01

Fair multilingual vandalism detection system for Wikipedia

Mykola Trokhymovych	mykola.trokhymovych@upf.edu
Muniza Aslam	muniza-ctr@wikimedia.org
Ai-Jou Chou	aiko@wikimedia.org
Ricardo Baeza-Yates	rbaeza@acm.org
Diego Saez-Trumper	diego@wikimedia.org

KDD'23, Long Beach, CA, August 8, 2023

Agenda

1. Introduction
2. System design
3. Data preparation
4. Model training
5. Results

Introduction

- Wikipedia content is frequently used for powering other websites and products.
- Wikipedia has, on average, around **16** pages edited per second.
- Not all edits are good-faith (Figure 1). Bad-faith ones should be identified and reverted.
- Although some models help patrollers (like ORES), there are still open problems like model performance, language coverage, and fairness.

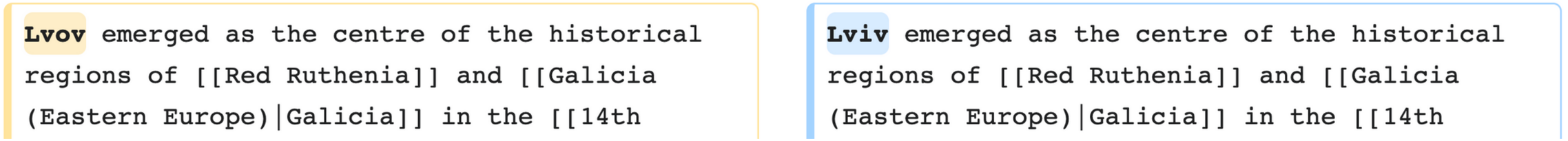

Figure 1. Enwiki revision 1149625753 adding biased narrative that is later reverted

04

Fairness challenge

- **Background**: Anonymous editors usually have a higher revert rate
- **Problem:** Models overfit, causing anonymous editors discrimination
- **Impact:** New/anonymous editors are not converting to active editors, so their number decreases in the long term.

anonymous users

05

Goal & contribution

- **Goal:** Create a model to help editors to identify edits that require patrolling.
- **Approach:** Use implicit annotations (reverts) to train the ML models
- **Contributions:**
  - Open-source multilingual model for content patrolling on Wikipedia, outperforming the state-of-the-art models;
  - Significantly increasing the number of languages covered in more than 60%;
  - Study of the biases of different models and discuss the trade-offs between performance and fairness;
  - Model inference productionalization and deployment;

System design

- Text features preparation:
  - Process wikitext and compare with parent revision
  - Extract mwedittypes* based features
  - Extract texts that were added, removed, and changed
- Masked Language Models (MLMs) features extraction:
  - Pass each of the texts that were added, removed, or changed to the pre-trained classification model
  - Apply mean and max pooling to the list of scores of each signal to extract the final unified feature set
- Final Classification
  - Combine all extracted features with user and revision metadata
  - Pass the features to the final classifier

