An Open Multilingual System for Scoring Readability of Wikipedia

Mykola Trokhymovych
mykola.trokhymovych@upf.edu

Indira Sen
indira.sen@uni-konstanz.de

Martin Gerlach
mgerlach@wikimedia.org

Про мене

Mykola Trokhymovych

trokhymovych.com | **mykola.trokhymovych@upf.edu**

- PhD-кандидат @ Universitat Pompeu Fabra
  - Працюю під науковим керівництвом Diego Saez-Trumper та Ricardo Baeza-Yates
- 2025 Google PhD Fellow
- Запрошений дослідник, Indiana University (OSoMe)
  - Досліджую виявлення контенту, згенерованого ШІ в соціальних мережах, разом із Fil Menczer
- Досвід дослідницької роботи з екосистемою Wikimedia більше 6 років.

Конференції:

- **ACL’25, ICWSM’25, ACL’24, KDD’23, CIKM’21**

Раніше:

🎓 Науки про дані @ УКУ; Системний аналіз @ ІПСА, КПІ

💼 Data Scientist / ML Engineer: Ciklum, Jooble, Surprise

Читабельність & Вікіпедія

Що таке читабельність?

Концепція читабельності має на меті визначити, наскільки легко читати певний текст. Зазвичай її визначають як сукупність усіх факторів, що впливають на розуміння читачем, швидкість читання та рівень зацікавленості.

(Dale and Chall, 1949)

Читабельність Вікіпедії

«[В англійській Вікіпедії] загальна читабельність погана».

(Lucassen et al., 2012)

«Більшість медичної інформації [в англійській Вікіпедії] залишається написаною на рівні, що перевищує здатність читати середньостатистичних дорослих».

(Brezar and Heilman, 2019)

Читабельність або як ВМС США навчили світ писати простіше

- Проблема:
  - Військові інструкції 1970-х були занадто складними.
- Рішення:
  - Формула, що рахує кількість слів у реченнях та складів у словах.
- Результат Flesch–Kincaid Grade Level (FKGL):
  - Відповідає класу в американській школі.
- Золотий стандарт:
  - FKGL 8-9 (зрозуміло широкій аудиторії).

Джерело: Wikipedia Commons

Вимірювання читабельності це складно

Поширені підходи:
формули для оцінки читабельності, наприклад, формула оцінки легкості читання Флеша (FRE)

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Вікіпедія існує більш ніж 300 мовами:

- Адаптація формул читабельності до кожної мови не масштабована/неможлива
- Відсутність багатомовних моделей/інструментів
- Нестача багатомовних даних для навчання і оцінки підходів
- Стандартні набори даних не є відкритими (наприклад, Newsela)
    - Переважно набори даних призначені для англійської мови (немає даних для більшості мов)

У цьому дослідженні ми прагнемо...

створити багатомовну модель, яка може автоматично оцінити читабельність тексту статей Вікіпедії різними мовами.

Новий багатомовний набір даних

WikiReaD (Wikipedia Readability Dataset)

З чого складається? — Пари енциклопедичних статей у двох рівнях читабельності (простий, складний) для 14 мов.

Яким чином створений? — Зіставлення статей Вікіпедії зі спрощеною версією.

- Simple English Wikipedia (>100K пар статтей)
- Дитячі енциклопедії (10-10K)
  - Vikidia (11 мов, серед них грецька, іспанська, французька, португальська, вірменська тощо.)
  - Klexikon (німецька)
  - Txikipedia (баскська)
  - Wikikids (нідерландська)

Новий багатомовний набір даних

WikiReaD (Wikipedia Readability Dataset)

