tenth international AAAI conference on web and social media
7. Preoţiuc-Pietro D, Volkova S, Lampos V, Bachrach Y, Aletras N (2015) Studying user income through language, behaviour and affect in social media. PLoS ONE 10(9):0138717
8. Lampos V, Aletras N, Geyti JK, Zou B, Cox IJ (2016) Inferring the socioeconomic status of social media users based on behaviour and language. In: European conference on information retrieval. Springer, Berlin, pp 689–695
9. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman ME et al(2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE 8(9):73791
10. Stier S, Breuer J, Siegers P, Thorson K (2019) Integrating survey data and digital trace data: key issues in developing an emerging field. Soc Sci Comput Rev
11. Settanni M, Azucar D, Marengo D (2018) Predicting individual characteristics from digital traces on social media: a meta-analysis. Cyberpsychol Behav Soc Netw 21(4):217–228
12. Chancellor S, De Choudhury M (2020) Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 3(1):1–11
13. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110(15):5802–5805
14. Bogolyubova O, Panicheva P, Tikhonov R, Ivanov V, Ledovaya Y (2018) Dark personalities on Facebook: harmful online behaviors and language. Comput Hum Behav 78:151–159
15. De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media
16. Reece AG, Danforth CM (2017) Instagram photos reveal predictive markers of depression. EPJ Data Sci 6(1):1
17. Organisation for Economic Cooperation and Development (2013) PISA 2012 Assessment and Analytical Framework Mathematics, Reading, Science, Problem Solving and Financial Literacy. OECD Publishing
18. Arendt JN (2005) Does education cause better health? A panel data analysis using school reforms for identification. Econ Educ Rev 24(2):149–160
19. Gottfredson LS, Deary IJ (2004) Intelligence predicts health and longevity, but why? Curr Dir Psychol Sci 13(1):1–4
20. Roth PL, BeVier CA, Switzer FS III, Schippmann JS (1996) Meta-analyzing the relationship between grades and job performance. J Appl Psychol 81(5):548
21. Olsson CA, McGee R, Nada-Raja S, Williams SM (2013) A 32-year longitudinal study of child and adolescent pathways to well-being in adulthood. J Happ Stud 14(3):1069–1083
22. Alyahyan E, Düştegör D (2020) Predicting academic success in higher education: literature review and best practices. Int J Educ Technol Higher Educ 17(1):3
23. Hellas A, Ihantola P, Petersen A, Ajanovski VV, Gutica M, Hynninen T, Knutas A, Leinonen J, Messom C, Liao SN (2018) Predicting academic performance: a systematic literature review. In: Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education, pp 175–199
24. Giunchiglia F, Zeni M, Gobbi E, Bignotti E, Bison I (2018) Mobile social media usage and academic performance. Comput Hum Behav 82:177–185
25. Lian D, Ye Y, Zhu W, Liu Q, Xie X, Xiong H (2016) Mutual reinforcement of academic performance prediction and library book recommendation. In: 2016 IEEE 16th international conference on data mining (ICDM). IEEE Press, New York, pp 1023–1028
26. Wang R, Harari G, Hao P, Zhou X, Campbell AT (2015) Smartgpa: how smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, pp 295–306

27. Kassarnig V, Bjerre-Nielsen A, Mones E, Lehmann S, Lassen DD (2017) Class attendance, peer similarity, and academic performance in a large field study. PLoS ONE 12(11)
28. Helal S, Li J, Liu L, Ebrahimie E, Dawson S, Murray DJ, Long Q (2018) Predicting academic performance by considering student heterogeneity. Knowl-Based Syst 161:134–146
29. Malik V (2019) The Russian panel study'trajectories in education and careers'. Longit Life Course Stud 10(1):125–144
30. Organisation for Economic Cooperation and Development (2014) PISA 2012 Results What Students Know and Can Do. Student Performance in Mathematics, Reading and Science. OECD Publishing
31. Sanders WL, Horn SP (1994) The Tennessee value-added assessment system (TVAAS): mixed-model methodology in educational assessment. J Pers Eval Educ 8:299–311
32. Schleicher A, Zimmer K, Evans J, Clements N (2009) Pisa 2009 assessment framework: key competencies in reading, mathematics and science. OECD Publishing (NJ1)
33. Kern ML, Eichstaedt JC, Schwartz HA, Dziurzynski L, Ungar LH, Stillwell DJ, Kosinski M, Ramones SM, Seligman ME (2014) The online social self: an open vocabulary approach to personality. Assessment 21(2):158–169
34. Kulkarni V, Kern ML, Stillwell D, Kosinski M, Matz S, Ungar L, Skiena S, Schwartz HA (2018) Latent human traits in the language of social media: an open-vocabulary approach. PLoS ONE 13(11):0201703
35. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
36. Smirnov I Predicting academic performance from short texts on social media. https://doi.org/10.17605/OSF.IO/9PBKR
37. Raghu M, Schmidt E (2020) A survey of deep learning for scientific discovery. arXiv preprint. arXiv:2003.11755
38. Schools of Saint Petersburg: Schools of Saint Petersburg. https://shkola-spb.ru/
39. Zeus: Zeus. http://zeus.volgamonitor.com/
40. Higher School of Economics: Quality of University Admission. https://ege.hse.ru/
41. Jackson M, Khavenson T, Chirkina T (2020) Raising the stakes: inequality and testing in the Russian education system. Soc Forces 98(4):1613–1635
42. Maaten LVD, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9:2579–2605
43. Open Data University Research Consortium. https://opendata.university/en/