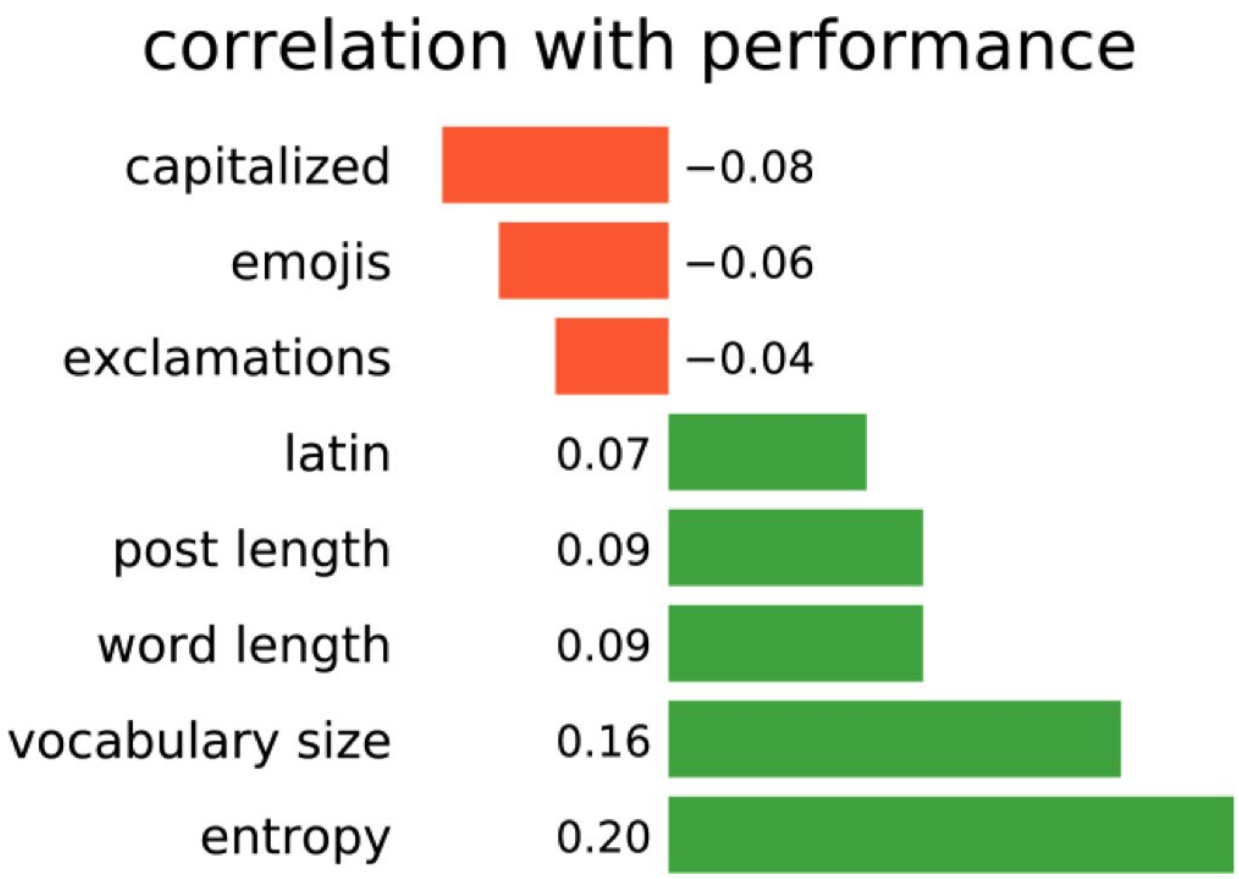
Artificial intelligence can predict students' educational outcomes based on tweets

October 22 2020



Pearson correlation between common text features and academic performance. Credit: Article I. Smirnov

Ivan Smirnov, Leading Research Fellow of the Laboratory of Computational Social Sciences at the Institute of Education of HSE

University, has created a computer model that can distinguish high academic achievers from lower ones based on their social media posts. The prediction model uses a mathematical textual analysis that registers users' vocabulary (its range and the semantic fields from which concepts are taken), characters and symbols, post length, and word length.