Dataset	#Pairs	Avg. #Sen.	Avg. #Char.
simplewiki-en	112,342	6.2/7.9	84.6/130.9
vikidia-en	1,991	6.4/14.3	83.3/142.8
vikidia-ca	234	5.2/9.7	79.3/145.2
vikidia-de	260	6.4/11.2	75.8/131.0
vikidia-el	39	6.0/11.8	96.8/134.9
vikidia-es	1,915	5.7/7.7	109.0/179.4
vikidia-eu	571	6.5/8.7	114.6/129.5
vikidia-fr	12,221	5.7/7.3	106.9/152.1
vikidia-hy	485	14.3/11.4	105.3/115.1
vikidia-it	1,662	4.5/6.0	84.6/152.6
vikidia-oc	12	4.2/7.1	77.0/105.6
vikidia-pt	809	5.7/11.8	97.3/157.9
vikidia-ru	125	5.8/11.2	83.8/110.6
vikidia-scn	10	3.8/4.7	50.9/86.3
klexikon-de	2,255	17.7/8.9	73.9/136.9
txikipedia-eu	1,162	7.3/8.4	107.4/126.4
wikikids-nl	12,090	8.0/7.5	83.7/112.0

Приклад

Простий (Simple Wikipedia)

Заголовок: Прапор Таїланду

Речення:
The Flag of Thailand has five stripes red, white, and blue.

Red-white-blue stand for nation-religion-king, an unofficial motto of Thailand.

The flag was made official on 28 September 1917 by a royal decree.

The Thai name for the flag is ธงไตรรงค์ ("Thong Trairong"), which means "tricolour flag"

Складний (English Wikipedia)

Заголовок: Прапор Таїланду

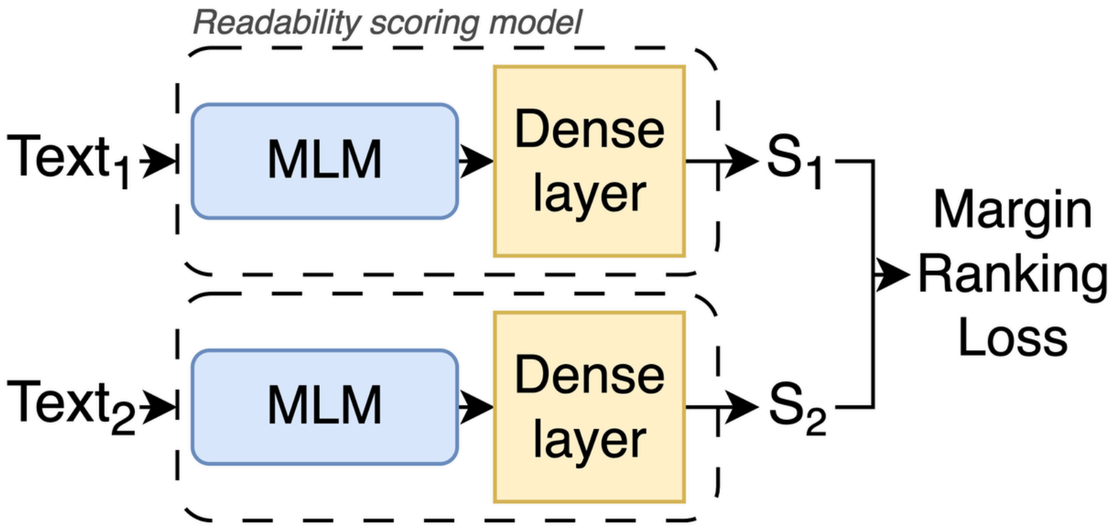
Речення:
The flag of Thailand (Thai: ธงไตรรงค์; thong trai rong, meaning 'tricolour flag') shows five horizontal stripes in the colours red, white, blue, white and red, with the central blue stripe being twice as wide as each of the other four.

The design was adopted on 28 September 1917, according to the royal decree issued by Rama VI.

Since 2016, that day is a national day of importance in Thailand celebrating the flag.

. . . .

Нова модель оцінювання читабельності



- xlm-roberta-longformer як базова модель
- Підтримка близько 100 мов
- Архітектура Siemease для навчання:
  - Навчання базується на попарному порівнянні, але все ще може оцінювати окремі тексти
- Loss: Margin Ranking Loss
  - Ми вчимо модель порівнювати
- Fine-tuning лише з використанням даних англійською мовою:
  - Без додаткового навчання на інших мовах! (zero-shot)

Точність моделі

Метрика оцінювання: Точність ранжування (RA)
Пояснення: RA це відсоток пар, які правильно ранжовані

Setup 1: Оцінювання точності моделі для англійської мови.

→
- Simple Wikipedia (тренувальні дані): RA=0.976
- Vikidia (out-of-corpus): RA=0.991

Setup 2: **Оцінювання для неанглійських текстів** (zero-shot cross-lingual transfer)

→
- RA > 0.8 для всіх мов наявних в даних
- RA > 0.9 для 10/15 датасетів

Setup 3: **Порівняння з попередніми роботами** (e.g. Lee&Vajjala, ACL 2022)

→
Перевершує референсні бенчмарки
- VikidiaFr: 0.978 (наш) vs 0.811 (NPRM)
- OneStopEnglish: 0.974 (наш) vs 0.878 (NPRM)

Точність моделі

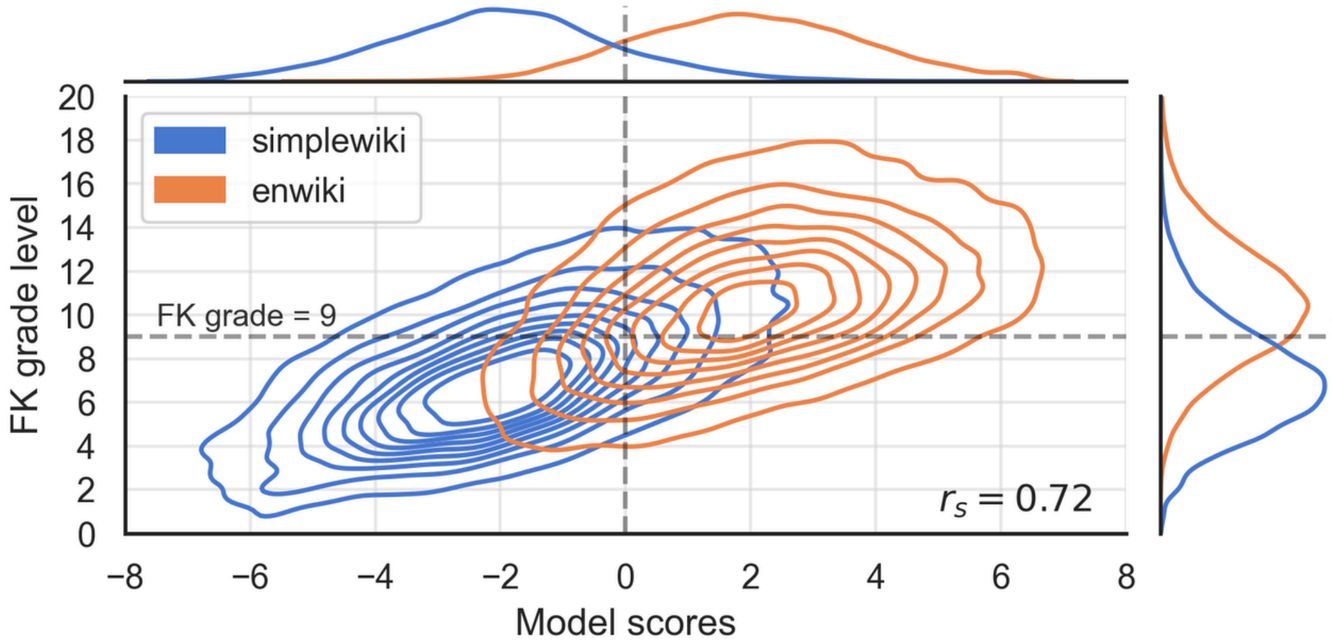
WikiReaD (Wikipedia Readability Dataset)

