Universitat Pompeu Fabra Barcelona

Як ШІ може допомагати боротися з вандалізмом у Вікіпедії та Вікіданих

Микола Трохимович
mykola.trokhymovych@upf.edu

Kyiv, 15 November, 2025

Не кожна зміна покращує Вікіпедію – деякі її спотворюють.

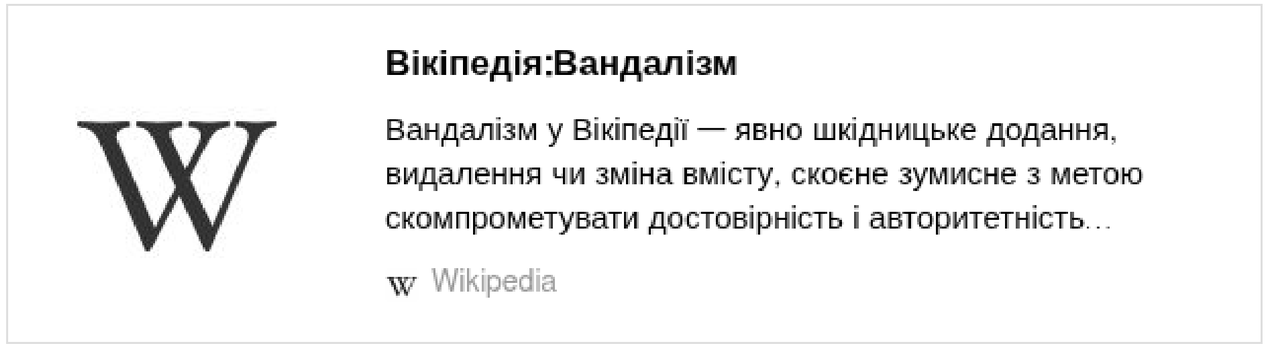
Example of edit (a.k.a. revision) reverting a bad-faith one (revision_id = 1149625753).

```
Lvov emerged as the centre of the historical regions of [[Red Ruthenia]] and [[Galicia (Eastern Europe)|Galicia]] in the [[14th
```

```
Lviv emerged as the centre of the historical regions of [[Red Ruthenia]] and [[Galicia (Eastern Europe)|Galicia]] in the [[14th
```

Хоча існували моделі, що допомагали патрульним (наприклад, ORES), залишалися відкриті проблеми, такі як:

- Продуктивність моделей;
- Покриття мов;
- Неупередженість.



AUGUST 6 - 10
KDD2023
LONG BEACH, CA

RESEARCH-ARTICLE

Fair Multilingual Vandalism Detection System for Wikipedia

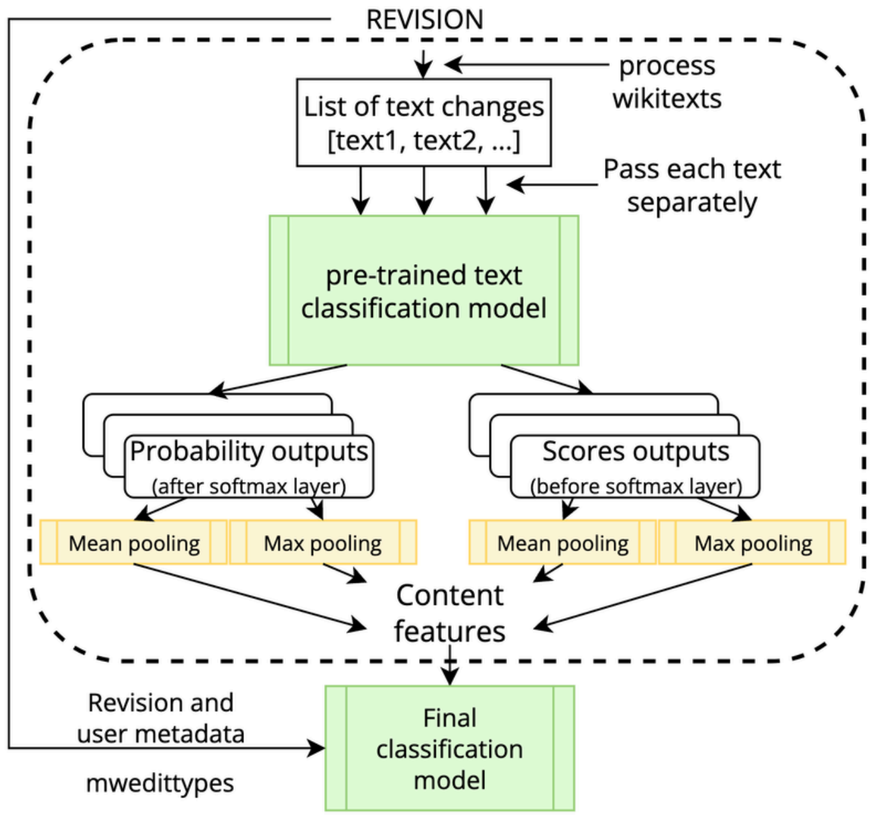
Authors: Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, Diego Saez-Trumper