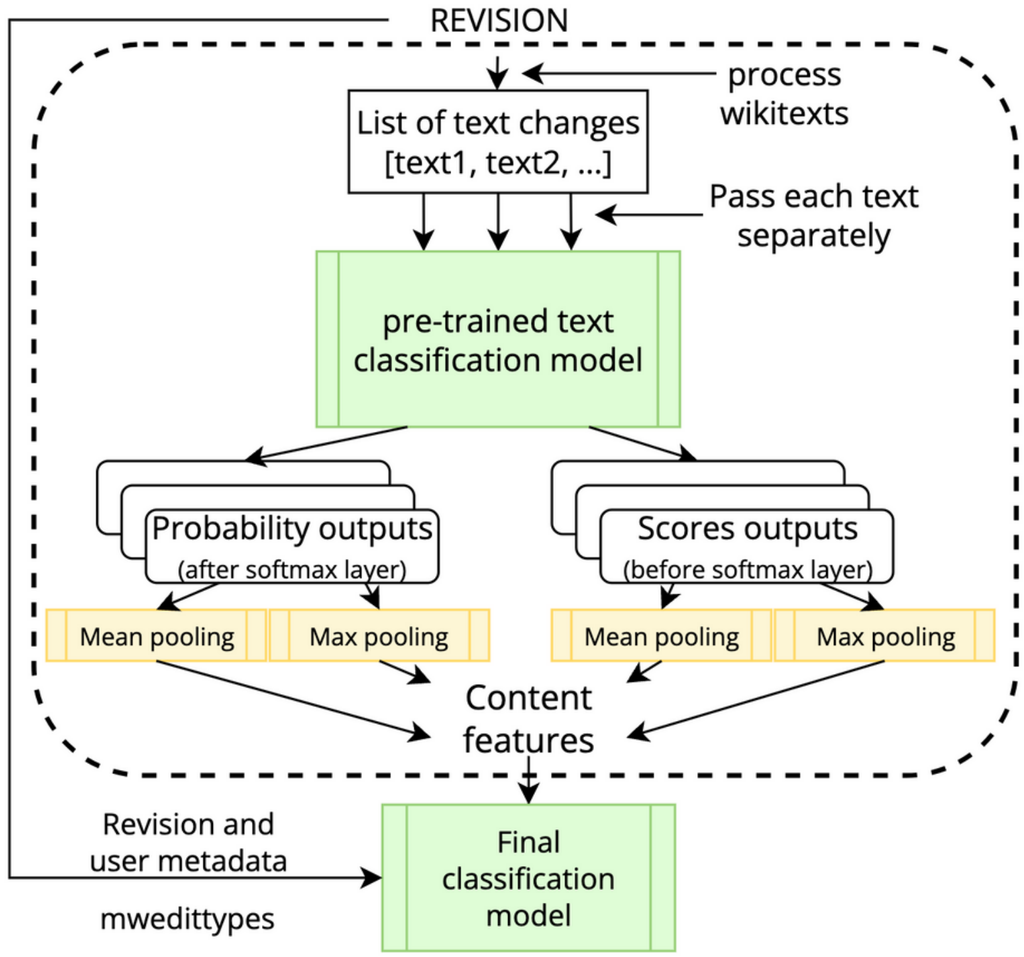

Figure 2. System design. Inference

Data preparation

Main characteristics of collected data:

- Using **mediawiki_history** and **mediawiki_wikitext_history**
- Collecting data for **47** most edited languages, except Kazakh, Portuguese, and Simple Wikipedia
- Snapshot dated **2022-07**
- The observation period is 2022-01-01 – 2022-07-01
- Additional only anonymous users dataset (IP edits)

Dataset	$\text{train}^{\text{anon}}$	$\text{train}^{\text{all}}$	test
Number of samples	3,693,571	8,586,362	1,079,265
Observation period	6 months	6 months	1 week
Anonymous rate	1.0	0.17	0.19
Revert rate	0.28	0.08	0.07

Data filtering

Filters applied:

