

СИБИРСКИЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

SIBERIAN
FEDERAL
UNIVERSITY

Технологии автоматической обработки и анализа текстовой информации (NLP)

ГФ25-01Б

Аксаментова Елена - презентация

Аркадьева Ксения – презентация, спикер

Идрисова Самира - мд

Карасова Вероника – поиск информации, биб

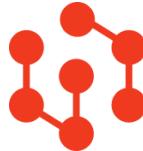


Что такое NLP?

Технологии автоматической обработки и анализа текстовой информации (NLP) (Natural Language Processing) или обработка естественного языка – это методы и инструменты, позволяющие компьютерам понимать, интерпретировать и извлекать смысл из текстовых данных.

Прежде чем компьютер сможет «понять» и проанализировать текст, его необходимо преобразовать в машиночитаемый формат. Этот процесс лежит в основе всех NLP-технологий

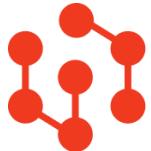




Ключевые задачи и технологии

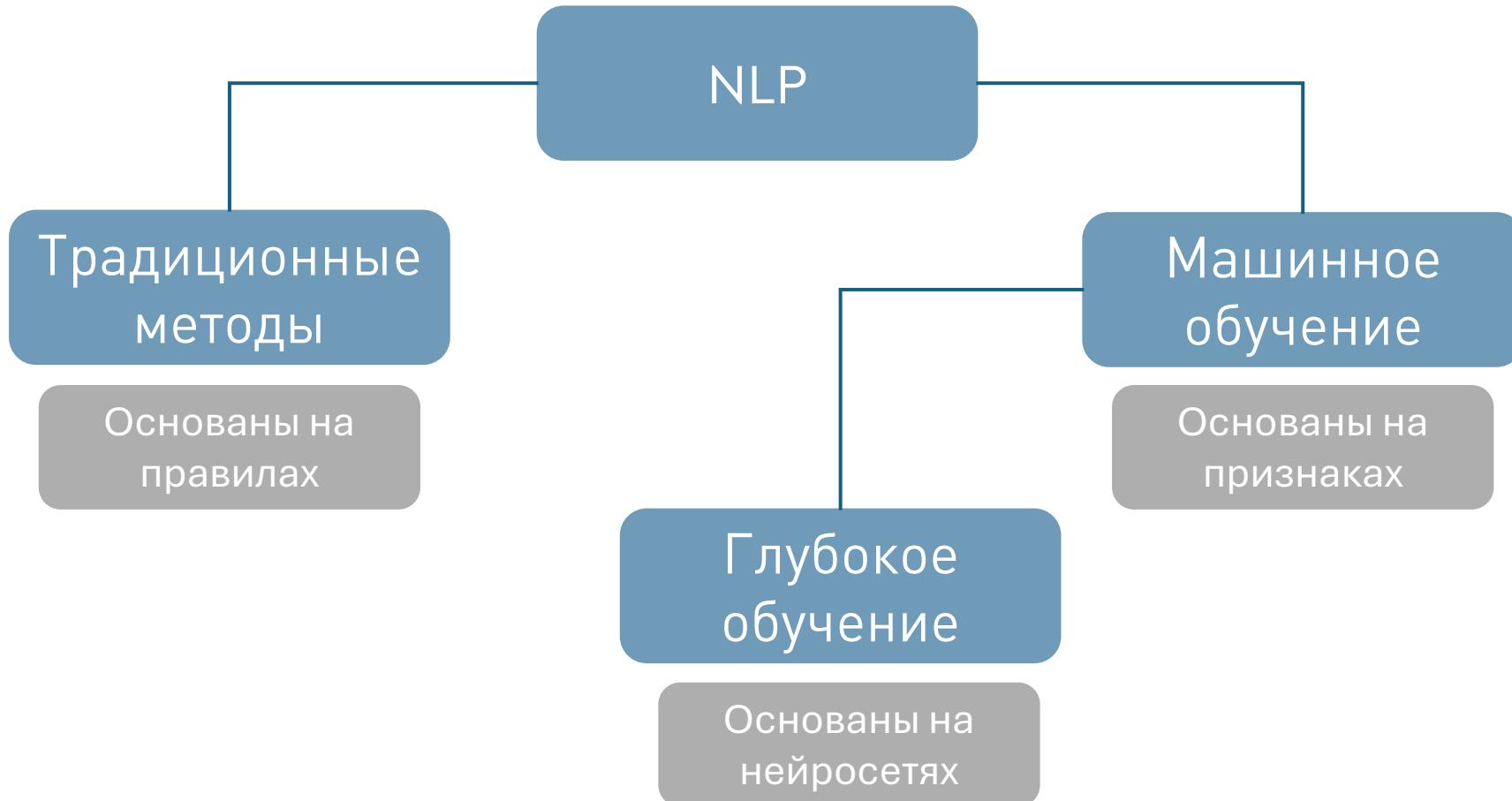
Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Классификация текстов:** Отнесение текста к категории.
Пример: Определение, является ли электронное письмо спамом.
- **Извлечение именованных сущностей:** Выявление и классификация сущностей (люди, организации, места).
Пример: "Джон Смит работает в Google" -> Имя: Джон Смит, Организация: Google.
- **Определение тональности:** Оценка эмоциональной окраски текста: позитивная, негативная, нейтральная.
Пример: "Мне понравился этот фильм!" -> Позитивный.
- **Машинный перевод:** Автоматический перевод с одного языка на другой.
- **Суммаризация текста:** Создание краткого содержания.
- **Вопросно-ответные системы:** Поиск ответов на вопросы в тексте.
- **Диалоговые системы (чат-боты):** Поддержка диалога и выполнение команд.