Every word has its own rating (a kind of IQ). Scientific and cultural topics, English words, and words and posts that are longer in length rank highly and serve as indicators of good academic performance. An abundance of emojis, words or whole phrases written in capital letters, and vocabulary related to horoscopes, driving, and military service indicate lower grades in school. At the same time, posts can be quite short—even tweets are quite informative. The study was supported by a grant from the Russian Science Foundation (RSF), and an article detailing the study's results was published in *EPJ Data Science*.

Foreign studies have long shown that users' social media behavior—their posts, comments, likes, profile features, user pics, and photos—can be used to paint a comprehensive portrait of them. A person's social media behavior can be analyzed to determine their lifestyle, personal qualities, individual characteristics, and even their mental health status. It is also very easy to determine a person's socio-demographic characteristics, including their age, gender, race, and income. This is where profile pictures, Twitter hashtags, and Facebook posts come in.

A user's likes on Facebook can reveal their religious and political views, sexual orientation, personal qualities, and level of life satisfaction. A user's comments on Facebook can reveal their level of aggressiveness, while a user's tweets on Twitter can indicate whether they suffer from depression. Blog posts also speak volumes about their authors' personalities. Even pictures and images (for example on Instagram) are a great source for digital psychometrics.

But digital traces provide rich material not only at the individual level but also at the aggregated level. For example, you can calculate the electoral preferences of city residents using data from Google Street View, a function that allows you to view panoramas of city streets and see what kinds of vehicles residents drive. Digitized books can be used to ascertain the level of national subjective wellbeing in different countries.

However, very few studies have focused on composite, more complex characteristics, such as a student's academic success and performance in school, which are dependent upon many factors. 'In our new study, we tried to predict the performance of high school and university students based on their VK and Twitter posts,' says Ivan Smirnov. 'Learning ability is a very complex human characteristic. It is influenced not only by character traits, but also by psychological well-being. Alas, in contrast to academic success, which is available in the public domain, there are no mechanisms within educational institutions for measuring the latter.'

It would seem intuitively clear that if a student posts about quanta, string theory, Newton, Shakespeare, and Nabokov on social media, they are at the very least motivated to learn. If you were to look at that student's report card, you'd probably see A's and B's. And if a teenager is writing posts about horoscopes or car accidents that are littered with spelling errors, chances are they are not an especially strong student. But to keep intuition from becoming a cognitive bias, it is always best to prove it with numbers. For example, it is possible calculate mathematically what words are 'smarter'.

However, the most important aspect of using digital data is that many things about adolescents are difficult to determine using traditional research methods, such as surveys and interviews. These research methods often contain personal questions, which teenagers may avoid answering or not answer truthfully. Teenagers tend to be more reserved

and are therefore more difficult for scientists to study. But digital data can provide deeper insight into this age group, and reveal hitherto unknown aspects of their life.

Smirnov's study used a representative sample of data from HSE University's longitudinal cohort panel study, 'Educational and Career Trajectories' (TrEC). The study traces the career paths of 4,400 students in 42 Russian regions from high schools participating in PISA (the Program for International Students Assessment). The study data also includes data about the students' VK accounts (3,483 of the student participants consented to provide this information).

"Since this kind of data, in combination with digital traces, is difficult to obtain, it is almost never used," Smirnov says. Meanwhile, this kind of dataset allows you to develop a reliable model that can be applied to other settings. And the results can be extrapolated to all other students—high school students and middle school students.

Posts from publicly viewable VK pages were used as a training sample—this included a total of 130,575 posts from 2,468 subjects who took the PISA test in 2012. The test allowed the researcher to assess a student's academic aptitude as well as their ability to apply their knowledge in practice. The study included only publicly visible VK posts from consenting participants.

When developing and testing the model from the PISA test, only students' reading scores were used an indicator of academic aptitude, although there are three tests in total: reading, mathematics, and science. 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.' The exam has six proficiency levels. Students who score a 2 are considered to meet only the basic, minimum level, while those who score a 5 or 6 are

considered to be strong students.