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

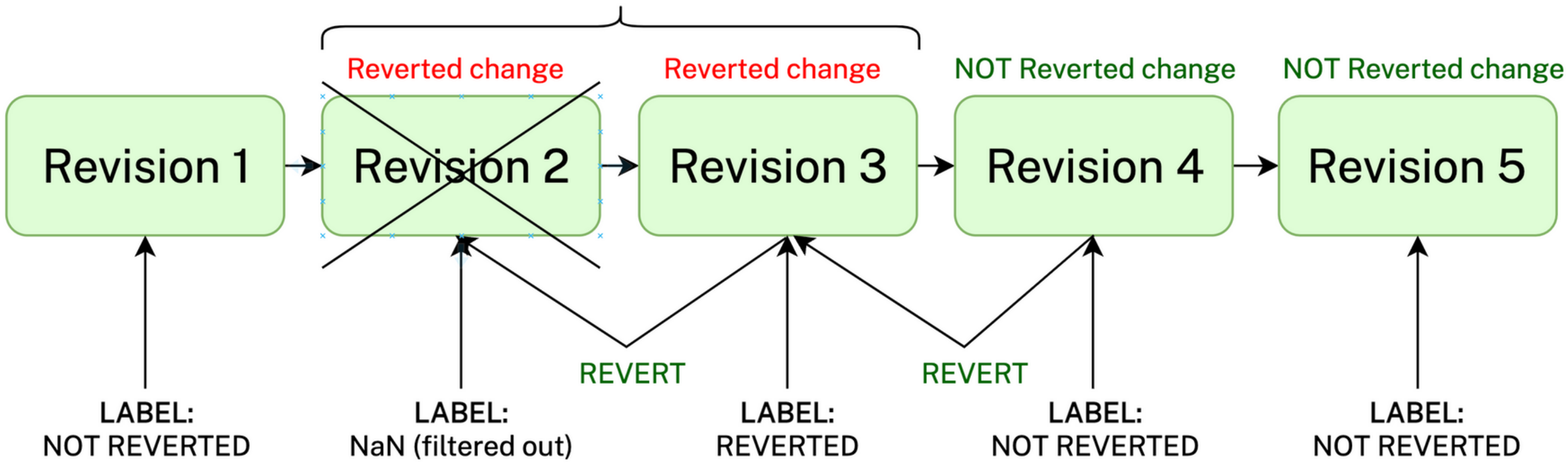

Figure 3. Revision-wars filtering logic

Text processing

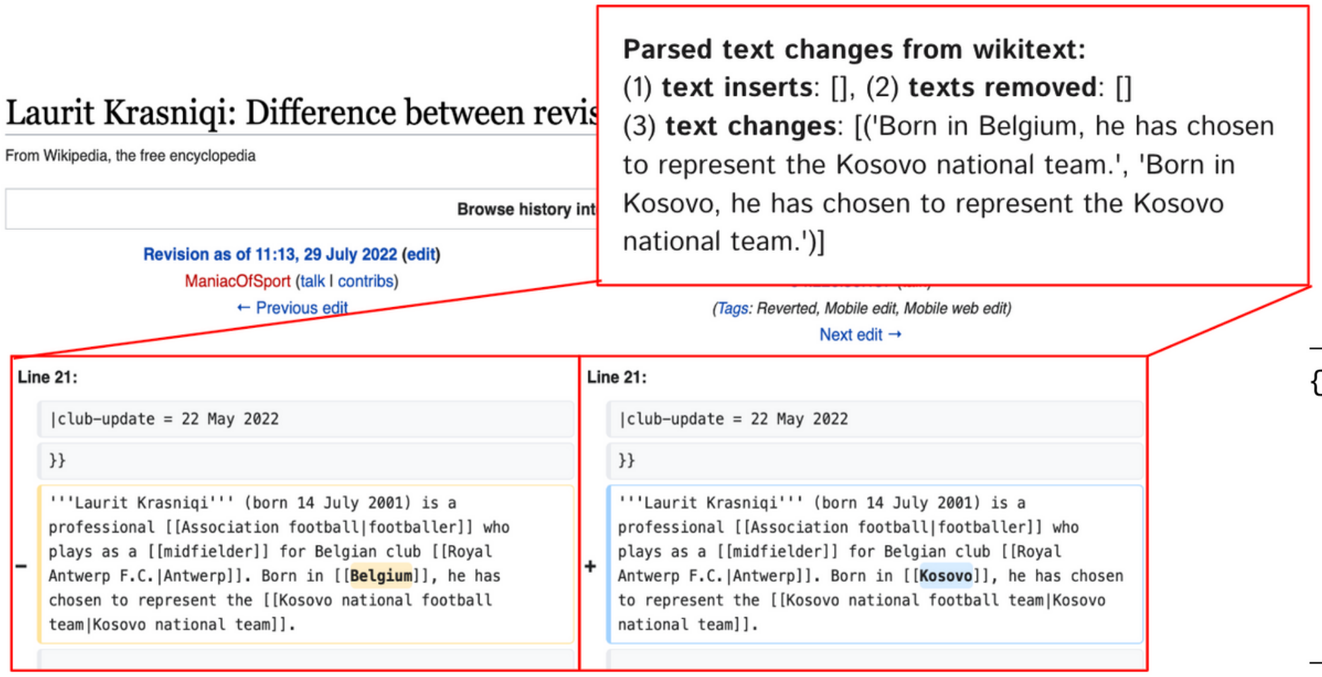

Figure 4. Text content changes extraction

```
{'change_Media': 0, 'insert_Media': 0, 'move_Media': 0,
     'remove_Media': 0, 'change_Punctuation': 0,
     'insert_Punctuation': 3, 'move_Punctuation': 0,
     'remove_Punctuation': 0, ..., 'change_Whitespace': 0,
     'insert_Whitespace': 9, 'move_Whitespace': 0,
     'remove_Whitespace': 0, 'change_Word': 0,
     'insert_Word': 9, 'move_Word': 0, 'remove_Word': 0 }
```

