Universitat Pompeu Fabra Barcelona

Advancing Open Knowledge

NLP for Accessibility, Fairness, and Knowledge Integrity in Collaborative Platforms

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

Internet, June 4, 2025

About me

Mykola Trokhymovych

trokhymovych.com | mykola.trokhymovych@upf.edu

- PhD student @ Universitat Pompeu Fabra
  - Co-advised by
    - Diego Saez-Trumper
    - Ricardo Baeza-Yates
- Research collaborator @ Wikimedia Foundation

Previously:

- MSc. Data Science @ Ukrainian Catholic University
- Data scientist / ML engineer @ Ciklum, Jooble, Surprise

Agenda

1. Wikipedia ecosystem
2. Readability
3. Vandalism detection
4. Discussion

Why Wikipedia ecosystem?

🌍 **Largest Human-Curated Knowledge Base**

- 300+ languages, 60M+ articles, 100M+ Wikidata triples;
- Maintained entirely by volunteers.

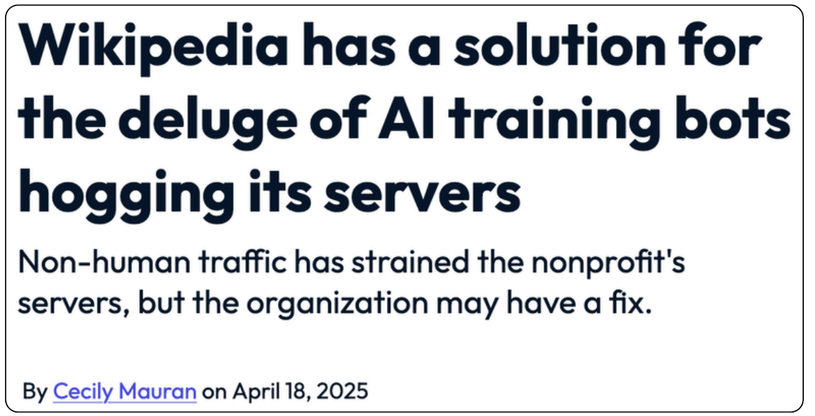
📖 **Open, Collaborative, Transparent**

- Fully open-source and publicly editable;
- Rich edit history, versioning, and discussion pages.

🤖 **Real-World Impact**

- Search engines, voice assistants, and education;
- Data backbone for LLMs and knowledge graphs.

wired



mashable

Wikipedia AI Strategy

> “Not too long ago, we were asked when we're going to replace Wikipedia's human-curated knowledge with AI.
>
> The answer? We're not.”

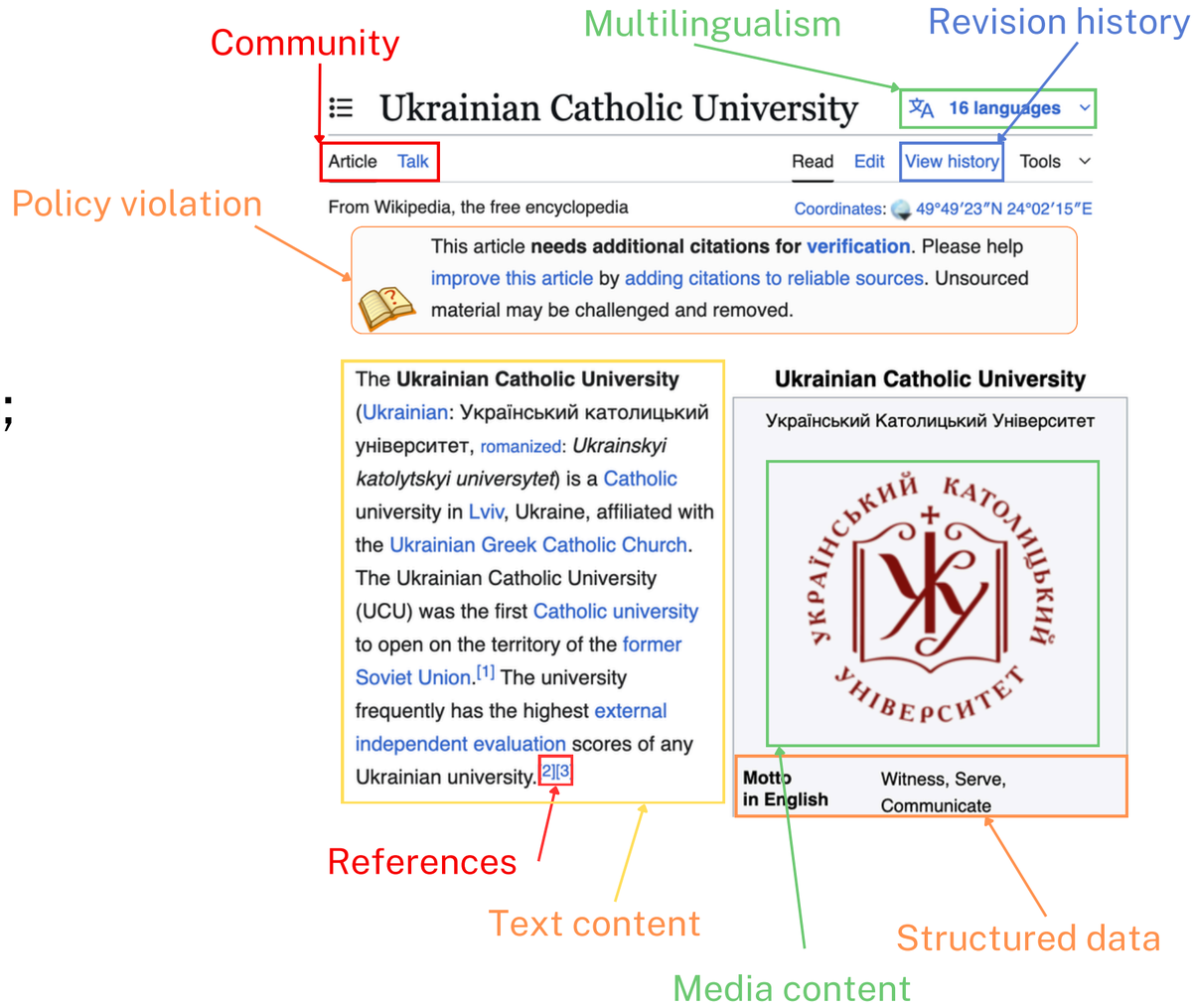
Our new AI strategy puts Wikipedia’s humans first

By Chris Albon and Leila Zia · 30 April 2025

Share

NLP Applications in the Wikipedia

- 🛡️ **Moderation tools**:
  - Talk page tone detection;
  - Vandalism & spam detection.
- ✏️ **Editing tools**:
  - Readability evaluation;
  - Grammar/misspelling correction;
  - Detection of policy violations;
  - Source verification.
- 📚 **Reader Tools:**
  - Article summaries;
  - Text to speech;
  - Articles search and recommendation.



