ОБНАРУЖЕНИЕ ВРЕДОНОСНЫХ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ В СЕТИ ИНТЕРНЕТ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Выступающий: Браницкий А.А., с.н.с., к.т.н.

Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук

РусКрипто

Солнечногорск, 19 марта 2020 г.

Актуальность

- Сеть Internet содержит огромное количество разнородной информации, смысловое наполнение которой может носить вредоносный характер
- Обнаружение подобных источников информации важно, т.к. их распространение и использование может приводить к серьезным негативным последствиям
- Задачу обнаружения вредоносной информации можно рассматривать как частный случай задачи категоризации смыслового наполнения информационных объектов (ИО), в которой необходимо выявить факт принадлежности ИО к одной из тех категорий, определенных системным оператором как содержащих нелегитимный контент
- Требуется разработать методику, направленную на повышение показателей эффективности при категоризации смыслового наполнения ИО на примере задачи классификации Web-страниц

Подходы к категоризации ИО

Авторы работы (год)	**Математический аппарат или прием**
Hayes P.J., Andersen P.M., Nirenburg I.B., Schmandt L.M. (1990) [1]	Продукционные правила, создаваемые вручную оператором-экспертом (система CONSTRUE)
Apté C., Damerau F., Weiss S.M. (1994) [2]	Автоматическая генерация правил классификации на основе ДНФ
Salton G., Buckley C. (1988) [3]	Использование массива вещественнозначных коэффициентов, составленных из относительных и абсолютных частот вхождения определенных слов в классифицируемый текст: (1) частота слова (term frequency, tf), (2) обратная частота документа (inverse document frequency, idf), (3) важность слова (term discrimination, td=tf$\times$idf)
Fattah M.A. (2017) [4]	Приписывание каждому слову его интегрального веса, включающего вероятность появления слова как в рамках определенной категории, так и внутри всей коллекции документов
Lewis D.D., Ringuette M. (1994) [5]	Байесовский классификатор и дерево решений
Joachims T. (1998) [6]	Машина опорных векторов
Johnson R., Zhang T. (2015) [7]	Два типа конволюционных нейронных сетей: (1) прямого распространения сигнала и (2) с преобразованием мешка слов (bag-of-words) на конволюционном слое
Ghareb A.S., Bakar A.A., Hamdan A.R. (2016) [8]	Генетический алгоритм

