Факультет компьютерных наук

Лаборатория анализа семантики

Москва
2025

Поймай бота: сервис для различения текстов, написанных людьми и сгенерированных ботами

НУГ «Поймай бота»:
Сериков Артём Игоревич, Коган Александра Сергеевна

Руководитель:
Громов Василий Александрович, профессор ДАДиИИ

Актуальность

Современные генеративные модели способны создавать настолько последовательные и логичные тексты, что их **практически невозможно отличить** от написанных людьми

Это свойство ИИ-моделей открывает широкие возможности для **подмены** авторских текстов сгенерированными


The Washington Post

Opinion
Megan McArdle

AI is an existential threat to colleges. Can they adapt?

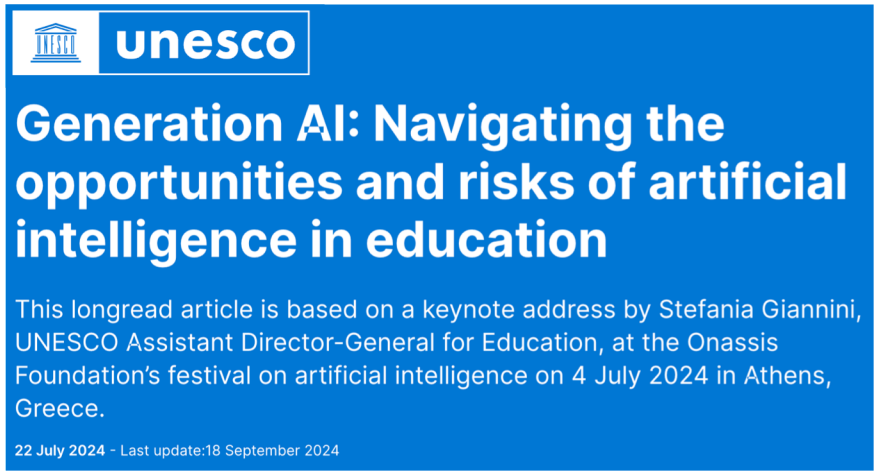
Schools should worry about threats to education if students use artificial intelligence to cheat.

September 30, 2024


The New York Times

Teachers Worry About Students Using A.I. But They Love It for Themselves.

Educators are increasingly using generative A.I. in their own work, even as they express profound hesitation about the ethics of student use.


unesco

Generation AI: Navigating the opportunities and risks of artificial intelligence in education

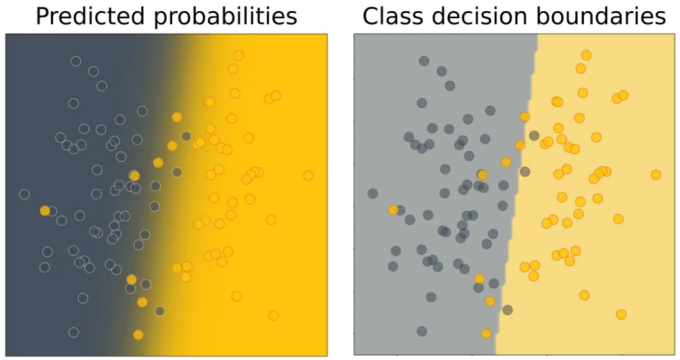
This longread article is based on a keynote address by Stefania Giannini, UNESCO Assistant Director-General for Education, at the Onassis Foundation's festival on artificial intelligence on 4 July 2024 in Athens, Greece.

22 July 2024 - Last update:18 September 2024

Существующие решения

Машинное обучение

Графовые методы — Временные ряды — Классический ML



Недостаток: ML-алгоритмы не анализируют лингвистические характеристики текстовых данных

Глубинное обучение

Антиплагиат — GigaCheck (Сбер) — Turnitin

OpenAI Quietly Shuts Down Its AI Detection Tool

Dashing the hopes of educators, OpenAI decommissions its AI Classifier due to poor accuracy.

