Факультет компьютерных наук

Научно-учебная лаборатория моделей и методов вычислительной прагматики

Москва
2024

Предсказание образовательных успехов студентов на основе информации из профиля в социальной сети ВКонтакте

Докладчик: Сергей Горшков,
Приглашенный преподаватель, аспирант ФКН ВШЭ

Исследовательские цели

- Создать цифровой профиль пользователя на основе открытых данных из ВКонтакте, используя различные типы данных.
- Разработать подход для создания векторного представления пользователя на основе открытых данных, которое может быть использовано как описание признаков или для дальнейшего обучения на конкретные задачи.
- Исследовать значимость отдельных факторов для различных задач.
- Анализ образовательных данных.
- Выполнить кластеризацию и мультикластеризацию по различным типам данных.

Актуальность, новизна и заинтересованные стороны

- Аналогичные работы не были найдены; как правило, используются опросы и/или подходы на основе графов, при этом почти отсутствует анализ контента. Максимум, что встречается — это анализ комментариев.
- Изначально данное решение предназначалось для использования в образовательном контексте для построения индивидуальных образовательных траекторий и оказания психологической поддержки студентам, создания менторских программ и выявления скрытых предпочтений и склонностей к изучению новых, нестандартных комбинаций предметов.
- Улучшение таргетинга в рассылках, рекламе и поиске со стороны HR-агентств.
- Использование подходов цифровыми экосистемами, которые обладают более широким доступом к данным для кросс-доменных рекомендаций.
- Социологические и психологические исследования на основе открытых данных.

Постановка конкретной задачи и целей исследования

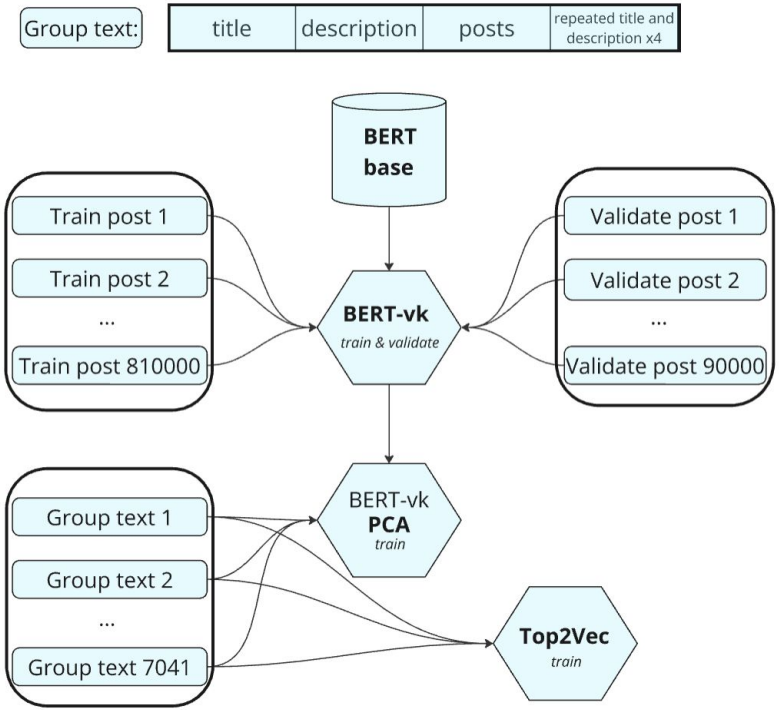
Выявление отлично учащихся студентов на основе их подписок в социальной сети ВКонтакте с применением различных методов обработки естественного языка

1. Обучить модель для получения векторных представлений текста, адаптированных к формату постов в группах ВКонтакте.
2. Создать цифровой профиль пользователя на основе его подписок в социальной сети ВКонтакте и сформировать профиль "успешного студента".
3. Предсказать уровень успеваемости студентов — относятся ли они к высокоуспевающим (студенты с оценками "A") или низкоуспевающим (студенты с оценками "C").
4. Проанализировать наиболее значимые факторы, влияющие на академическую успеваемость.