Методика обнаружения вредоносных ИО в сети Интернет

Методика обнаружения вредоносных ИО в сети Интернет

Шаги методики

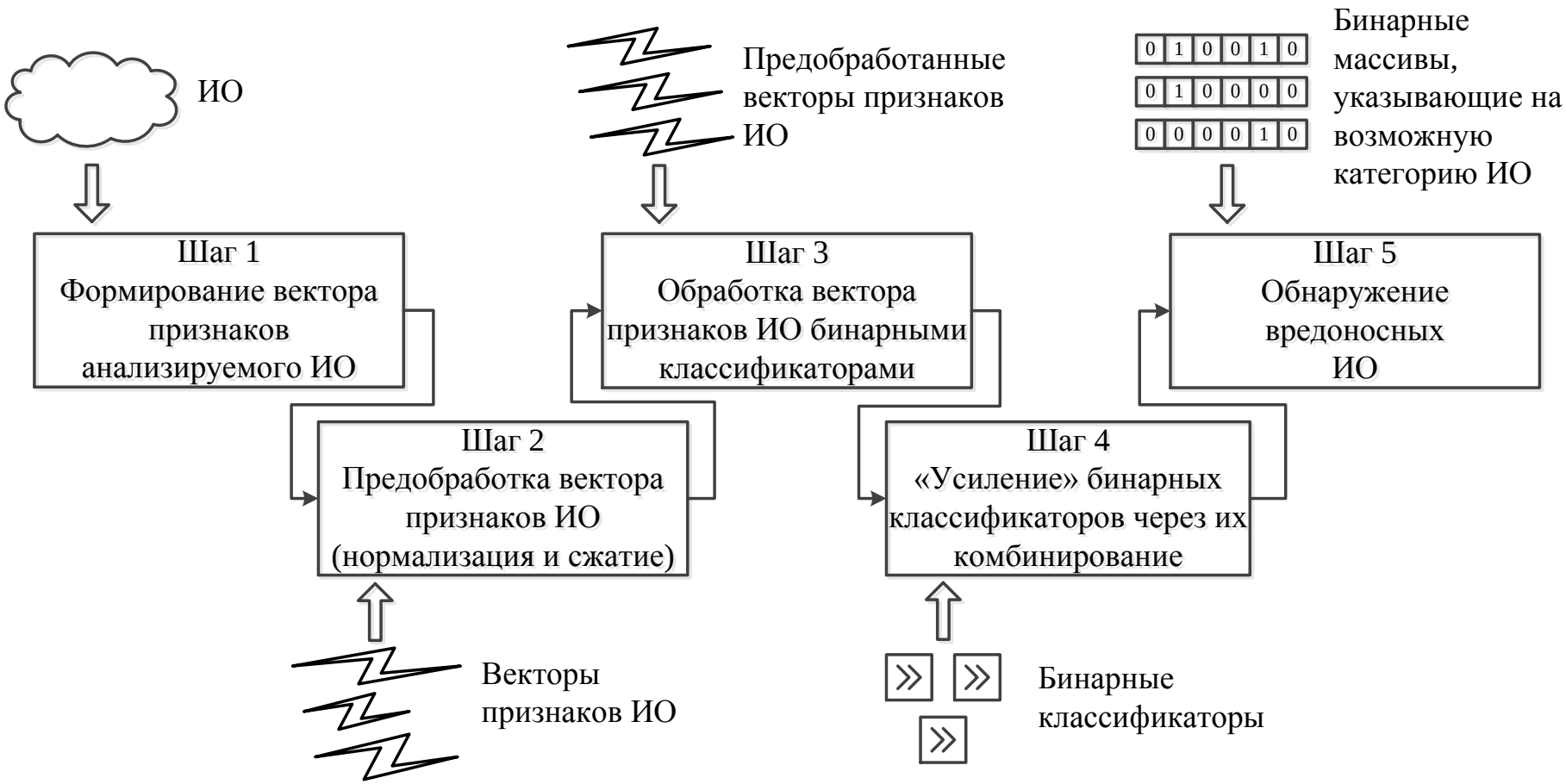
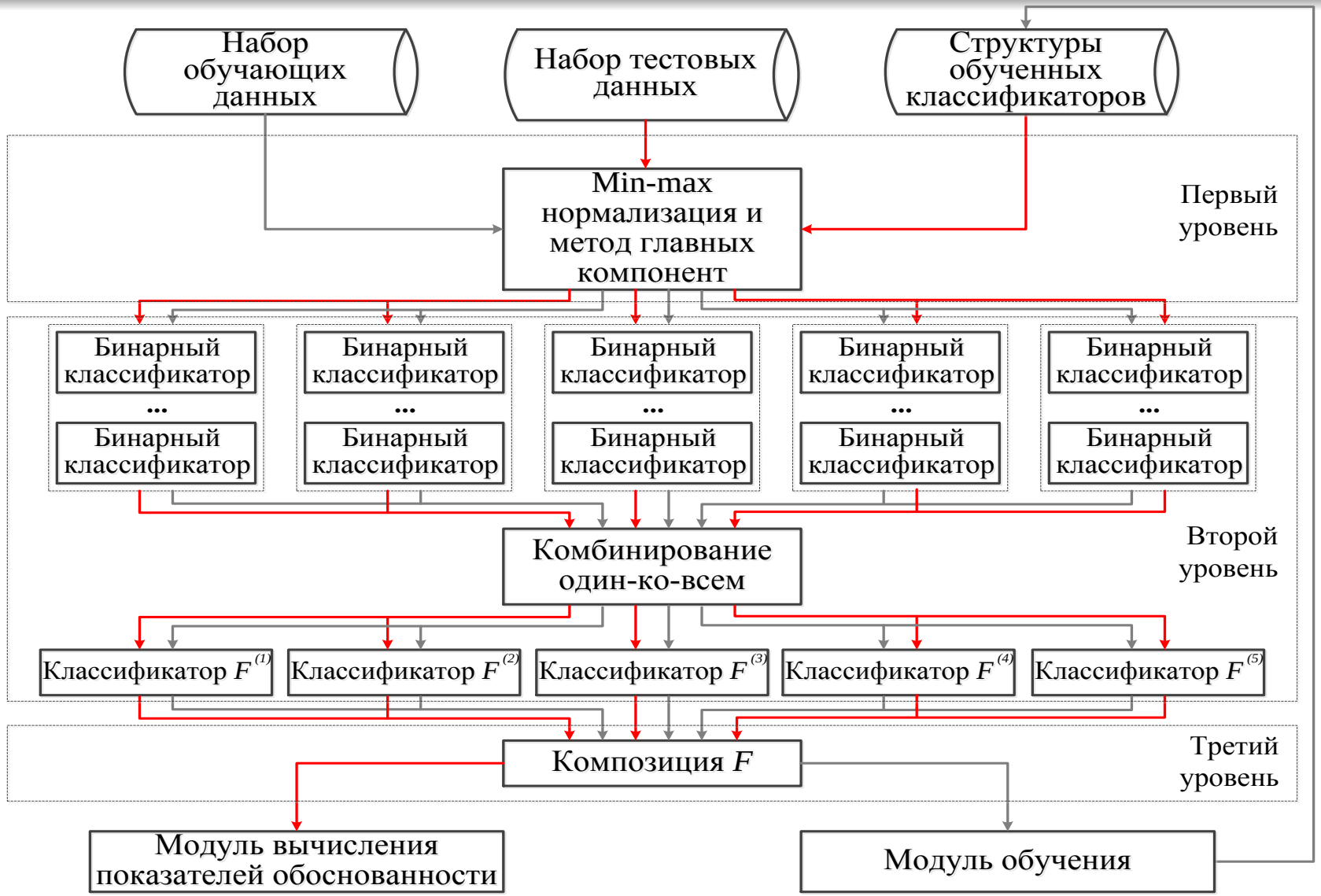


