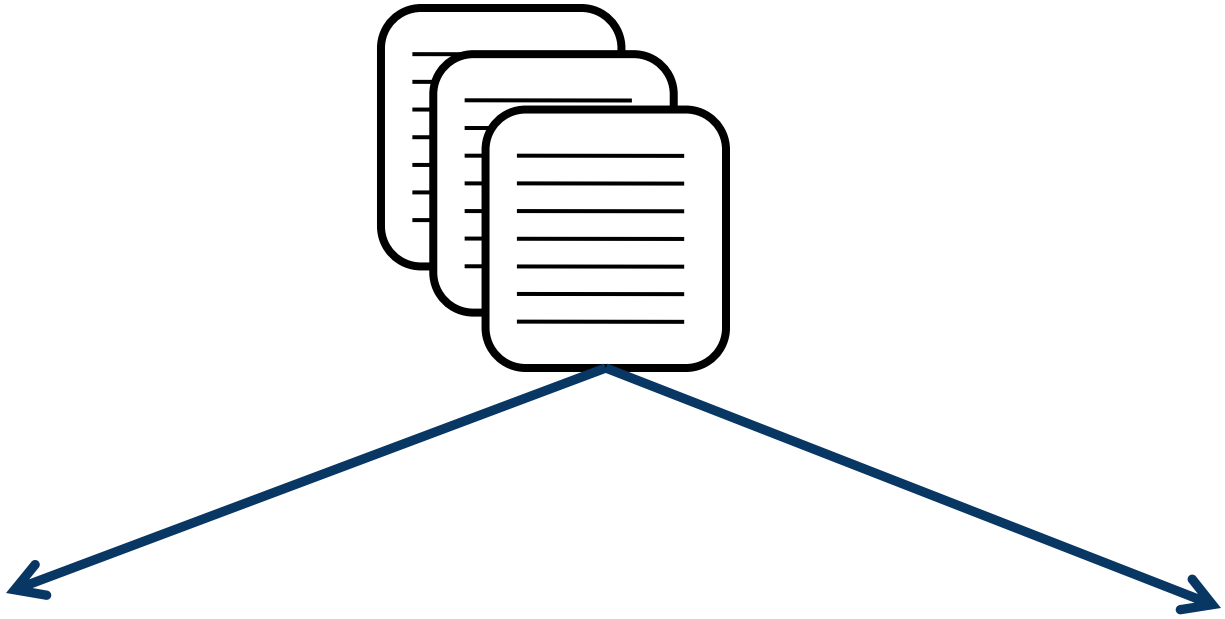
Алгоритм классификации текстов на русском языке по возрасту и гендерному признаку автора

Выполнила: *К.С. Туманова, гр.545*
Научный Руководитель: *д.ф.-м.н., проф. Б.А. Новиков*
Рецензент: *к.ф-м.н., доцент К.В. Вяткина*

Автоматическое профилирование автора текста

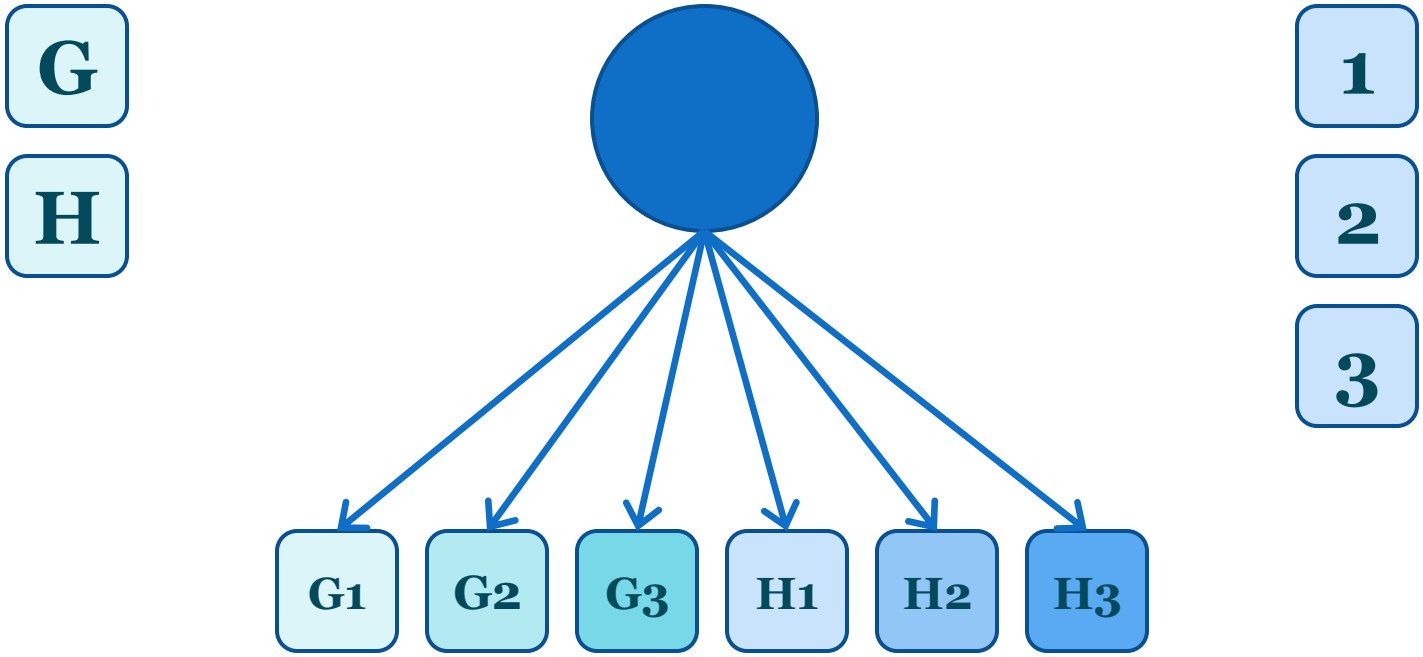


«Мужчины» **«Женщины»**

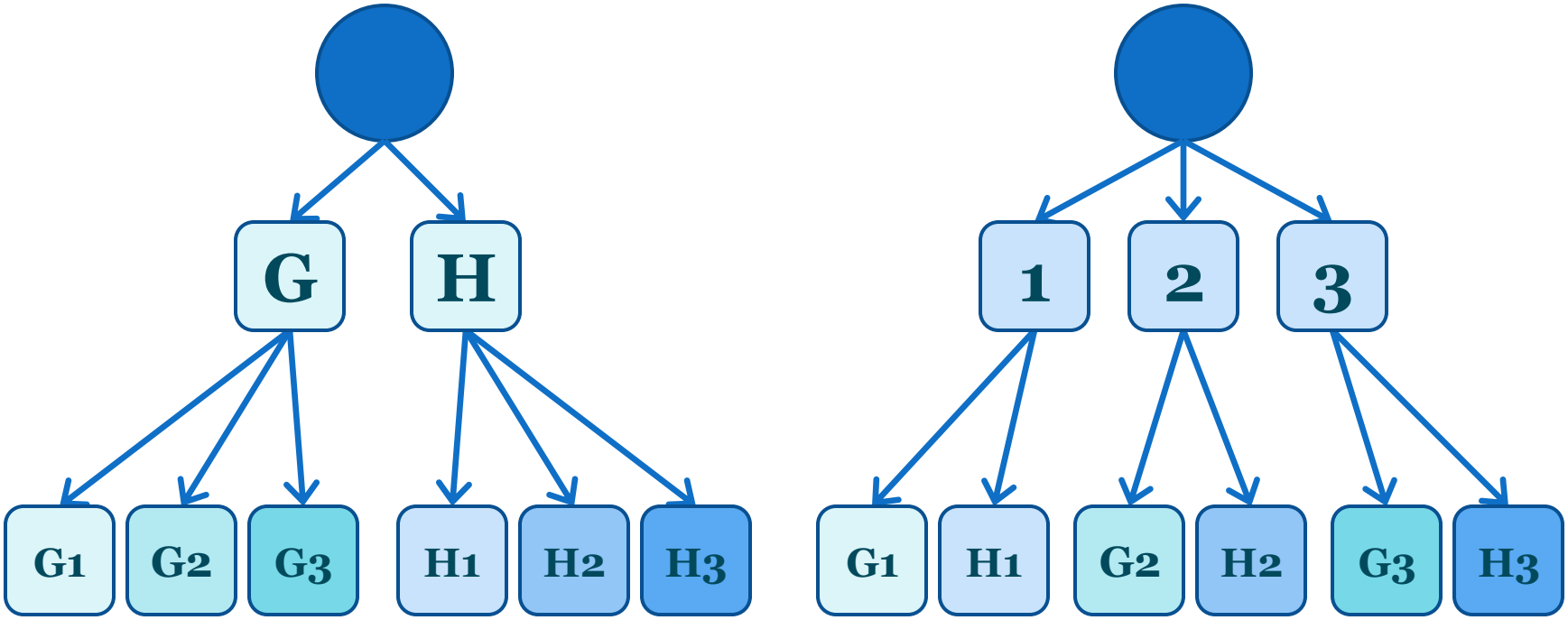
Постановка задачи

- Профилирование автора по нескольким признакам одновременно
- Классификация текстов на русском языке по возрасту и гендерному признаку автора
- Выбор характеристик
  - Характеристики должны отражать глубинные особенности письменной речи

Плоская классификация



Иерархическая классификация



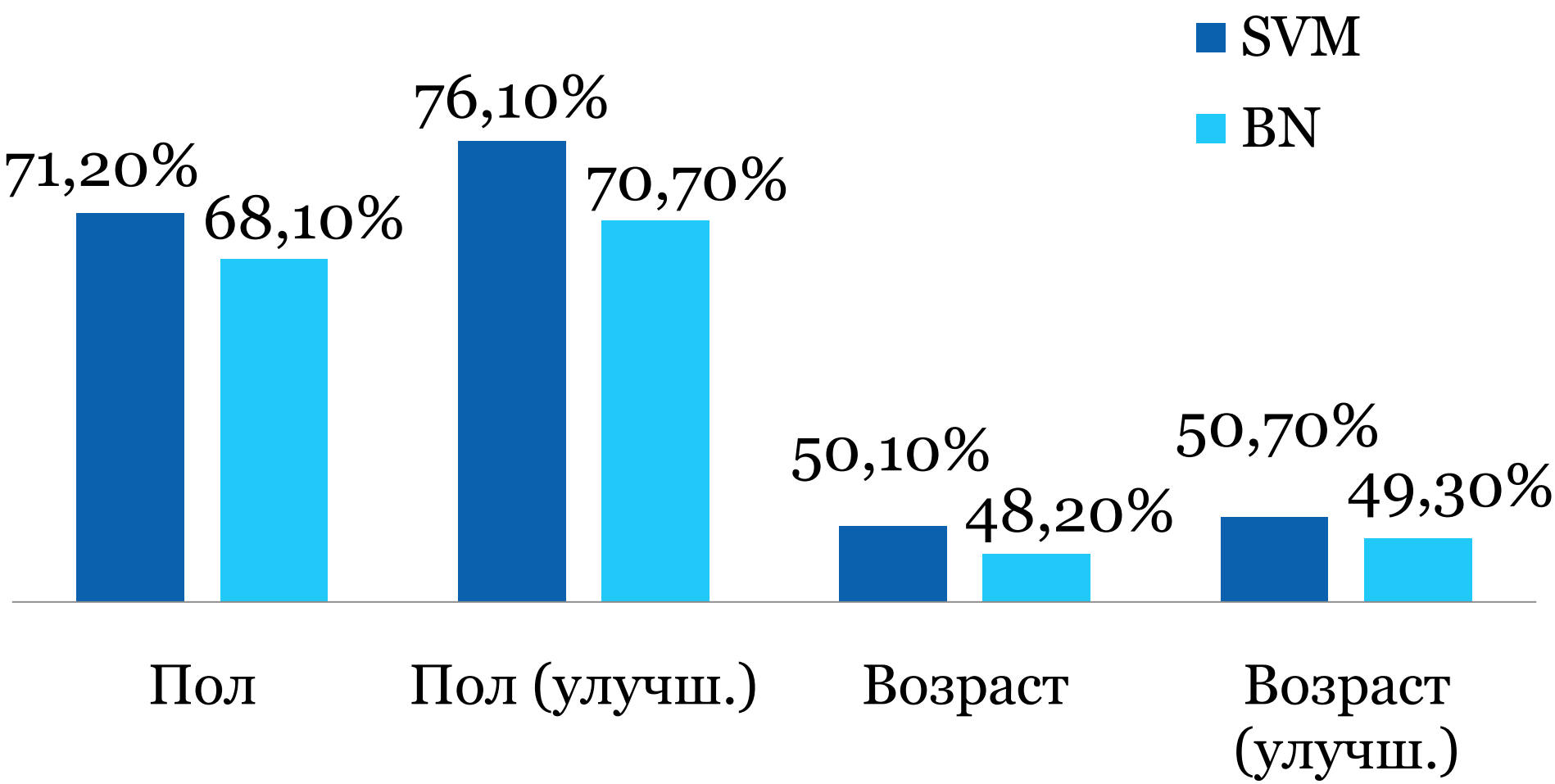
Эксперименты (1)

- Классификация текстов на русском языке по полу и возрасту автора
- База – алгоритм классификации по одному признаку
- Корпус текстов на основе блогов (351 текст)
- 4 возрастные группы
  (до 18, от 20 до 27, от 30 до 37, старше 40)
- 129 характеристик
  - частота использования знаков пунктуации, частей речи и их сочетаний, речевых оборотов и фразеологизмов, смайликов
  - длина предложений и слов
  - словарный запас

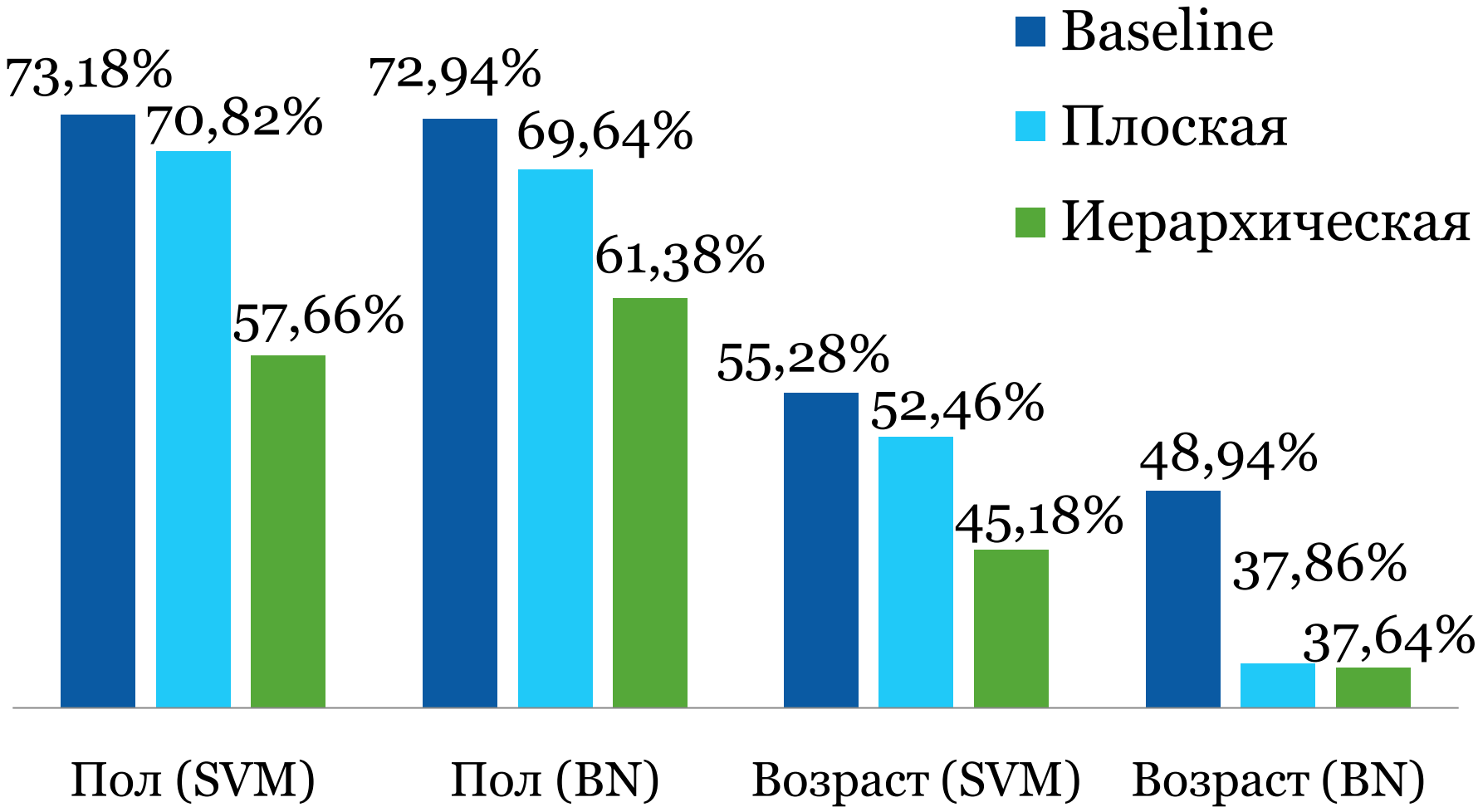
Эксперименты (2)

- Алгоритмы классификации
  - Метод опорных векторов (SVM)
  - Байесовские сети (BN)
- Протокол тестирования
  - Train Set – 75%, Test Set – 25%
  - 5 – fold cross validation
- Точность (Accuracy)

Классификация по одному признаку



Классификация по двум признакам



Результаты

- Предложено два универсальных подхода для автоматического профилирования автора по нескольким признакам одновременно
- Реализованы основанные на предложенных подходах алгоритмы
- Создана экспериментальная среда и подготовлен корпус текстов на русском языке
- Проведен анализ результатов экспериментов
- Исследован ряд характеристик для классификации текстов по стилю
- Создана основа для исследований в области профилирования авторов в применении к русским текстам