

Graph-Linguistic Fusion:

Using Language Models for Wikidata Vandalism Detection

Authors:

Mykola Trokhymovych

Lydia Pintscher

Diego Saez-Trumper

Ricardo Baeza-Yates

What is Wikidata?

Label Madrid (Q2807) **Item Identifier (QID)**

Description municipality and capital of Spain

▼ In more languages
Configure

Language	Label	Description
English	Madrid	municipality and capital of Spain
Russian	Мадрид	столица и крупнейший город Испании
Spanish	Madrid	capital y municipio más poblado de España
Catalan	Madrid	capital d'Espanya

Statements

Property country

Value Spain

Qualifier start time 23 January 1516

Collapsed references ▶ 1 reference

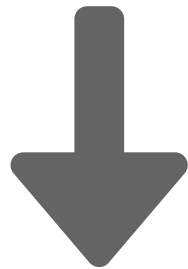
- Structured Knowledge Storage
 - ~100M+ Wikidata triples
 - ~8 edits per second
- Richly Annotated Records:
 - Texts in 300+ languages
 - Timestamps
 - Numeric values
 - Geographic coordinates
 - And more
- Powering countless services on the Web

Vandalism in Wikidata

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 anthem 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



Vandalism

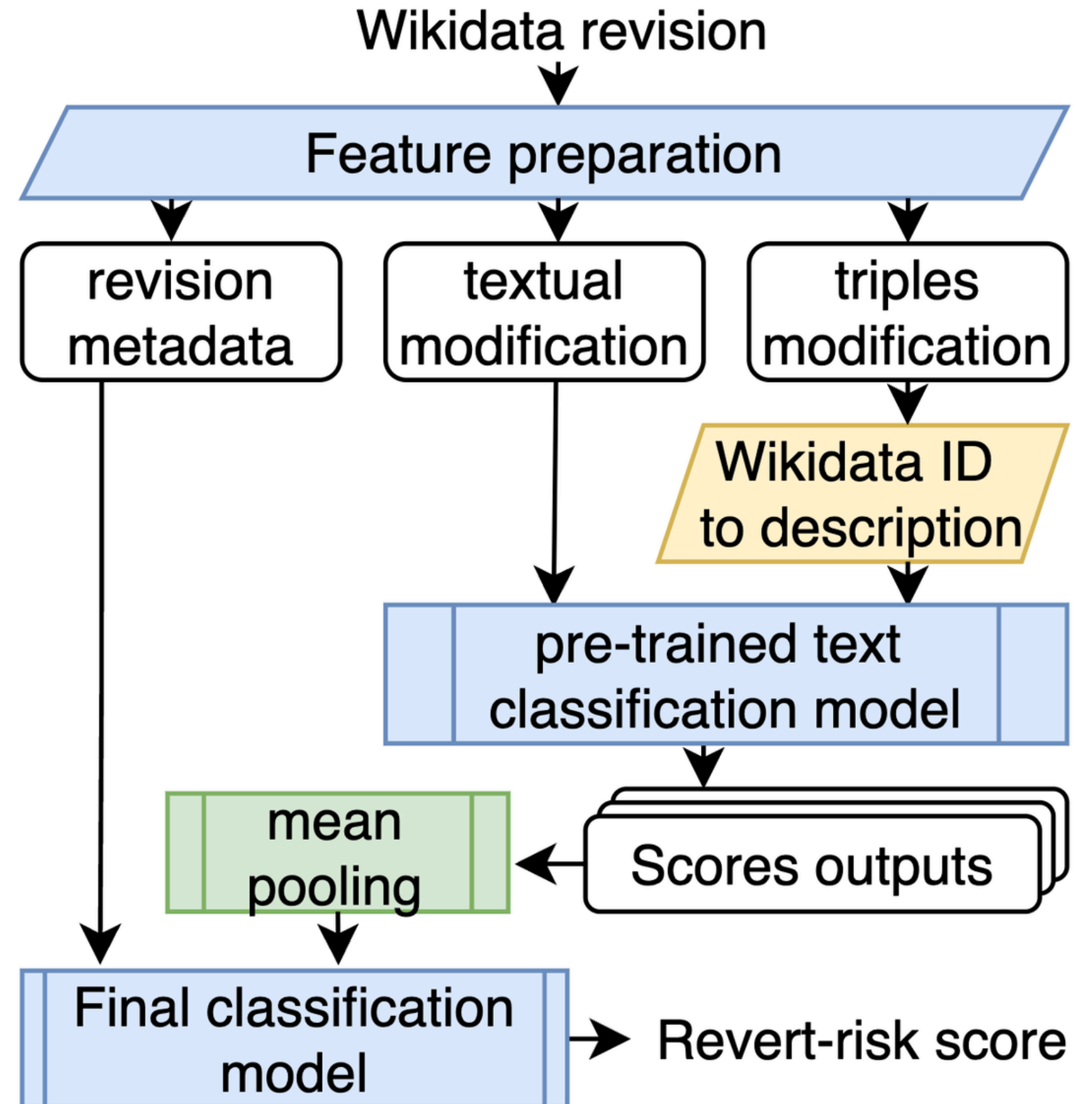
On Wikipedia, vandalism has a very specific meaning: editing (or other behavior) deliberately intended to obstruct or defeat the project's purpose, which is to create a free...

