**Proposal для выпускной квалификационной работы по курсу "Компьютерная лингвистика"**

**Тема:**  
**Автоматизированный анализ эмотивности и тематики в политических речах немецких политиков на основе методов машинного обучения и дистрибутивной семантики**

**1. Актуальность и научная значимость**

Политическая коммуникация играет ключевую роль в формировании общественного мнения. Современные выступления политиков содержат сложные риторические стратегии, включая эмоциональные призывы, модальные конструкции и манипулятивные техники. Однако ручной анализ таких речей трудоемок и подвержен субъективности.

Данная работа направлена на разработку автоматизированного подхода к анализу:

* **эмоциональной тональности** (эмотивность, модальность, оценочность),
* **тематической структуры** (ключевые слова, тематические кластеры),
* **эволюции риторики** (динамика изменений во времени).

В основе исследования лежит корпус речей немецких политиков за всё время существования Бундестага, что позволяет проследить исторические тенденции в политическом дискурсе.

**2. Цели и задачи**

**Цель:**  
Разработка модели анализа политических речей, сочетающей методы компьютерной лингвистики и машинного обучения для выявления эмоциональных паттернов и тематической динамики.

**Задачи:**

1. Сбор и предобработка текстовых данных (корпус речей немецких политиков).
2. Применение методов **дистрибутивной семантики** (Word2Vec, BERT, GermanBERT) для векторного представления текста.
3. Классификация **эмотивной лексики** (анализ модальности и эмоциональной окраски).
4. Тематическое моделирование (**LDA, BERTopic**) для выявления ключевых тем.
5. Сравнительный анализ речей разных политиков и партий (**кластеризация, снижение размерности**).
6. Визуализация результатов (**графики, облака слов, временные тренды**).

**3. Методология**

В работе будут использованы следующие методы:

**3.1. Машинное обучение**

* **Регрессия и классификация** (оценка тональности, предсказание эмоционального окраса).
* **Валидация моделей** (кросс-валидация, метрики точности).
* **Гиперпараметрическая оптимизация** (GridSearch, Bayesian Optimization).
* **Кластеризация** (K-means, DBSCAN для группировки схожих речей).
* **Ансамбли и градиентный бустинг** (XGBoost, CatBoost для улучшения качества классификации).
* **Снижение размерности** (PCA, t-SNE для визуализации).

**3.2. Компьютерная лингвистика**

* **Дистрибутивная семантика** (Word2Vec, FastText, BERT-эмбеддинги).
* **Анализ эмотивности** (лексико-семантические маркеры, модальные конструкции).
* **Тематическое моделирование** (LDA, NMF, BERTopic).
* **Обработка естественного языка (NLP)** (SpaCy, Hugging Face Transformers).

**3.3. Нейросетевые архитектуры**

* **Fine-tuning предобученных языковых моделей** (GermanBERT, GPT-3 для генерации аннотаций).
* **Архитектуры для классификации текста** (CNN, LSTM, Transformer-based модели).

**4. Ожидаемые результаты**

1. **Модель классификации эмоциональной тональности** политических речей.
2. **Тематические кластеры**, отражающие ключевые дискурсы немецких партий.
3. **Визуализация динамики изменений** в риторике (например, сдвиг в сторону популизма или усиление эмоциональных обращений).
4. **Сравнительный анализ** речей разных политиков (например, Ангелы Меркель vs. Олафа Шольца).

**5. Практическая значимость**

Разработанный инструмент может быть полезен:

* **Политологам и социологам** – для анализа эволюции политического дискурса.
* **Журналистам** – для быстрой оценки тональности выступлений.
* **Гражданским активистам** – для выявления манипулятивных техник в речах.

**6. План работы**

1. **Обзор литературы** (анализ существующих подходов к NLP в политических исследованиях).
2. **Сбор и предобработка данных** (парсинг, лемматизация, удаление шума).
3. **Эксперименты с моделями** (сравнение методов классификации и кластеризации).
4. **Интерпретация результатов** (анализ ошибок, выявление значимых паттернов).
5. **Написание и оформление работы**.

**7. Заключение**

Данное исследование сочетает методы компьютерной лингвистики и машинного обучения для анализа политических речей, что позволяет автоматизировать процесс выявления эмоциональных и тематических тенденций. Результаты могут быть использованы в академических и прикладных исследованиях политического дискурса.