Какие технологические подходы используются в NLP?

Технологии автоматической обработки
и анализа текстовой информации (NLP)





Что такое традиционные методы NLP?

Технологии автоматической обработки и анализа текстовой информации (NLP)

Традиционные методы NLP — это подходы, основанные на жестких лингвистических правилах и словарях, создаваемых экспертами вручную. Компьютер следует явно прописанным инструкциям для обработки текста

Почему традиционные методы ещё используют в NLP:

- Точность для шаблонов - идеально подходят для номеров, дат, email
- Объяснимость - понятная логика вместо "чёрного ящика" нейросетей
- Работа без данных - не нужны тысячи примеров для обучения
- Дёшево и быстро - простые задачи решаются за часы, а не недели
- Надёжность - всегда работают одинаково, нет случайных ошибок
- Юридические требования - в банках и медицине нужна прозрачность решений



Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

Суть традиционного метода

Обработка языка по жёстким правилам, созданным лингвистами вручную. Компьютер не учится, а лишь выполняет инструкции.

Как работает

- Эксперт анализирует язык и составляет правила (словари, грамматики, шаблоны)
- Программист кодирует эти правила в систему
- Компьютер применяет правила к тексту, ища точные совпадения
- Результат всегда предсказуем, но не адаптируется к новым случаям

Аналогия

Дать ребёнку инструкцию: «Всех зверей с усами называй кошками». Он будет так называть и кошку, и хорька, и тигра — не понимая разницы, а лишь слепо следуя правилу



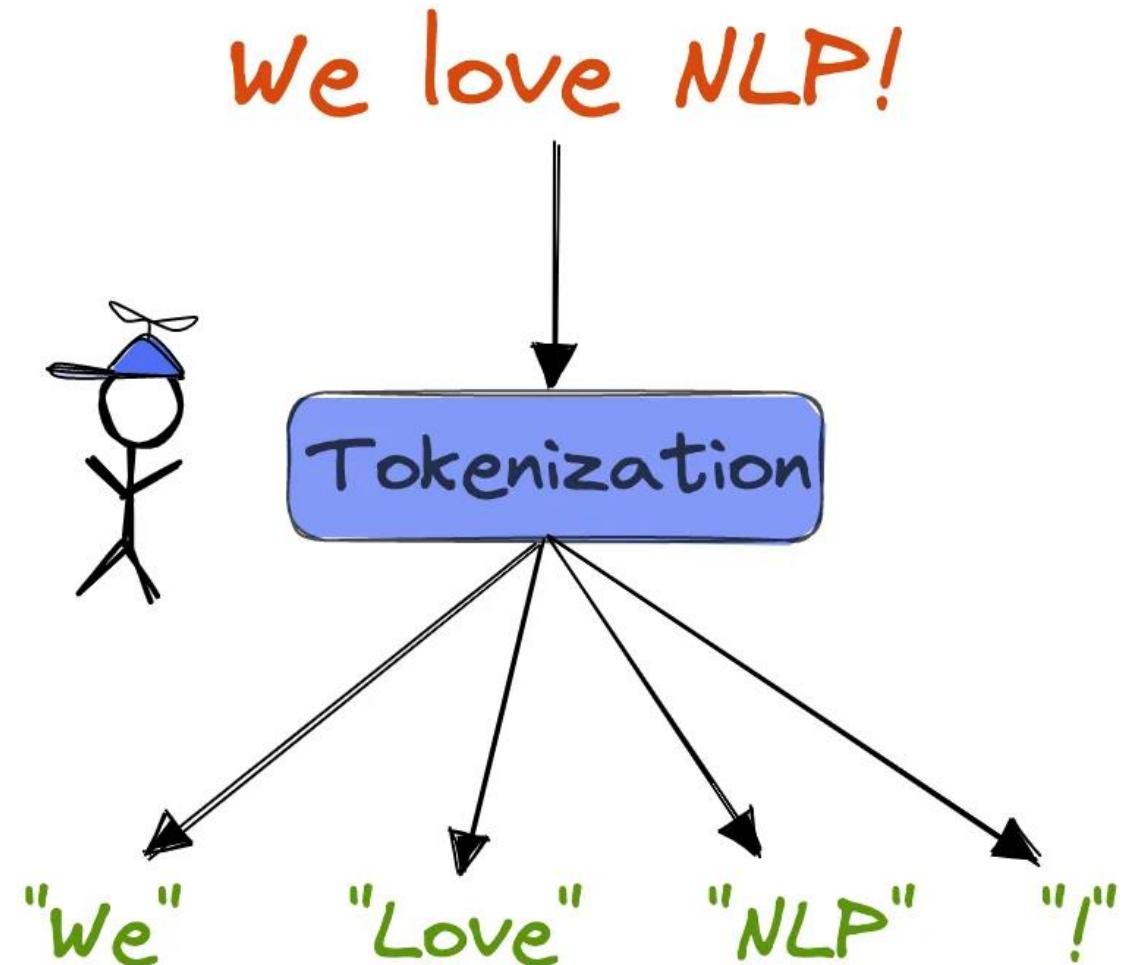
Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- Токенизация

Разбиение на отдельные слова
(токены)

Пример: «Книга лежит на столе» ->
[«Книга», «лежит», «на», «столе»]





Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

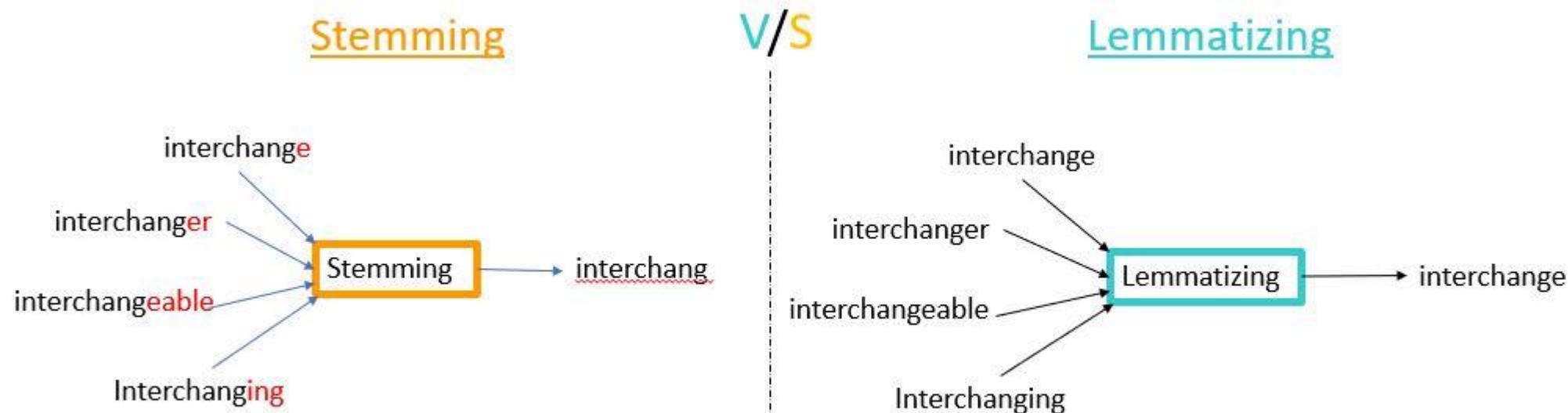
• Стемминг/Лемматизация

Приведение слов к корневой форме.

Стемминг удаляет окончания.

