Предсказание результатов арбитражных судов с помощью методов машинного обучения

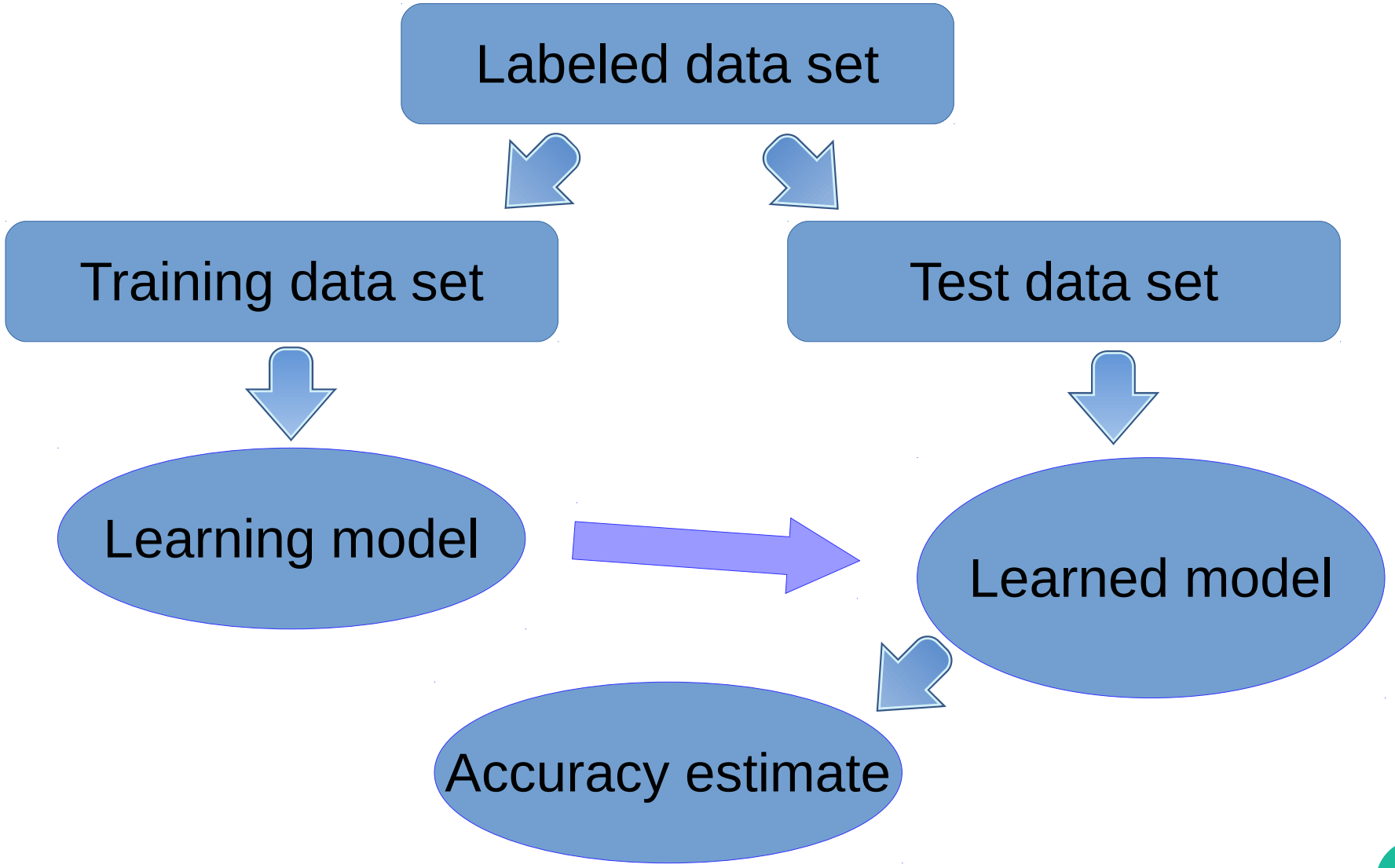
PREDICTION

Выполнил: Ёров Собир
Руководитель: Шпильман А. А.