AI Detection Tools Falsely Accuse International Students of Cheating

Stanford study found AI detectors are biased against non-native English speakers

Недостаток: нейронные сети адаптированы только к ботам, на которых были обучены

Данные

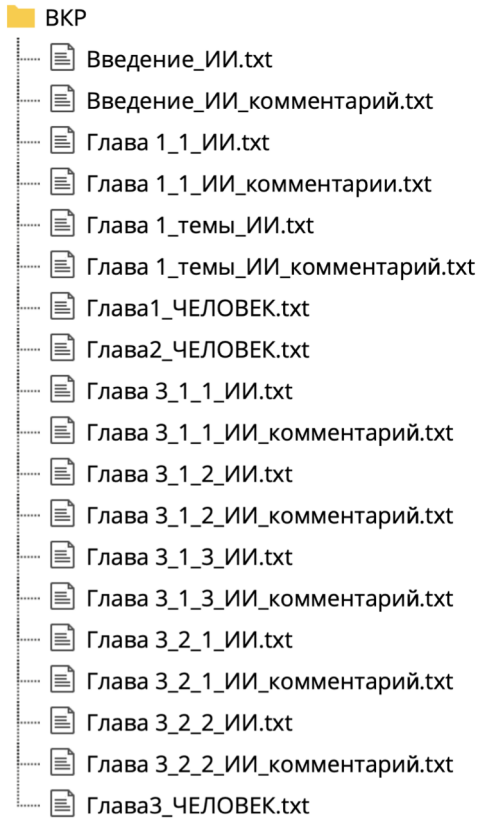
Обучение и оценка качества проводилась на корпусе ВКР объёмом в 12k работ

Введение_ИИ.txt

```
<В современном обществе проблема правильного питания становится все более актуальной, особенно для
студентов-первокурсников, проживающих в общежитиях московских ВУЗов. Иногородние студенты,
переехавшие в Москву для получения высшего образования, часто сталкиваются с вызовами организации
своего рациона питания в новой среде. Исследования, посвященные иногородним студентам или студентам,
переехавшим в другой город, подчеркивают важность адаптации к новым условиям и разработку оптимальных
стратегий для поддержания здорового образа жизни (Huang et al., 2018; Vadeboncoeur et al., 2017).
```

Введение_ИИ_комментарий.txt

```
<Написать введение для академической работы по теме «Практики питания студентов-первокурсников
московских ВУЗов, проживающих в общежитиях». Объем текста должен быть от 200 до 500 слов. В тексте
использовать ссылки на исследования, посвященные иногородним студентам или студентам, которые
переехали в другой город. Текст должен описывать, чем важна тема исследования практик питания
иногородних студентов-первокурсников и почему эту тему необходимо исследовать. В тексте указать, чем
отличается именно опыт практик питания иногородних студентов в Москве.>
```

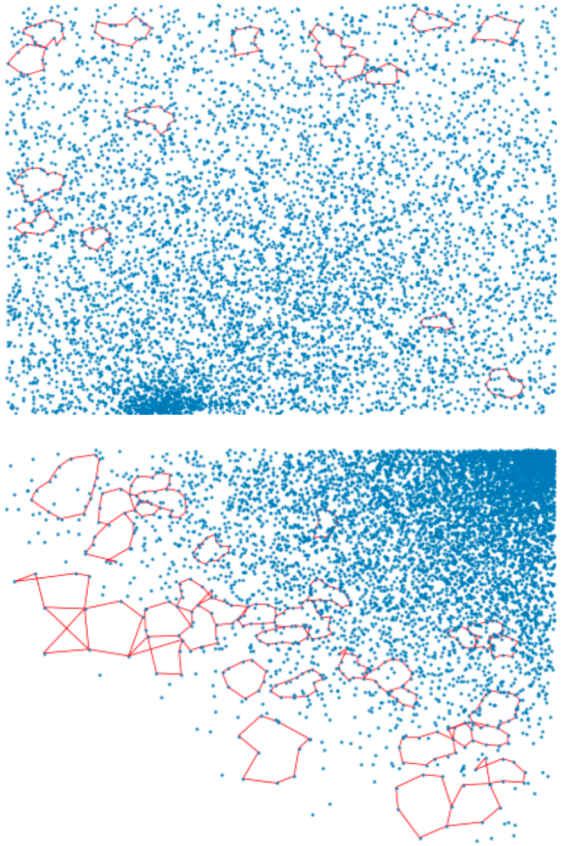


Модель на основе топологии естественных языков

В семантических пространствах находятся различные топологические структуры: существуют как плотные области, так и **разреженные**

Гипотеза:

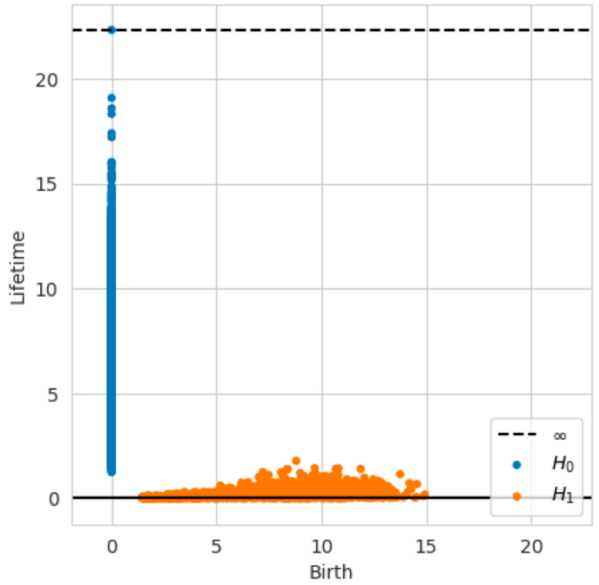
В текстах, написанных людьми, распределения n-грамм статически чаще сдвинуты к границам пустот, в то время как распределения сгенерированных ботами текстов концентрируются в плотных областях пространств

