АНАЛИЗ ОСНОВНЫХ МЕТОДОВ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ ТЕКСТОВ

Студент: Тэмуужин Янжинлхам ИУ7И-57б (ИУ7-53б)

Научный руководитель: Кивва Кирилл Андреевич

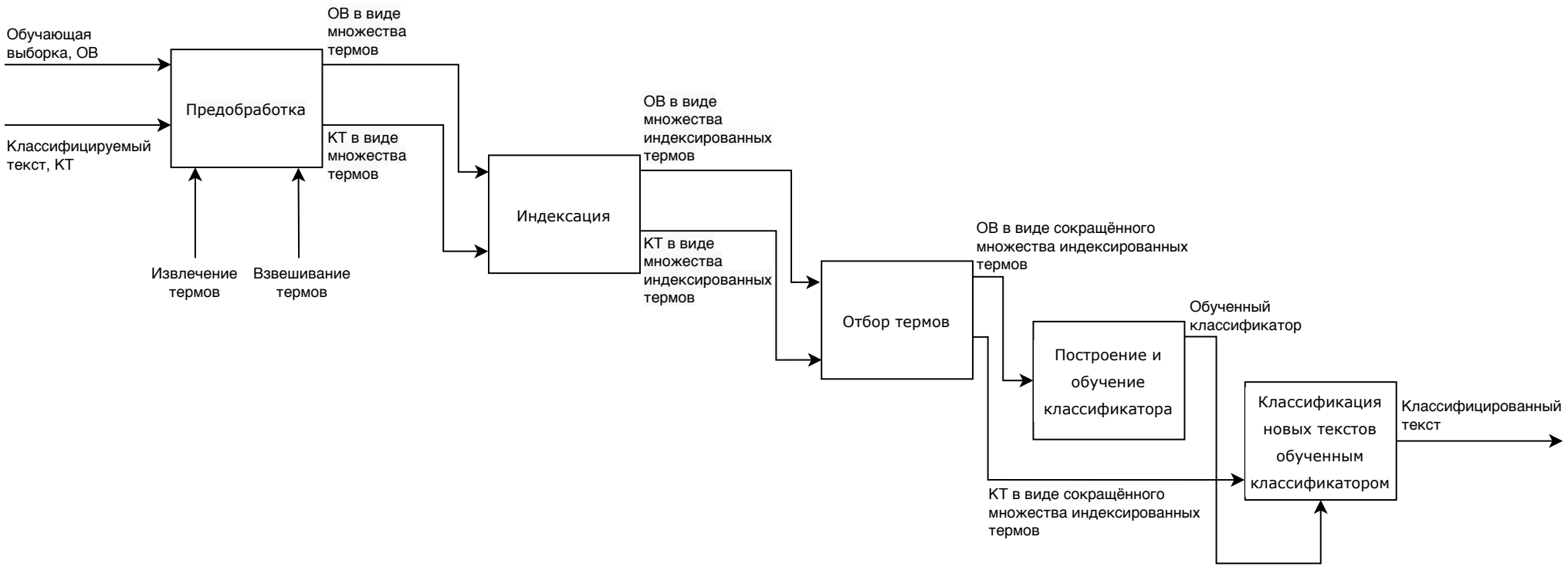
ЦЕЛЬ И ЗАДАЧИ

Цель: провести обзор основных методов решения задачи классификации текстов.

Задачи:

- изучить предметную область;
- изучить этапы процесса классификации текстов;
- изучить существующие подходы к решении задач классификации текстов;
- изучить основные методы решения задачи классификации текстов.

ЭТАПЫ ПРОЦЕССА КЛАССИФИКАЦИИ ТЕКСТОВ



ОСНОВНЫЕ ТИПЫ СИСТЕМ КЛАССИФИКАЦИИ ТЕКСТА

1. **Системы на основе правил (Rule-based systems).**
   - текст категоризуется по содержанию с помощью написанных вручную лингвистических правил
2. **Системы на основе машинного обучения с обучением (Supervised machine learning based systems).**
   - для классификации текстов используется предобученный классификатор
3. **Гибридные системы (Hybrid systems).**
   - сочетание двух вышеуказанных методов

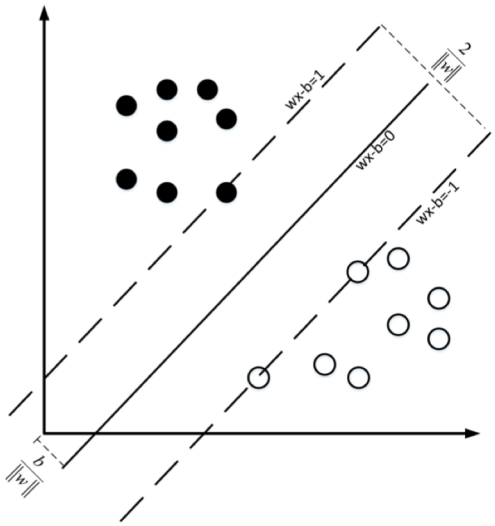
АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ, ИСПОЛЬЗУЕМЫЕ В КЛАССИФИКАЦИИ ТЕКСТОВ

Метод k ближайших соседей (K-Nearest Neighbor, KNN)

Документ **d** считается принадлежащим тому классу, который является наиболее распространенным среди **k** соседей данного документа.

- низкое качество классификации
- высокая интерпретируемость и простая реализация

Метод опорных векторов (Support Vector Machines, SVM)



Поиск разделяющей гиперплоскости, максимально удаленной от любой точки обучающих данных.

- хорошо работает при документах с низкой степенью шума
- низкое качество классификации при больших документах

АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ, ИСПОЛЬЗУЕМЫЕ В КЛАССИФИКАЦИИ ТЕКСТОВ

Алгоритм Наивного Байеса (Naive Bayes, NB)

Вычисление апостериорной вероятности класса по теореме Байеса. Слова w_i документа предполагаются независимыми

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \cdot \prod_{i \in n} P(w_i|c).$$

$$P(c) = \frac{N_c}{N_{doc}} \qquad P(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in D} count(w, c)}$$

где $P(c)$ – априорная вероятность того, что документ относится к классу c, $P(w_i|c)$ – вероятность найти слово w_i документа в классе c.

- высокая скорость работы
- не учитывается взаимодействие признаков (слов) документа
- устойчивость к шуму в исходных данных

Логистическая регрессия (Logistic Regression, LR)

Вероятность вычисляется как функция $softmax$ от взвешенной суммы признаков z со смещением b

$$z = \sum_{i=1}^{n} w_i \cdot x_i + b = \overrightarrow{w} \cdot \overrightarrow{x} + b$$

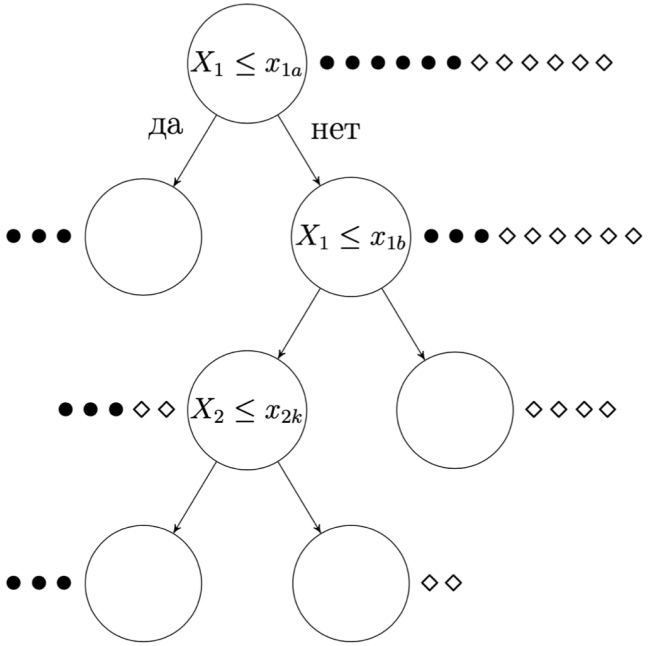
$$softmax(z_i) = \frac{exp(z_i)}{\sum_{j=1}^{K} exp(z_j)}, 1 \leq i \leq K.$$

$$P(y_k = 1|x) = \frac{exp(\overrightarrow{w_k} \cdot \overrightarrow{x} + b_k)}{\sum_{j=1}^{K} exp(\overrightarrow{w_j} \cdot \overrightarrow{x} + b_j)}$$

- высокое качество классификации
- часто используется в задачах классификации академических документов
- хорошо работает при бинарной классификации или определении тональности документа

АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ, ИСПОЛЬЗУЕМЫЕ В КЛАССИФИКАЦИИ ТЕКСТОВ

Метод деревьев решений (Decision Tree, DT)



В каждом узле условия разбиения подбираются так, чтобы максимизировать снижение энтропии (Entropy, E) или прирост информации (Information Gain, IG).

$$E = -\sum_{i=1}^{K} p_i \log_2(p_i) \qquad IG(Q) = E_{parent} - \sum_{i=1}^{q} \frac{N_i}{N} E_i$$

- высокая достоверность и скорость при классификации документов различных областей
- устойчивость к шуму

Случайный лес (Random Forest, RF)

Использование большого набора (ансамбля) деревьев решений, созданных на случайном подвыборке данных. Набор таких деревьев-классификаторов образует лес.

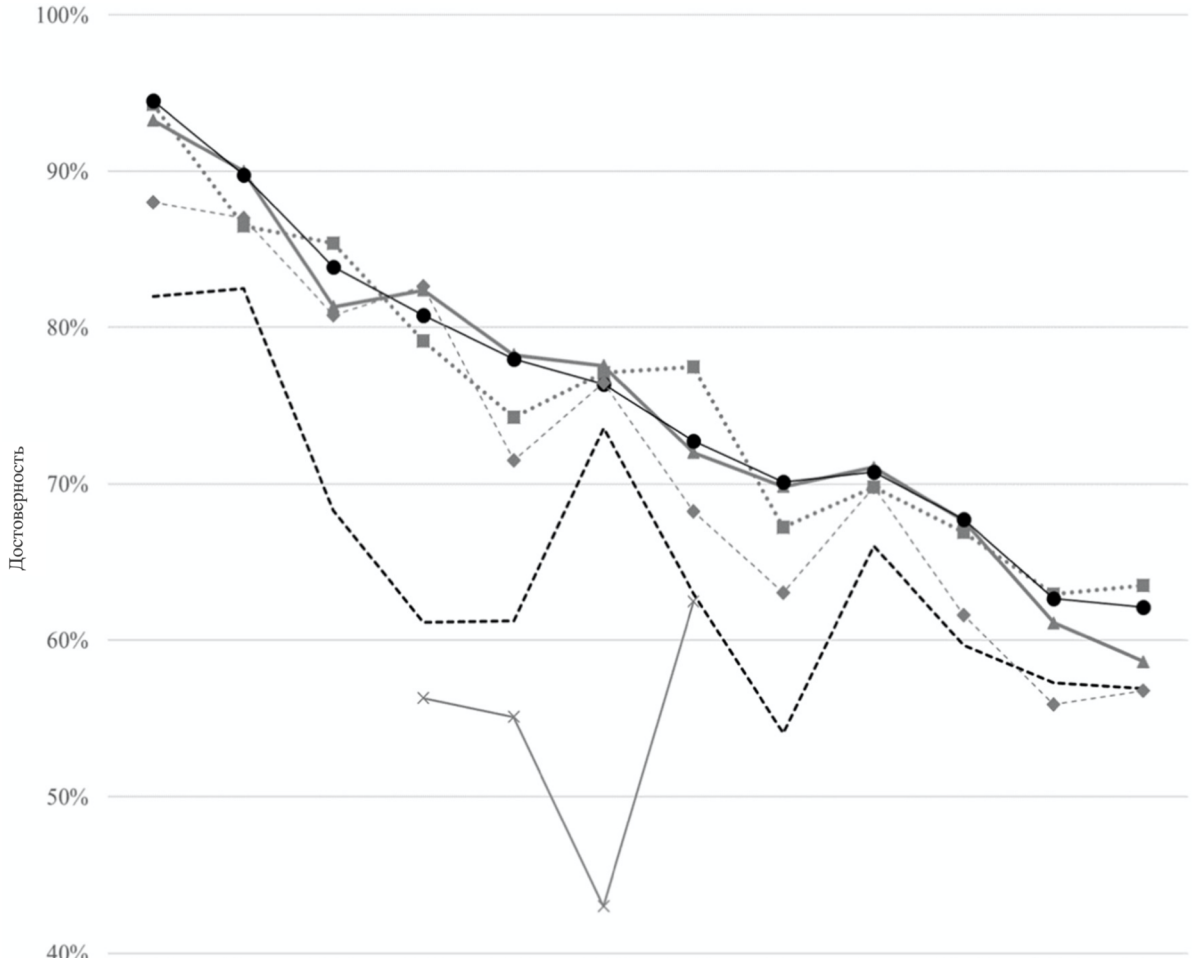
- наиболее высокая достоверность классификации (по сравнению с DT) документов различных областей
- низкая скорость

СРАВНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ КЛАССИФИКАЦИИ ТЕКСТОВ

Критерии	**kNN**	**LR**	**NB**	**SVM**	**DT**	**RF**
Достоверность	низкая	высокая	средняя	средняя	высокая	высокая
Скорость	низкая	средняя	высокая	средняя	средняя	низкая
Устойчивость к шуму	нет	нет	да	нет	да	да
Интерпретируемость	лёгкая	сложная	лёгкая	сложная	лёгкая	сложная

- kNN – метод k ближайших соседей
- LR – логистическая регрессия
- NB – наивный байесовский алгоритм
- SVM – метод опорных векторов
- DT – дерево решений
- RF – случайный лес

РЕЗУЛЬТАТ ЭКСПЕРИМЕНТА ПО КЛАССИФИКАЦИИ ДОКУМЕНТОВ, ОХВАТЫВАЮЩИХ РАЗЛИЧНЫЕ СОЦИАЛЬНЫЕ СЕТИ, ПЛАТФОРМЫ ЭЛЕКТРОННОЙ КОММЕРЦИИ И КОРПОРАТИВНЫЕ БЛОГИ

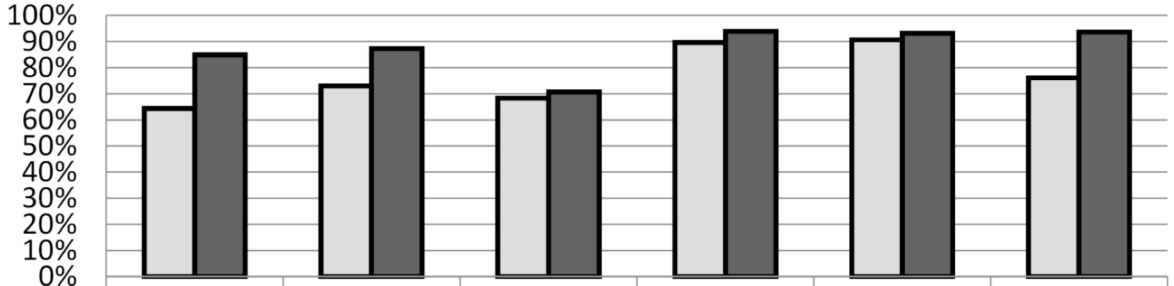


	SMS	YTU	ROT	AMR	YEL	AMT	IMD	FBK	CBP	TWS	TWC	CBC
ANN	93.3	90.0°	81.3	82.4°	78.3°	77.6°	72.0	69.8°	71.0°	67.7°	61.1	58.6
kNN	82.0	82.5	68.3	61.1	61.3	73.6	63.0	54.0	66.0	59.7	57.3	56.9
NB	94.3°	86.5	85.4°	79.1	74.3	77.1°	77.5°	67.2	69.8	66.9°	63.0°	63.5°
RF	94.5°	89.8°	83.9	80.8	78.0°	76.4	72.8	70.1°	70.8°	67.7°	62.7°	62.1
SVM	88.0	87.0	80.8	82.6°	71.5	76.4	68.3	63.0	69.8	61.6	55.9	56.8
DICT				56.3	55.1	43.0	62.5					