 Wikipedia



System design

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



Challenges:

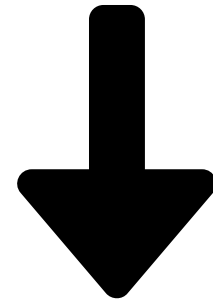
- **Anonymous and new users:** Anonymous and new editors usually have a higher revert rate.
- **Content complexity and noisy target:**
 - Wikidata item represented as a complex nested structure of dictionaries or lists.
 - Target is community-generated → noisy
- **Productionalization:** The model should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

```
{
  "type": "item",
  "id": "Q105644691",
  "labels": {
    "es": {
      "language": "es",
      "value": "Eva Mora"
    },
    "ca": {
      "language": "ca",
      "value": "periodista y reportera"
    },
    "gl": {
      "language": "gl",
      "value": "periodista e xpositiva"
    },
    "nl": {
      "language": "nl",
      "value": "journalist en verslaggever"
    }
  },
  "snaktype": "value",
  "property": "P31",
  "hash": "ad7d38a03cd0b1c561d670ec8f5fab1132465a7c779b55",
  "numeric-id": 5,
  "id": "Q5",
  "type": "wikibase-entityid",
  "rank": "normal"
},
{
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P21",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q105644691",
    "numeric-id": 5,
    "rank": "normal"
  },
  "hash": "943348a1c561d670ec8f5fab1132465a7c779b55",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P106",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q42909",
    "numeric-id": 42909,
    "rank": "normal"
  },
  "hash": "54bde8861506b0925a",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P569",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "f77e401c3294559f5f96c8fc18f1e3369306e",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P19",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q1985727",
    "numeric-id": 1985727,
    "rank": "normal"
  },
  "hash": "936ab6668-9996fe8c5105",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P131",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5756",
    "numeric-id": 5756,
    "rank": "normal"
  },
  "hash": "f84ab7930700:00:00Z",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P580",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q4323630",
    "numeric-id": 4323630,
    "rank": "normal"
  },
  "hash": "f77e401c3294559f5f96c8fc18f1e3369306e",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P582",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "f77e401c3294559f5f96c8fc18f1e3369306e",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P69",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5854722",
    "numeric-id": 5854722,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P512",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "7a8ce79aa6d1b7406d20d9af18542697",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P580",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q45892697",
    "numeric-id": 45892697,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P512",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "7a8ce79aa6d1b7406d20d9af18542697",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P69",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5854722",
    "numeric-id": 5854722,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P512",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "7a8ce79aa6d1b7406d20d9af18542697",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P69",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5854722",
    "numeric-id": 5854722,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P512",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "7a8ce79aa6d1b7406d20d9af18542697",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P69",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5854722",
    "numeric-id": 5854722,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P512",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "7a8ce79aa6d1b7406d20d9af18542697",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P69",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5854722",
    "numeric-id": 5854722,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P512",
  "snaktype": "value",
  "value": {
    "type": "time",
    "value": "+1974-10-21T00:00:00Z",
    "timezone": "UTC",
    "precision": 0,
    "before": 0,
    "after": 0,
    "precision": 9,
    "calendar": "Gregorian"
  },
  "hash": "7a8ce79aa6d1b7406d20d9af18542697",
  "datatype": "time",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a19-5d75-8ec1eec2160b",
  "property": "P69",
  "snaktype": "value",
  "value": {
    "type": "item",
    "id": "Q5854722",
    "numeric-id": 5854722,
    "rank": "normal"
  },
  "hash": "5d67c53b1e3369306e",
  "datatype": "wikibase-entityid",
  "type": "statement",
  "id": "Q105644691$8558b1cc-4b29-6a1
```

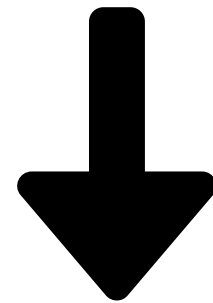
Anonymous editors have a higher revert rate.



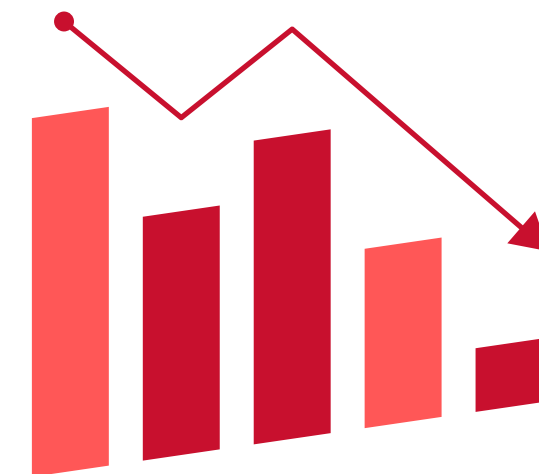
anon. VS. all



Models overfit and learn to mistrust anonymous users.



Anonymous editors are less likely to stay and contribute.



Risk of the number of active editors dropping.

Content processing

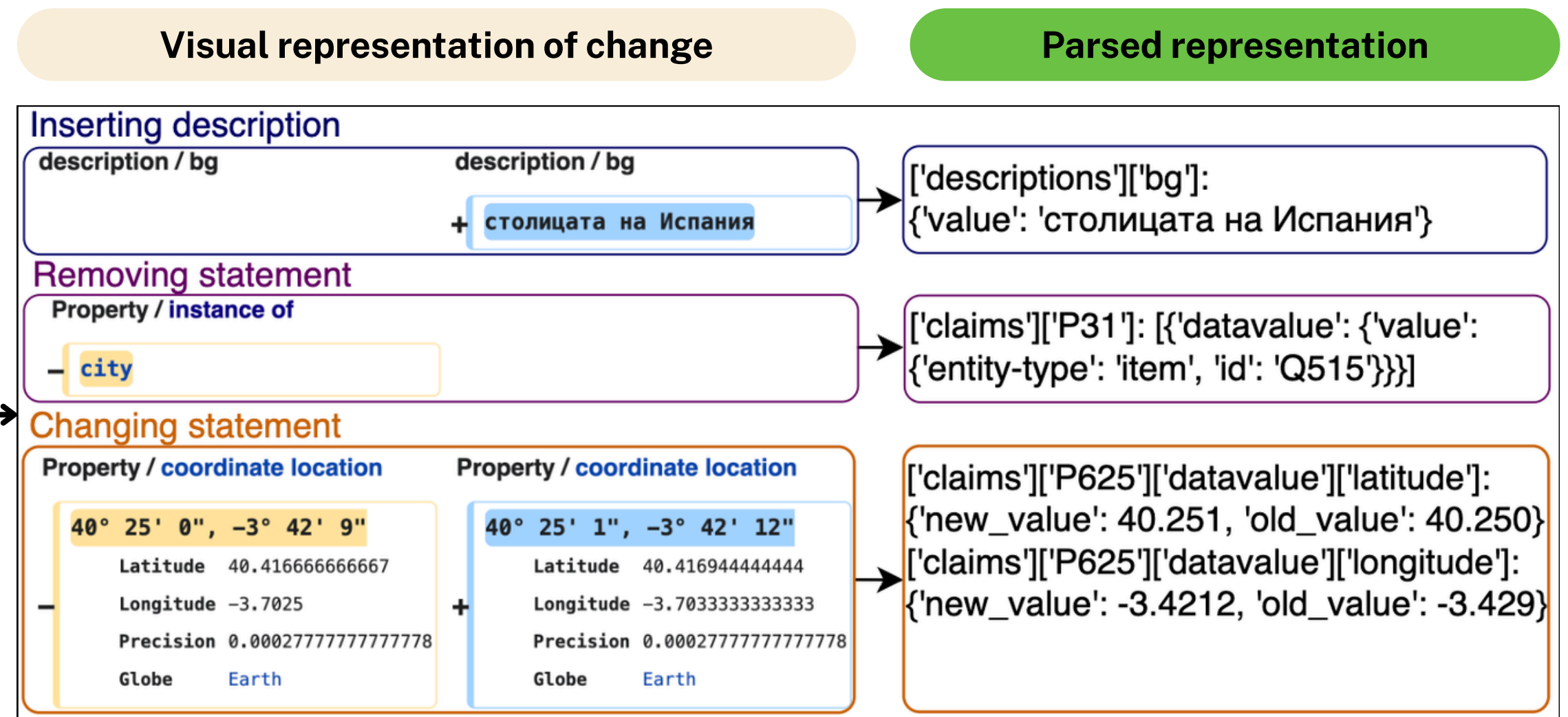
DeepDiff

DeepDiff v 8.2.0

downloads **21M/month** python **3.8 | 3.9 | 3.10 | 3.11 | 3.12 | 3.13** license **MIT** Unit Tests **passing** codecov **97%**

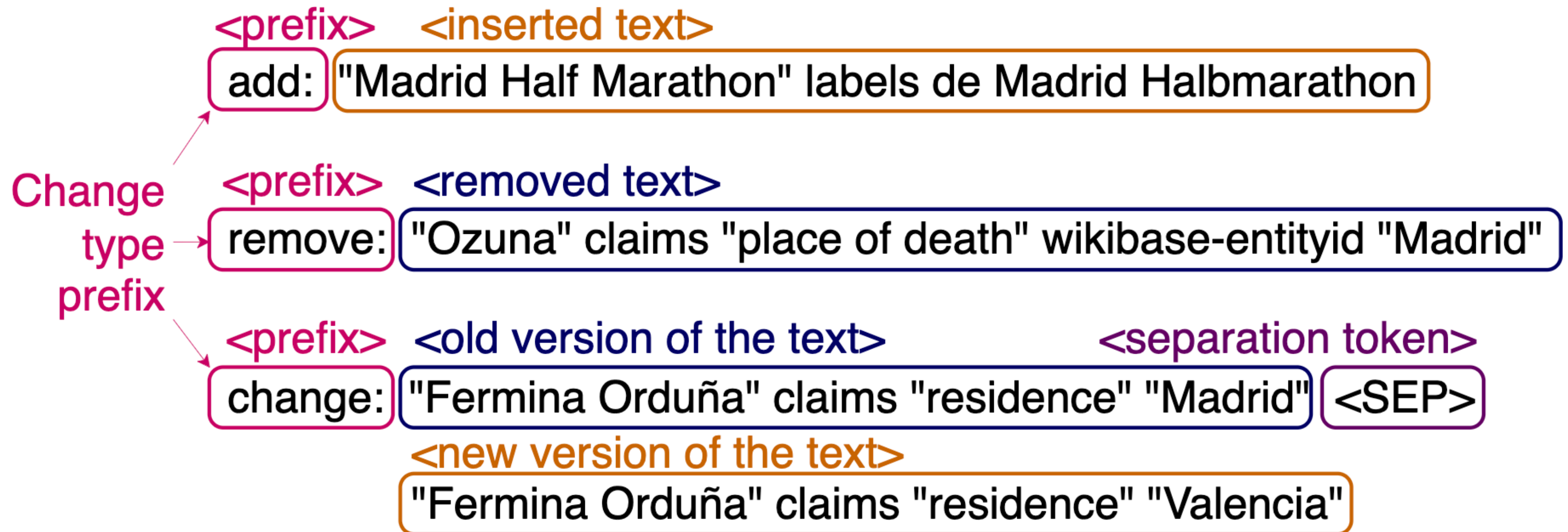
Modules

- [DeepDiff](#): Deep Difference of dictionaries, iterables, strings, and ANY other object.
- [DeepSearch](#): Search for objects within other objects.
- [DeepHash](#): Hash any object based on their content.
- [Delta](#): Store the difference of objects and apply them to other objects.
- [Extract](#): Extract an item from a nested Python object using its path.
- [commandline](#): Use DeepDiff from commandline.



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

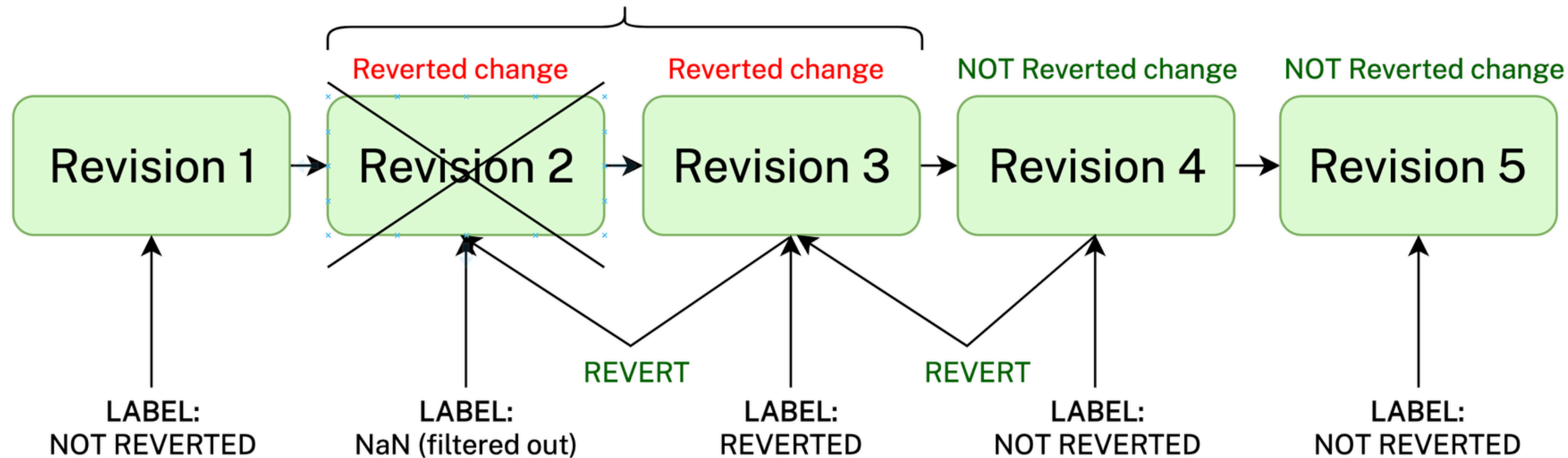
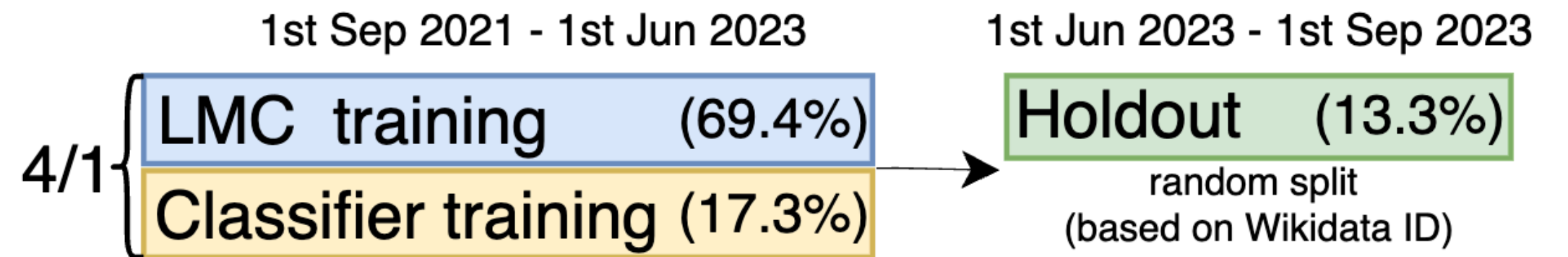


Figure: Revision-wars filtering logic

Dataset

Characteristic	Training	Hold-out
# of samples	4,197,231	645,264
Period	21 months	3 months
Anon. rate	10.7%	8.3%
Revert rate	7.9%	6.2%

Experiment



System evaluation

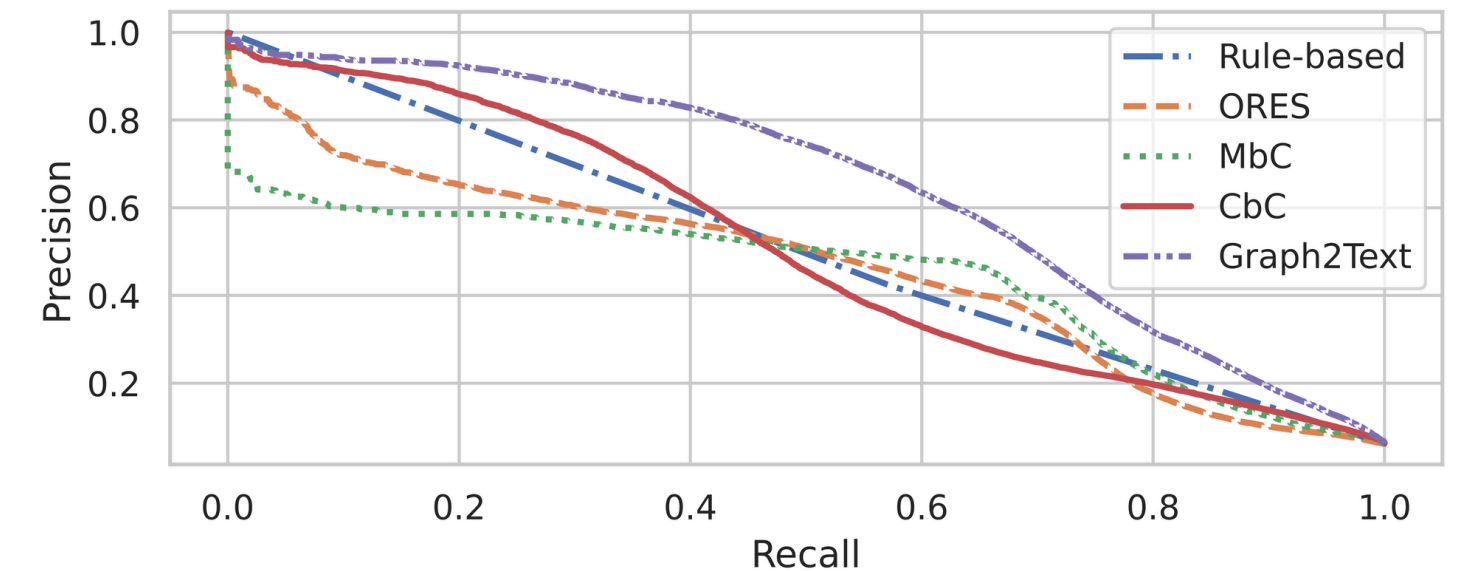
System performance on holdout testing set

Model	AUC	CI	FR@99	FR@90	FR@70
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

System performance on expert-labeled data

Model	AUC	CI	FR@99	FR90	FR70
ORES	0.885	[0.879, 0.892]	0.593	0.799	0.881
Graph2Text	0.932	[0.926, 0.937]	0.698	0.846	0.918

The precision/recall curves for models.



- **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)

System evaluation

System fairness based on holdout testing set

Model	DIR ^{anon}	DAUC ^{anon}	DIR ^{new}	DAUC ^{new}
ORES	5.69	0.035	1.37	-0.193
MbC	4.09	0.097	1.15	-0.155
CbC	2.07	-0.04	1.08	-0.027
Graph2Text	4.43	-0.01	1.24	-0.096

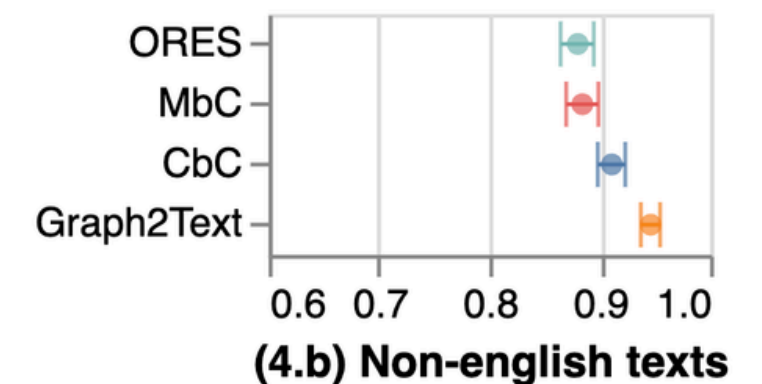
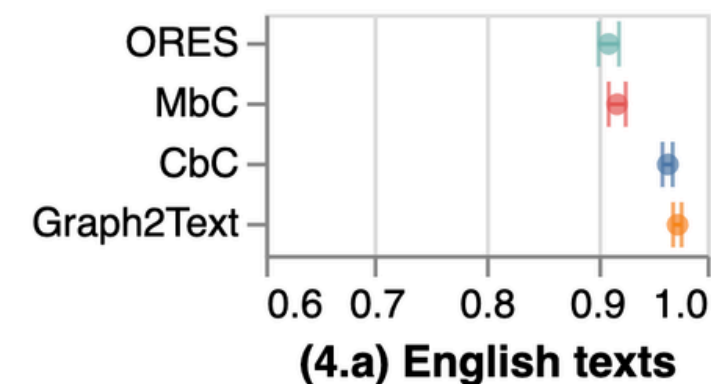
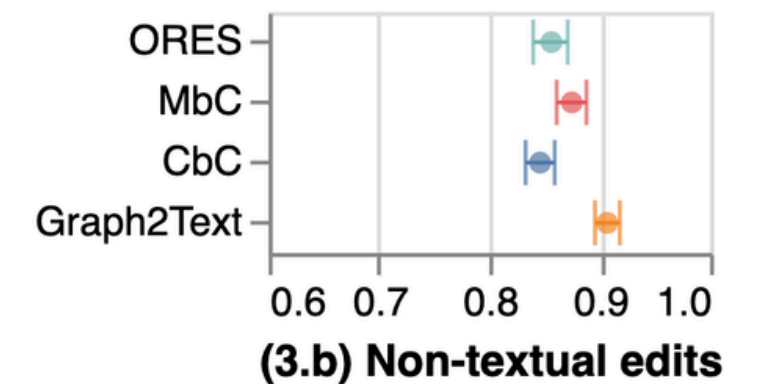
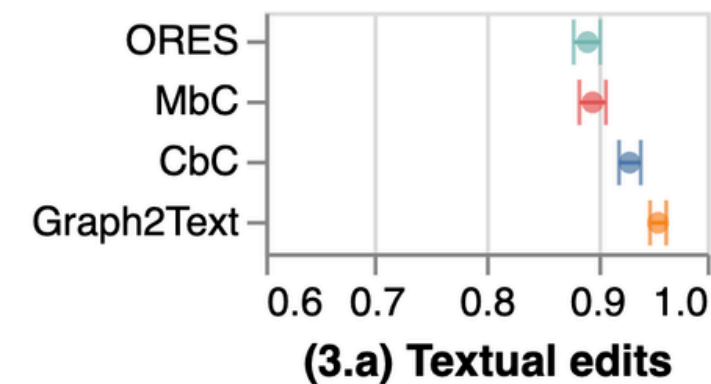
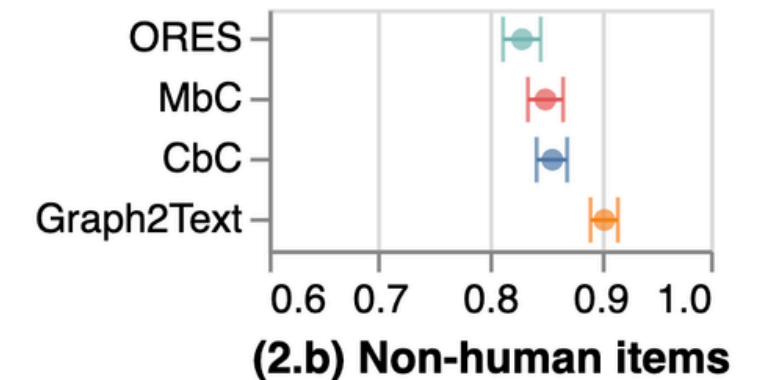
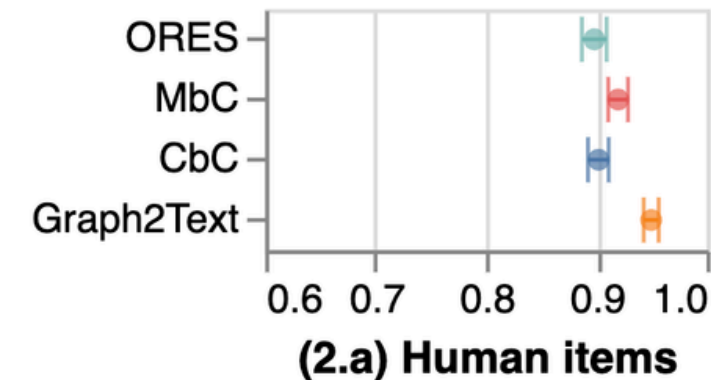
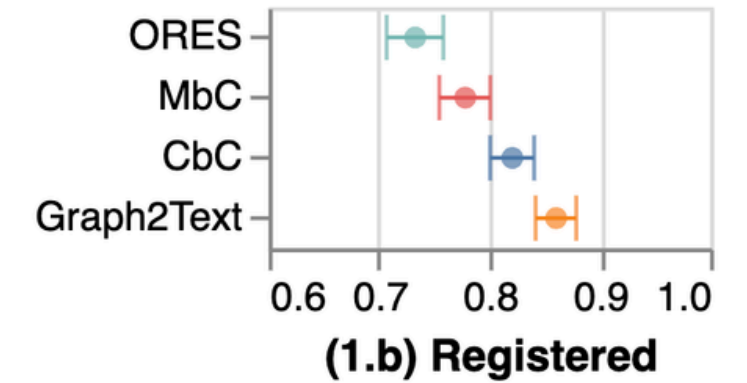
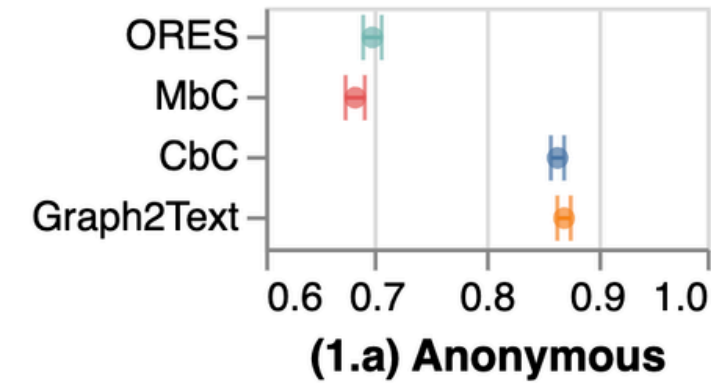
Disparate Impact Ratio (DIR)