Authors Info & Claims

KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining • Pages 4981 - 4990
https://doi.org/10.1145/3580305.3599823

Published: 04 August 2023 Publication History

Fair Multilingual Vandalism Detection System for Wikipedia



Authors: Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, Diego Saez-Trumper
KDD'23 Industry Track

Мета:

Створити модель, яка допомагає редакторам визначати правки, що потребують патрулювання.

Підхід:

- Використовувати історичні дані про скасування змін для навчання моделей машинного навчання.

Результат:

- Відкрита, багатомовна модель для патрулювання контенту на Вікіпедії;
- Перевершує попередні SOTA-моделі за продуктивністю та неупередженістю;
- Впроваджена і працює для 47 мов.

Wikimedia API Portal

Анонімні редактори мають вищий відсоток скасованих змін.

anon. *28%* VS. all *8%*

Моделі перенавчаються і починають дискримінувати анонімних користувачів.

Disparate Impact Ratio (DIR)

$$DIR = \frac{Pr(\hat{Y}=1|D=unprivilaged)}{Pr(\hat{Y}=1|D=privilaged)}$$ ***= 20***

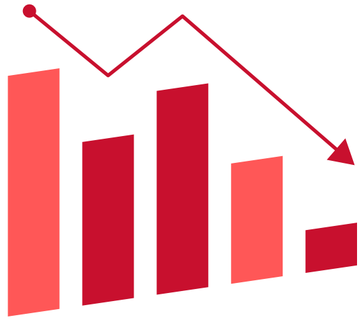
***for ORES**

Pr - probability

$\hat{Y}$ - predicted value,

D - a group of users (anon. or registered)

Анонімні редактори рідше залишаються та продовжують робити внесок.



Ризик зменшення кількості активних редакторів.

Вандалізм поза Вікіпедією

Malicious changes to Wikidata:

Example of a revision (ID: 593195479) vandalizing the Wikidata entry for Bulgaria. Original triple IDs are mapped to their corresponding English labels.

before	Q219 (Bulgaria)	P85 (anthem)	Q182115 (Mila Rodino)
after	Q219 (Bulgaria)	P85 (anthem)	Q28572509 (Despacito)

Який гімн Болгарії? 🇧🇬

TL ; DR

Why does Siri think the national anthem of Bulgaria is 'Despacito?' / A cursory investigation with no clear answer in sight

by **Nick Statt**

Oct 5, 2017, 1:15 AM GMT+2

theverge

Graph-Linguistic Fusion: Using Language Models for Wikidata Vandalism Detection

Mykola Trokhymovych, Lydia Pintscher, Ricardo Baeza-Yates, Diego Sáez Trumper

Abstract

We introduce a next-generation vandalism detection system for Wikidata, one of the largest open-source structured knowledge bases on the Web. Wikidata is highly complex: its items incorporate an ever-expanding universe of factual triples and multilingual texts. While edits can alter both structured and textual content, our approach converts all edits into a single space using a method we call Graph2Text. This allows for evaluating all content changes for potential vandalism using a single multilingual language model. This unified approach improves coverage and simplifies maintenance. Experiments demonstrate that our solution outperforms the current production system. Additionally, we are releasing the code under an open license along with a large dataset of various human-generated knowledge alterations, enabling further research.