Readability & Wikipedia

What is Readability?

The readability concept aims to capture how easy it is to read a given text. It is usually defined as the sum of all factors that affect a reader's understanding, reading speed, and level of interest.

(Dale and Chall, 1949)

Readability of Wikipedia

"[in English Wikipedia the] overall readability is poor. "

(Lucassen et al., 2012)

"Most of the health information [in English Wikipedia] remains written at a level above the reading ability of average adults."

(Brezar and Heilman, 2019)

Measuring readability is challenging

Common approaches are readability formulas
e.g., Flesch Reading Ease (FRE) Score Formula

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

Wikipedia exists in more than 300 languages:

- Adapting readability formulas to each language is not scalable/feasible
- Lack of multilingual models/tools
- Scarcity of multilingual ground-truth data
  - Standard datasets are not open (e.g., Newsela)
  - Mostly datasets are for English (no data for most languages)

In this research we aim to . . .

build a multilingual model that can automatically access the readability of the text of Wikipedia articles across languages.

An Open Multilingual System for Scoring Readability of Wikipedia

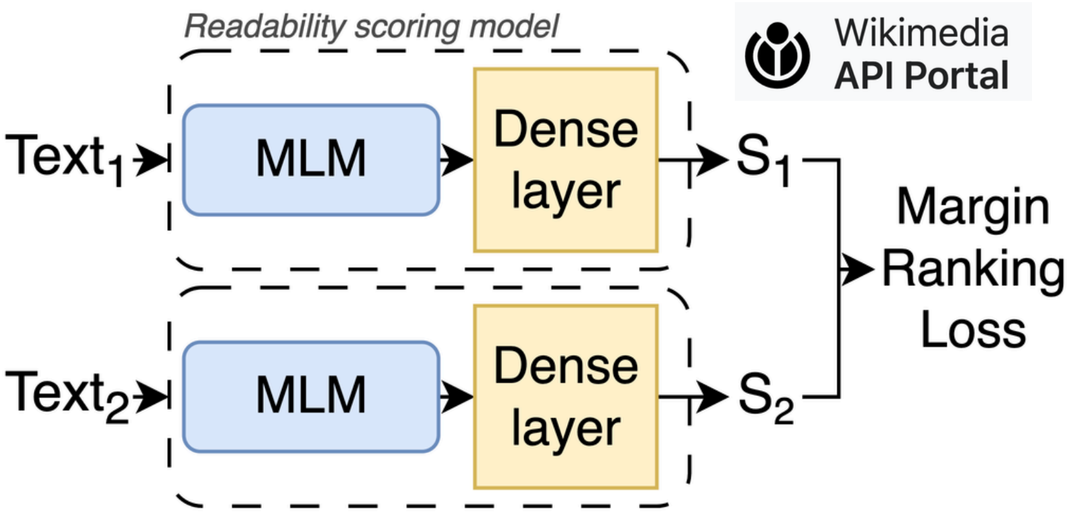
Data

WikiReaD (Wikipedia Readability Dataset)

What? Pairs of encyclopedic articles in two readability levels (simple, hard) for 14 languages. (>100K samples)

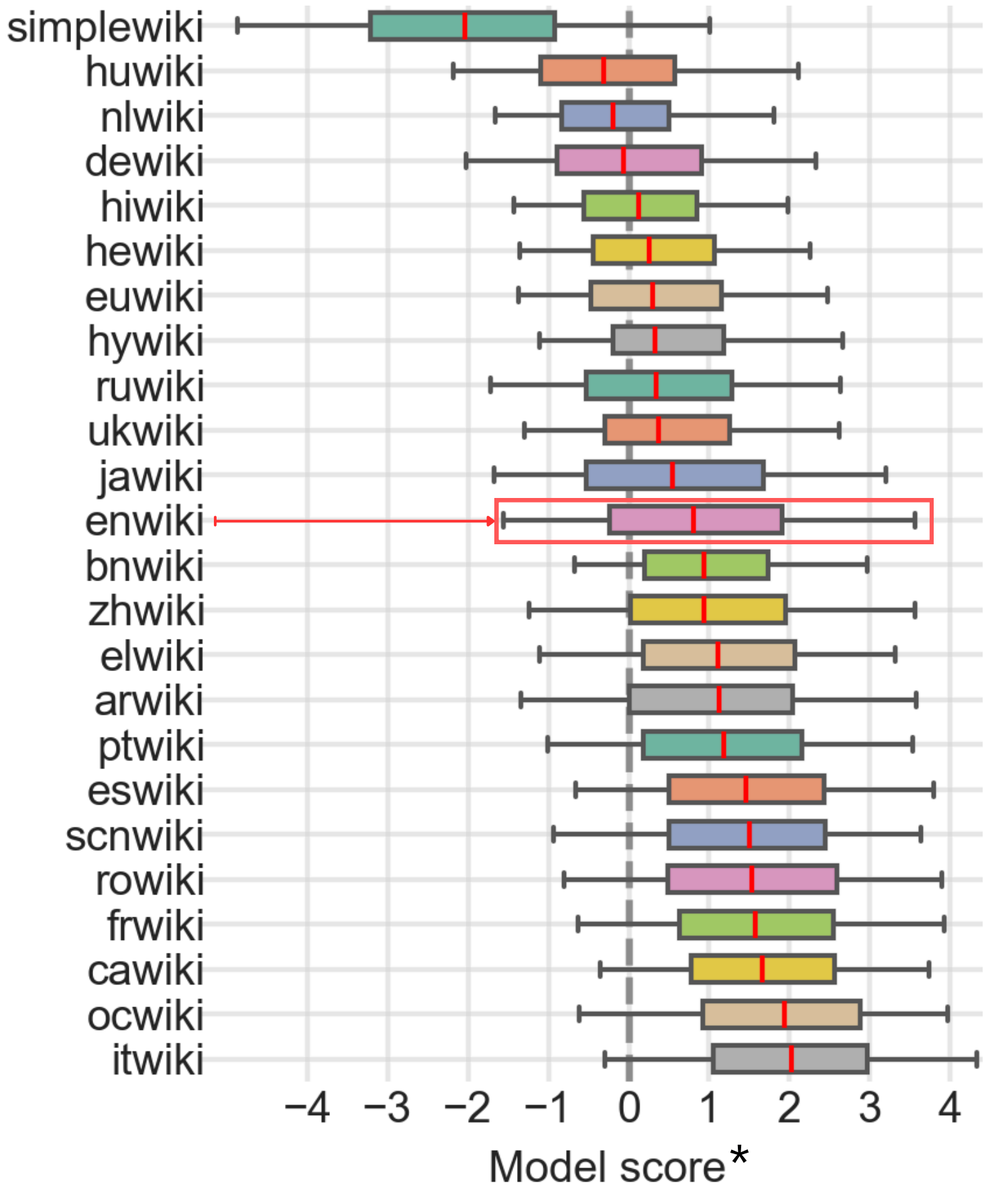
How? Matching Wikipedia articles with a simplified version.