Dataset	NS	±CI	FRE	±CI	LFC	±CI	LFR	±CI	LMC	±CI	TRank	±CI	SRank	±CI
simplewiki-en	0.543	0.007	0.868	0.005	0.937	0.003	0.945	0.003	0.965	0.002	**0.976**	0.002	0.972	0.002
vikidia-en	0.814	0.017	0.935	0.011	0.979	0.006	0.981	0.006	0.979	0.006	**0.991**	0.004	0.985	0.005
vikidia-ca	0.782	0.054	0.906	0.038	0.94	0.031	0.932	0.033	0.936	0.032	**0.962**	0.025	0.936	0.032
vikidia-de	0.735	0.054	0.815	0.048	0.888	0.039	0.869	0.042	0.908	0.036	**0.938**	0.03	0.919	0.034
vikidia-el	0.718	0.144	0.718	0.144	0.744	0.14	0.795	0.129	0.897	0.096	**0.923**	0.086	0.897	0.097
vikidia-es	0.573	0.023	0.842	0.017	0.883	0.015	0.892	0.014	0.879	0.015	**0.911**	0.013	0.909	0.013
vikidia-eu	0.541	0.042	0.673	0.04	0.639	0.04	0.623	0.041	0.63	0.04	**0.818**	0.032	0.736	0.037
vikidia-fr	0.553	0.009	0.84	0.007	0.82	0.007	0.845	0.006	0.849	0.007	**0.923**	0.005	0.918	0.005
vikidia-hy	0.394	0.045	0.594	0.045	0.534	0.045	0.598	0.044	0.637	0.044	**0.802**	0.036	0.761	0.039
vikidia-it	0.569	0.024	0.83	0.019	0.919	0.013	0.94	0.012	0.925	0.013	**0.958**	0.01	0.952	0.01
vikidia-oc	0.667	0.273	0.667	0.271	0.75	0.25	0.667	0.27	0.917	0.159	**1.0**	0.0	0.917	0.161
vikidia-pt	0.761	0.03	0.869	0.024	0.938	0.017	0.934	0.017	0.921	0.019	**0.960**	0.014	0.938	0.017
vikidia-ru	0.728	0.08	0.608	0.087	0.736	0.078	0.776	0.074	0.736	0.079	**0.880**	0.058	0.760	0.077
vikidia-scn	0.4	0.314	0.6	0.309	0.6	0.308	0.8	0.254	0.6	0.31	0.9	0.191	**1.0**	0.0
klexikon-de	0.114	0.013	0.984	0.005	0.999	0.002	0.995	0.003	0.991	0.004	**0.999**	0.002	0.996	0.003
txikipedia-eu	0.512	0.029	0.707	0.027	0.689	0.027	0.698	0.027	0.67	0.027	**0.81**	0.023	0.762	0.025
wikikids-nl	0.427	0.009	0.795	0.007	0.831	0.007	0.834	0.007	0.85	0.007	0.897	0.006	**0.907**	0.005