PDF

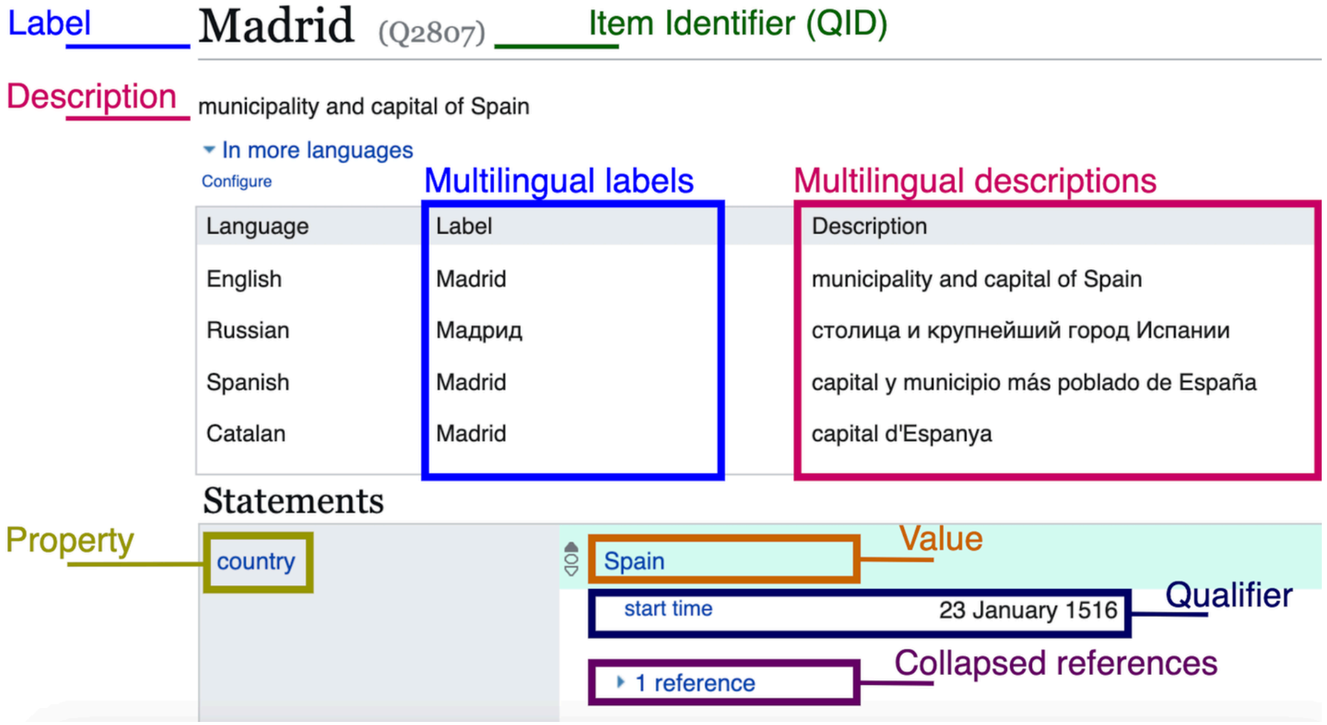
Cite

Search

Fix data

Anthology ID: 2025.acl-industry.21
Volume: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)
Month: July
Year: 2025
Address: Vienna, Austria
Editors: Georg Rehm, Yunyao Li
Venue: ACL

Що ми знаємо про Wikidata?



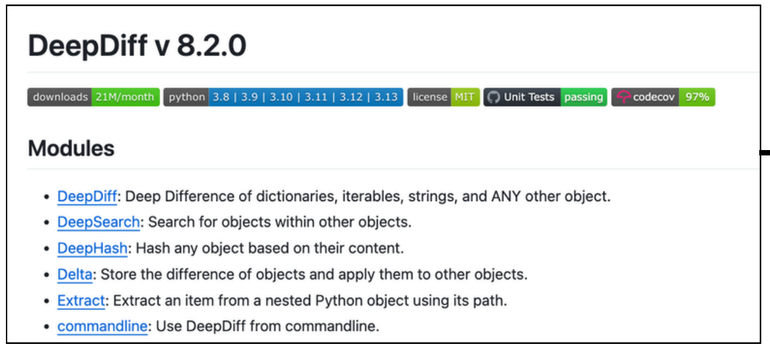
- Структуроване сховище знань
  - ~100 млн+ триплетів у Wikidata
  - ~8 правок на секунду
- Різноманіття типів даних:
  - Тексти більш ніж 300 мовами
  - Мітки часу
  - Числові значення
  - Географічні координати
  - І багато іншого...
- Живить численні сервіси в Інтернеті

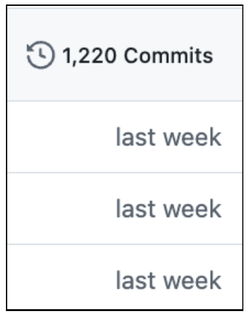
Challenges: 

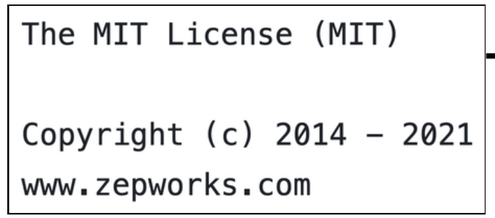
- **Анонімні та нові користувачі:** Анонімні та нові редактори зазвичай мають вищий відсоток скасованих змін.
- **Впровадження:** Модель повинна працювати при обмежених обчислювальних ресурсах.
- **Складність контенту:** Представлення Wikitext для елемента Wikidata є складною вкладеною структурою словників або списків.

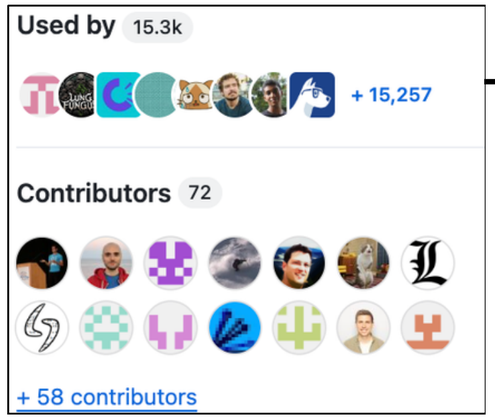
{"type":"item","id":"Q105644691","labels":{"es":{"language":
n":{"language":"en","value":"Eva Mora"},"ca":{"language":"ca
iptions":{"es":{"language":"es","value":"periodista y report
nd reporter"},"ca":{"language":"ca","value":"periodista i re
n verslaggever"},"gl":{"language":"gl","value":"periodista e
{"snaktype":"value","property":"P31","hash":"ad7d38a03cdd40c
m","numeric-id":5,"id":"Q5"},"type":"wikibase-entityid"}},"t
9","rank":"normal"}],"P21":[{"mainsnak":{"snaktype":"value",
avalue":{"value":{"entity-type":"item","numeric-id":6581072,
d":"Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b","rank":
ash":"943348a1c561d670ec8f5fab1132465a7c779b55","datavalue":
ype":"wikibase-entityid"}},"type":"statement","id":"Q1056446
{"snaktype":"value","property":"P106","hash":"54bde886150c59
m","numeric-id":42909,"id":"Q42909"},"type":"wikibase-entity
6b0925a","rank":"normal"}],"P569":[{"mainsnak":{"snaktype":"
ff","datavalue":{"value":{"time":"+1974-10-21T00:00:00Z","ti
\/www.wikidata.org\/entity\/Q1985727"},"type":"time"}},"type
ank":"normal"}],"P19":[{"mainsnak":{"snaktype":"value","prop
e":{"value":{"entity-type":"item","numeric-id":2807,"id":"Q2
31":[{"snaktype":"value","property":"P131","hash":"936ab6c53
tem","numeric-id":5756,"id":"Q5756"},"type":"wikibase-entity
-6668-9996fe8c5105","rank":"normal"}],"P69":[{"mainsnak":{"s
01e3369306e","datavalue":{"value":{"entity-type":"item","num
e":"statement","qualifiers":{"P512":[{"snaktype":"value","pr
alue":{"value":{"entity-type":"item","numeric-id":43236300,"
ue","property":"P580","hash":"f77e401c3294559f5f96c8fc18f16d
one":0,"before":0,"after":0,"precision":9,"calendarmodel":"h
[{"snaktype":"value","property":"P582","hash":"f84ab79307da7
00:00:00Z","timezone":0,"before":0,"after":0,"precision":9,"
ime"}}]},"qualifiers-order":["P512","P580","P582"],"id":"Q10
ak":{"snaktype":"value","property":"P69","hash":"5d67c53bf40
m","numeric-id":5854722,"id":"Q5854722"},"type":"wikibase-en
e","property":"P512","hash":"7a8ce79aa6d1b7406d20d9af185477a
2697,"id":"Q45892697"},"type":"wikibase-entityid"}}],"P580":

Парсинг ознак

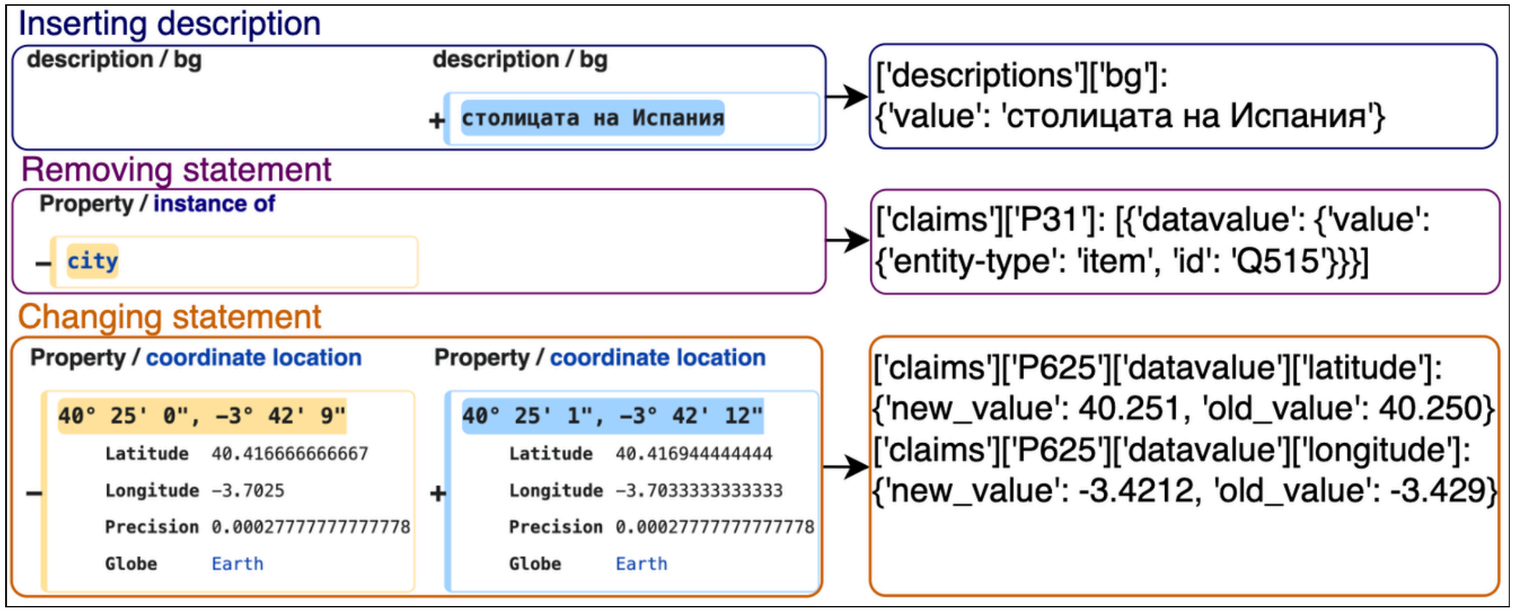








DeepDiff



Деталі обробки тексту

Три різні сигнали – одна мала мовна модель (SLM).

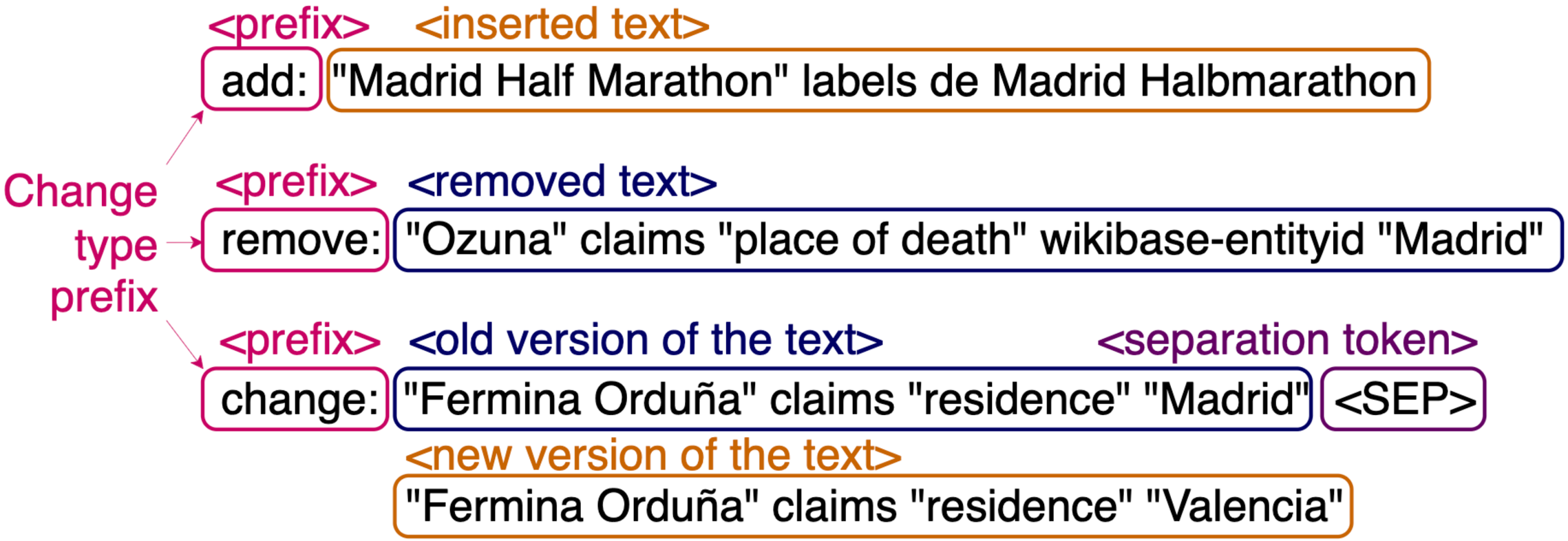
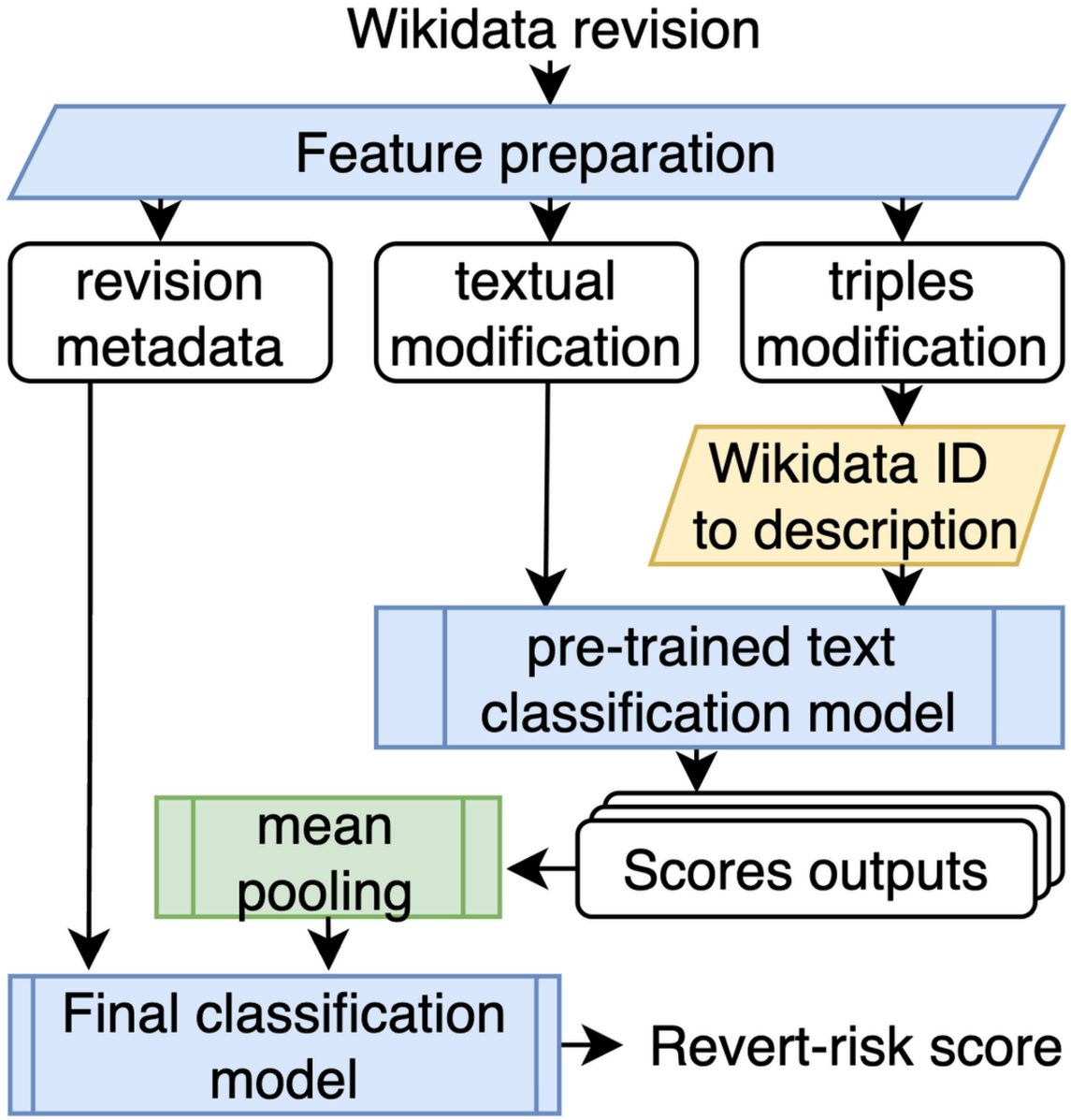


Схема системи



Оцінка якості системи

Table 3: System performance on holdout testing set.

Model	AUC	CI	FR@0.99	FR@0.9	FR@0.7
Rule-based	0.760	[0.74, 0.78]	0.0	0.0	0.92
ORES	0.859	[0.84, 0.87]	0.45	0.88	0.94
MbC	0.880	[0.87, 0.89]	0.55	0.89	0.94
CbC	0.876	[0.86, 0.89]	0.60	0.82	0.93
Graph2Text	**0.924**	[0.91, 0.93]	**0.71**	**0.91**	**0.96**

- **Rule-based**: all revisions by anonymous users are reverted;
- **ORES**: previous SOTA model;
- **MbC**: Metadata-based Classifier (MbC)
- **CbC**: Content-based Classifier (CbC)
- **Graph2Text**: proposed model (CbC + MbC)

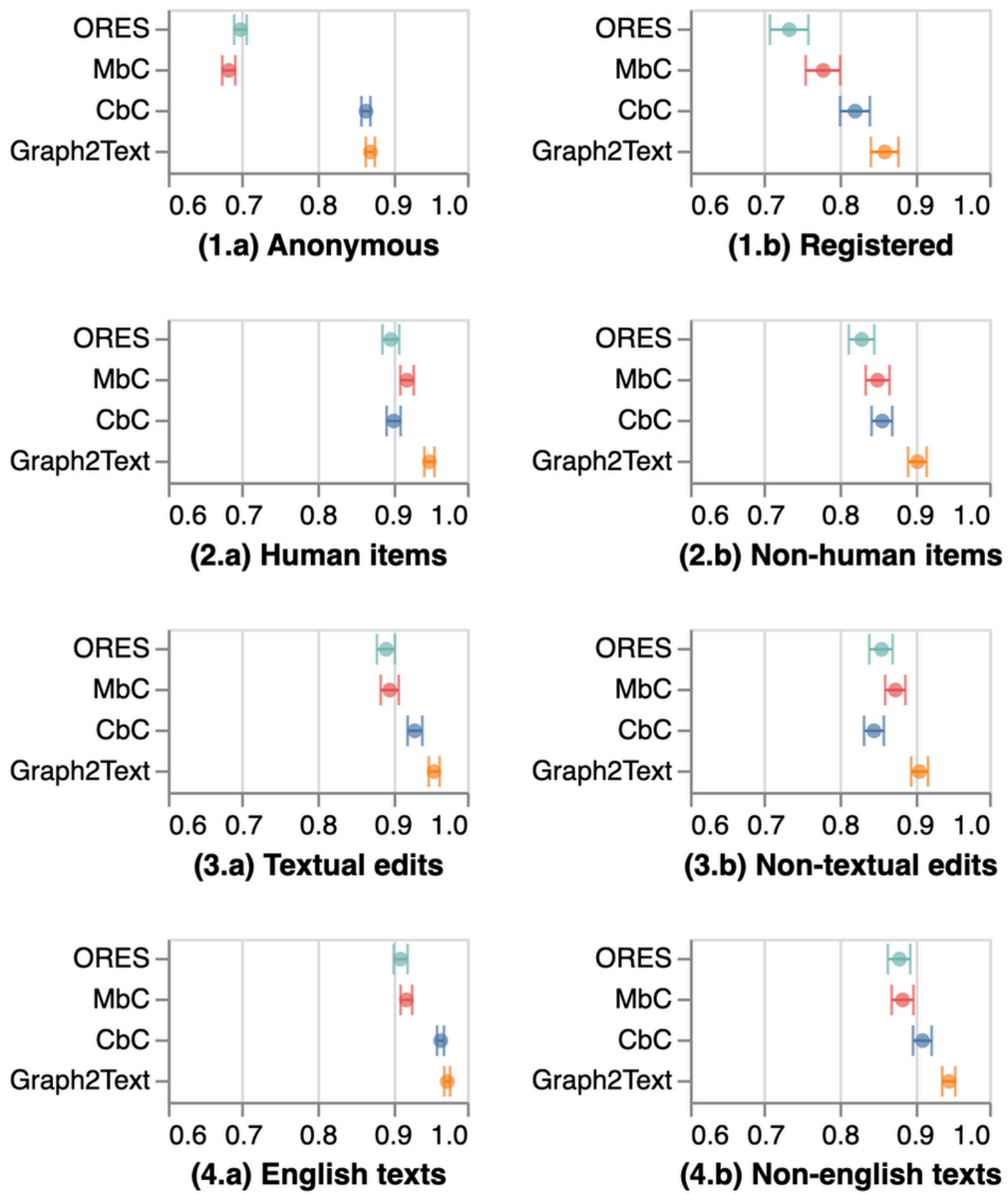
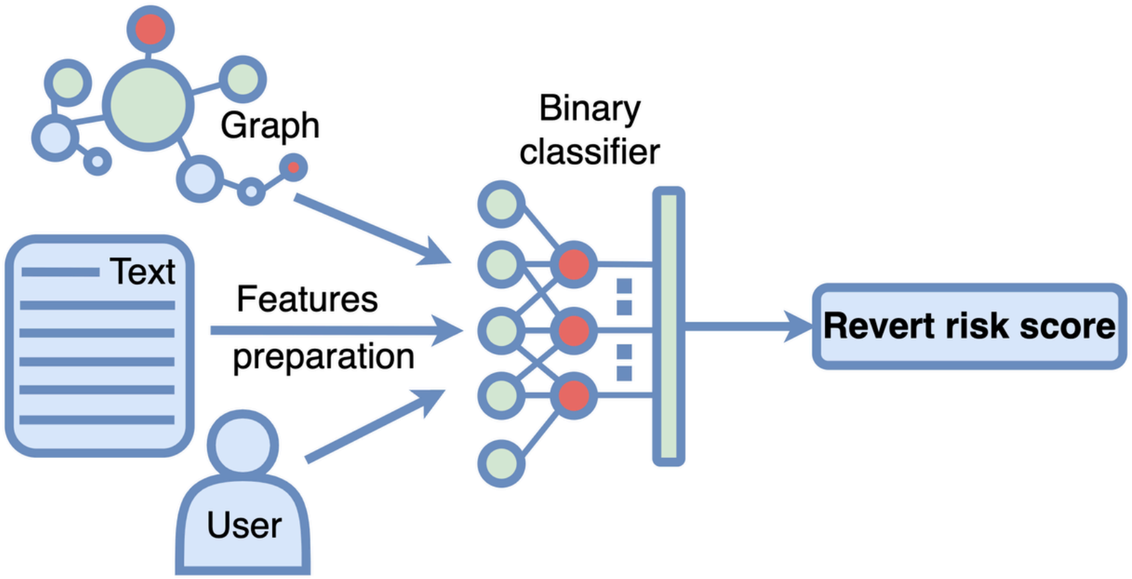


Figure: Models performance (AUC)

Набуті знання

- Ми можемо оцінювати якість змін контенту, не аналізуючи сам контент.
- **Сильні базові моделі** потребують швидкого процесингу ознак (наприклад, метадані користувача).
- **Сильні ознаки контенту допомагають зменшити упередженість.**
- Впровадження становить **80% роботи.**



Дякую!

Mykola Trokhymovych
trokhymovych.com
mykola.trokhymovych@upf.edu