Model



- xlm-roberta-longformer
- Siemease architecture for training:
- Loss: Margin Ranking Loss
- Fine-tuning only using data in English
- Ranking Accuracy > 0.8 for all languages and > 0.9 for 10/15 datasets.
- Outperform reference benchmarks not used for training (VikidiaFr, OneStopEnglish)

Authors: Mykola Trokhymovych, Indira Sen, Martin Gerlach.
ACL'24 Main Track



Wikipedia is difficult to read across most languages

In English Wikipedia, more than half the articles exceed the average readability level of American adults (~grade level 8-9) (Brezar&Heilman 2019)

Overall readability in most Wikipedias are similar to English Wikipedia

* Model score of 0 corresponds to average readability level of American adults (~grade level 8-9)

Authors: Mykola Trokhymovych, Indira Sen, Martin Gerlach.

ACL'24 Main Track

Not every edit improves Wikipedia — some distort it.

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

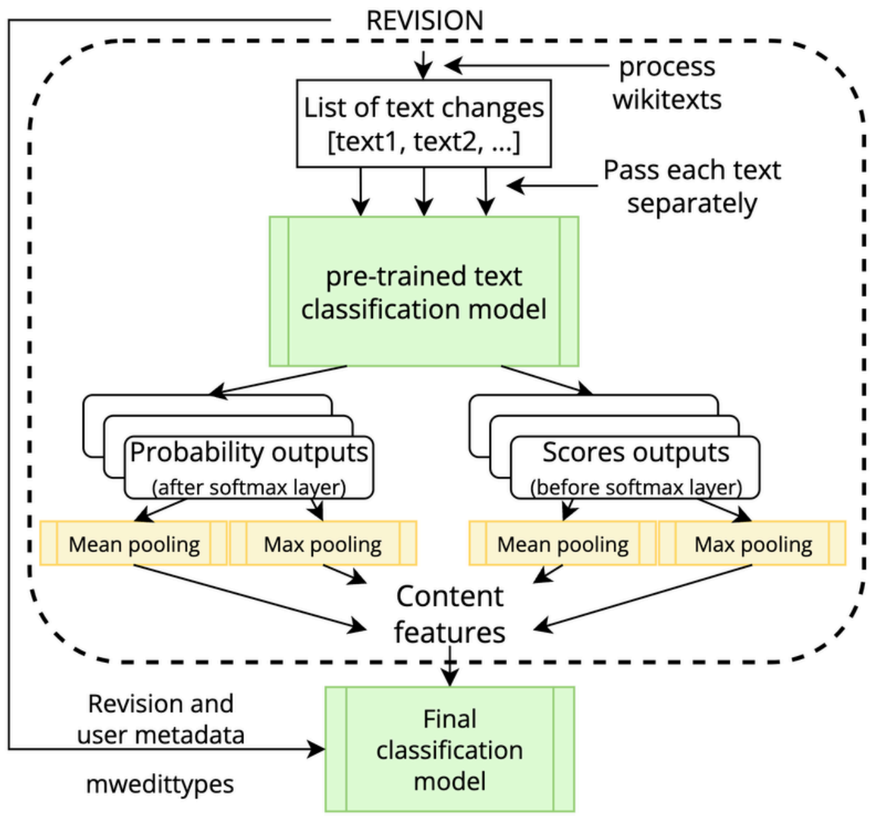
```
Lvov emerged as the centre of the historical
regions of [[Red Ruthenia]] and [[Galicia
(Eastern Europe)|Galicia]] in the [[14th
```

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

Although some models help patrollers (like ORES),
there are still open problems like:

- Model performance;
- Language coverage;
- Fairness.

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

Goal:

Create a model to help editors identify edits that require patrolling.

Approach:

- Use implicit annotations (reverts) to train the ML models.

Result:

- An open-source, multilingual model for content patrolling on Wikipedia;
- Outperforming the previous SOTA models in performance and fairness;
- Productionalized and is working in production for 47 languages.

Wikimedia API Portal

Anonymous editors have a higher revert rate.

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

Models overfit and learn to mistrust anonymous users.

Disparate Impact Ratio (DIR)

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

= 20

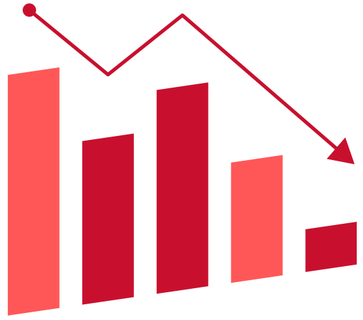
***for ORES**

Pr - probability

$\hat{Y}$ - predicted value,

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

Anonymous editors are less likely to stay and contribute.



Risk of the number of active editors dropping.

Vandalism beyond Wikipedia

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)

What is the athem of Bulgaria? 🇧🇬

TL ; DR

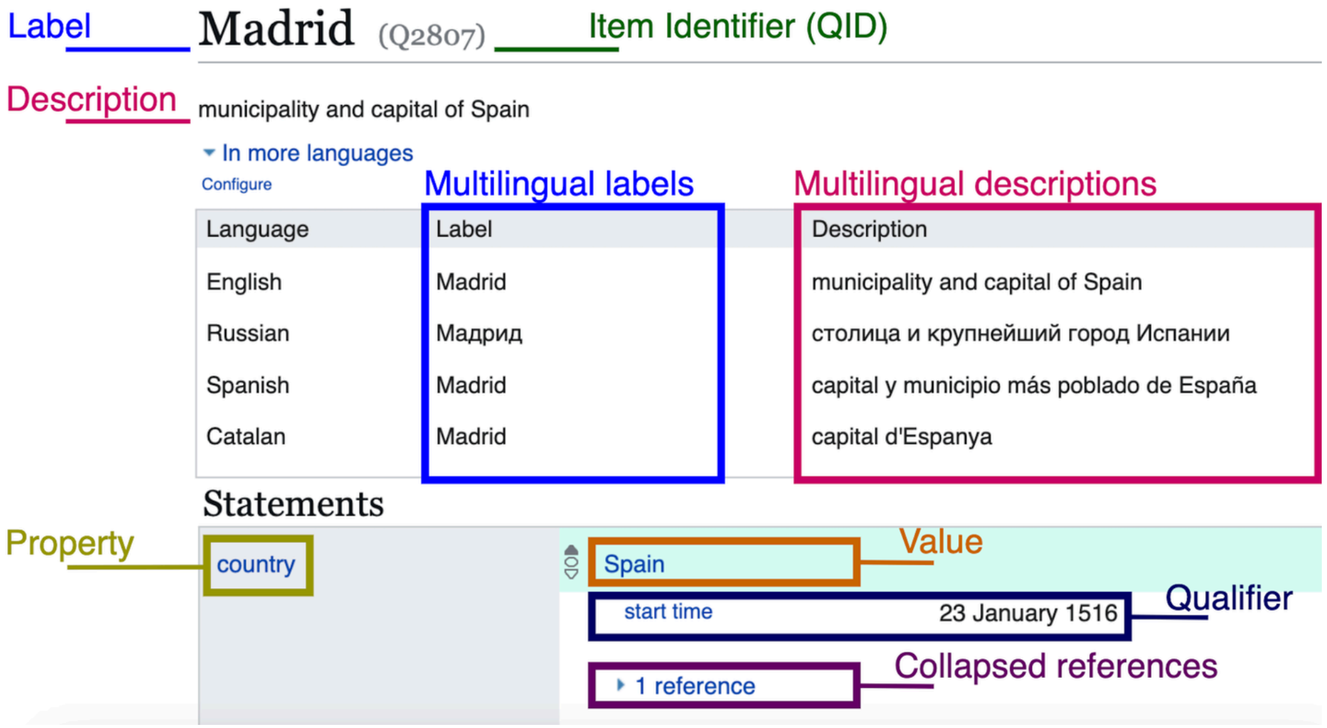
Why does Siri think the national anthem of Bulgaria is 'Despacito?' / A cursory investigation with no clear answer in sight

by **Nick Statt**

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

theverge

Wikidata vandalism detection



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

Authors: Mykola Trokhymovych, Lydia Pintscher,
Diego Saez-Trumper, Ricardo Baeza-Yates
ACL'25 Industry Track

Goal:

Create a model to help Wikidata editors identify edits that require patrolling

Approach:

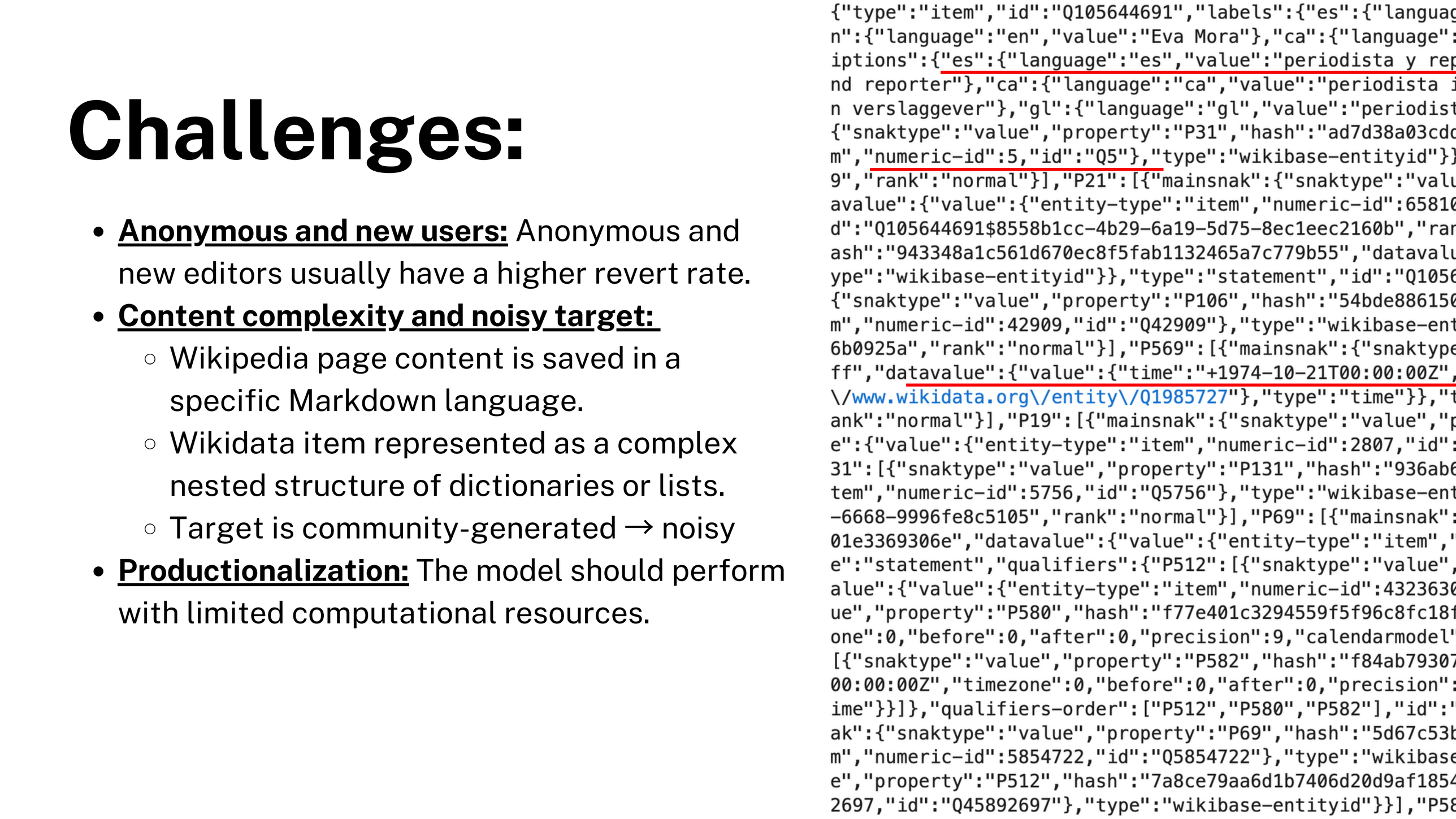
- Use implicit annotations (reverts) to train the ML models.

Result:

- Introduced **Graph2Text**, enabling content processing with a single LM.
- Outperformed prior state-of-the-art in performance and fairness.
- Improved content processing for better productization.

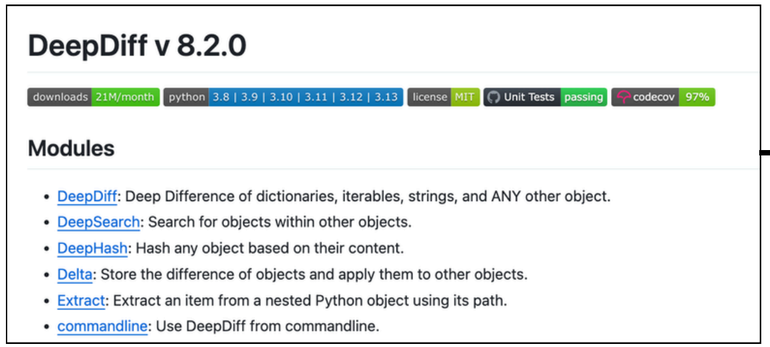
Challenges:

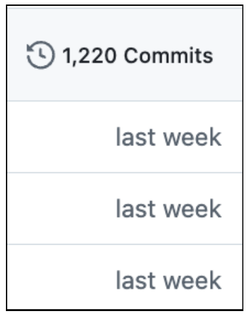
- **Anonymous and new users:** Anonymous and new editors usually have a higher revert rate.
- **Content complexity and noisy target:**
  - Wikipedia page content is saved in a specific Markdown language.
  - Wikidata item represented as a complex nested structure of dictionaries or lists.
  - Target is community-generated → noisy
- **Productionalization:** The model should perform with limited computational resources.

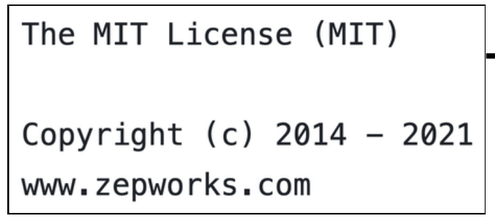


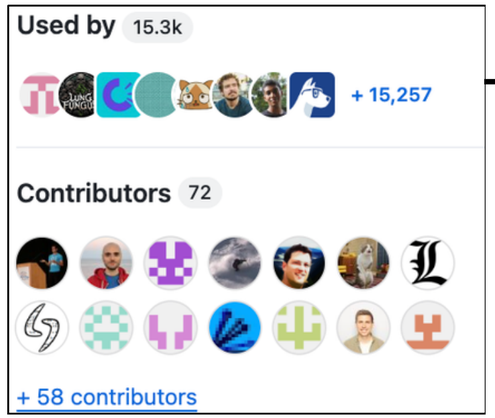
{"type":"item","id":"Q105644691","labels":{"es":{"languag
n":{"language":"en","value":"Eva Mora"},"ca":{"language":
iptions":{"es":{"language":"es","value":"periodista y rep
nd reporter"},"ca":{"language":"ca","value":"periodista i
n verslaggever"},"gl":{"language":"gl","value":"periodist
{"snaktype":"value","property":"P31","hash":"ad7d38a03cdc
m","numeric-id":5,"id":"Q5"},"type":"wikibase-entityid"}}
9","rank":"normal"}],"P21":[{"mainsnak":{"snaktype":"valu
avalue":{"value":{"entity-type":"item","numeric-id":65810
d":"Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b","ran
ash":"943348a1c561d670ec8f5fab1132465a7c779b55","datavalu
ype":"wikibase-entityid"}},"type":"statement","id":"Q1056
{"snaktype":"value","property":"P106","hash":"54bde886150
m","numeric-id":42909,"id":"Q42909"},"type":"wikibase-ent
6b0925a","rank":"normal"}],"P569":[{"mainsnak":{"snaktype
ff","datavalue":{"value":{"time":"+1974-10-21T00:00:00Z",
\/www.wikidata.org\/entity\/Q1985727"},"type":"time"}},"t
ank":"normal"}],"P19":[{"mainsnak":{"snaktype":"value","p
e":{"value":{"entity-type":"item","numeric-id":2807,"id":
31":[{"snaktype":"value","property":"P131","hash":"936ab6
tem","numeric-id":5756,"id":"Q5756"},"type":"wikibase-ent
-6668-9996fe8c5105","rank":"normal"}],"P69":[{"mainsnak":
01e3369306e","datavalue":{"value":{"entity-type":"item","
e":"statement","qualifiers":{"P512":[{"snaktype":"value",
alue":{"value":{"entity-type":"item","numeric-id":432363
ue","property":"P580","hash":"f77e401c3294559f5f96c8fc18f
one":0,"before":0,"after":0,"precision":9,"calendarmodel"
[{"snaktype":"value","property":"P582","hash":"f84ab79307
00:00:00Z","timezone":0,"before":0,"after":0,"precision":
ime"}}]},"qualifiers-order":["P512","P580","P582"],"id":"
ak":{"snaktype":"value","property":"P69","hash":"5d67c53b
m","numeric-id":5854722,"id":"Q5854722"},"type":"wikibase
e","property":"P512","hash":"7a8ce79aa6d1b7406d20d9af1854
2697,"id":"Q45892697"},"type":"wikibase-entityid"}}],"P58

Content processing

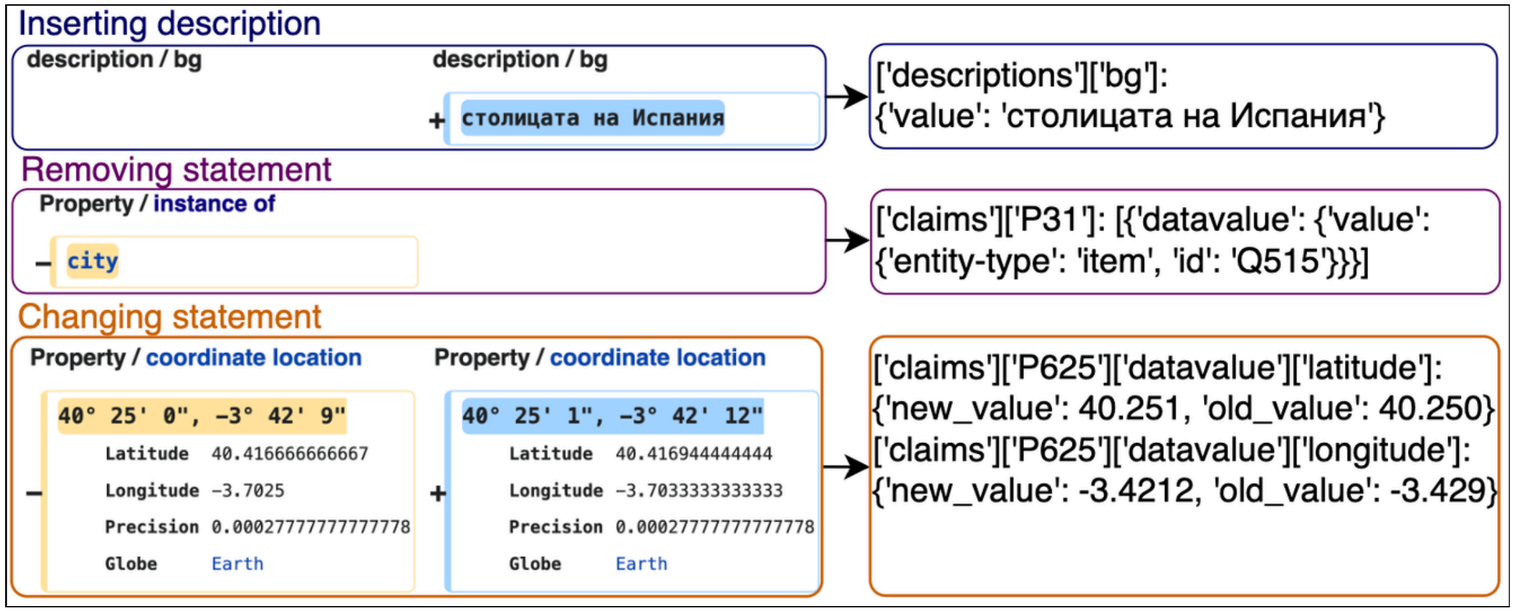






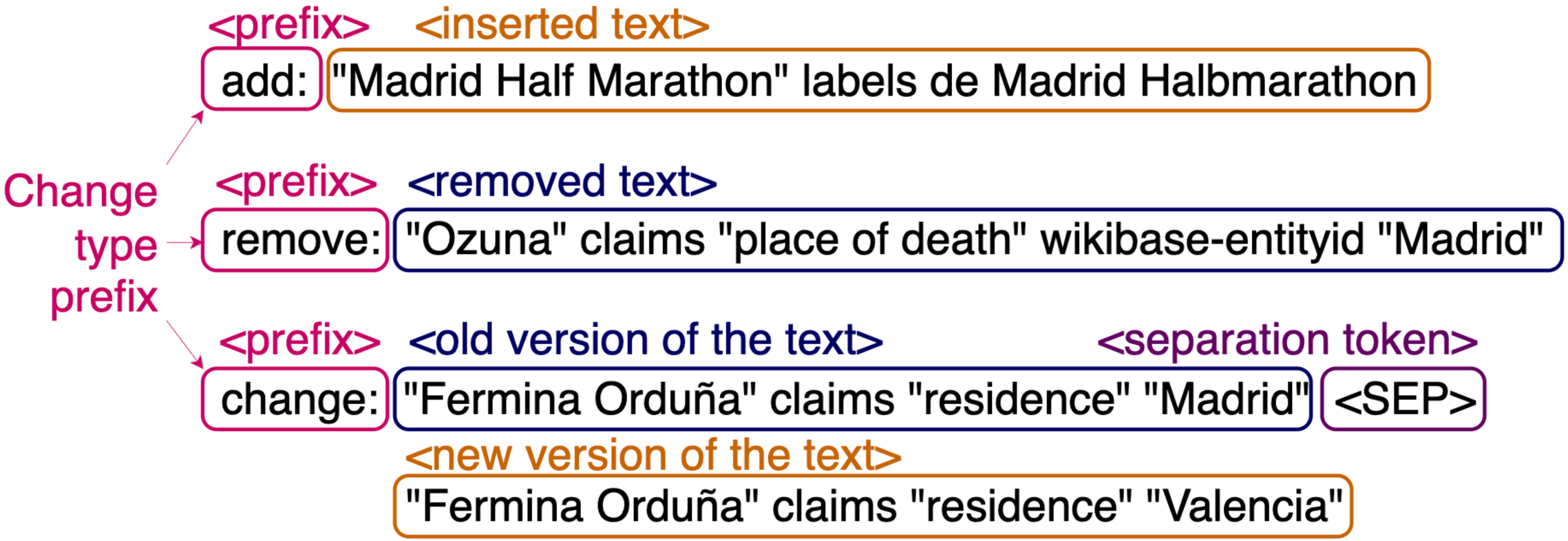


DeepDiff



Text processing

Three different signals - one Small language model (SLM).



Target preparation

Filters applied:

- Filter for "revision-wars" (leave only those reverted revisions that were not later reverted)
- Filter for "self-reverts"
- Filter revisions created by bots

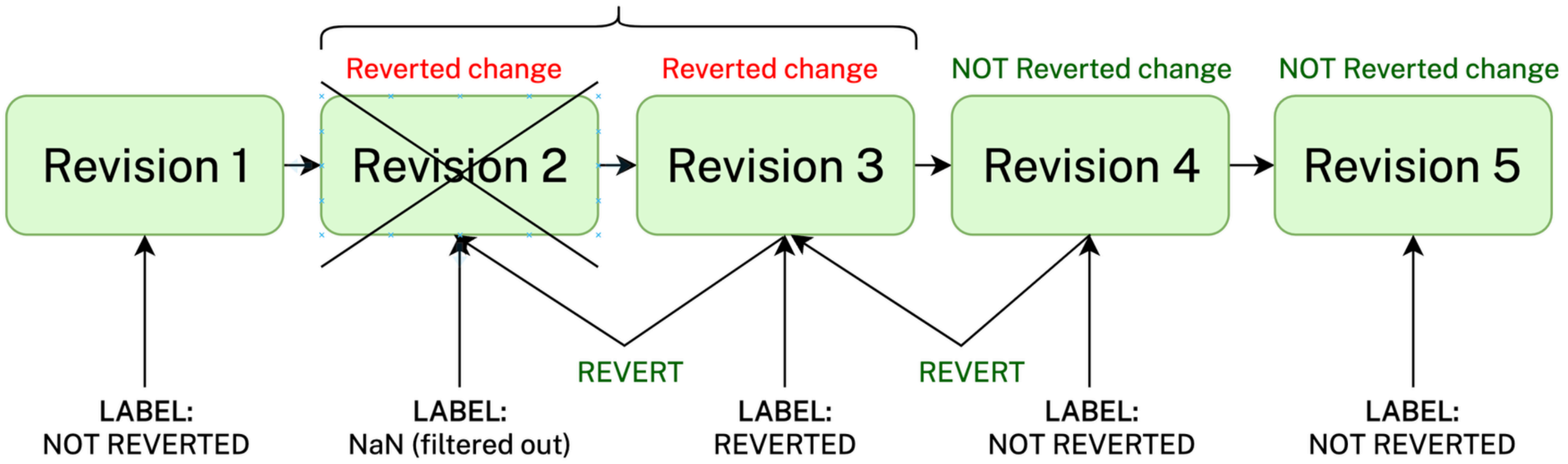
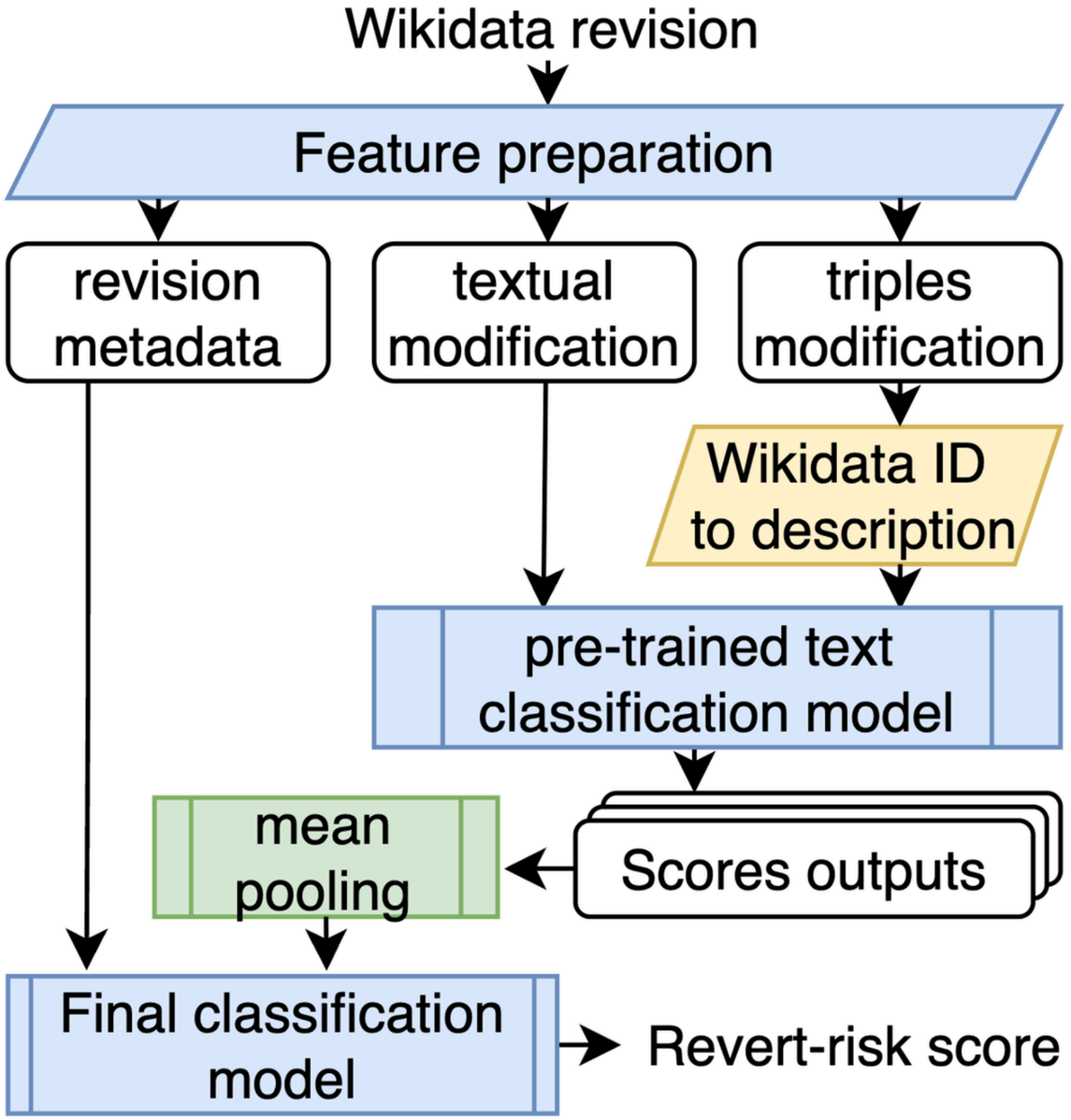


Figure: Revision-wars filtering logic

System design

- Graph-based changes processing
- Text-based changes processing
- The final Classification Model includes:
  - text features;
  - graph features;
  - revision metafeatures.



System evaluation

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)

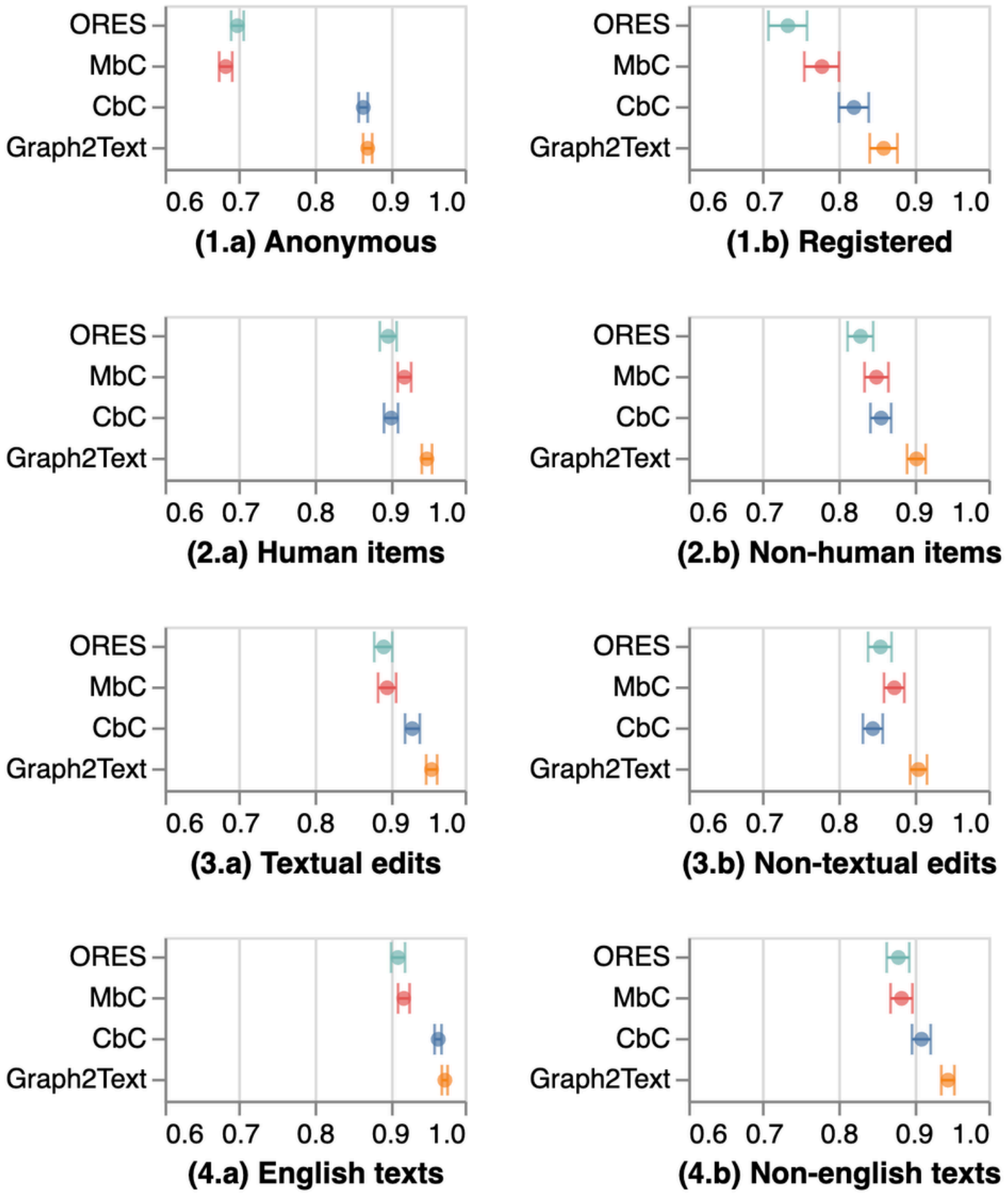
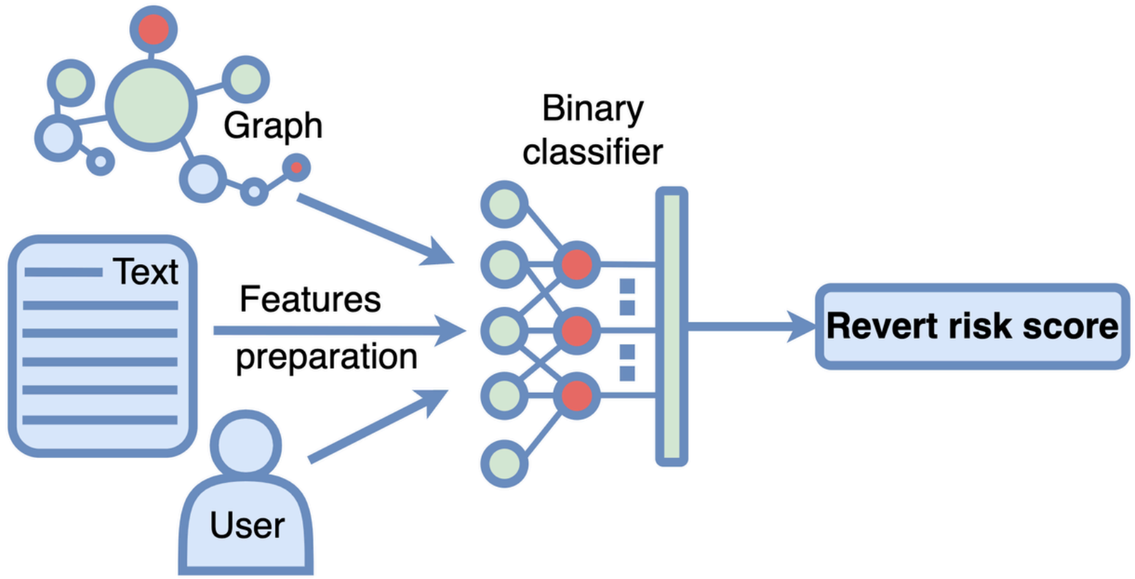


Figure: Models performance (AUC)

Learnings

- **Do we need content?** We can access the quality of the content changes without analyzing the content itself.
- **Strong baselines** need easy-to-extract features (e.g., edit speed, user metadata).
- **Strong content features** help to reduce bias.
- **Quality > Quantity**: Data filtering is the key
- **Productization is 80%** of the work.



Thank you!

Q&A

Mykola Trokhymovych

trokhymovych.com
mykola.trokhymovych@upf.edu

Acknowledgments

Advisors:

- Diego Saez-Trumper
- Ricardo Baeza-Yates

Collaborators:

- Indira Sen
- Martin Gerlach
- Muniza Aslam
- Ai-Jou Chou
- Lydia Pintscher