Анализ внутренних представлений языковых моделей для поиска сгенерированных документов

Анастасия Евгеньевна Вознюк
Научный руководитель: к.ф.-м.н. А. В. Грабовой

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных

17 мая 2024

Цель работы

Проблема: Исследуется проблема детекции машинно-сгенерированных документов. Стандартные методы детекции не являются устойчивыми к смене генератора или смене домена, поэтому нужно предложить более устойчивые методы. Кроме того, дополнительной сложностью является смешение авторов в документе уже не на уровне фрагментов но на уровне небольших изменений.

Цель работы: Предлагается исследовать методы анализа внутренних представлений языковых моделей, а также внутренних размерностей текстов при обработке текстового документа. Дополнительно, планируется сравнить разницу в представлениях в зависимости от домена.

Общая постановка задачи

Определим документ как конечную последовательность символов из заданного алфавита $\mathbf{W}$. Пространство документов:

$$\mathbb{D} = \left\{ \left[t_j \right]_{j=1}^{n} \quad | \quad t_j \in \mathbf{W}, n \in \mathbb{N} \right\}.$$

Дан набор из N документов

$$\mathbf{D} = \bigcup_{i=1}^{N} D^i, D^i \in \mathbb{D}.$$

Определим множество авторов, тексты которых встречаются в наборе $\mathbf{D}$:

$$\mathbf{C} = \{0, \ldots, k-1\}.$$

Тогда задача классификации автора документа записывается как:

$$\phi : \mathbb{D} \to \mathbf{C} \tag{1}$$

Задача 1: Деткция преобразований текстов документов

Определим два операции:

1. Удалить DELETE$(D, i_{\text{start}}, i_{\text{end}}) = D_{<i_{\text{start}}} \cdot D_{>i_{\text{end}}}$
2. Вставить INSERT$(D, t, i_{\text{start}}) = D_{<i_{\text{start}}} \cdot t \cdot D_{\geq i_{\text{start}}}$

Любые преобразования над текстом можно выразить через эти две операции. Так, самыми распространенными операциями являются: а) *Polish*, б) *Complete*, в) *Rewrite*.
Для экспериментов был взят датасет искуственных текстов с различными дополнительными преобразованиями RAID[1].

[1] RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors

Признаки латентного пространства модели

$\boldsymbol{\theta}$ - параметры нашей модели f.

$$y_k = f(D, y_1, \ldots, y_{k-1} \mid \boldsymbol{\theta})$$

$$\boldsymbol{y} = [y_1, \ldots, y_n]^T$$

Maxiumum Sequence Probability

$$\text{MSP}(\boldsymbol{y}|D, \boldsymbol{\theta}) = 1 - P(\boldsymbol{y}|D, \boldsymbol{\theta}) \tag{2}$$

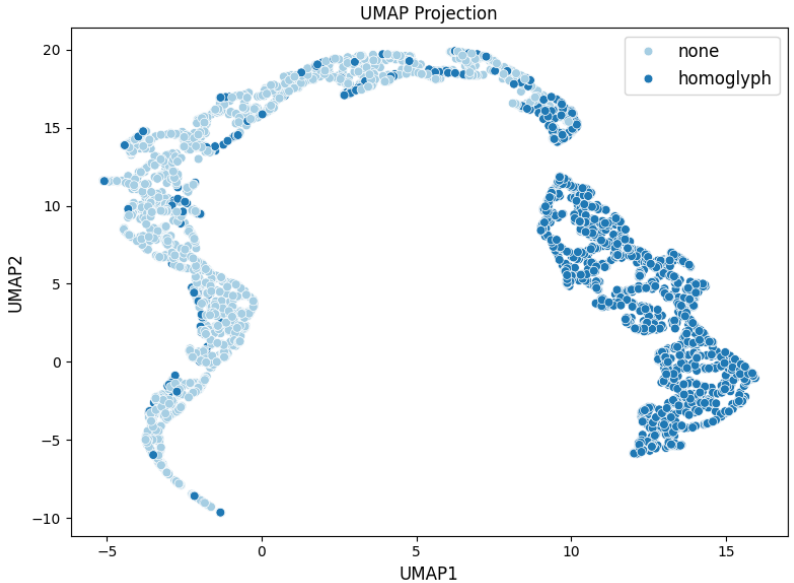
Perplexity

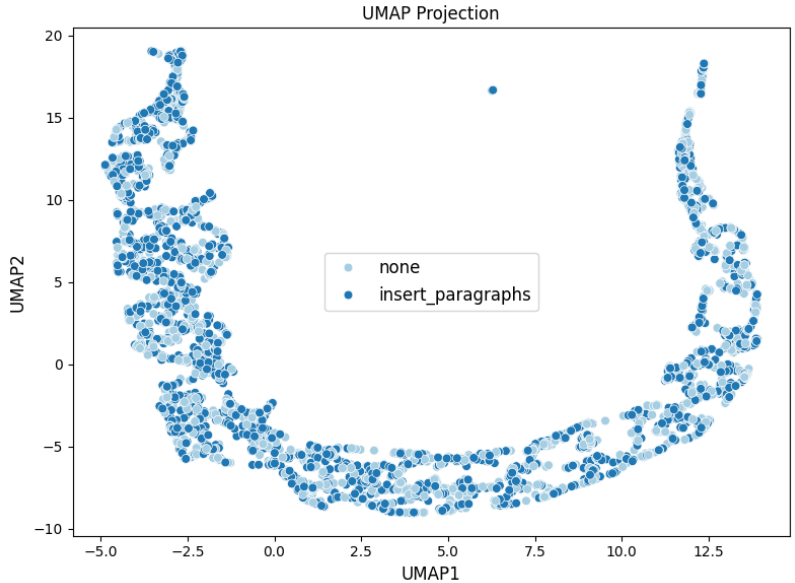
$$P(\mathbf{y}|D, \boldsymbol{\theta}) = \exp\{-\frac{1}{|D|} \log P(\mathbf{y}|D, \boldsymbol{\theta})\} \tag{3}$$

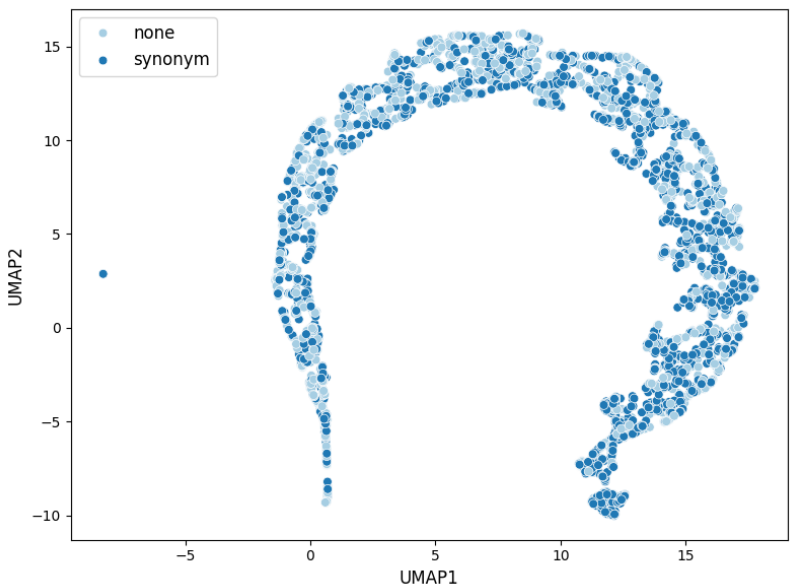
Mean Token Entropy

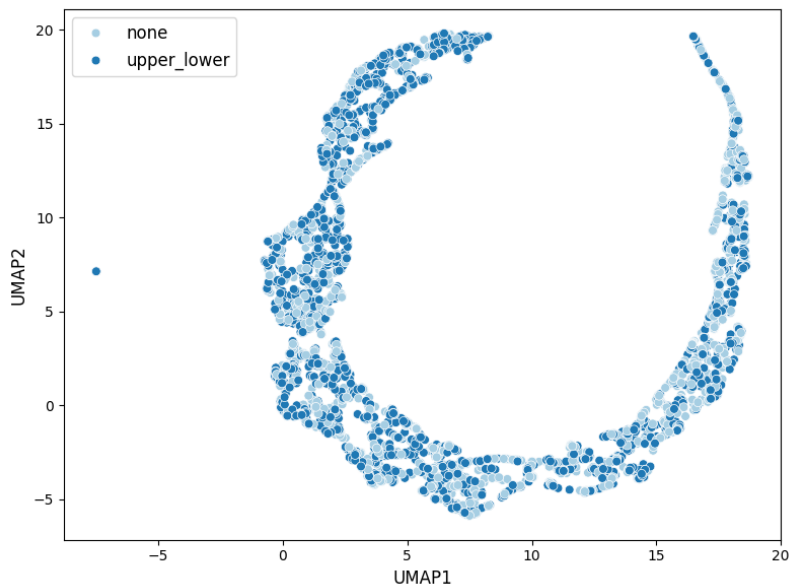
$$\mathcal{H}_T(\mathbf{y}, D; \boldsymbol{\theta}) = \frac{1}{|D|} \sum_{l=1}^{|D|} \mathcal{H}\left(y_l \mid \mathbf{y}_{<l}, D, \boldsymbol{\theta}\right) \tag{4}$$

RQ1: Можно ли разделить между собой тексты с преобразованием и без с помощью этих признаков?

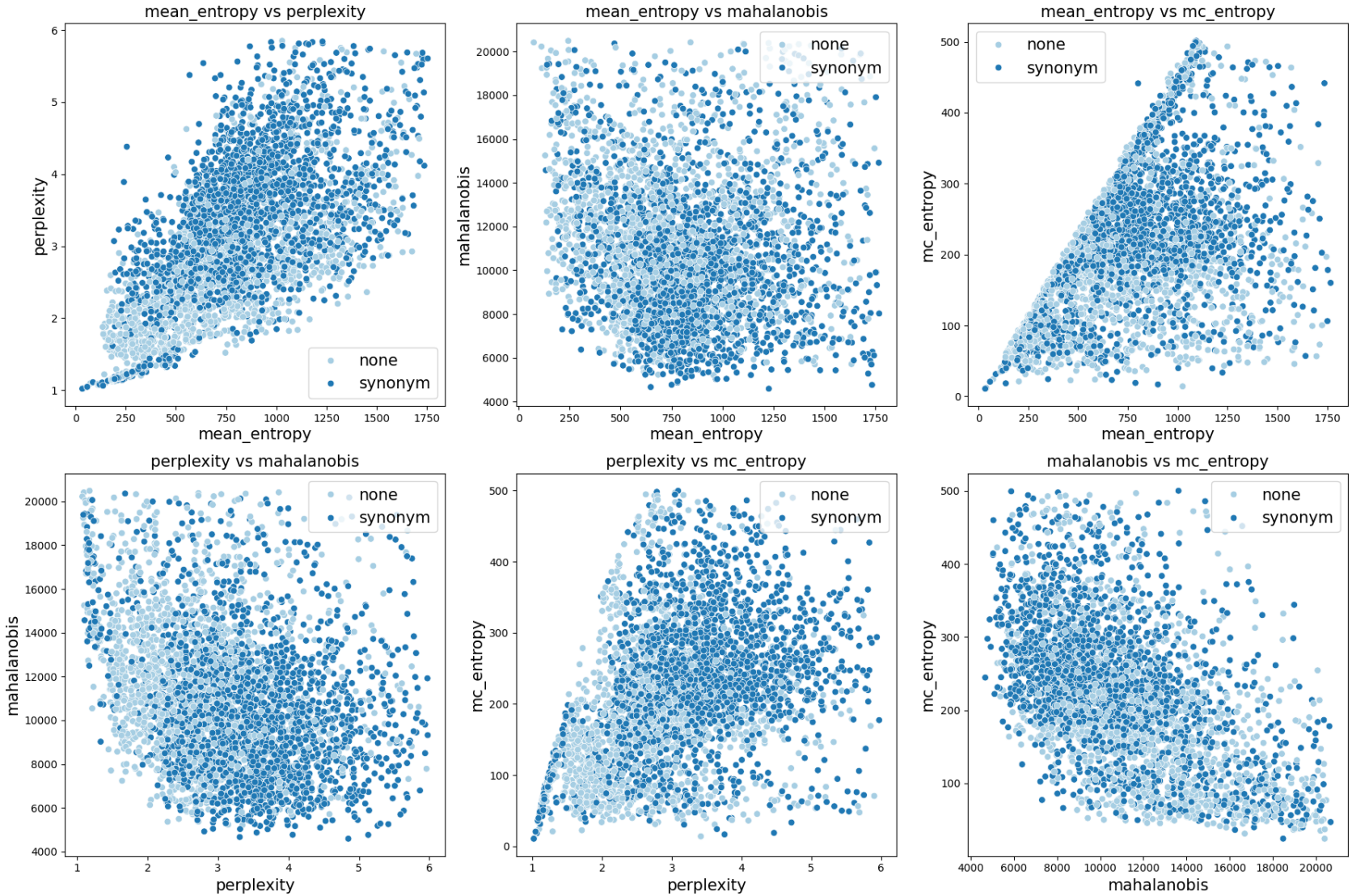




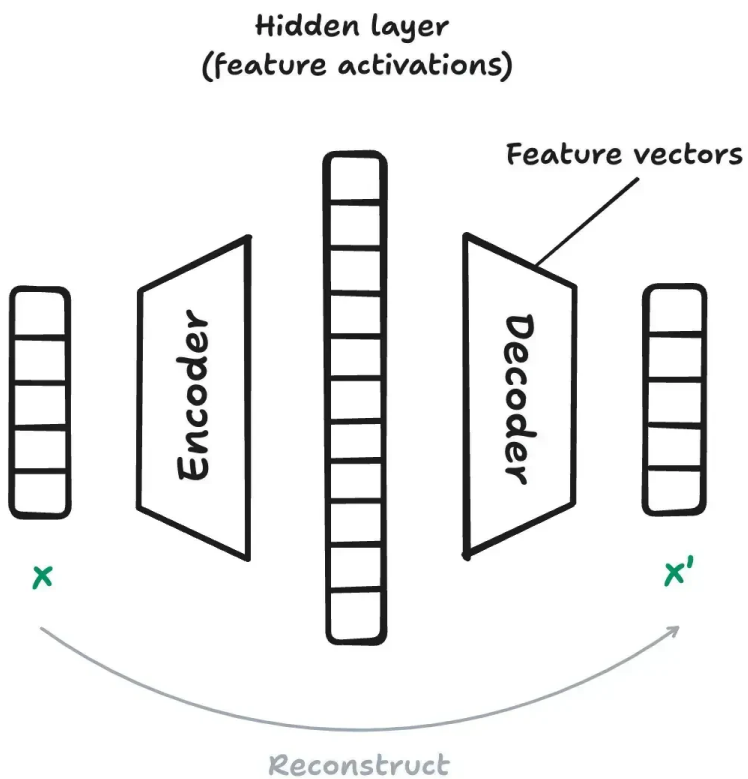




RQ1: Можно ли разделить между собой тексты с преобразованием и без с помощью этих признаков?



Задача 2: Анализ внутренних представлений с помощью Sparse Autoencoders



Пусть есть активации языковой модели $\mathbf{x}$, SAE декомпозирует их и потом конструирует обратно с некоторой функцией σ:

$$f(\mathbf{x}) = \sigma(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}})$$

$$\hat{x}(f) = \mathbf{W}_{\text{dec}} f(\mathbf{x}) + \mathbf{b}_{\text{dec}}$$

$\hat{x}(f(\mathbf{x}))$ – отображение обратно в $\mathbf{x}$; $f(\mathbf{x}) \in \mathbb{R}^M$ $(M \gg d)$ – sparse вектор признаков.

Результаты применения SAE

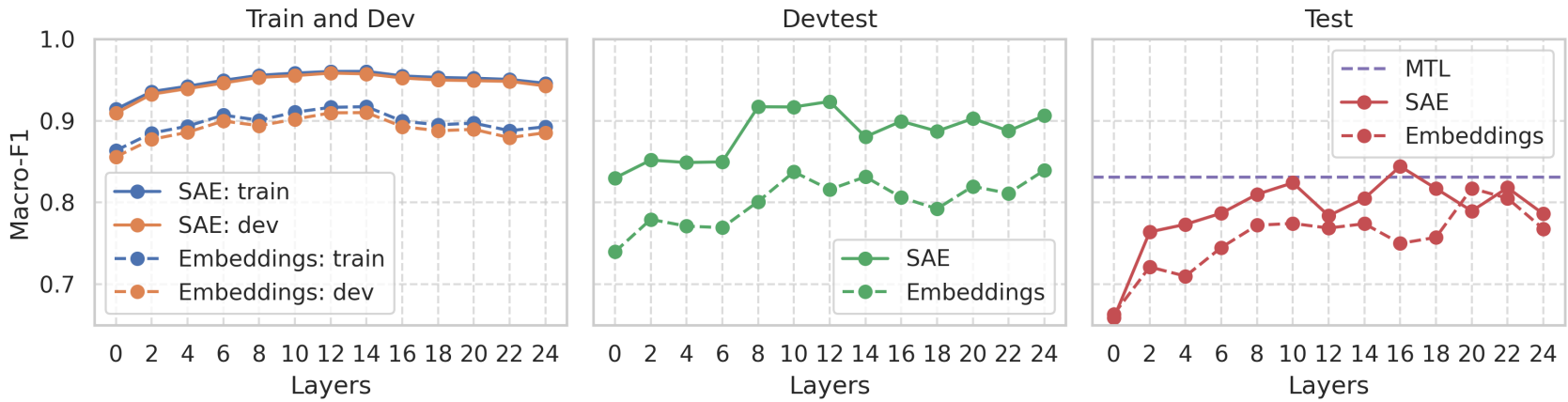


Рис.: Macro F1 при обучении XGBoost на внутренних представлениях и признаках, полученных с помощью SAE на различных поднаборам датасета COLING[2]

[2]Feature-Level Insights into Artificial Text Detection with Sparse Autoencoders

Результаты анализа по доменами SAE

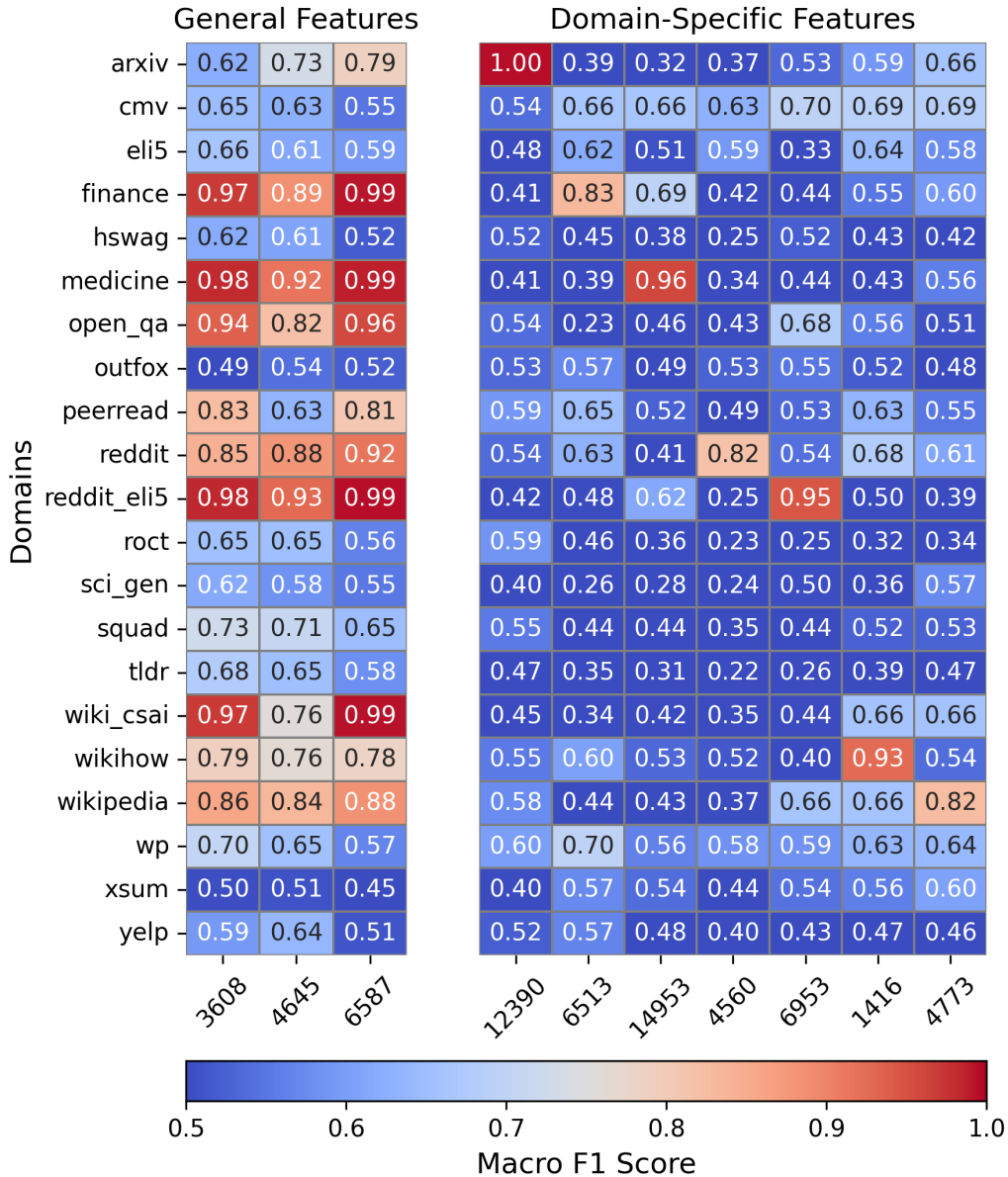


Рис.: F1 Macro by the domains subsets for some general and domain-specific features[3]

Итоги НИР за семестр и планы на следующий семестр

Результаты

1. Принята 1 работа на A конференцию, еще с одной работой было выступление на конференции AAAI;
2. Проведены эксперименты с подсчетом признаков над латентным пространством для выявления текстовых преобразований;
3. Получены результаты для применения SAE для детекции искуственных текстов; проведена интерпретация полученных признаков;

Планы

1. Применить SAE для различных преобразований над текстами (уже в работе, проводим эксперименты);
2. Выработать теоретическое обоснование методов;

Список работ автора по теме НИР

Препринты

1. Feature-Level Insights into Artificial Text Detection with Sparse Autoencoders. *принято на ACL Findings 2025*

Выступления с докладом

1. Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts// *AAAI 2025 Preventing and Detecting LLM Misinformation (PDLM), 3 марта 2025*