Figure 5. Example of mwedittypes features

Model training

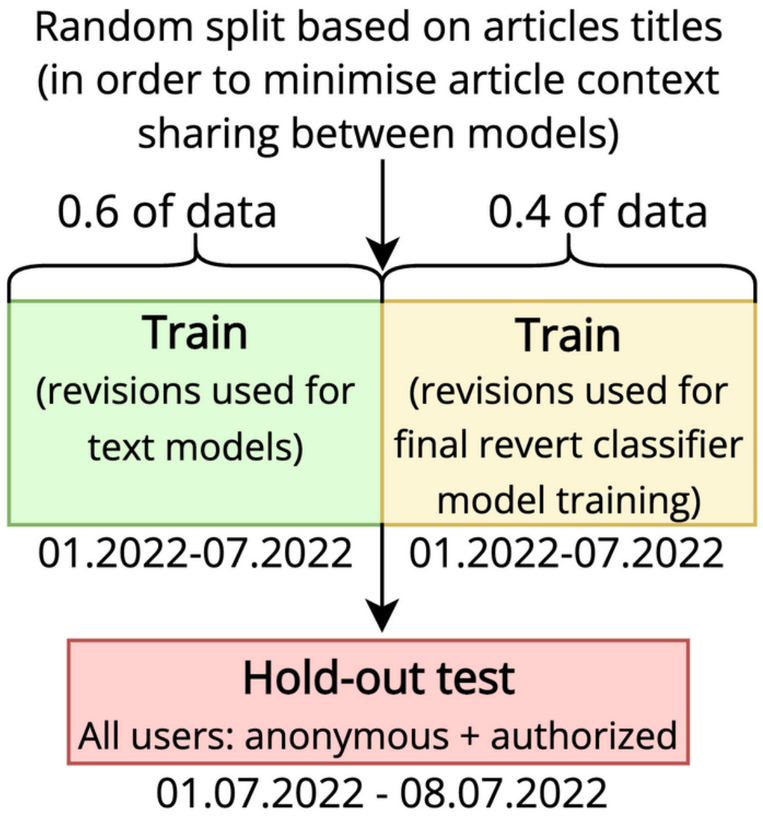

Figure 6. Data splitting logic

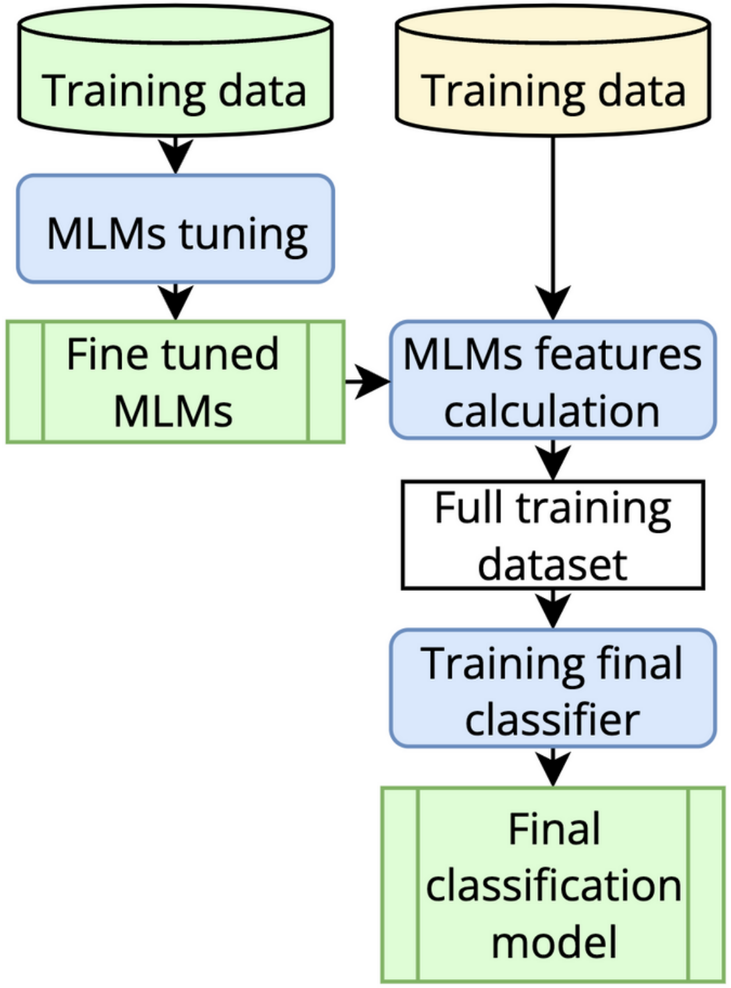

Figure 7. Training process pipeline

Performance metrics

Table: System performance on a test set of all users

Model	AUC	Pr@R0.75
Rule-based	0.75	0.07
ORES	0.84	0.22
Multilingual^{anon}	0.77	0.14
Multilingual^{anon} + MLM	0.79	0.15
Multilingual^{all}	0.82	0.18
Multilingual^{all} + MLM	0.84	0.20
Multilingual^{all} + user features	0.87	0.27
Multilingual^{all} + MLM & user features	**0.88**	**0.28**

Table: System performance on a test set of anonymous users

Model	AUC	Pr@R0.75
Rule-based	0.50	0.24
ORES	0.70	0.31
Multilingual^{anon}	0.77	0.40
Multilingual^{anon} + MLM	**0.80**	**0.44**
Multilingual^{all}	0.75	0.38
Multilingual^{all} + MLM	0.78	0.42
Multilingual^{all} + user features	0.76	0.39
Multilingual^{all} + MLM & user features	0.79	0.43

Infobox:

Multilingual^{all}
all users revisions metadata

Multilingual^{anon}
anonymous revisions metadata

MLM
Masked language models features

AUC
Area Under the ROC Curve

Pr@R0.75
Precision at Recall level 0.75

Performance metrics

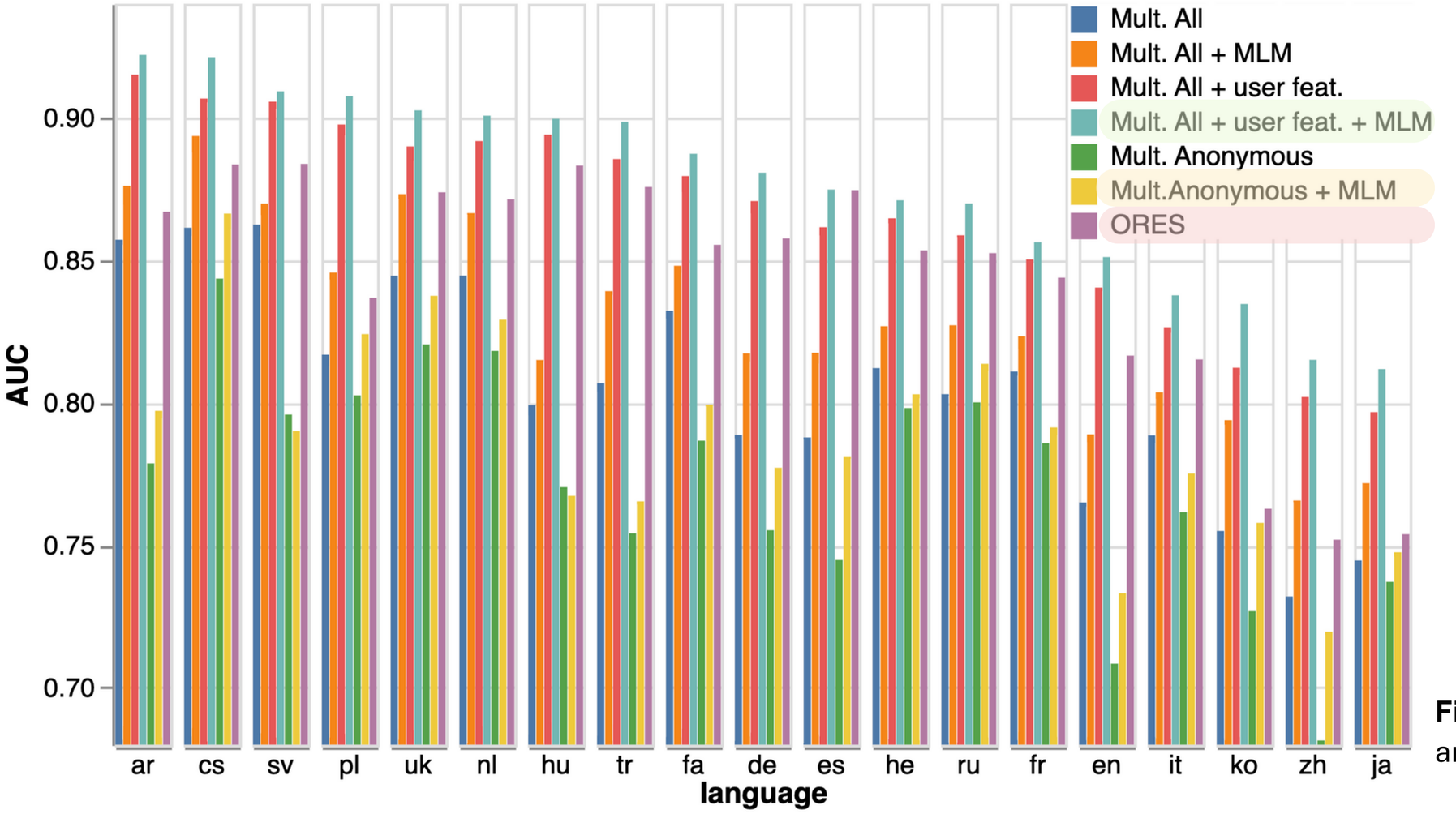

Figure 8. AUC score per model and language.

Fairness metrics

Table: Fairness metrics evaluation

Model	DIR	AUC diff
ORES	20.02	-0.043
Multilingual^{anon}	1.98	0.073
Multilingual^{anon} + MLM	2.06	0.084
Multilingual^{all}	2.91	0.010
Multilingual^{all} + MLM	3.08	0.017
Multilingual^{all} + user features	9.36	-0.035
Multilingual^{all} + MLM & user features	9.54	-0.017

Disparate Impact Ratio (DIR)

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

Pr - probability

$\widehat{Y}$ - predicted value,

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

DIR_{base} = 7.93, where for DIR_{base} we use Y (real value) instead of $\widehat{Y}$

AUC diff - difference between AUC scores of an unprivileged group (anon. users) and privileged (registered users)

Future work

- **Incorporating Non-Textual Changes**
  - Only ~56% of revisions have at least some changes in the text, so there is a need to analyze changes in media, tables, and other non-textual elements.
- **Comparative Analysis of Language Models:**
  - The study only explored one language model, so testing various language models is needed to determine how different they perform in detecting vandalism.
- **Increasing Language Diversity:**
  - Expanding the number of supported languages in the analysis
- **Temporal Evolution Analysis:**
  - Instead of focusing on fixed-time analysis, future research could take a longitudinal approach, tracking the evolution of Wikipedia pages and their revisions over an extended period.

Thank you!

Do you have any questions?

Contact:

Github: