Обнаружение мультиавторности в тексте с помощью ансамбля моделей

А. О. Фаттахов[1]

[1]Московский Физико-Тенический Институт

Численные методы обучения по прецедентам (В. В. Стрижов)/Группа 594, весна 2018

Цель исследования

Проблема

Имеется набор текстов, для каждого из которых необходимо определить, написан он одним автором или несколькими

Цель работы

Построить ансамбль моделей, классифицирующих тексты, исследовать вклад каждой модели в итоговое качество ансамбля

Литература

- Kuznetsov, M.; Motrenko, A.; Kuznetsova, R.; Strijov, V. Methods for IntrinsicPlagiarism Detection and Author Diarization Notebook for PAN at CLEF 2016.
- han, J. Style Breach Detection: An Unsupervised Detection Model Notebook for PANat CLEF 2017
- Safin, K.; Kuznetsova, R. Style Breach Detection with Neural Sentence EmbeddingsNotebook for PAN at CLEF 2017

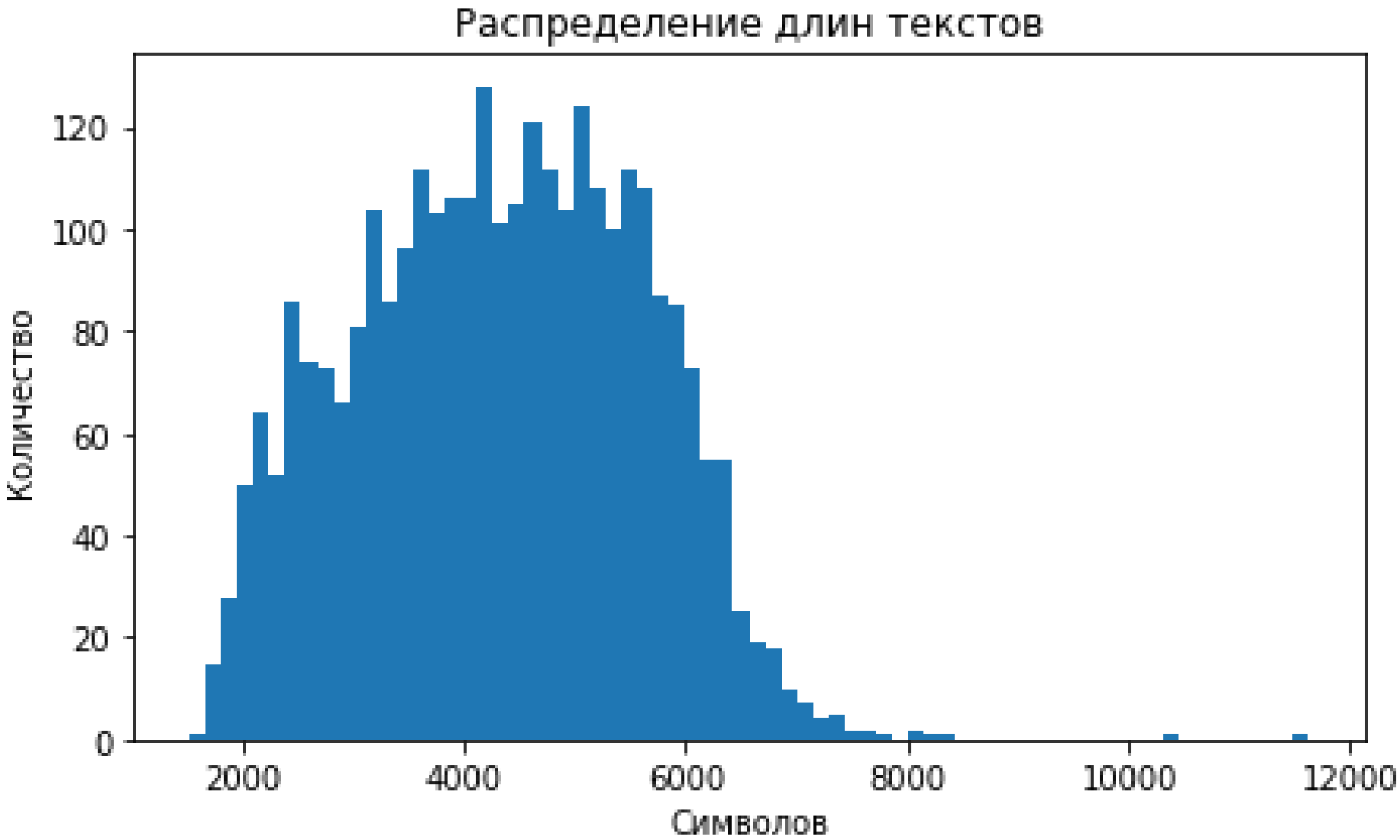
Постановка задачи

Дана выборка $D = \{(d_k, t_k)\}$, где d_k — текстовый документ, а $t_k \in \{0, 1\}$ - является ли текст мультиавторным. Требуется построить модель классификации $f : d_k \mapsto y_k$, минимизирующую функцию ошибки S

$$f = \frac{1}{|D|} \sum_{k=1}^{|D|} \arg\min_f S(y_k, t_k) \tag{1}$$

$$S = -\frac{1}{|D|} \sum_{k=1}^{|D|} t_k \log y_k + (1 - t_k) \log (1 - y_k) \tag{2}$$

Описание выборки



Описание стекинга моделей

Базовые модели

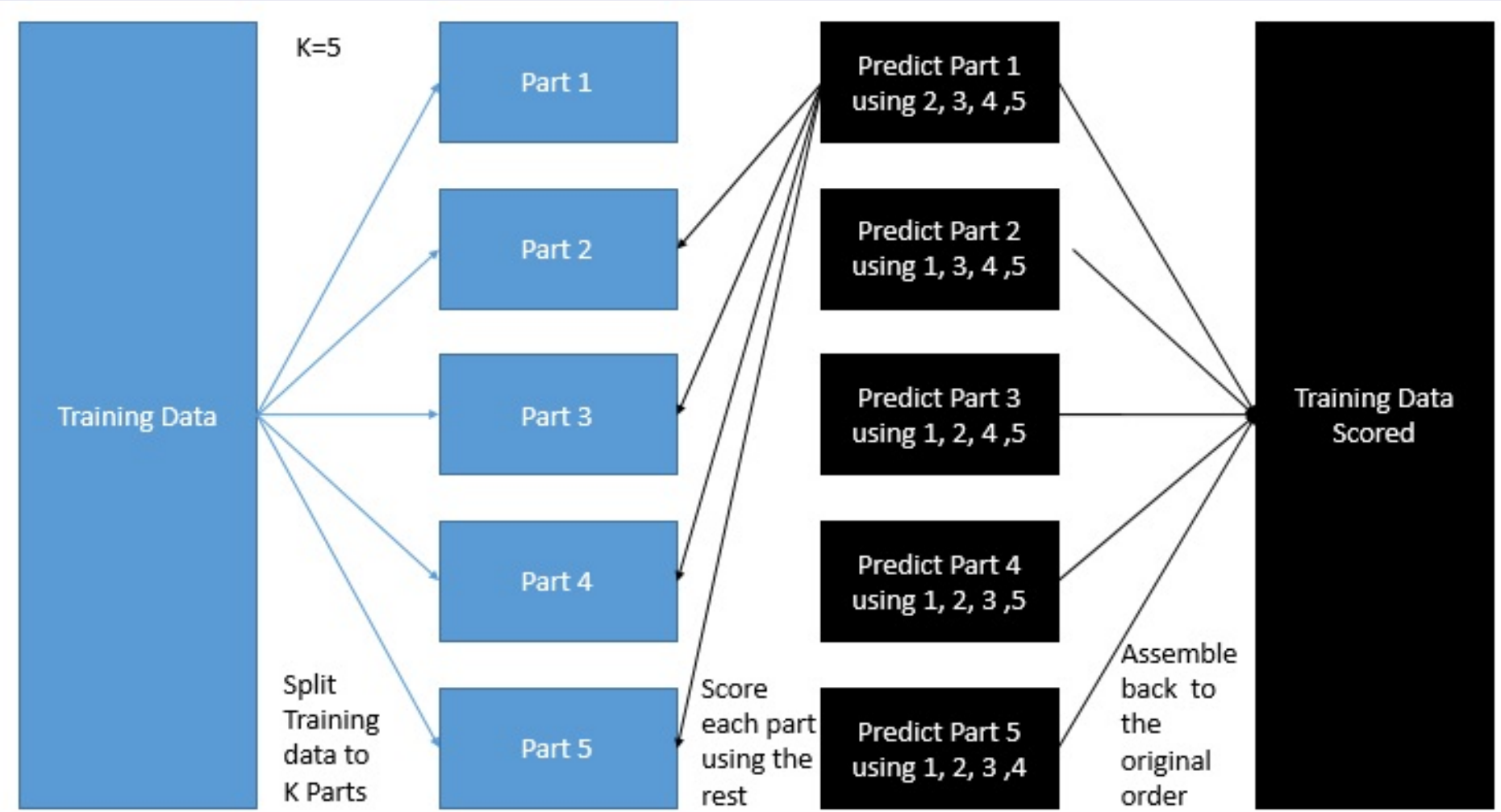
Строятся базовые модели f_i из разных семейств

Модель второго уровня

Предсказания моделей f_i на тренировочную выборку используются как признаки для модели второго уровня

Описание стекинга моделей

Предсказания базовых моделей должны быть out-of-fold



Список базовых алгоритмов

- Градиентный бустинг деревьев
  - Признаки, связанные с частотами слов
  - Признаки, связанные с частотами символов
- Модель с использованием функции стиля
- Рекуррентная нейронная сеть LSTM с векторными представлениями слов
- Сверточная нейронная сеть с векторными представлениями символов

Признаки для градиентного бустинга деревьев

Word-level CountVectorizer

Текст d_i отображается в d_i посредством подсчета частот вхождений слов и их n-грамм в текст d_i

Char-level CountVectorizer

Текст d_i отображается в d_i посредством подсчета частот вхождений символов и их n-грамм в текст d_i

Lightgbm

На полученных признаках d_i настраиваются параметры модели f, где в качестве f берется градиентный бустинг деревьев

Модель на основе функции стиля

Разбиение текста на сегменты

Для каждого текста d_k формируется разбиение на сегменты $s_{k\,i}$. В простом случае $s_{k\,i}$ - это предложения данного текста.

Функция стиля

Для каждого сегмента $s_{k\,i}$ текста строится вектор признаков $\mathrm{s}_{k\,i}$. Считается статистика $\sigma(\mathrm{s}_{k\,i})$, которая сглаживается скользящим средним:

$$\sigma(\mathrm{s}_{k\,i}) = ||\mathrm{s}_{k\,i} - \mathrm{s}_{k\,ave}|| \tag{3}$$

Модель

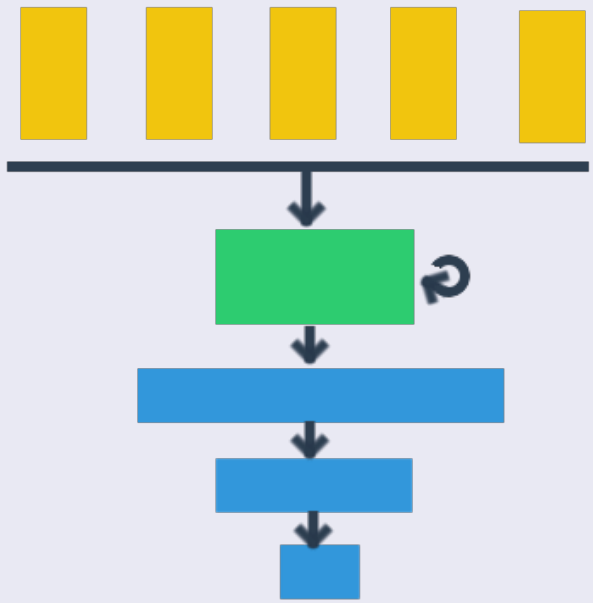
Если в тексте наблюдаются скачки функции стиоля выше некоторого порога, то текст классифицируется как мультиавторный

Нейронная сеть LSTM

Эмбеддинги слов

В качестве отображения $g : w_k \mapsto \mathrm{w_k}$, где w_k - это слово текста, берется предобученные вектора Google News

Архитектура



Вычислительный эксперимент

Таблица: Качество моделей

	модель1	модель2	модель3	модель4	Ансамбль
Accuracy	0.726	0.669	0.638	0.661	0.730
std	0.013	0.016	0.001	0.021	0.010

Вычислительный эксперимент

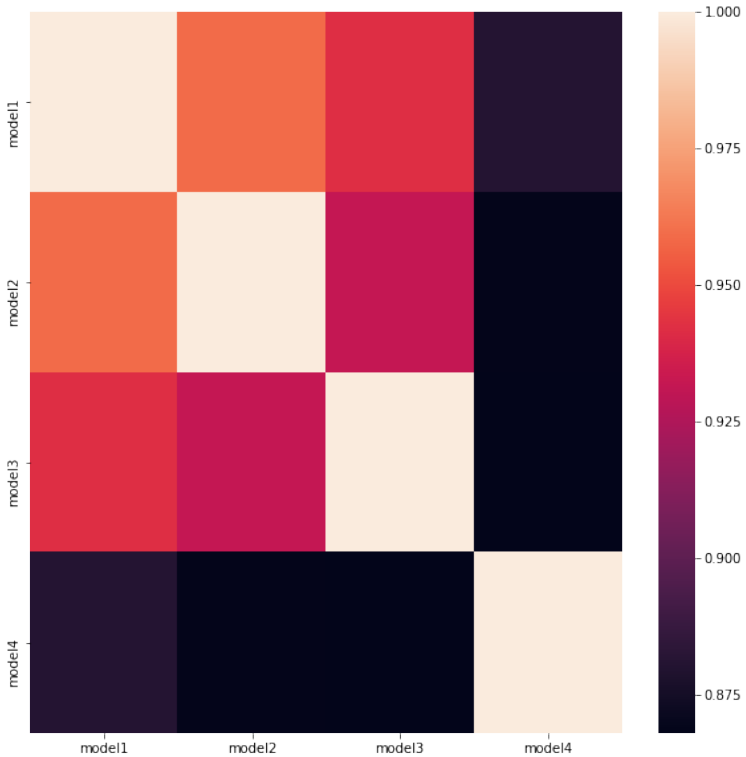


Рис.: Корелляция предсказаний моделей

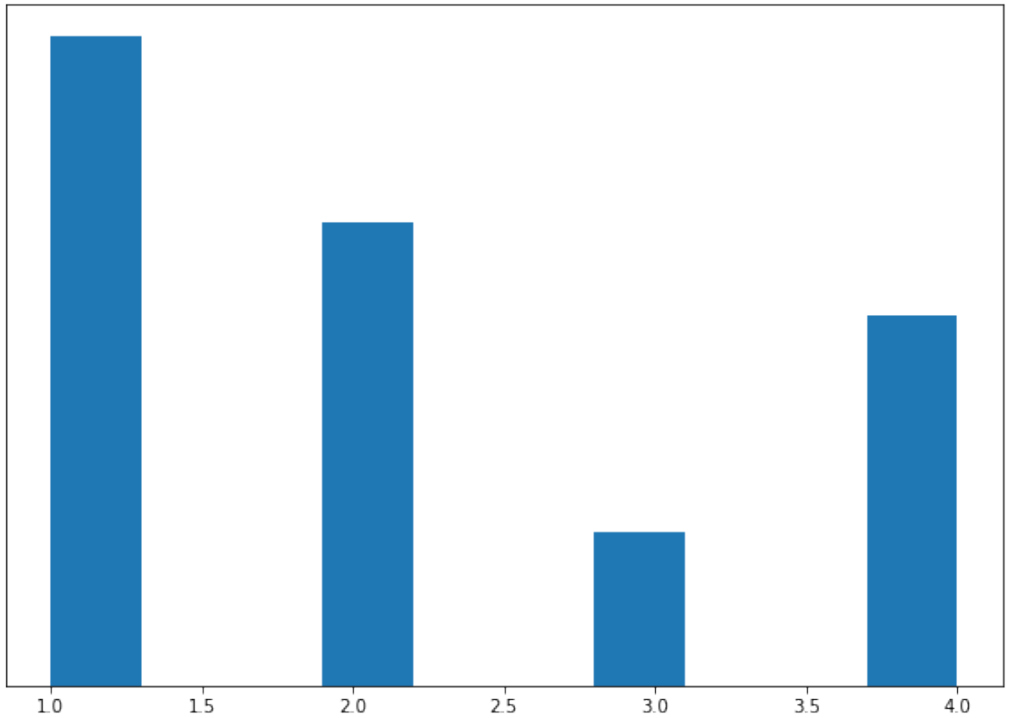


Рис.: Вклад каждой модели в ансамбль