Финансирано от Европейския съюз
СледващоПоколениеЕС

План за възстановяване и устойчивост

Република България

# ВЪТРЕШЕН ДОКЛАД

от гл. ас. д-р Ивелина Стоянова
до проф. д-р Светла Коева, ръководител на проекта,

за текуща самооценка на техническата готовност на **продукта**

## ГОЛЯМ НАБОР ОТ ЕЗИКОВИ ДАННИ IfGPT,

планиран резултат от **Работен пакет 4. Колекциониране и предварителна обработка на данни за фина настройка на големи езикови модели**
по проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“, договор No ПВУ – 55 от 12.12.2024 г. /BG-RRP-2.017-0030-C01/
от Механизма за възстановяване и устойчивост за изпълнение на инвестиция по C2I2 „Повишаване на иновационния капацитет на Българската академия на науките (БАН) в сферата на зелените и цифровите технологии“ от Плана за възстановяване и устойчивост

**Вътрешен доклад ID: IFGPT-2025-REPORT-003**

Име и версия на продукта: **IfGPT Dataset version 0.2**
Тип на продукта: **Набор от езикови данни**
Текущо ниво на техническа готовност: **TRL 4**
Дата: **15.07.2025 г.**

**Извършени дейности**

Извършени са следните дейности по компилиране на голям набор от данни IfGPT за български език:

**1. Изработен е общ модел на метаданните на IfGPT:**

**Identifier** – уникален идентификатор на документа във всички колекции, създаден с езиков код **bg** като префикс;

**Licence** – условията за използване, например лиценз CC BY-SA 4.0;

**PublicationDate** – датата на оригиналната публикация на документа (ако е налична) в ISO 8601 формат;

**DocumentTitle** – четимо за човек заглавие на документа;

**Source** – името на организацията, която е публикувала изходния документ, т.е. списание, издател, блог, уебсайт и т.н.;

**Medium** – дали документът е текст, звуков файл, изображение или видео;

**Url** – оригиналният интернет адрес, от който е извлечен документът, ако е приложимо;

**Domain** – класификация на документите по предварително зададени (24) специфични тематични области; всеки текст може да се класифицира в до шест домейна;

**Keywords** – извлечени ключови думи, които са специфични за документа; могат да бъдат изброени до шест ключови думи;

**NumberWords** – общият брой думи в документа;

**NumberSentences** – общият брой изречения в документа;

**NumberTokens** – общият брой токъни в документа;

**PersonallyIdentifiableInformation** – масив от стойности, показващи изреченията, които съдържат чувствителна информация, и процента на токъните, съдържащи такава информация, спрямо общия брой токъни в тези изречения;

**BiasedInformation** – масив от стойности, показващи изреченията, които съдържат потенциална информация, показваща пристрастия, и процента на токъните, съдържащи такава информация, спрямо общия брой токъни в тези изречения;

Следните метаданни са незадължителни за документите:

**Author** – име на автора на текста в изходния документ;

**Style** – литературният стил на текста в документа, избран от предварително дефиниран списък: Художествен, Административен, Правен, Журналистически и т.н.;

**Type** – уточнява типа на изходния документ (напр. книга, глава, есе, вестникарска статия, блог публикация и т.н.);

**Subdomain** – допълнителна класификация на документите в по-тесни категории, напр. научни домейни за областта на науката или културни домейни за областта на културата; поддомейнът е свързан с конкретен домейн;

**TranslatedDocument** – дали документът е оригинално създаден на български или е бил преведен;

**CollectionDate** – датата на събиране на документа в ISO 8601 формат;

**LicenseLink** – връзката към лицензa на уебсайта на източника, ако е налична;

**NumberParagraph** – общият брой абзаци в документа;

**TaskCategories** – приложенията (избрани от предварително дефиниран списък), за които данните в документа са разработени или са подходящи, напр. за въпроси и отговори.

**2. Изработена е Neo4j графова база данни за оптимално представяне на метаданните и взаимовръзките между тях.**

**3. Създаден е първоначален (лабораторен) модел на продукта, в който са интегрирани основните компоненти** – текстови данни и метаданни. Метаданните са организирани в графова база данни Neo4j. В набора от данни са включени данни от съществуващи корпуси и масиви от данни за български език, чиито формат на текстовете и на метаданните са прехвърлени в единния формат на IfGPT; текущата версия на IfGPT съдържа 390 хил. документа и 227 млн. токъна. **Следните налични корпуси за български език са приведени във формата на общия набор от текстови данни IfGPT:**

- MARCELL (25 хил. текста и 45 млн. токъна),
- CURLICAT (113 хил. текста и 35 млн. токъна),
- Административен корпус на БНК (17 хил. текста и 79 млн. токъна),
- Уикипедия статии от БНК (89 хил. текста и 41 млн. токъна),
- Субтитри от БНК (146 хил. текста и 27 млн. токъна).

**4. Проведени са автоматични тестове за валидиране на наличните данни и са приложени процедури за оценка и подобряване на качеството на данните** с цел тяхното ефективно приложение за фина настройка на големи езикови модели:

а) **Изчистване на текстовете от стандартни компоненти**, свързани със структурата, оформлението и навигацията в онлайн страници.

б) **Дедупликиране** – отстраняване на повторенията (както точни съвпадения, така и еднотипни и подобни текстове); предстои и разработването на процедури за откриване на смислово подобни текстове, които също натоварвам модела, без да допринасят за ефективната фина настройка. За целта предварително са проучени редица съществуващи подходи за прилагането на тези процедури с фокус върху езиково независимите методи, а някои от тях са адаптирани и приложени за български език – например, MinHash и Locality Sensitive Hashing за дедупликиране.

в) **Идентифициране на лични данни и/или чувствителна информация** в текстовите единици; все още тези процедури се разработват и усъвършенстват. Понастоящем за всяко изречение се дава стойност за процентното съдържание (като брой токъни) на такава информация в цялото изречение.

г) **Допълнителни процедури за установяване и неутрализиране на езикови и културни пристрастия в данните (bias)**. Предстои усъвършенстване на тези процедури.

Понастоящем за всяко изречение се дава стойност за процентното съдържание (като брой токъни) на такава информация в цялото изречение.

**5. Снабдяване на текстовите единици с възможно най-подробни метаданни** относно произхода и съдържанието им, което е съществено за следващите етапи на работа при разработването на методи за фина настройка на модели за специфични тематични области. От особено значение са:

- Тематична област и допълнителни тематични области (до 6 области от предварително дефинирани категории) – по този начин текстовете са класифицирани и лесно могат да се извличат подкорпуси от данни за специфични цели.
- Ключови думи – които предоставят допълнителни възможности за класификация.
- Възможни приложения на данните – все още предстои за голяма част от наличните данни да бъдат определени възможните им приложения, например за въпроси и отговори, за анализ на мнение и на лично отношение (opinion mining, sentiment analysis) и др.

**6. Проведени са функционални тестове в контролирана среда върху различни части от данните за извличане на подмножества от данните въз основа на описанията на метаданните.** Тестовете потвърждават валидността на концепцията за организация на данните и работоспособността на интегрираната система. Демо: https://ifgpt.dcl.bas.bg/ifgpt-dataset/

**7. Идентифицирани са технологични ограничения и критични точки**:

- При увеличаване на броя текстови единици (над 100,000 текстови единици) търсенето и извличането на подмножества от данните в реално време става бавно и ненадеждно. Предвиждат се подобрения в подходите за достъп до базата данни.
- Направено е двуфазово извличане на масиви от данни – 1) извличане на метаданните (JSON или CSV формат), а след това 2) извличане на изреченията или целите текстови единици. Извличането на отделни изречения е по-добър подход (тъй като може да се зададат и параметри за дължина на изреченията или извличане на изречения по дадени лексикални и други параметри, критерии за наличие/неналичие на чувствителна информация или пристрастия), но предлага предизвикателства, свързани със скоростта, тъй като изисква обхождане на повече данни.

**8. Към момента данните са налични в следните формати**:

- Описанието на метаданните на текущата версия на IfGPT са публично достъпни: https://ifgpt.dcl.bas.bg/wp-content/uploads/2025/10/metadata-2025-09-30.json
- Данните са налични в две хранилища:
  https://huggingface.co/datasets/DCL-IBL/IfGPT-Dataset
  https://github.com/DCL-IBL/IfGPT-Dataset

- За демонстриране на концепцията и възможностите за използване на метаданните при извличането на подмножества се предоставя онлайн демо за търсене в метаданните, което дава като резултат описание на селектираните текстови единици (JSON формат): https://ifgpt.dcl.bas.bg/ifgpt-dataset/.

**Получени резултати**

Получени са следните резултати:

1. Разработен е **цялостен концептуален модел на продукта** – организация на данните и метаданните в графова база данни, възможностите за достъп и използване на базата от метаданни и др.

2. Дефинирани са ключовите технически параметри на продукта – формат на данните, формат и обхват на метаданните.

3. **Първоначален (лабораторен) и напълно функционален модел на продукта, в който са интегрирани основните компоненти** – текстови данни и метаданни. Метаданните са организирани в графова база данни Neo4j. Данните са изчистени, дедупликирани и са преминали оценка за наличие на лични и/или чувствителни данни и пристрастия.

4. **Проведени функционални тестове в контролирана среда върху различни части от данните за извличане на подмножества от данните въз основа на описанията на метаданните.** Отчетени са проблемните области и са планирани стратегии за преодоляването на предизвикателствата.

5. Публично предоставяне и описание на данните:
https://huggingface.co/datasets/DCL-IBL/IfGPT-Dataset
https://github.com/DCL-IBL/IfGPT-Dataset

**Доказателства и аргументиране на ниво на техническа готовност TRL 4**

Конкретните доказателства за постигнато ниво TRL 4 включват:

1. Настоящия **вътрешен доклад** за извършените дейности и получените до момента резултати по компилирането на големия набор от данни IfGPT.

2. Статия: Koeva, S., I. Stoyanova, J. Kralev. *Advancing NLP for Low-Resource Languages (LowResNLP 2025)* в рамките на международната конференция *Recent Advances in Natural Language Processing (RANLP 2025)* (in print).

Предал: Ивелина Стоянова
Дата: 15.07.2025

Приел: Светла Коева