$$DIR = \frac{Pr(\hat{Y}=1|D=unprivileged)}{Pr(\hat{Y}=1|D=privileged)}$$

Pr - probability

$\hat{Y}$ - predicted value,

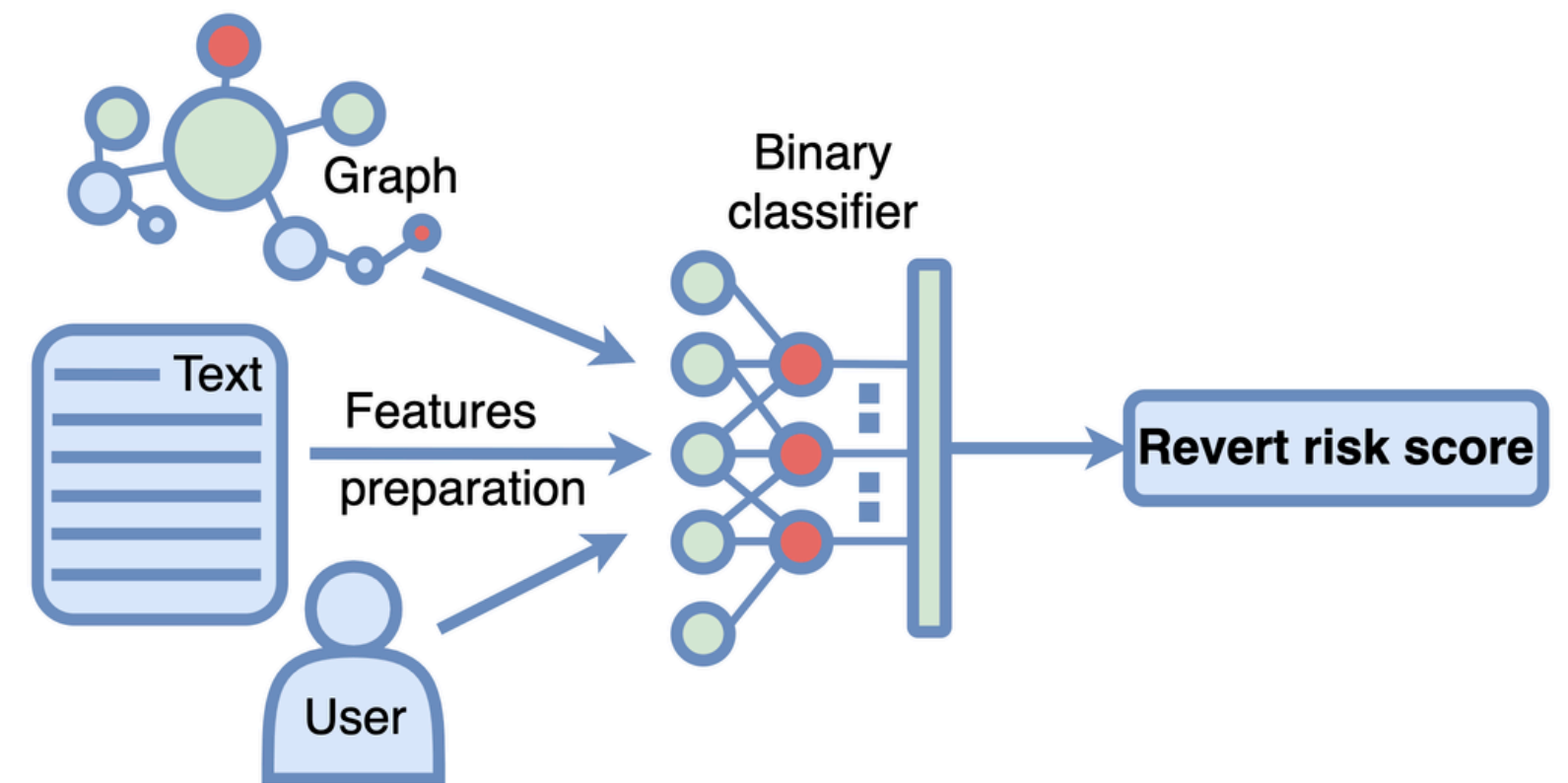
D - a group of users (anon. or registered)



Model performance (AUC) comparison across various Wikidata edit characteristics.

Summary

- **Next-generation vandalism detection system for Wikidata:**
 - **New open benchmark dataset** for vandalism detection in Wikidata, containing about 5M unique samples.
 - All content types are processed using a **single multilingual small language model**.
 - **System productionalization** addressing limitations imposed by **resource-constrained** infrastructure and product requirements.



Thank you!