Схема комбинирования классификаторов



Набор данных и признаки

Набор данных

№	Категория	Мощность
1	Adult	6748
2	Alcohol	2802
3	Beer	5881
4	Casino	3581
5	Cigarette	1905
6	Cigars	3809
7	Cults	3263
8	Dating	4672
9	Jew related	3446
10	Marijuana	5365

№	Категория	Мощность
11	Occults	5097
12	Prescription drugs	6003
13	White supremacy	399
14	Racist groups	4627
15	Religion	5281
16	Sport betting	3628
17	Violence	1892
18	Weapon	2448
19	Wine	4046
All		74893

Признаки

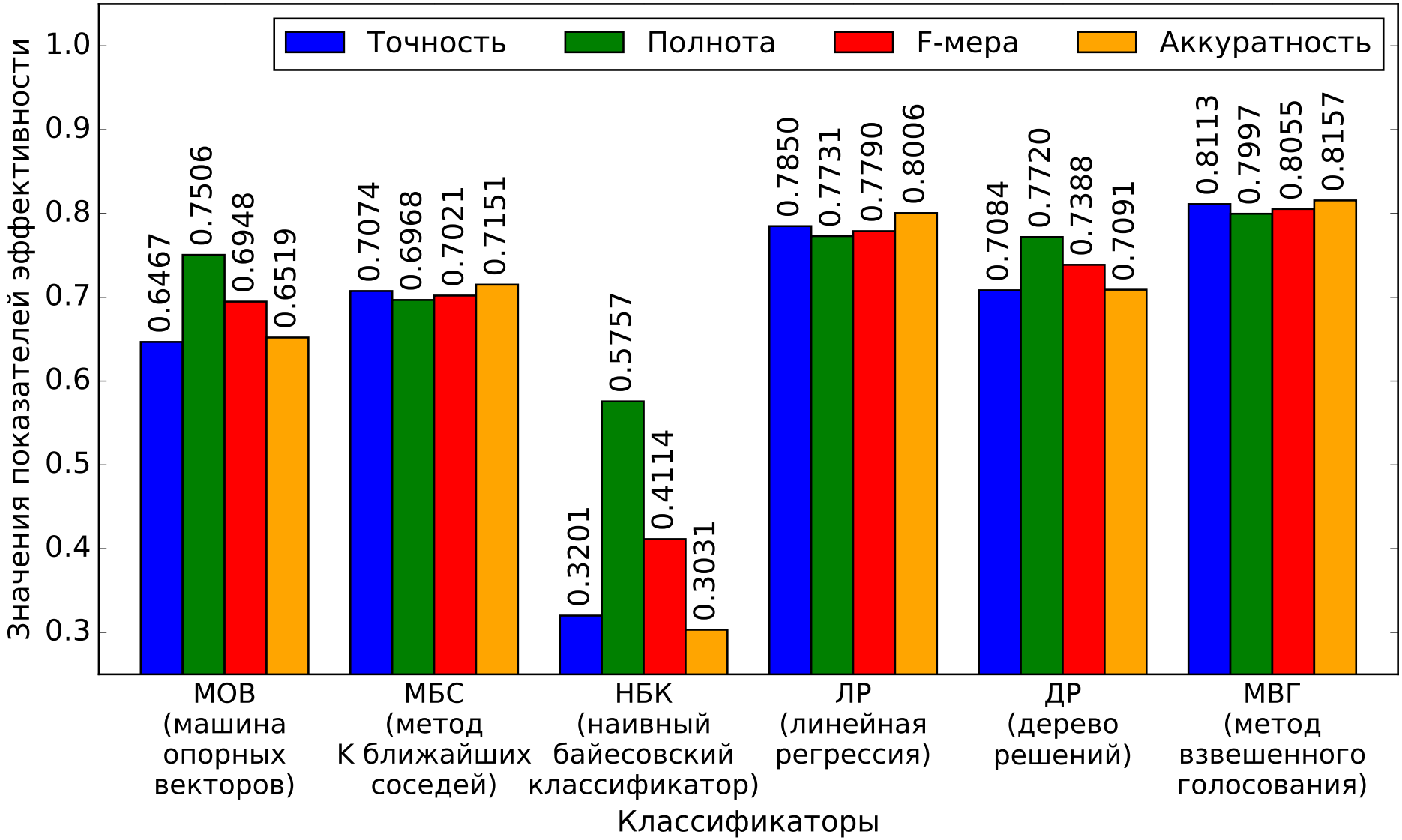
- 402 параметра:
  - ✓ относительное число html-тегов
  - ✓ количество ссылок
  - ✓ размер заголовка
  - ✓ ...
  - ✓ степень семантической схожести слов в терминах word2vec [9]
- 3 типа исходных данных:
  - ✓ строка url
  - ✓ структура документа (статистические сведения о html-тегах)
  - ✓ текст, полученный после удаления html-тегов

Оценка показателей эффективности

Эксперимент 1. Оценка показателей эффективности

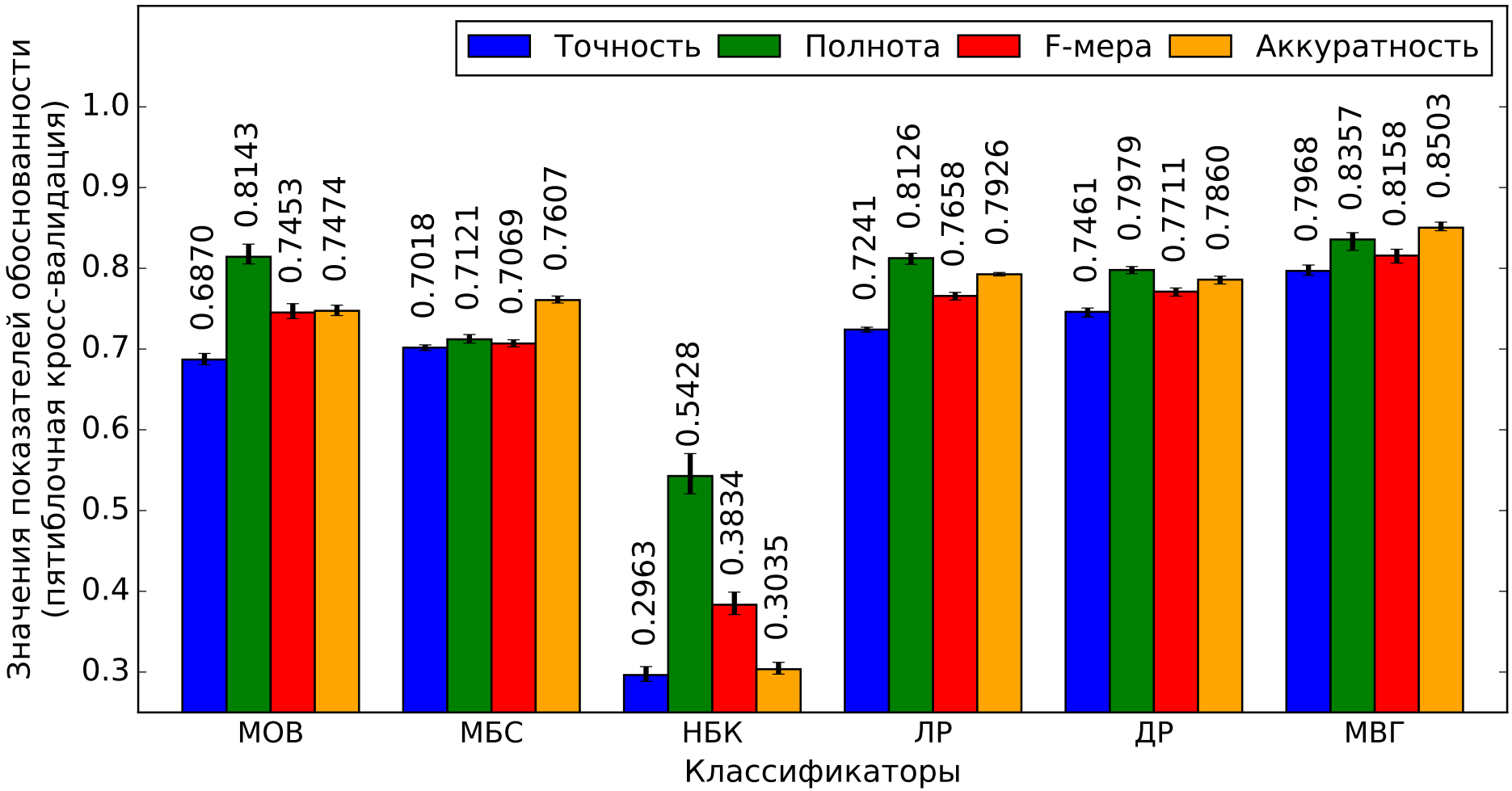
Показатели эффективности (56656 тестовых элементов)

Обучающая выборка — 18237 элементов, тестовая выборка — 56656 элементов.
$PR \Uparrow$ на 2.63%, $RC \Uparrow$ на 2.66%, $FM \Uparrow$ на 2.65%, $AC \Uparrow$ на 1.51%.



Показатели эффективности (пятиблочная кросс-валидация)

Пятиблочная кросс-валидация: 4 части — на обучение, 1 часть — на контроль.
$PR \Uparrow$ на 5.07%, $RC \Uparrow$ на 2.14%, $FM \Uparrow$ на 4.47%, $AC \Uparrow$ на 5.77%.

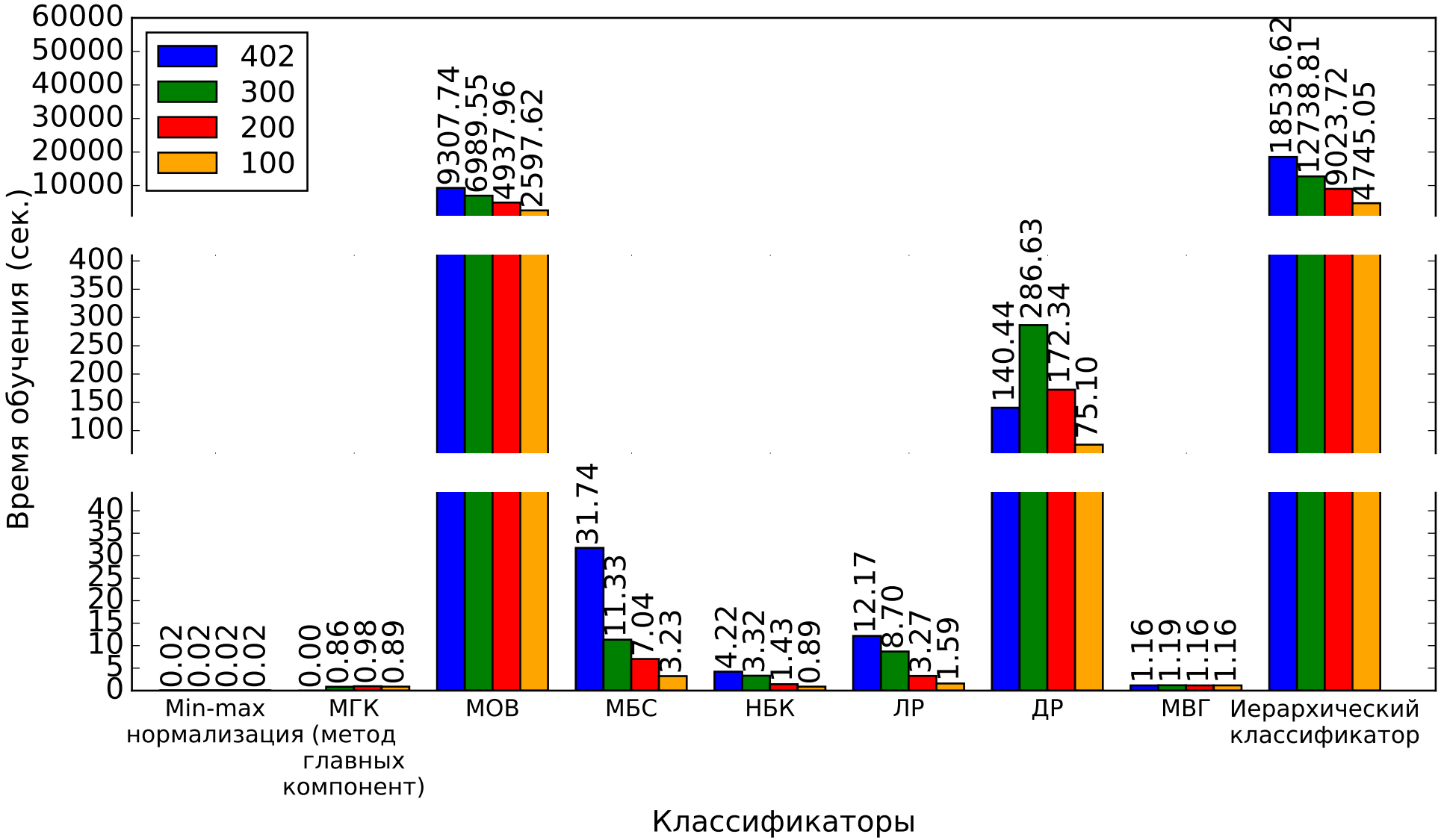


Оценка времени обучения

Эксперимент 2. Оценка времени обучения

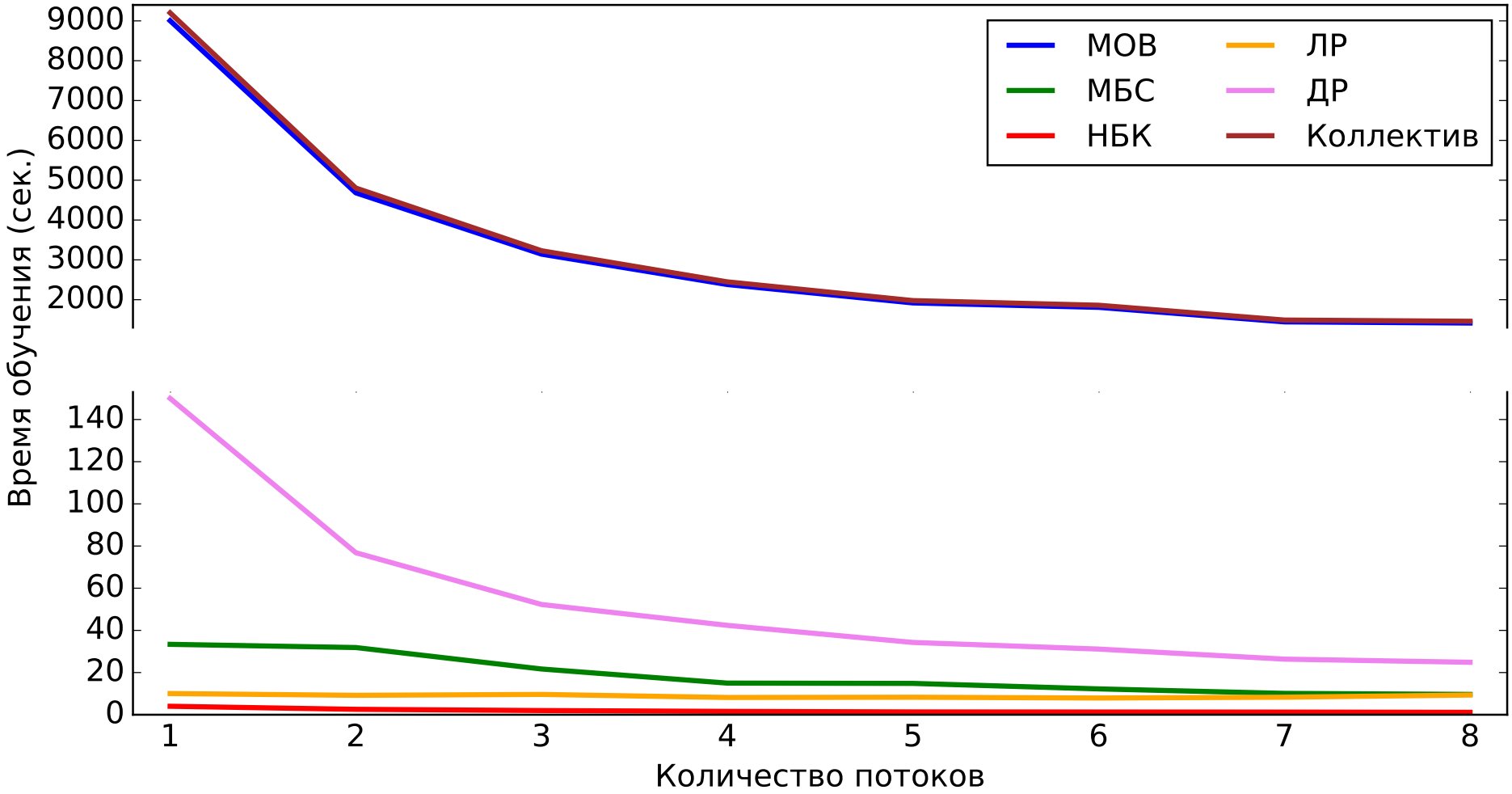
Зависимость времени обучения от числа главных компонент