Постановка задачи

- **Предсказание решения суда по составу дела**
  - Предсказание решения арбитражного суда по нарушению договора поставки

Постановка задачи



Этапы

Обработка текста:

- PyPdf
- Pdfminer
- Tokenizer

Векторное представление слов:

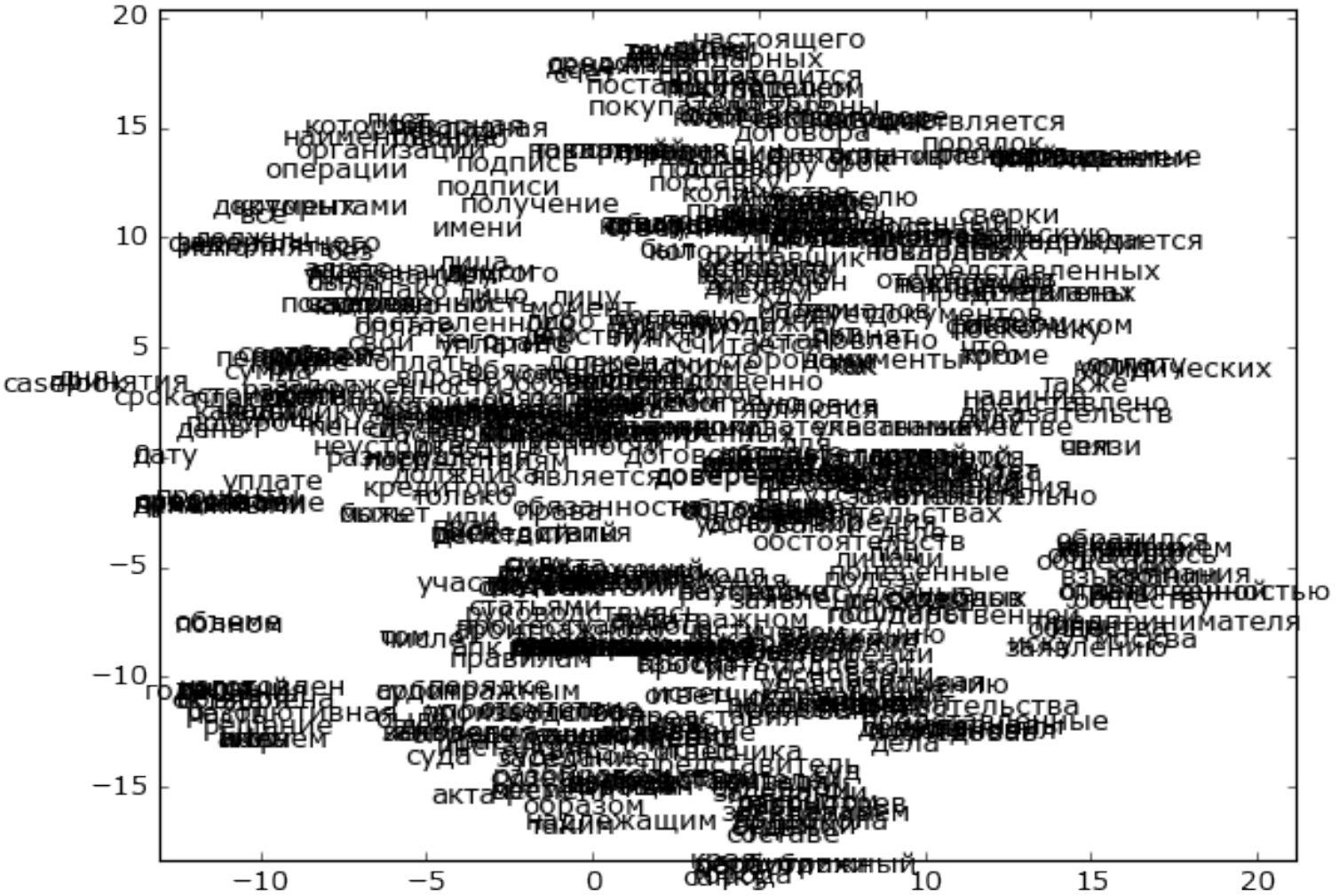
- Word2Vec

Модели для классификации:

- RandomForestClassifier
- BernoulliNB

Анализ моделей

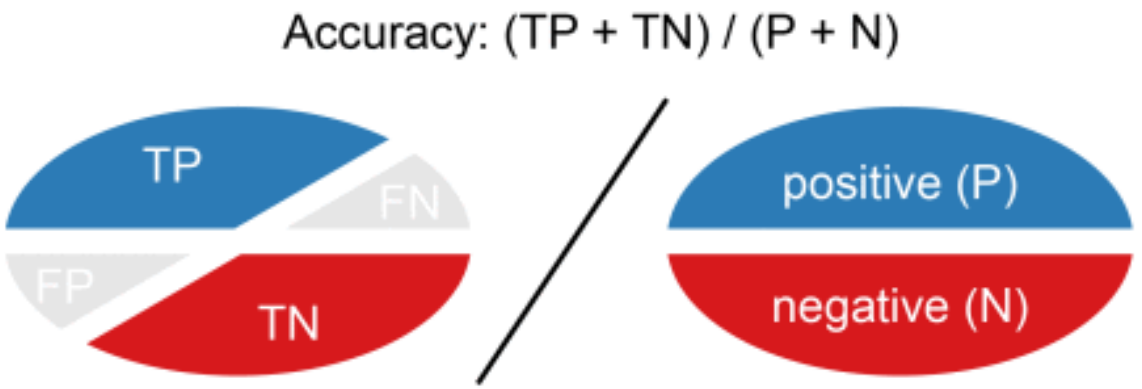
Word2Vec



Метрики качества классификации

- **Точность классификации (classification accuracy)**
- **Area under ROC Curve**
- **Матрица неточностей (confusion matrix)**
- **и т.д**

Classification accuracy



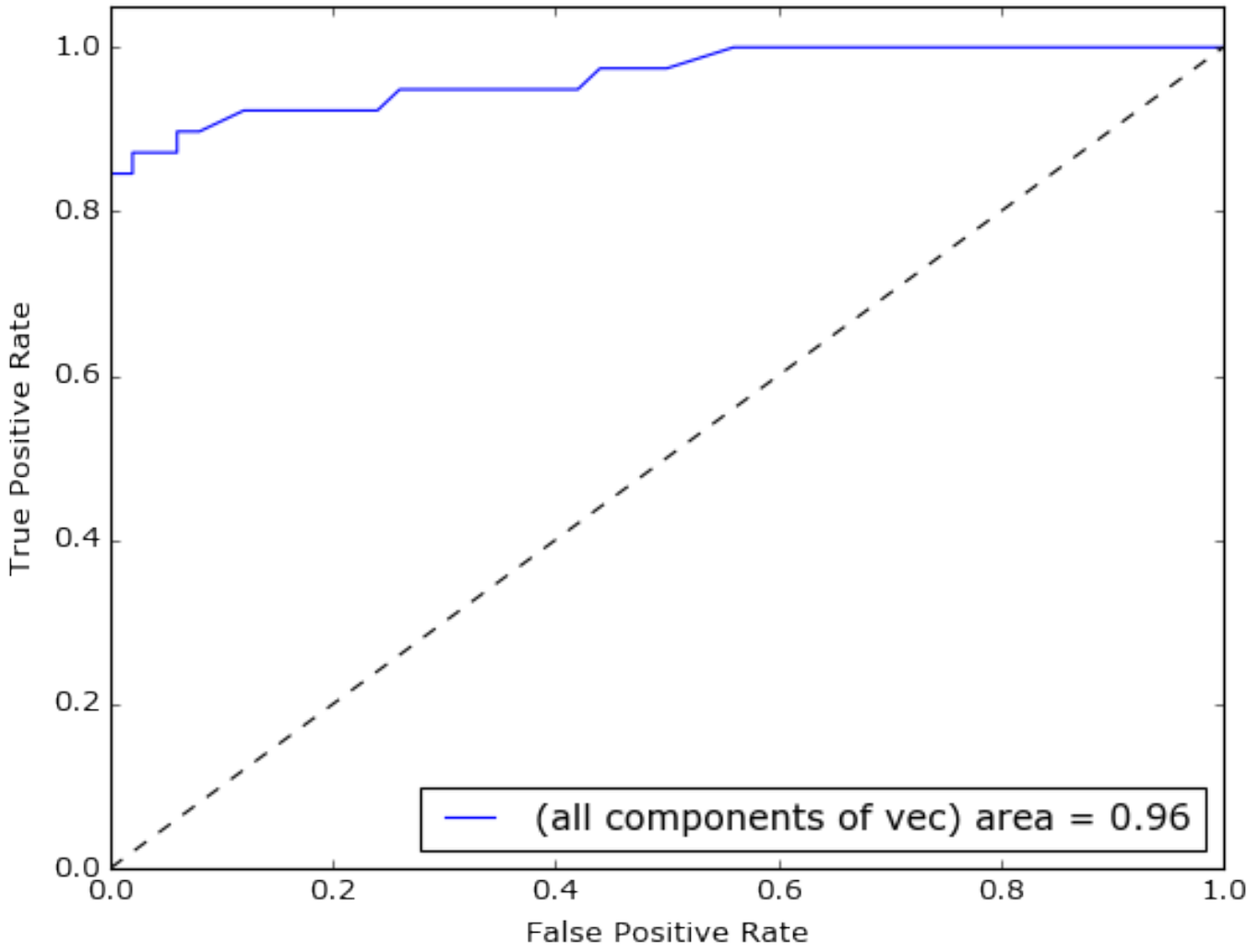
Размерность пространства признаков = 300:

0.93258426966292129

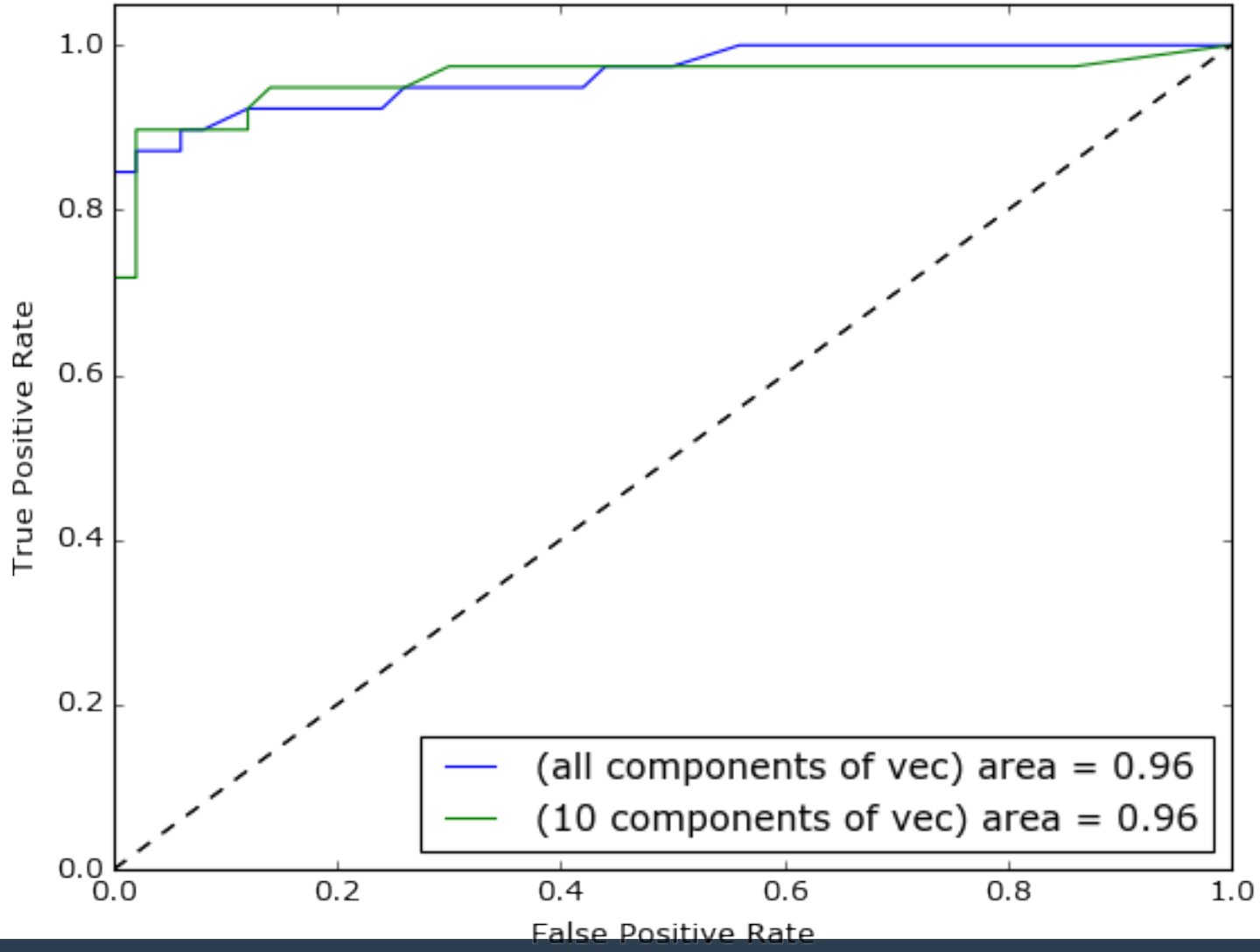
Модель с 10 наиболее важными компонентами вектора (feature_importance_):