In the study, unsupervised machine learning with word vector representations was performed on VK post corpus (totaling 1.9 billion words, with 2.5 million unique words). It was combined with a simpler supervised machine learning model that was trained in individual positions and taught to predict PISA scores.

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

By 'predict,' the researcher does not refer to future forecasting, but rather the correlation between the calculated results and the real scores students earned on the PISA exam, as well as their USE scores (which are publicly available online in aggregated form—i.e., average scores per school). In the preliminary phase, the model learned how to predict the PISA data. In the final model, the calculations were checked against the USE results of high school graduates and university entrants.

The final model was supposed to be able to reliably recognize whether a strong student or a weak student had written a particular social media post, or in other words, differentiate the subjects according to their academic performance. After the training period, the model was able to distinguish posts written by students who scored highly or poorly on PISA (levels 5-6 and levels 0-1) with an accuracy of 93.7%. As for the comparability of PISA and the USE, although these two tests differ, studies have shown that students' scores for the two tests strongly correlate with each other.

"The model was trained using PISA data, and we looked at the correlation between the predicted and the real PISA scores (which are

available in the TrEC study)," Smirnov explains. "With the USE things gets more complicated: since the model does not know anything about the unified exams, it predicted the PISA scores as before. But if we assume that the USE and PISA measure the same thing—academic performance—then the higher the predicted PISA results are, the higher the USE results should be." And the fact that the model learned to predict one thing and can predict another is quite interesting in itself, Smirnov notes.

However, this also needed to be verified, so the model was then applied to 914 Russian high schools (located in St. Petersburg, Samara and Tomsk; this set included almost 39,000 users who created 1.1 million posts) and one hundred of Russia's largest universities (115,800 people; 6.5 million posts) to measure the academic performance of students at these institutions.

It turned out that "predicted academic performance is closely related to USE scores," says Smirnov. "The correlation coefficient is between 0.49 and 0.6. And in the case of universities, when the predicted academic performance and USE scores of applicants were compared (the information is available in HSE's ongoing University Admissions Quality Monitoring study), then the results also demonstrated a strong connection. The correlation coefficient is 0.83, which is significantly higher than for high schools, because there is more data."

But can the model be applied to other social media sites? "I checked what would happen if, instead of posts on VK, we gave the model tweets written by the same users," Smirnov says. 'It turned out that the quality of the model does not significantly decrease.' But since a sufficient number of twitter accounts were available only for the university dataset (2,836), the analysis was performed only on this set.

It is important that the model worked successfully on datasets of

different social media sites, such as VK and Twitter, thereby proving that is can be effective in different contexts. This means that it can be applied widely. In addition, the model can be used to predict very different characteristics, from student academic performance to income or depression.

First, Smirnov highlighted the general textual features of posts in relation to the academic performance of their authors (Fig. 1). The use of capitalized words (-0.08), emojis (-0.06), and exclamations (-0.04) were found to be negatively correlated with academic performance. The use of the Latin characters, average post and word length, vocabulary size, and entropy of users' texts on the other hand, were found to positively correlate with academic performance (from 0.07 to 0.16, respectively).

It was also confirmed that students with different levels of academic performance have different vocabulary ranges. Smirnov explored the resulting model by selecting 400 words with the highest and lowest scores that appear at least 5 times in the training corpus. Thematic clusters were identified and visualized (Fig. 2).

The clusters with the highest scores (in orange) include:

- English words (above, saying, yours, must);
- Words related to literature (Bradbury, Fahrenheit, Orwell, Huxley, Faulkner, Nabokov, Brodsky, Camus, Mann);
- Concepts related to reading (read, publish, book, volume);
- Terms and names related to physics (Universe, quantum, theory, Einstein, Newton, Hawking);

