

# Методы детекции машинно-сгенерированных фрагментов в документах на основе смены стиля

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

Кафедра интеллектуальных систем ФПМИ МФТИ  
Специализация: Интеллектуальный анализ данных  
Направление: 01.03.02 Прикладные математика и информатика

15 июня 2024

## Цель работы

**Проблема:** Большие языковые модели сейчас повсеместно используются при написании документов в различных областях - например, в научных работах, новостных статьях и в студенческих эссе - и необходимо уметь отличать человеческий текст от искусственно-сгенерированных. Очень часто машинный текст смешивается в документе с человеческим, поэтому помимо классификации автора текста необходимо уметь выделять фрагменты отличающиеся между собой по стилю.

### Цели работы:

1. Предложить метод детекции фрагментов различного авторства в документах в случае смены авторов по фиксированным позициям с помощью марковской линейной цепочки.
2. Предложить метод детекции фрагментов различного авторства в документах в случае единственной смены авторов — смены стиля.

## Общая постановка задачи

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

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

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  $\mathbf{D}$ :

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

Рассматривается три подзадачи: классификация автора всего документа, детекция фрагментов в документах по заданным позициям в документе и детекция смены стиля в произвольной позиции в документе.

# Постановки подзадач

## Классификация автора документа

$$\phi : \mathbb{D} \rightarrow \mathbf{C}.$$

Метрика: точность или F1-мера.

## Детекция фрагментов

Для каждого документа  $d \in \mathbb{D}$  существует представление в виде разбиения на фрагменты различного авторства:

$$\mathbb{T} = \left\{ \left[ t_{s_j}, t_{f_j}, C_j \right]_{j=1}^J \mid t_{s_j} = t_{f_{j-1}}, \quad s_j \in \mathbb{N}_0, \quad f_j \in \mathbb{N}, \quad C_j \in \mathbf{C} \right\},$$

где  $J$  - количество фрагментов разного авторства в документе,  $t_{s_j}$  и  $t_{f_j}$  - начало и конец  $j$ -ого фрагмента, внутри которого все токены одного авторства,  $C_j$  - автор  $j$ -ого фрагмента.

## Постановки подзадач

Тогда модель детектора определяется как композиция отображений:

$$\phi : \mathbb{D} \rightarrow \mathbb{T} \quad \phi : \mathbf{g} \circ \mathbf{f},$$

где  $\mathbf{f}$  — отображение для выделения текстовых фрагментов, а  $\mathbf{g}$  — отображение для классификации получившихся фрагментов.  
Метрика: мера Жаккара для истинного разбиения и разбиения, полученного с помощью модели.

### Детекция смены стиля

Пусть для документа  $d \in \mathbb{D}$  известно, что

$$\exists l_D \in \mathbb{N}_0 \quad \mathbf{g}([t_0, t_l]) = 0, \quad \mathbf{g}([t_{l+1}, t_{|D|})) = 1$$

Метрика: средняя абсолютная ошибка между истинной позицией индекса смены и предсказанной позицией смены авторства.

# Методы решения

## Классификация автора документа

Пусть для  $d \in \mathbb{D}$  были получены векторные представления  $\mathbf{h} = (\mathbf{h}_{\text{CLS}}, \mathbf{h}_1, \dots, \mathbf{h}_n)$  с помощью BERT или другого энкодера. Тогда автором документа будет:

$$\hat{y} = \arg \max(\text{softmax}(\mathbf{W} \cdot \mathbf{h}_{\text{CLS}} + \mathbf{b})),$$

где  $\mathbf{W}, \mathbf{b}$  - обучаемые параметры модели.

## Детекция фрагментов

Пусть для документа  $d \in \mathbb{D}$  известны его параграфы  $\mathcal{P} = (p_1, \dots, p_n)$  и  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  — векторные представления параграфов.

$$p(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n) = p(\mathbf{y} | \mathbf{x}),$$

где  $y_i \in \mathbf{C}$  - метки авторов

# Методы решения

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp \Phi(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp(\Phi(\mathbf{x}, \mathbf{y}'))},$$

где  $\mathcal{Y}^n$  — все возможные последовательности меток длины  $n$ .

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \left( \log \phi_{\text{EMIT}}(y_i \rightarrow x_i) + \log \phi_{\text{TRANS}}(y_{i-1} \rightarrow y_i) \right),$$

Задача оптимизации:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}) = -\log(p(\mathbf{Y}|\mathbf{X})) = -\sum_{i=1}^n \log(p(\mathbf{y}^i|\mathbf{x}^i)) \rightarrow \min$$

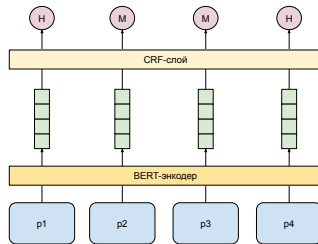


Рис.: Схема модели с марковской линейной цепочкой

## Детекция смены стиля

Введем функцию-детектор с параметром скользящего окна  $\ell$ , которая для токена в документе оценивает его вероятность быть токеном, в котором сменяются авторы:

$$\psi_\ell : \mathbb{D} \times \mathbb{N}_0 \quad \psi_\ell(\mathbf{D}, i) = \mathbb{P}(t_i = 1 | t_{i-\ell}, t_{i-\ell+1}, \dots, t_{i-1})$$

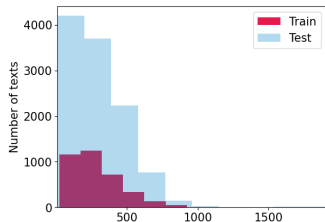
$$I_{\mathbf{D}} = \arg \max_{0 \leq i \leq |\mathbf{D}|} \psi_\ell(\mathbf{D}, i)$$



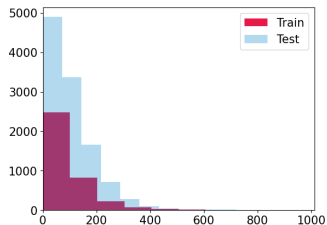
# Данные

1. Для бинарной классификации документов был взят набор текстами на русском языке с соревнования RuATD.
2. Для детекции фрагментов по параграфам был сгенерирован новый датасет на основе статей с сайта Medium.com. Брали тексты в 4-6 параграфов и некоторые параграфы заменяли на машинно-сгенерированные.
3. Для детекции смены стиля данные были взяты из набора данных с соревнования SemEval2024 Task 8 SubtaskC. Данные были дополнительно аугментированы для увеличение размера выборки и внесения разнообразия в позиции смены автора.

# Статистики исходных текстов



(a) Длины текстов



(b) Позиция смены автора

Рис.: Распределения статистик в текстах после токенизации

## Эксперименты с бинарной классификацией

Метод	Точность
TF-IDF	0.64223
BERT-base-multilingual	0.73430
RoBERTa-base	0.63847
XLM-RoBERTa-base	0.72661
XLM-RoBERTa-large	0.76777
DeBERTa-v3-base	0.72661
mDeBERTa-v3-base	0.76662
ruBERT	0.77288

**Таблица:** Точность бинарной классификации различных подходов. Цветом выделены модели, предобученные на корпусе русских текстов

# Эксперименты с детекцией фрагментов по параграфам

Метод	Точность
RoBERTa	0.89
RoBERTa-CRF	0.93

Таблица: Точность детекции

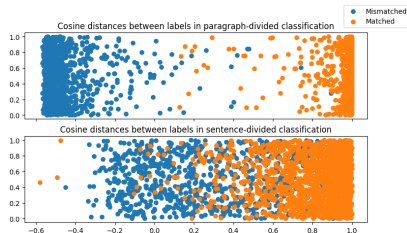


Рис.: Разделение векторных представлений параграфов и предложений

## Эксперименты со сменой стиля

Модель	Исходный датасет	Новый датасет
RoBERTa-base	31.56	30.71
RoBERTa-large	25.25	20.66
longformer-base	23.16	22.94
longformer-large	22.97	20.33
DeBERTaV3-base	16.12	13.98
DeBERTaV3-large	15.16	<b>13.38</b>
Top 1 соревнования	15.68	-

**Таблица:** Метрика MAE на исходных и новых (аугментированных) данных. Дополнительно приведено лучшее решение, получившее первое место по результатам соревнования