- AMT – заголовки отзывов о продуктах на Amazon
- AMR – отзывы продуктов на Amazon
- IMD – обзоры фильмов, опубликованные на сайте IMDB
- YEL – отзывы о ресторанах на сайте Yelp
- FBK – комментарии на Facebook
- CBC – комментарии в корпоративных блогах
- TWS – посты в Twitter
- YTU – комментарии Youtube
- SMS – текстовые сообщения, предоставляемые компаниями Almeida, Gomez Hidalgo
- ROT – обзоры фильмов, опубликованные на сайте Rotten Tomatoes
- CBP – посты в корпоративных блогах
- TWC – посты в Twitter

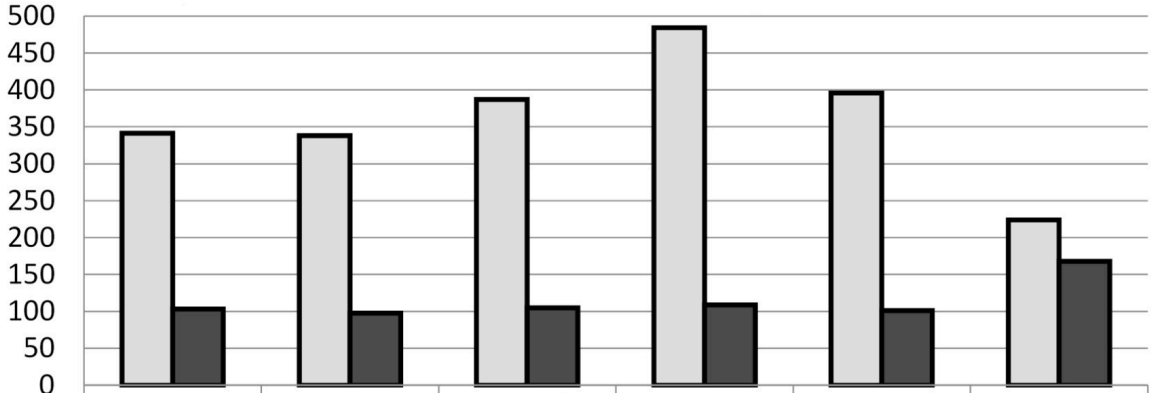
- kNN (метод k ближайших соседей)
- NB (наивный байесовский алгоритм)
- RF (случайный лес), DICT (4 методы на основе правил)
- ANN (искусственные нейронные сети)

РЕЗУЛЬТАТ ЭКСПЕРИМЕНТА ПО КЛАССИФИКАЦИИ УЧЕБНО-МЕТОДИЧЕСКИХ ДОКУМЕНТОВ НАУЧНО-ОБРАЗОВАТЕЛЬНОГО УЧРЕЖДЕНИЯ



	Полиномиальный Наивный Байес	Бернулли Наивный Байес	Метод k-ближайших соседей (k = 250)	Деревья решений	Логистическая регрессия	Нейронная сеть
До обработки	64,4%	73,1%	68,3%	89,7%	90,7%	76,1%
После обработки	84,9%	87,3%	70,7%	94,0%	93,3%	93,6%

Сравнение достоверности классификации



	Полиномиальный Наивный Байес	Бернулли Наивный Байес	Метод k-ближайших соседей (k = 250)	Деревья решений	Логистическая регрессия	Нейронная сеть
До обработки	341,253	337,991	387,09	484,079	395,827	224
После обработки	103,333	97,306	104,627	108,564	100,879	168

Сравнение времени обучение классификатора

ЗАКЛЮЧЕНИЕ

- изучена предметная область – классификация текстов
- изучены этапы процесса классификации текстов
- изучены существующие подходы к решении задач классификации текстов
- изучены основные методы решения задачи классификации текстов