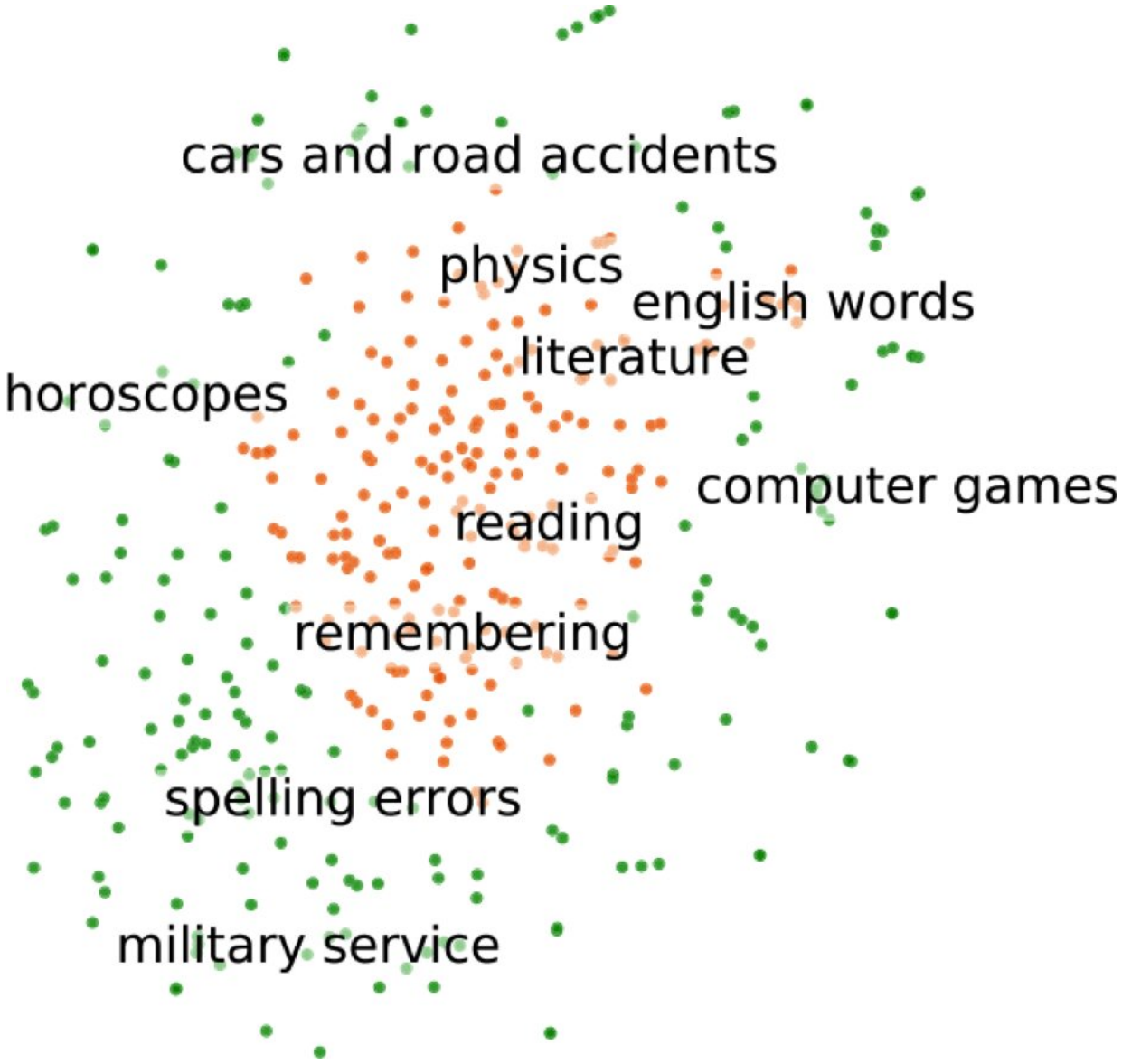
- Words related to thought processes (thinking, memorizing).

Clusters with low scores (in green) include misspelled words, names of

popular computer games, concepts related to military service (army, oath, etc.), horoscope terms (Aries, Sagittarius), and words related to driving and car accidents (collision, traffic police, wheels, tuning).

