# Автоматическая аннотация тем в вероятностном тематическом моделировании

Карпинская Анна Викторовна

417 группа ММП, ВМК МГУ

# Введение

- Вероятностное тематическое моделирование широко используется для анализа текстовых данных
- Проблема: сложность интерпретации результатов, так как модели часто возвращают только списки ключевых слов
- Цель исследования: создание связных, информативных и кратких текстовых аннотаций тем
- Методика объединяет:
  - Вероятностное тематическое моделирование
  - Методы извлечения ключевых фраз
  - Генеративные языковые модели

# Актуальность задачи

- Автоматическая аннотация тем упрощает анализ больших текстовых коллекций
- Использование генеративных моделей позволяет улучшить связность и читаемость текстов, формируя понятные аннотации
- Применение для:
  - Информационного поиска
  - Автоматического резюмирования
  - Анализа новостей

# Постановка задачи

- **Входные данные:** документы $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, разбитые по темам $\mathcal{T} = \{T_1, T_2, \ldots, T_K\}$.
- **Цель:** Для каждой темы $T_k$ сгенерировать связное описание $S_k$.
- **Методы:**
  - Извлечение репрезентативных фрагментов
  - Выделение ключевых фраз
  - Генерация текста
- **Критерии**: кратное, связное и содержательное описание $S_k$

# Алгоритм аннотации

1. **Извлечение репрезентативных фрагментов:**
   - Лемматизация текстов и удаление стоп-слов
   - Оценка релевантности предложений теме
   - Выбор $N_{top}$ предложений с максимальной релевантностью
2. **Извлечение ключевых фраз:**
   - Выделение устойчивых словосочетаний
   - Оценка сходства фраз с центроидным вектором темы
   - Отбор наиболее релевантных ключевых фраз
3. **Генерация текстов:**
   - Формирование промпта из ключевых фраз и фрагментов
   - Использование T5 для генерации аннотаций

# Метрики оценки качества

- **ROUGE** Используется для измерения точности $P$, полноты $R$ и F1-метрики на основе пересечения n-грамм между предсказанием и эталоном:

$$P = \frac{\text{Совпадающие } n\text{-граммы}}{\text{Общее количество } n\text{-грамм в предсказании}},$$

$$R = \frac{\text{Совпадающие } n\text{-граммы}}{\text{Общее количество } n\text{-грамм в эталоне}},$$

$$F_1 = \frac{2PR}{P + R}$$

# Метрики оценки качества

- **BLEU** Основан на точности совпадений n-грамм между генерируемым текстом и эталоном:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

где BP — штраф за длину, $w_n$ — веса для n-грамм, $P_n$ — доля совпадений n-грамм

# Метрики оценки качества

- **METEOR** Учитывает лемматизацию, синонимы и порядок слов:

$$\text{METEOR} = 10 \cdot \frac{\text{Совпадения}}{\text{Средняя длина текста}}$$

- **Семантическая близость**
  Измеряет косинусное сходство между эмбеддингами генерируемого текста и эталона:

$$\text{Sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

# Данные

Для экспериментов использовался корпус **BBC News**, включающий текстовые данные пяти категорий. Распределение данных по темам представлено в таблице:

| **Категория** | **Количество документов** | **Пример содержания** |
|---|---|---|
| Business | 510 | Экономика, финансы, корпоративные новости |
| Entertainment | 386 | Кино, музыка, шоу-бизнес |
| Politics | 417 | Политические события, выборы |
| Sport | 511 | Новости спорта, результаты матчей |
| Tech | 401 | Технологические разработки, гаджеты |