0.943820224719

Area under ROC Curve



Area under ROC Curve



Матрица неточностей (Confusion matrix)

Модель на векторах размера 300:

49	1
5	34

Модель на 10 важных компонент векторов:

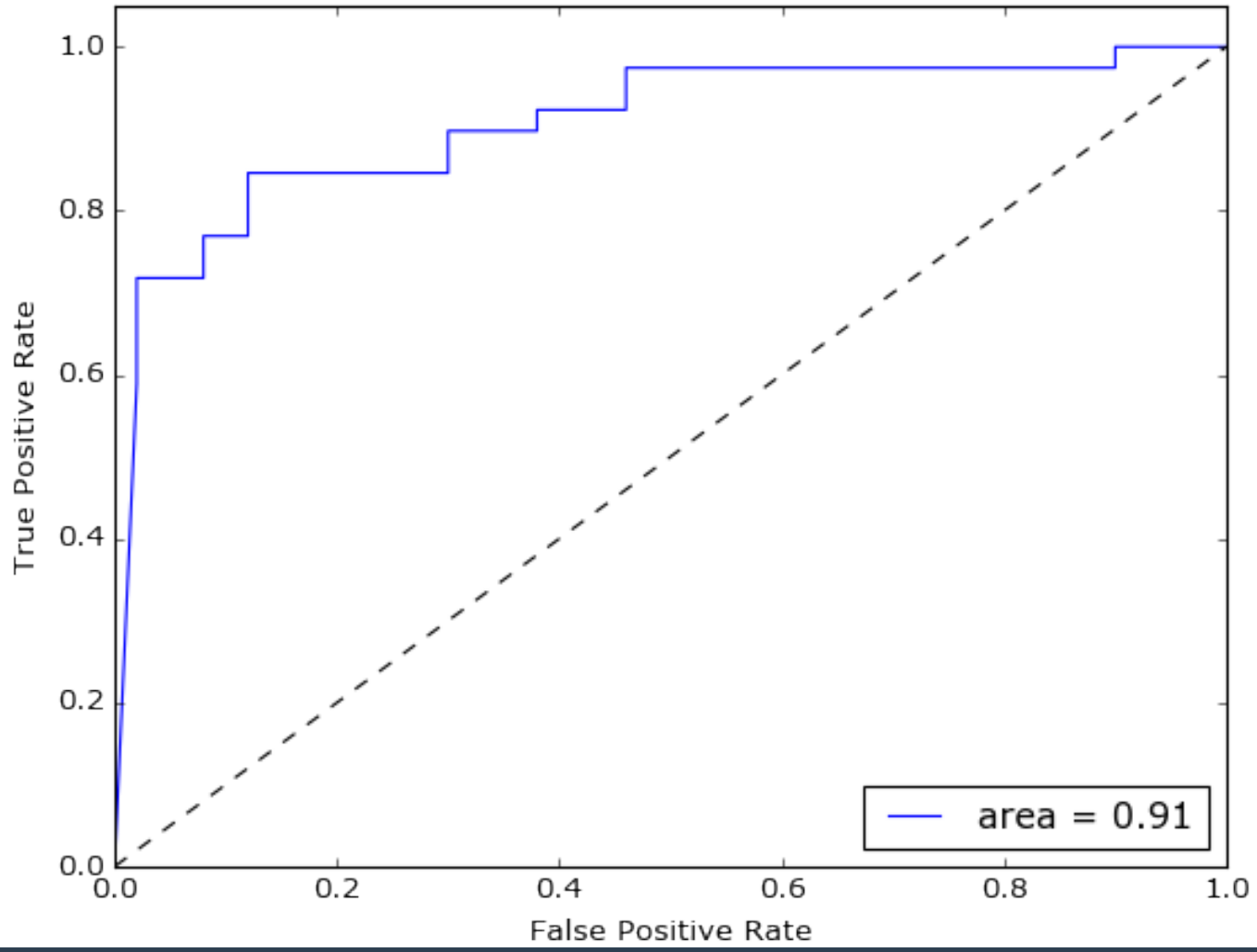
49	1
4	35

$$Precision = \frac{TP}{TP + FP} = \frac{49}{50} = 0.98$$

$$Recall = \frac{TP}{TP + FN} = 0.907$$

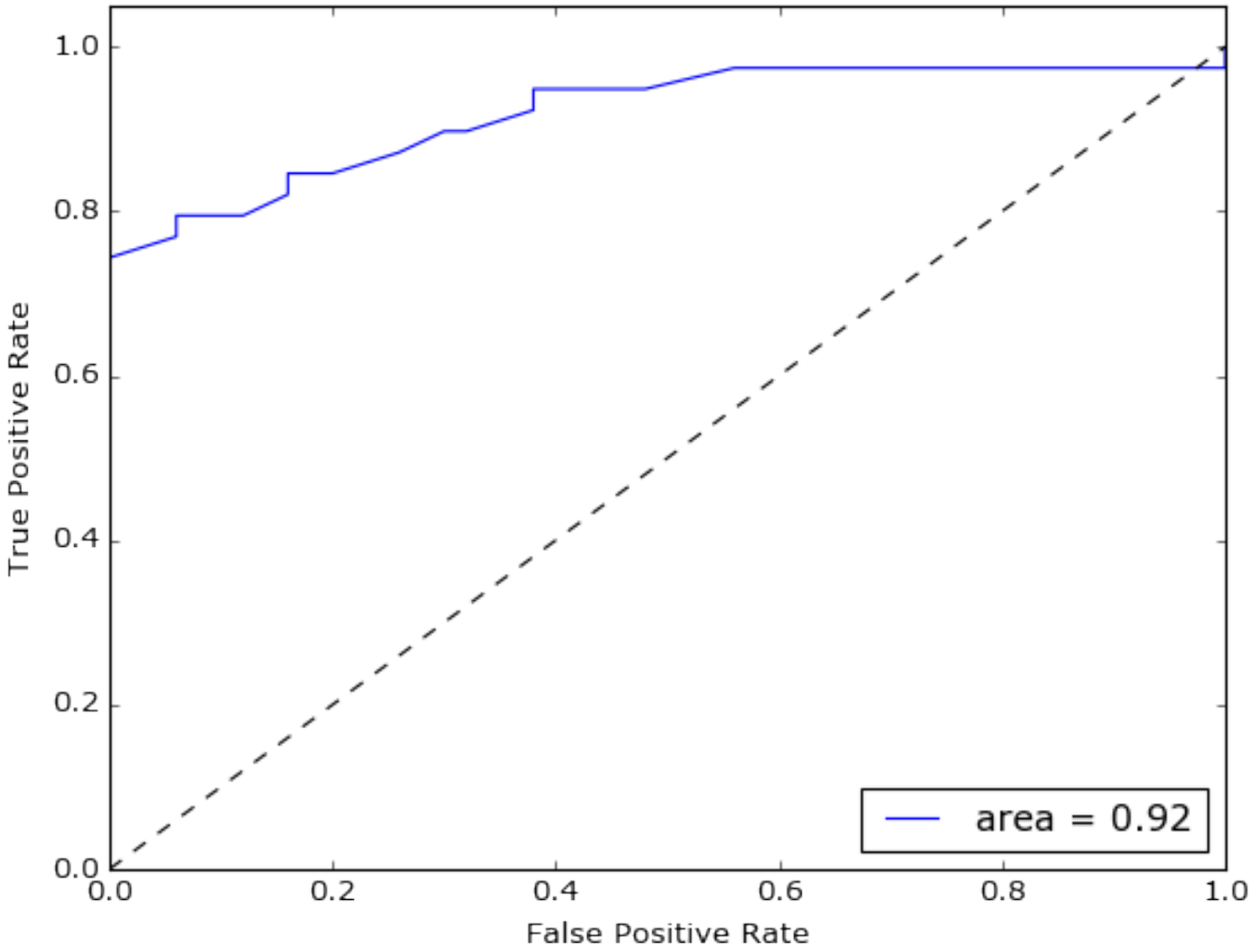
BernoulliNB (Размерность пространства признаков = 300)

Accuracy score: 0.8314606741573034



BernoulliNB (10 наиболее значимые компоненты векторов)

Accuracy score: 0.8426966292134831



Технологии

- **Python**
  - Sklearn
  - Pdfminer
  - Word2Vec

Что сделали и что хотим сделать

- **Обучили следующие классификаторы:**
  - Random Forest
  - BernoulliNB

  Анализировали результаты, выявили зависимости.

- **Хотим расширить выборку и обучить нейронную сеть**
- **Хотим сделать crawler по kad.arbitr.ru**

Спасибо за внимание!

- Код доступен по ссылке:
  - https://github.com/YorovSobir/research.git