Использование персональных данных

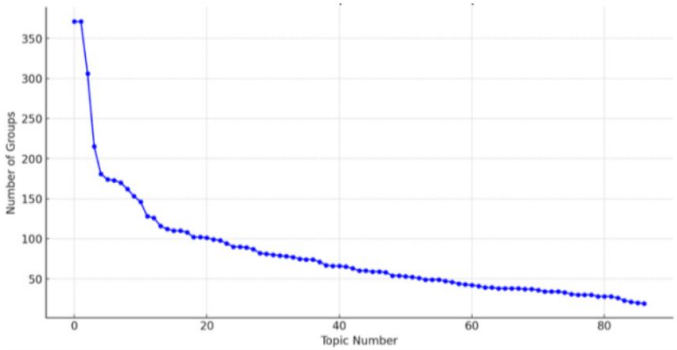
- Была использована информацию о 4513 студентах Томского государственного университета. Участники предоставили согласие на обработку своих персональных данных, в частности, идентификатора (id) своей персональной страницы ВКонтакте, и разрешили сбор общедоступной информации с их страниц, в том числе с использованием автоматизированных средств. Мы объяснили участникам цель исследования и заверили их, что данные будут надежно храниться и не передаваться третьим лицам. Кроме того, было указано, что данные будут анонимизированы перед публикацией результатов опроса.
- Для каждого студента мы использовали только пол, информацию о факультете и уровне образования (бакалавриат, специалитет, магистратура), список id групп, на которые студент подписан во ВКонтакте, а также оценки студента.
- Было получено этическое одобрение от институционального наблюдательного совета на проведение исследований

Обучение моделей для эмбеддингов и тематического моделирования



Обучение моделей для эмбеддингов и тематического моделирования

Распределение числа документов в темах после тематического моделирования. Самые популярные темы: Юмор, мемы, о кино, мода, любовные эмоции, музыкальные события, путешествия, подарки, знаменитости, политика, домашние животные

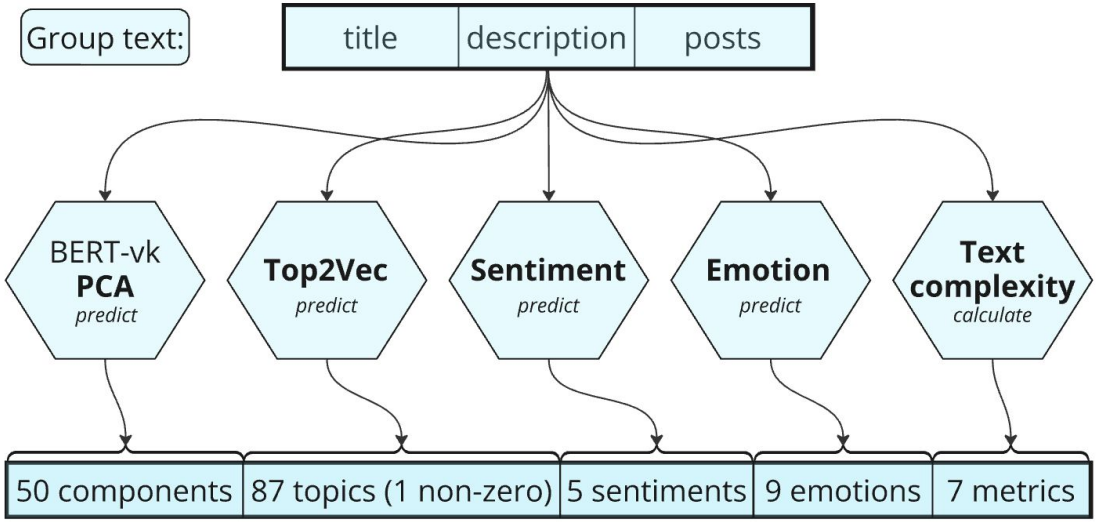


EMBEDDING QUALITY METRICS

Model	MLM Accuracy	Perplexity
baseBERT	0.893	3.342
baseBERT + domain adaptation	**0.927**	**2.496**
mBERT	0.886	4.560
mBERT + domain adaptation	0.922	2.614
ruBERT	0.855	8.506
ruBERT + domain adaptation	0.917	4.598

Результаты доменной адаптации моделей

Создание векторного представления для сообществ



Создание векторного представления для пользователя

Algorithm 1: High-level steps for constructing a digital student profile.

input: DS: a dataset containing user interaction data with VK group_ids

output: UIV: user interest vectors

for each user u **in** DS:

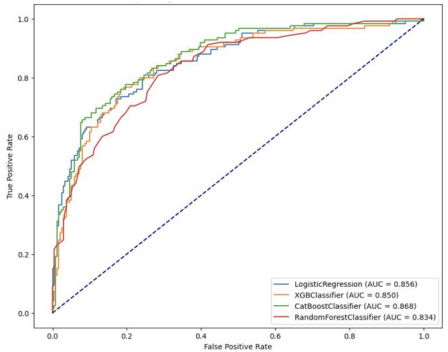
1. **initialize** user interest vector UIV_u to zero.
2. **for each** community with index i of N **in** user u's ranked community list:
   a. **assign** weight $\omega(i)$as follows:
      if $i \leq 3$ **then** $\omega(i) \leftarrow 1$
      if $i > 3$ **then** $\omega(i) \leftarrow \frac{2}{\sqrt{i+1}}$
   b. **for each** vector v component $v[j]$ with index j:
      if v from *PCA-transformed Embeddings* or *Sentiment/Emotion Analysis* **then:**
      $$UIV_u[j] += \frac{\omega(i) * v[j]}{N}$$
      if v from *Topic Modeling* **then**:
      $$UIV_u[j] += \omega(i) * v[j]$$
      if v from *Text Complexity Metrics* **then**:
      $$UIV_u[j] = \max(UIV_u[j], \omega(i) * v[j])$$
3. **return** the constructed user interest vector $UIV_u[j]$ for the user u.

Разделение наиболее и наименее успевающих студентов

- P: Perfect Achievers (GPA = 5.0): 647 students
- AA: Above Average Performers (GPA ≥ 4.8): 1177 students
- BA: Below Average Performers (GPA < 4.0): 1171 students
- L: Low Performers (GPA < 3.8): 945 students
- NP: Non-Perfect Performers (GPA < 5.0): 3798 students

ROC-AUC Scores Obtained On Cross-validation For Binary Classification Under Different Conditions

Parameter Groups	Logistic Regression	XGBoost	CatBoost	Random Forest
P vs BA	0.8563	0.8499	**0.8671**	0.8342
P vs L	0.8413	0.8653	**0.8765**	0.8392
AA vs BA	0.8484	**0.8740**	0.8737	0.8496
AA vs L	0.8571	0.8542	**0.8619**	0.8542
P vs NP	0.7396	0.7407	**0.7455**	0.7405

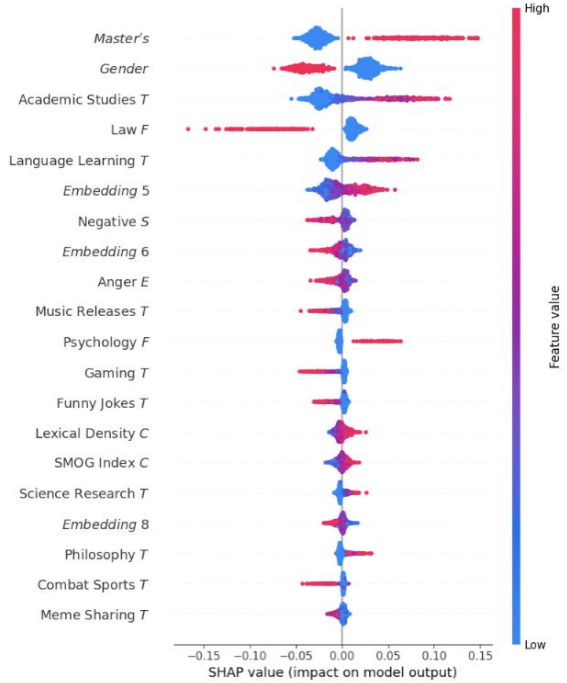


Наиболее значимые признаки

TOP 15 MOST SIGNIFICANT FEATURES FOR DIFFERENT ALGORITHMS ON THE P VS BA TASK

	Logistic Regression	Random Forest	XGBoost	CatBoost
1	- Law F	+ Academic Studies T	+ Academic Studies T	+ Academic Studies T
2	+ Academic Studies T	- Law F	- Law F	- Law F
3	- Music Releases T	+ Language Learning T	- Side-Job Listings T	+ Language Learning T
4	+ Language Learning T	+ SMOG Index C	+ Psychology F	- Negative S
5	+ Psychology F	+ Philosophy T	- Negative S	- Anger E
6	- Music Events T	+ Lexical Density C	+ Enthusiasm E	- Music Releases T
7	- Gaming T	+ University Life T	- Astrology T	+ Psychology F
8	- Footbal T	- Relatable Humor T	+ University Life T	- Gaming T
9	+ Philosophy T	- Anger E	+ Language Learning T	- Funny Jokes T
10	+ Science Research T	+ Coleman-Liau Index C	+ Science Research T	+ Lexical Density C
11	- Music Trends T	+ Photography T	- Music Releases T	+ SMOG Index C
12	+ University Life T	- Meme Sharing T	- Gaming T	+ Science Research T
13	- Side-Job Listings T	- Russian Football T	- Music Events T	+ Philosophy T
14	- Funny Jokes T	+ Classic Literature T	+ SMOG Index C	- Combat Sports T
15	- Video Games T	- Music Events T	- Football T	- Meme Sharing T

Наиболее значимые признаки



- Темы, положительно связанные с академической успеваемостью, включают: учебные занятия, городская жизнь, изящные искусства, декор дома, изучение языков, владение языками, домашние животные, философия, поэтическая меланхолия, программирование, научные исследования, путешествия, студенческая жизнь, волонтерство, похудение.
- Темы, отрицательно связанные с академической успеваемостью, включают: аниме, астрология, бодибилдинг, автозапчасти, боевые виды спорта, модные товары, футбол, смешные шутки, видеоигры, распространение мемов, музыкальные события, музыкальные релизы, музыкальные тренды, юмор о повседневной жизни, вакансии подработок, татуировки, видеоигры.

Основные выводы

- Алгоритм показал высокую эффективность в разделении студентов с высокой и низкой успеваемостью, достигнув ROC-AUC выше 0.86. Это свидетельствует о том, что алгоритм может быть полезным инструментом для выявления потенциально успешных студентов на основе открытых данных. Предсказание академической успеваемости возможно на основе подписок студентов.
- Алгоритм гибкий и может применяться к новым данным без необходимости повторного обучения. Анализ ключевых признаков с использованием SHAP значений подтвердил согласованность с существующими исследованиями.
- Созданные векторные представления пользователей могут быть применены к любому профилю ВКонтакте, что расширяет возможности их использования. Тем не менее, результаты не должны использоваться для дискриминации, а лишь как дополнительный сигнал в сложных ситуациях.

Вторая задача: использование других модальностей

Предсказание вероятности отчисления студентов

1. Определить подходящие методы анализа контента в социальных сетях, которые учитывают образовательные аспекты и эффективно характеризуют студентов.
2. Разработать комплексный цифровой профиль студентов на основе контента, которым они делятся в социальных сетях, с учетом временных аспектов.
3. Создать предсказательную модель вероятности отчисления и проанализировать значимость различных факторов, связанных с этим.

Схема сбора данных

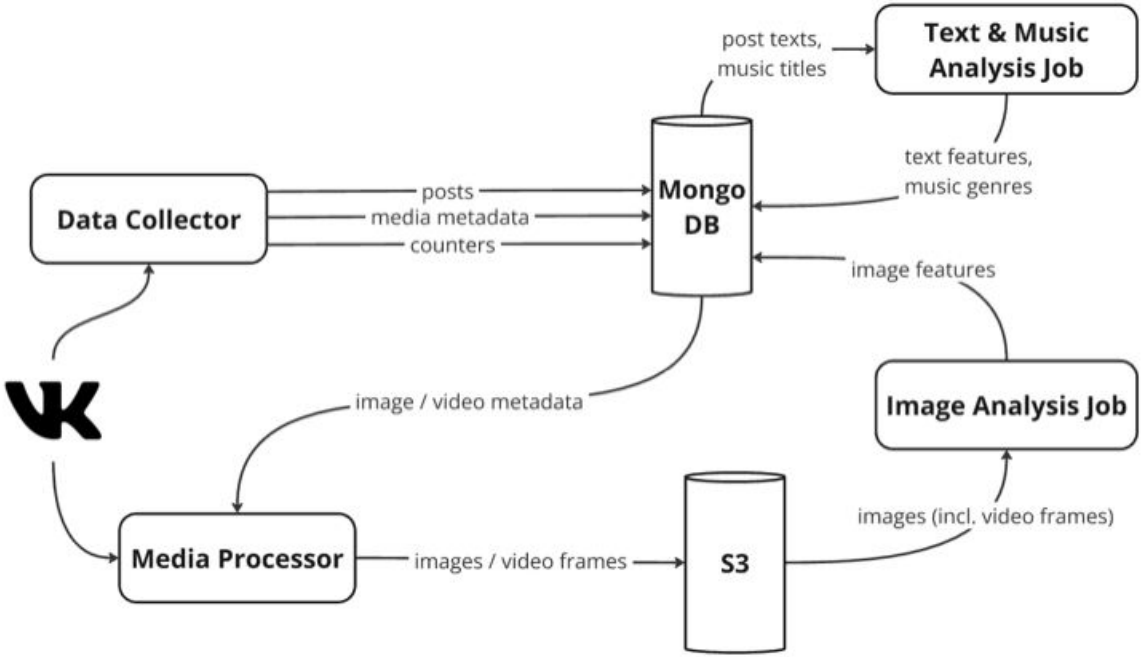
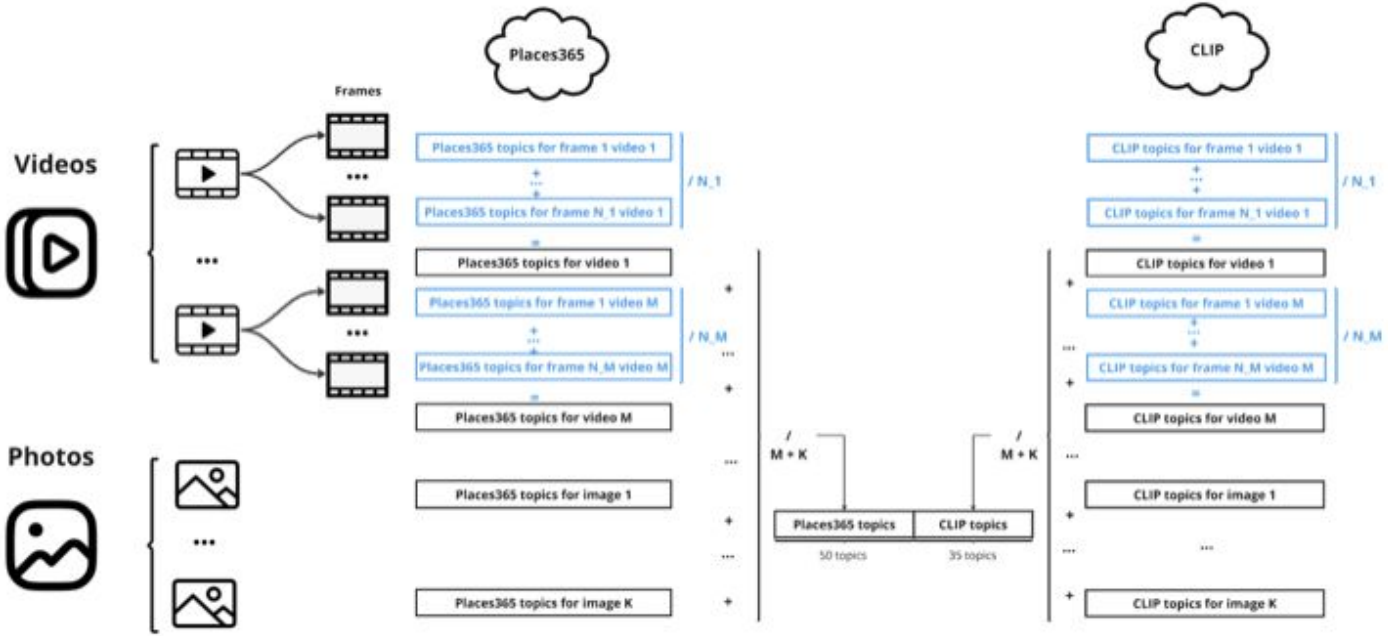
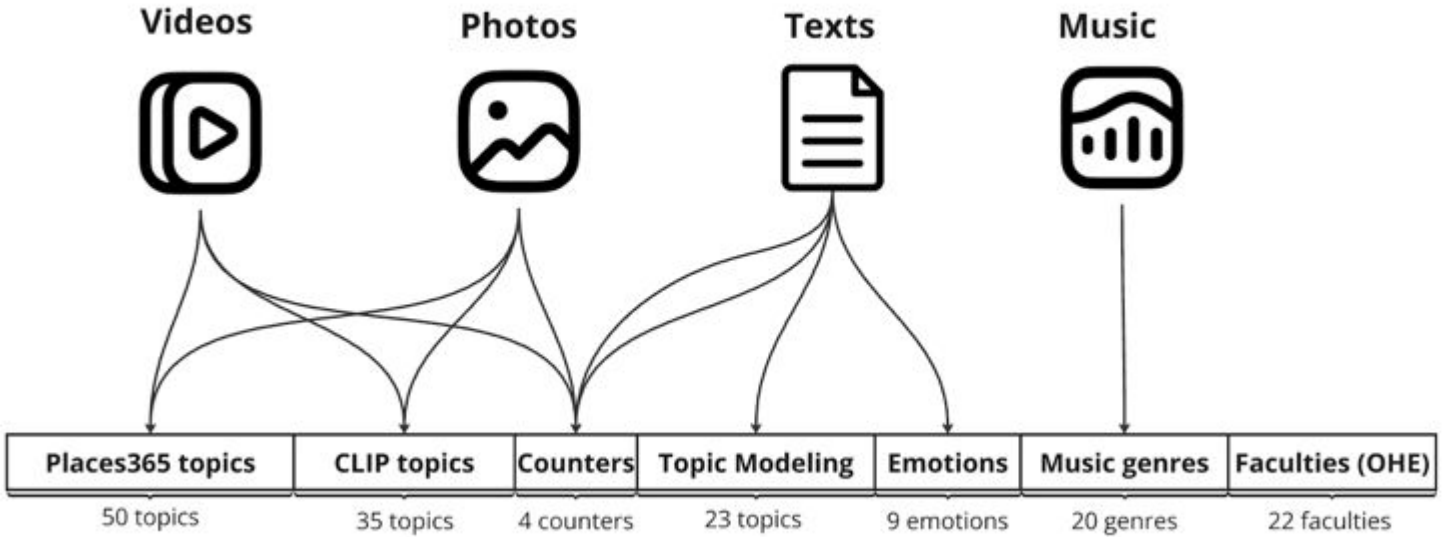


Схема извлечения признаков для фото и видео



Составление вектора пользователя с использованием всех данных



Результаты

ROC-AUC AND BALANCED ACCURACY FOR VARIOUS MODELS AND FEATURE REPRESENTATIONS

Model	Used content	ROC-AUC	Balanced Accuracy
Random Forest	Only new content	0.675	0.603
LightGBM	Only new content	0.673	0.613
CatBoost	Only new content	0.671	0.617
Random Forest	Only historical content	0.753	0.703
LightGBM	Only historical content	0.766	0.702
CatBoost	Only historical content	0.761	0.718
Random Forest	All content	0.755	0.703
LightGBM	All content	0.764	0.715
CatBoost	All content	0.761	0.718
Random Forest	Concatenation of new and historical content	0.775	0.728
LightGBM	Concatenation of new and historical content	0.780	0.731
CatBoost	Concatenation of new and historical content	**0.788**	**0.738**

Спасибо!