Final Report

**Название:** Text Completion in Kyrgyz using mGPT

**Участники:** Азимов Давлет, Эмильбеков Темирлан, Эрмеков Кутманшер, Кондубаев Аскар, Мурхалидинов Элдар

### **1. Введение**

В рамках курса по машинному обучению наша команда работала над задачей автоматического дополнения текстов на кыргызском языке. Это одна из ключевых задач обработки естественного языка (NLP), позволяющая использовать модели автодополнения в различных сферах: генерация новостей, чат-боты, интеллектуальные системы помощи, редакторы и другие.

Мы использовали **предобученную трансформерную модель**, способную генерировать текст на кыргызском языке. Проект направлен на демонстрацию возможностей таких моделей при работе с языками с низким уровнем цифровой представленности, как киргизский.

### **2. Цель проекта**

Разработка системы автодополнения текстов на кыргызском языке с использованием модели ai-forever/mGPT-1.3B-kirgiz без дополнительной дообучки.

### **3. Датасет**

Наш датасет построен на базе **постов из кыргызских новостных источников и медийных платформ**. Он включает разнообразные тексты: новости, общественно-политические события, прогнозы погоды, интервью и прочее. Это позволило обеспечить тематическое и лексическое разнообразие обучающего материала.

Каждый текст в датасете разбивается на:

* **Prompt (ввод):** первые 8 слов
* **Completion (продолжение):** оставшаяся часть текста

Такой подход позволяет обучить модель предугадывать логическое продолжение текста.

**Работа над датасетом** выполнена: **Эрмеков Кутманшер и Эмильбеков Темирлан**

### **4. Используемая модель**

Мы использовали модель ai-forever/mGPT-1.3B-kirgiz — мультилингвальную модель автодополнения, обученную в том числе на киргизском языке. Эта модель построена на архитектуре GPT и предназначена для генерации текста на естественных языках.

Модель была загружена из библиотеки **Hugging Face Transformers** и применена без дополнительной дообучки (zero-shot режим).

**Работа с моделью и код** выполнена: **Азимов Давлет**

### **5. Предобработка данных**

Обработка текста включала:

* Загрузка текста из all\_texts.txt
* Токенизация с помощью HuggingFace Tokenizer
* Разделение строк на prompt и completion
* Конвертация в формат datasets.Dataset для удобства генерации

### **6. Генерация текста**

Модель получает на вход начало текста (prompt) и генерирует логичное продолжение. Пример:

**Prompt:** *"Бүгүнкү аба ырайы"* **Completion:** *"...жылы болот деп күтүлүүдө. Батыштан шамал согуп, жаан-чачын болушу мүмкүн."*

Такая генерация применима в контексте автоматической генерации новостей, интерактивных интерфейсов и контент-автоматизации.

### **7. Результаты**

Несмотря на отсутствие дообучки, модель показала **удовлетворительные результаты генерации**:

* Синтаксически корректные тексты
* Тематическая связность
* Актуальное лексическое наполнение

**Ограничения:**

* Иногда появляются бессмысленные или повторяющиеся фрагменты
* Нет встроенной оценки качества (нужно внедрить метрики, например, Perplexity или BLEU)

### **8. Выводы**

Проект продемонстрировал, что трансформерные модели, даже без дообучки, могут **дополнять тексты на кыргызском языке** на приемлемом уровне. Вводит хорошо текст, грамматически верно и в формальном стиле, отталкиваясь от датаста