Порівняння з попередніми роботами

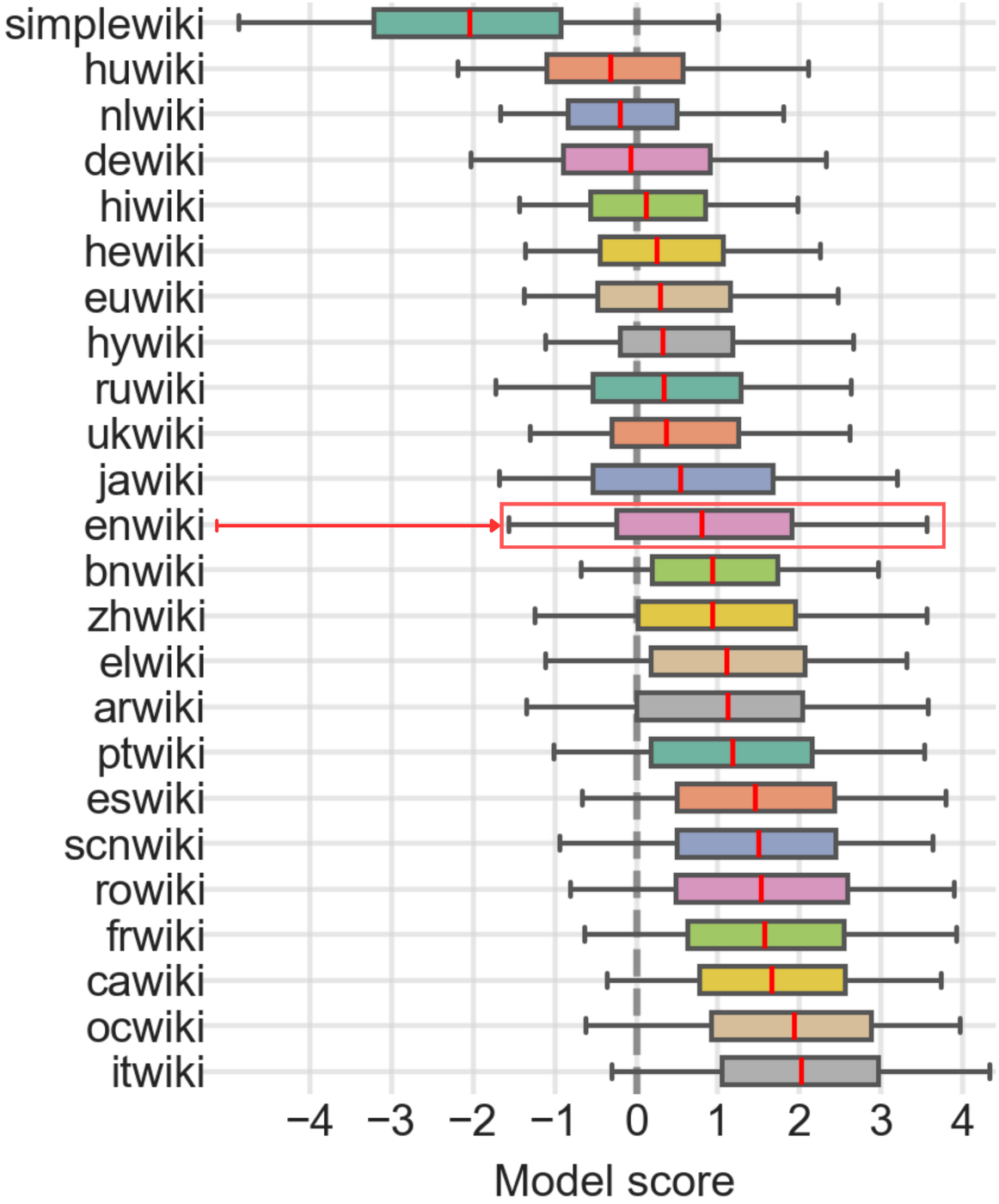
Dataset	NS	±CI	FRE	±CI	LFC	±CI	LFR	±CI	LMC	±CI	TRank	±CI	$NPRM$	±CI
VikidiaEn	0.966	0.005	0.948	0.006	0.888	0.008	0.946	0.006	0.965	0.005	**0.984**	0.003	0.975	0.004
VikidiaFr	0.952	0.005	0.899	0.008	0.878	0.008	0.888	0.008	0.75	0.011	**0.978**	0.004	0.811	0.010
OSE	0.794	0.059	0.915	0.04	0.889	0.046	0.873	0.048	0.942	0.034	**0.974**	0.023	0.878	0.048

Оцінки моделі корелюють з існуючими формулами читабельності



Оцінка моделі близька до 0 розділяє тексти від simplewiki (простих) до enwiki (складних).

Це відповідає рівню за шкалою Флеша-Кінкейда (FKGL) приблизно 9.



Вікіпедію важко читати більшістю мов

В англомовній Вікіпедії понад половина статей перевищує середній рівень читабельності для дорослих американців (~8-9 клас) (Brezar&Heilman 2019)

Загальна читабельність у більшості Вікіпедій подібна до англійської Вікіпедії

Висновки

- Випустили **новий відкритий багатомовний набір даних**, що охоплює **14 мов**
- Розробили **багатомовну модель** для оцінки читабельності тексту
- Представили перший систематичний огляд **стану читабельності статей Вікіпедії** поза межами англійської мови
- Надали відкритий **API endpoint** моделі для використання читачами, редакторами та дослідниками

WikiReaD dataset

API endpoint

Дякую!