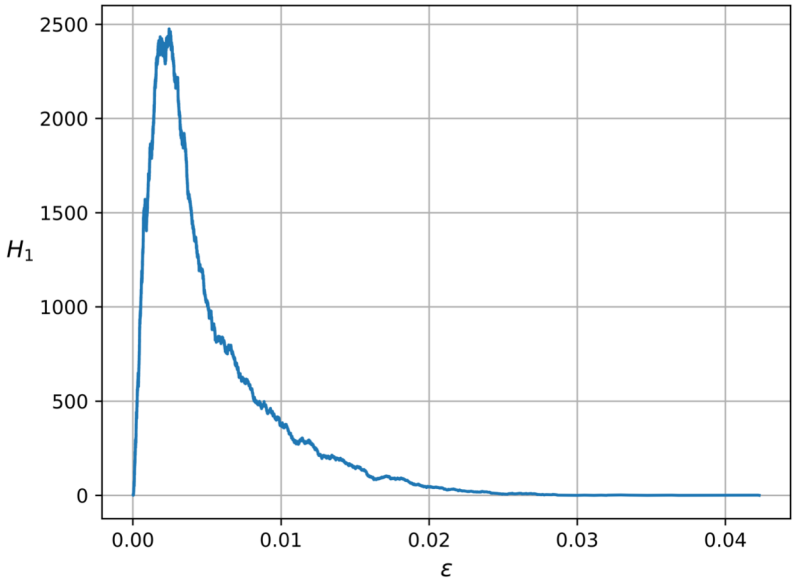


Поиск пустот

Семантические пространства являются топологическими ⟹ к ним применимы методы **топологического анализа данных**

Пустоты были найдены с помощью вычисления персистентных гомологий, из которых были отобраны наиболее устойчивые



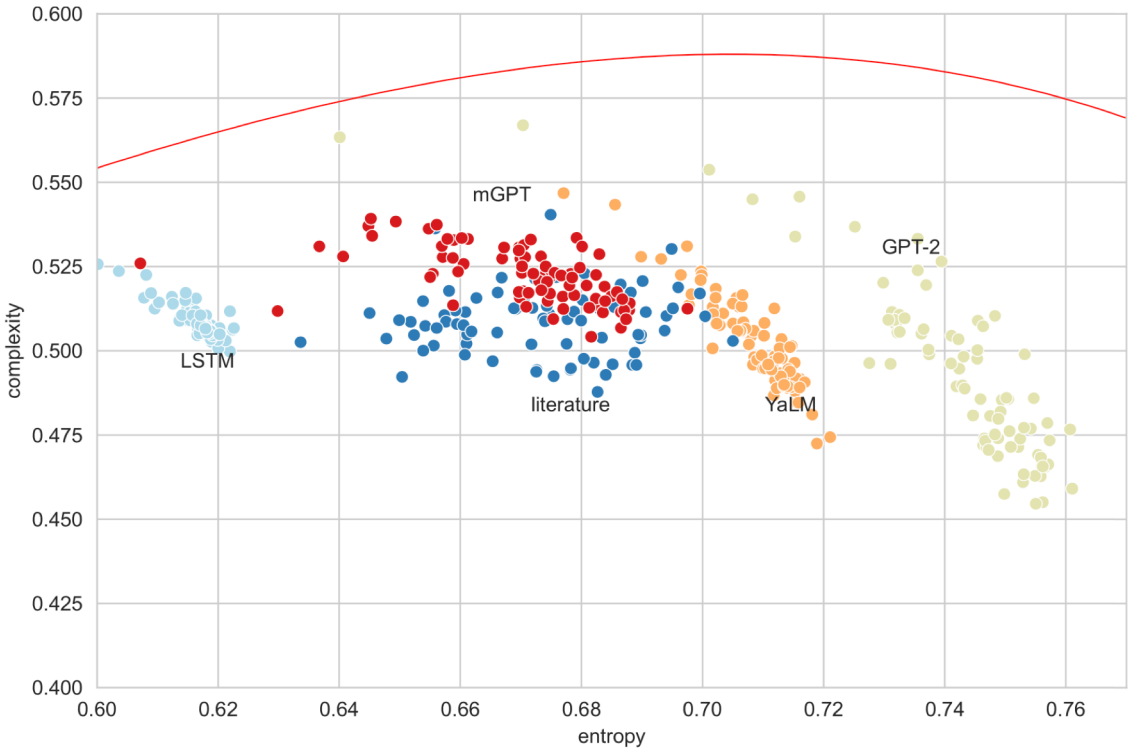


Модель на основе плоскости «Энтропия-Сложность»

Последовательность токенов можно представить в виде временных рядов и анализировать методами теории информации, вычисляя энтропию Шеннона и сложность

Гипотеза:
Ряды сгенерированных текстов являются регулярными, а ряды, порождённые человеческими текстами – хаотичные

Регулярные ряды концентрируются в левом нижнем углу, "белый шум" – в правом нижнему углу, хаотические ряды – посередине (у вершины допустимой области)

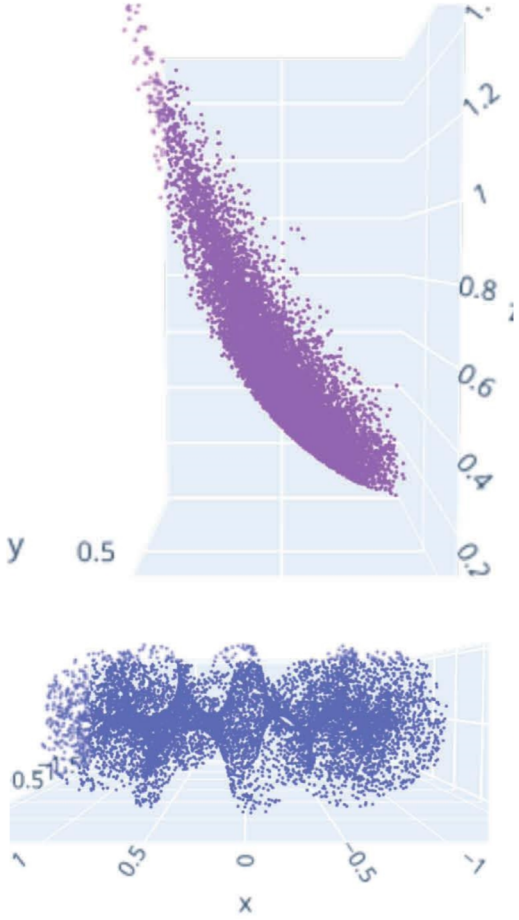


Модель на основе внутренней размерности

Идея:
Определить число параметров, необходимое для описания исследуемого множества в окрестности частной точки

Гипотеза:
Внутренняя размерность естественных текстов выше, чем у сгенерированных

Недостаток:
Повышенная температура генерации модели увеличивает внутреннюю размерность сгенерированных текстов



Качество классификации

Оценка качества проводилась отдельно для следующих ОП:

- Социология публичной сферы и цифровая аналитика
- Философия
- Политология
- Экономика

ОП	Качество
Социология	0.92
Философия	0.92
Политология	0.79
Экономика	0.71

Доля правильных ответов интегральной модели составляет **в среднем 80%**

Направления развития проекта

Ближайшие планы:

- Включение в тестирование технических ОП, реализуемых в НИУ ВШЭ
- Добавление новых признаков AI-моделей (например, сложных сетей)
- Определение минимального объёма текста, для которого возможно корректно установить применение генерации
- Увеличение количества языков, поддреживаемых системой

Используемые материалы

1. UNESCO. Технологии искусственного интеллекта в образовании. Руководство для лиц, ответственных за формирование политики. 2022
2. Gromov, V., Borodin, N. S., Yerbolova, A. S. A Language and Its Dimensions: Intrinsic Dimensions of Language Fractal Structures
3. Gromov, V. A., Dang, Q. N., Kogan A. S. Spot the bot: the inverse problems of NLP. 2024
4. Bellegarda J. R. Latent Semantic Mapping: Principles and Applications. Bellegarda, Jerome R, 2007
5. Zhu X. Persistent Homology: An Introduction and New Text Representation for Natural Language Processing. 2013
6. Edelsbrunner H., Harer J. L. Computational Topology: An Introduction. Edelsbrunner, Herbert and Harer, John L, 2022

Факультет компьютерных наук

Лаборатория анализа семантики

Москва
2025

Поймай бота:
сервис для различения текстов, написанных людьми и сгенерированных ботами