Лемматизация использует морфологический
анализ

Пример: Стемминг: "бегущий" -> "бег",
Лемматизация: "бегущий" -> "бежать"





Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Векторное представление текста**

Преобразование текста в числовой вектор

1. Bag of words (BoW):

учитывает частоту слов, не учитывая порядок

Пример: «кот спит», «собака спит», -> [«кот»: 1, «спит»: 2, «собака»: 1]





Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Векторное представление текста**

Преобразование текста в числовой вектор

2. TF-IDF:(Term Frequency-Inverse Document Frequency)

Оценивает важность слова с учётом его частоты в
документе и редкости в коллекции

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

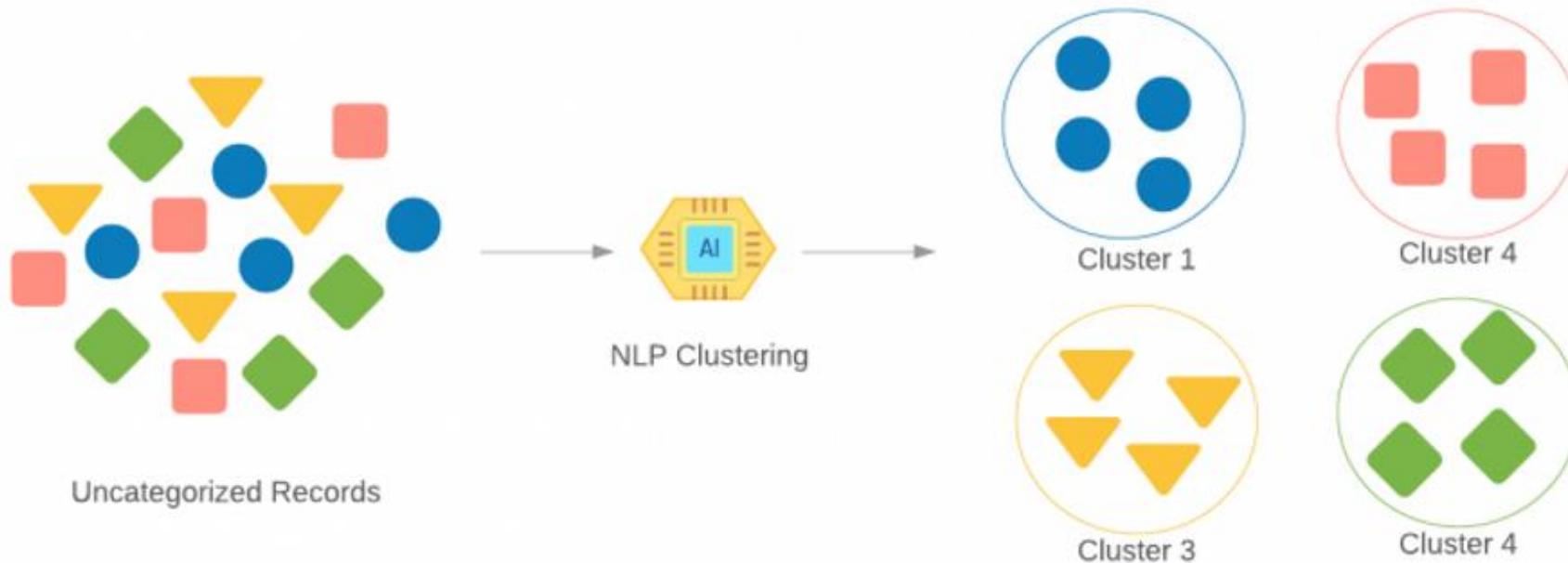


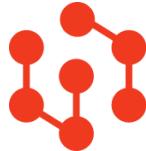
Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Кластеризация**

Разбиение документов на тематические группы



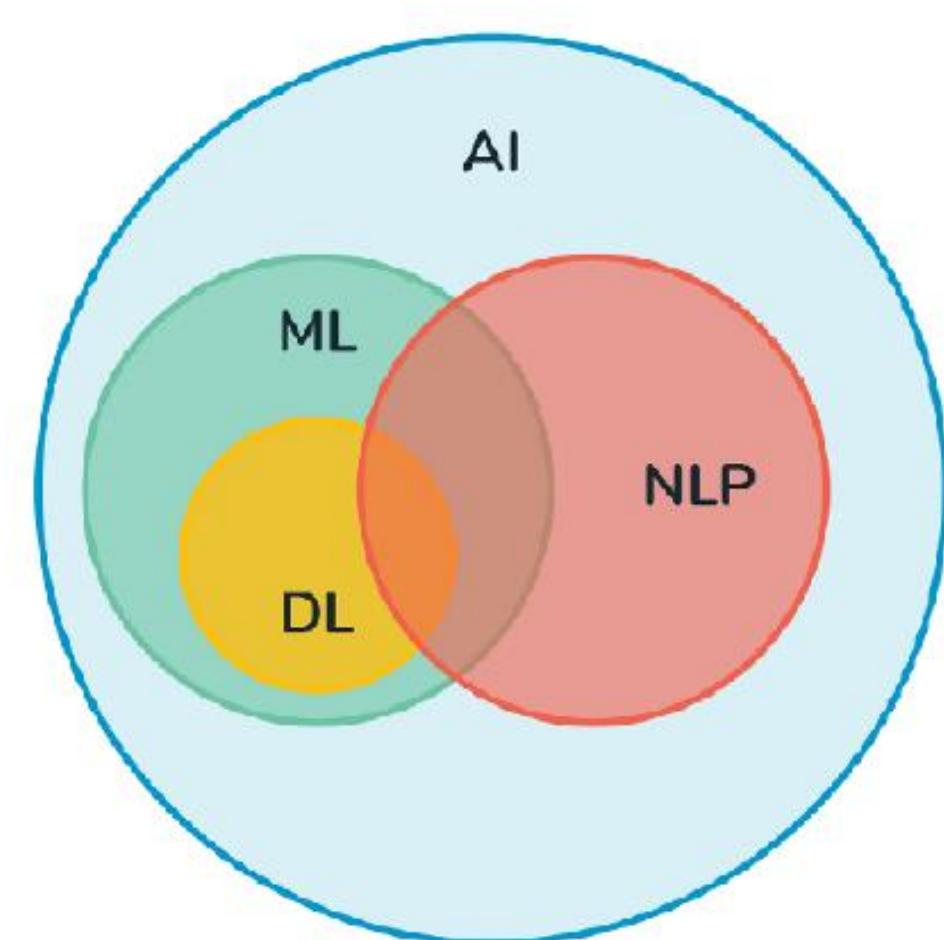


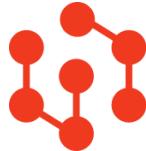
Машинное и глубокое обучение

Машинное обучение (Machine Learning, ML), и глубокое обучение (Deep Learning, DL) являются фундаментальными частями современного NLP

- NLP — это цель (научить компьютер понимать язык).
- Машинное обучение — это основной инструмент для достижения этой цели.
- Глубокое обучение — это самый современный и мощный инструмент в арсенале ML, который произвел революцию в NLP, позволив компьютерам понимать тонкости и контекст человеческого языка на невиданном ранее уровне

Технологии автоматической обработки и анализа текстовой информации (NLP)





Машинного обучения

Технологии автоматической обработки
и анализа текстовой информации (NLP)

Суть машинного обучения:

Компьютер учится решать задачу на примерах, а не по заранее написанным правилам.

Как работает:

- Показываем алгоритму много данных с «правильными ответами»
- Алгоритм сам находит закономерности и паттерны
- Используем найденные закономерности для прогнозов на новых данных

Аналогия:

Показываем компьютеру 1000 кошек и собак → он сам учится отличать их
→ распознаёт новых животных.

Методы:

- **Наивный Байес** - классификация спама
- **SVM** - определение тематики текстов
- **Логистическая регрессия** - анализ тональности
- **Случайный лес** - категоризация документов



Глубокое обучение

Технологии автоматической обработки и анализа текстовой информации (NLP)

Суть глубокого обучения

Автоматическое извлечение сложных паттернов из данных с помощью многослойных нейросетей

Как работает

- Данные последовательно проходят через слои нейросети
- Каждый слой выделяет признаки от простых к сложным
- Модель самостоятельно учится оптимальным преобразованиям
- Система обобщает знания на новые данные

Аналогия

Ребёнок изучает кошек не по правилам, а рассматривая тысячи животных и самостоятельно формируя целостный образ, что позволяет узнавать кошек в любых условиях

Методы:

- **Сверточные сети (CNN)** — классификация, анализ тональности, выявление паттернов.
- **Рекуррентные сети (RNN) и LSTM** — обработка последовательностей, учёт контекста, моделирование зависимостей.
- **Трансформеры (BERT)** — решение разнообразных задач, достижение точности, применение механизма внимания.



Заключение

Технологии автоматической обработки
и анализа текстовой информации (NLP)

Что узнали об NLP:

- **Традиционные методы** — работают по четким правилам, но не умеют учиться. Для простых задач
- **Машинное обучение** — учится на примерах, но требует ручной настройки. Для стандартных задач
- **Глубокое обучение** — само учится и находит сложные закономерности. Для сложных задач