Таблица: Распределение документов по темам в корпусе BBC News

# Анализ результатов

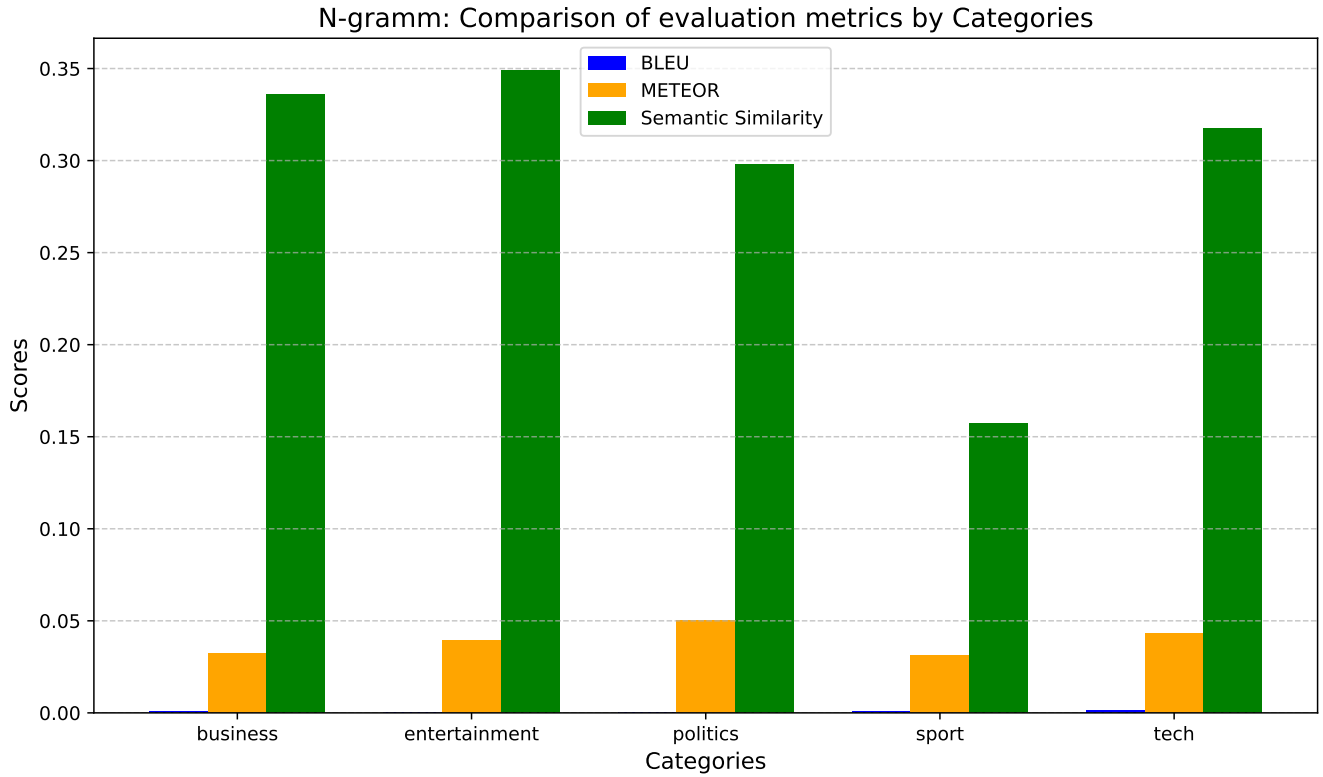


Рис.: Сравнение метрик BLEU, METEOR и Semantic Similarity

- Семантическая близость значительно выше BLEU и METEOR
- Алгоритм улавливает контекст текста, даже если синтаксически структура отличается

# Анализ результатов: ROUGE-метрики

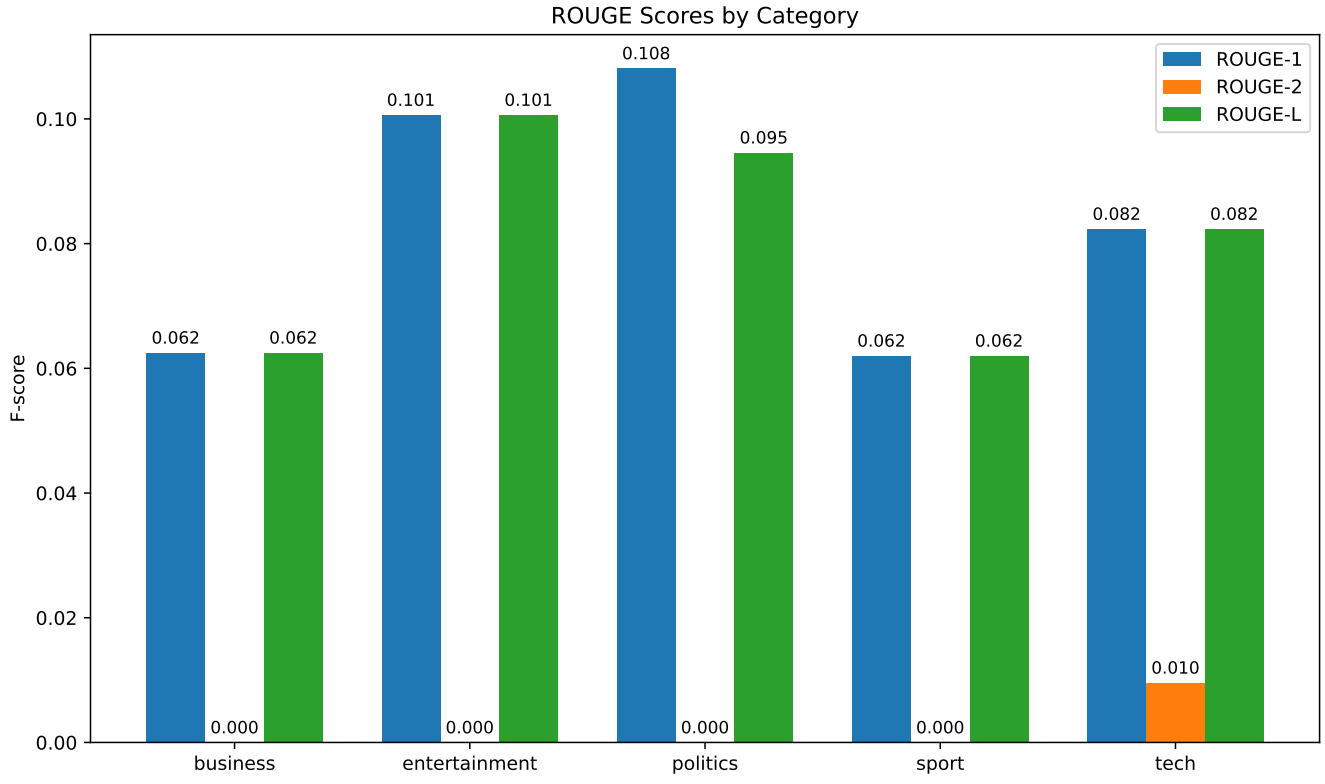


Рис.: Сравнение ROUGE-метрик

- ROUGE-1 выше для категорий politics и tech
- ROUGE-2 ниже, что свидетельствует о сложностях передачи длинных последовательностей

# Заключение

- Разработан алгоритм аннотации, объединяющий LDA и генеративные модели.
- Использование метрик, таких как ROUGE и семантическая близость, подтвердило высокую релевантность и читаемость генерируемых текстов
- Результаты демонстрируют потенциал метода для автоматического анализа текстовых данных
- Перспективы:
  - Увеличение масштабируемости метода для больших корпусов
  - Улучшение обработки категорий с высокой вариативностью тем

# Конец

**Спасибо за внимание!**