$402 \to 300$: $t_{\text{обуч.}} \Downarrow$ в 1.46, $402 \to 200$: $t_{\text{обуч.}} \Downarrow$ в 2.05, $402 \to 100$: $t_{\text{обуч.}} \Downarrow$ в 3.9



Зависимость времени обучения от числа потоков

$1 \to 2: \ t_{\text{обуч.}} \Downarrow$ в 1.92, $1 \to 4: \ t_{\text{обуч.}} \Downarrow$ в 3.76, $1 \to 8: \ t_{\text{обуч.}} \Downarrow$ в 6.31

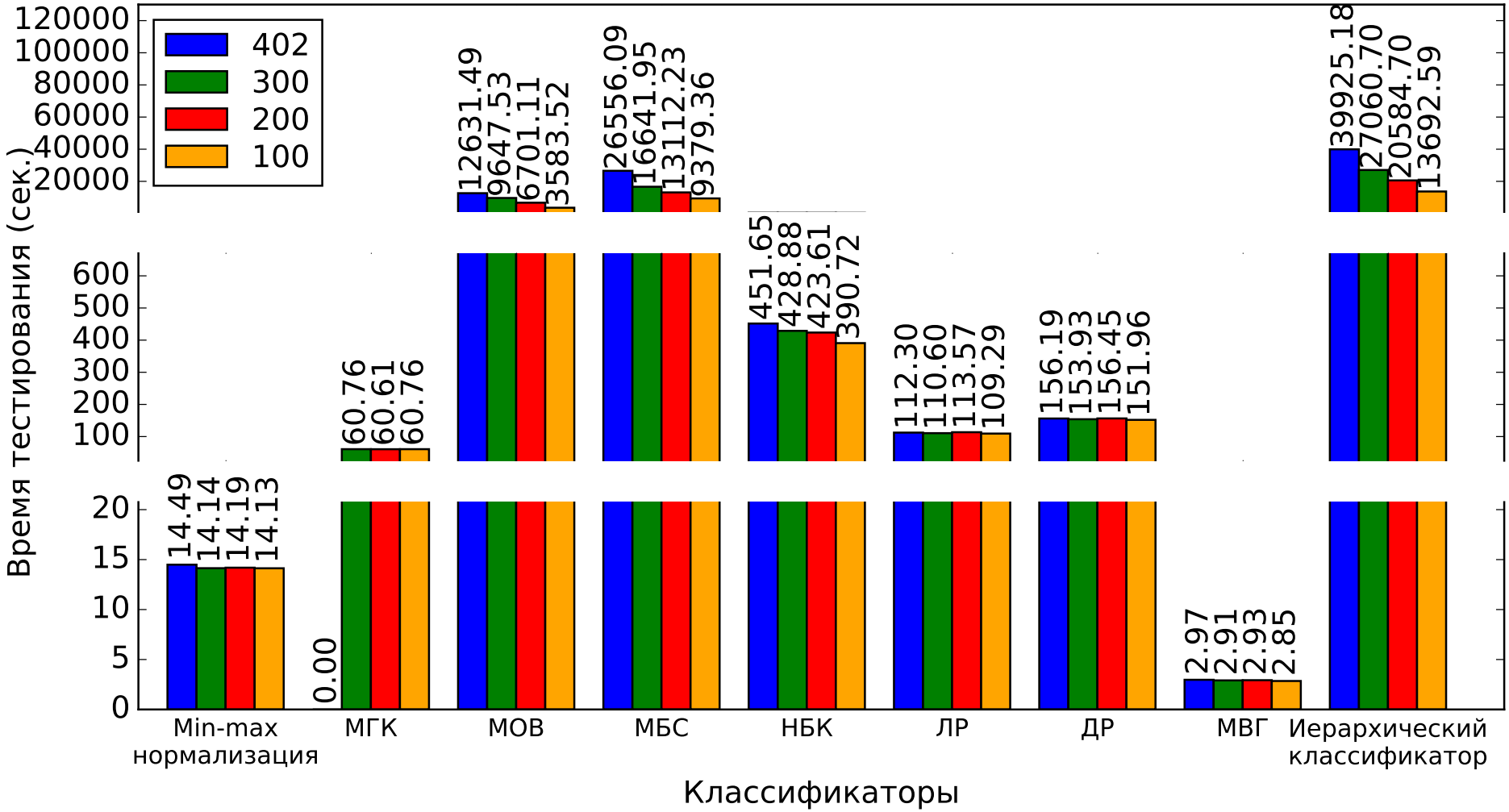


Оценка времени анализа объектов

Эксперимент 3. Оценка времени анализа объектов

Зависимость времени анализа объектов от числа главных компонент

$402 \to 300: \ t_{\text{тест.}} \Downarrow$ в 1.48, $402 \to 200: \ t_{\text{тест.}} \Downarrow$ в 1.94, $402 \to 100: \ t_{\text{тест.}} \Downarrow$ в 2.92



Заключение

- Разработана методика обнаружения вредоносных ИО в сети Интернет
- Выполнена экспериментальная оценка показателей эффективности пяти классификаторов машинного обучения и их композиции на базе метода взвешенного голосования
- Направление дальнейших исследований — расширение списка вычисляемых параметров

Благодарности. Работа выполнена при финансовой поддержке проекта РНФ № 18-11-00302 в СПИИРАН

Список литературы

[1] Hayes P.J., Andersen P.M., Nirenburg I.B., Schmandt L.M. TCS: a shell for content-based text categorization // Sixth Conference on Artificial Intelligence Applications. — 1990. — Pp. 320–326.

[2] Apté C., Damerau F., Weiss S.M. Automated learning of decision rules for text categorization // ACM Transactions on Information Systems (TOIS). — 1994. — Vol. 12, no. 3. — Pp. 233–251.

[3] Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information processing & management. — 1988. — Vol. 24, no. 5. — Pp. 513–523.

[4] Fattah M.A. A novel statistical feature selection approach for text categorization // Journal of information processing systems. — 2017. — Vol. 13, no. 5. — Pp. 1397–1409.

[5] Lewis D.D., Ringuette M. A comparison of two learning algorithms for text categorization // Third annual symposium on document analysis and information retrieval. — 1994. — Vol. 33. — Pp. 81–93.

[6] Joachims T. Text categorization with support vector machines: learning with many relevant features // European conference on machine learning. Springer, Berlin, Heidelberg. — 1998. — Pp. 137–142.

[7] Johnson R., Zhang T. Effective use of word order for text categorization with convolutional neural networks // Proceeding of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015. — Pp. 103–112.

[8] Ghareb A.S., Bakar A.A., Hamdan A.R. Hybrid feature selection based on enhanced genetic algorithm for text categorization // Expert Systems with Applications. — 2016. — Vol. 49. — Pp. 31–47.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space // arXiv. — 2013. https://arxiv.org/pdf/1301.3781.