Smirnov calculated the coefficients for all 2.5 million words of the vector model and made them available for further study. Interestingly, even words that are rarely found in a training dataset can predict academic performance. For example, even if the name 'Newt' (as in the Harry Potter character, Newt Scamander) never appears in the training dataset, the model might assign a higher rating to posts that contain it. This will happen if the model learns that words from novel series are used by high-achieving students, and, through unsupervised learning, 'intuit' that that the name 'Newt' belongs to this category (that is, the word is closely situated to other concepts from Harry Potterin the vector space).

Why is this model important?



Thematic clusters: t-SNE representation of the words with the highest and lowest scores from the training data set. Credit: Article I. Smirnov

The proposed model can be applied to a wide variety of areas, such as literature, food, politics, and more. For example, education researchers are interested in understanding what distinguishes successful schools from average schools. But if, say, you look at the features of schools

with high USE scores, this does not tell us much, but it is clear that strong, better prepared students study in these schools.

"It would be good for researchers to look at the schools that show the greatest increase in scores," says Ivan Smirnov. "And theoretically, our method can be used to measure this increase and then look at the factors associated with it on the school level." In general, since the model 'does not depend on a particular language, source of texts, or target variable,' he notes, "it could be applied to a wide variety of settings."

Notably, there has never been such a representative study in Russia that correlates student academic data and their social media activity. However, there have been studies—also conducted by Ivan Smirnov—that approach it. Three years ago, he showed that the pages and groups high school students follow on VK correlate not only with their interests, but also with their academic performance.

In a previous study, Smirnov built a model that allows one to determine students' PISA scores based on the pages they follow on social media. As it turned out then, strong and weak students have different interests, which is reflected in the pages they choose to follow. Thus, stronger students often visit pages about science, technology, and culture, while poorer students are more likely to visit humor and horoscope pages. As the researcher commented at the time, strong students "seek out content online that is educational rather than entertaining."

Digital trace analysis is extremely popular, and studies are always of great interest, but a central issue concerning how they are conducted—their ethics—has yet to be resolved. Is it morally acceptable to use data from social media? One has to weigh the pros and cons and decide what outweighs the other, says Smirnov.

While all of the information is public, there are important factors that

could weigh against its use. "For example, not all users may understand that information about them is available, and not all users have the technical skills to download information, which creates inequality," says Ivan Smirnov. "No one thinks twice about viewing someone else's profile, because this is something everyone can do, but only in theory can everyone download information."

"Probably, not all people think that when they go to someone's VK page, they are doing the same thing as researchers or private companies: they send a request to the servers of the social network and save information about the person on their computer," says Smirnov. "They just do it not through their own script, like we do, but with the help of a ready-made program—a browser." Technically, however, these actions are identical; the differences are only in scale. Although it is also quite possible to view the pages of several thousand people without any scripts.

There are also factors which can tip the scales in favor of the 'pro' side. "Unlike private companies—some of which download the entire VK site every day—our activities are non-commercial and public," the researcher notes. "We are trying to gain new useful knowledge about the world, and everything we do is publicly available."

And if a user, for example, has read a piece on IQ.HSE and becomes angry that they are being 'watched,' this means that they have benefited from the study. "Now they know that their data on VK can be used, and they will go and protect by using the site's privacy settings. Not only we, researchers, but also malicious agents, will not be able to access it," says Ivan Smirnov.

However, Smirnov's study does not violate any moral norms or personal boundaries. First, it only used publicly accessible posts. Secondly, the VK administration specifically warns that "personal information posted by the User may become available to other Site Users and Internet users,

be copied and disseminated by such users."

"Our studies were assessed by an ethical commission and accepted to leading journals. [This methodology] is now considered the norm," says Ivan Smirnov.

Moreover, the methods used in the study are not designed to work at the individual level. "Our findings [in a number of recent projects] are general in nature: parents mention their sons more often than their daughters on social media, and students with similar academic performance levels are more likely to be friends with each other," says Ivan Smirnov. "And when I say that we can use this method to track student progress, I don't mean at the individual level."

More information: Ivan Smirnov, Estimating educational outcomes from students' short texts on social media, *EPJ Data Science* (2020). DOI: 10.1140/epjds/s13688-020-00245-8

Provided by National Research University Higher School of Economics