Fair multilingual vandalism detection system for Wikipedia

This paper introduces a new generation of systems designed to help the Wikipedia community deal with vandalism on the platform.

Universitat Pompeu Fabra Barcelona

WIKIMEDIA FOUNDATION

Authors

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

Introduction

Wikipedia is a crucial web resource, frequently empowering websites and products. With around **16** pages edited per second, the platform is ever-changing. However, not all edits are made in good faith, requiring identifying and reversing bad-faith changes. While models like ORES help patrollers to fight vandalism, some challenges persist, like improving model performance, fairness, and language coverage to improve Wikipedia's knowledge integrity.

Objective

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

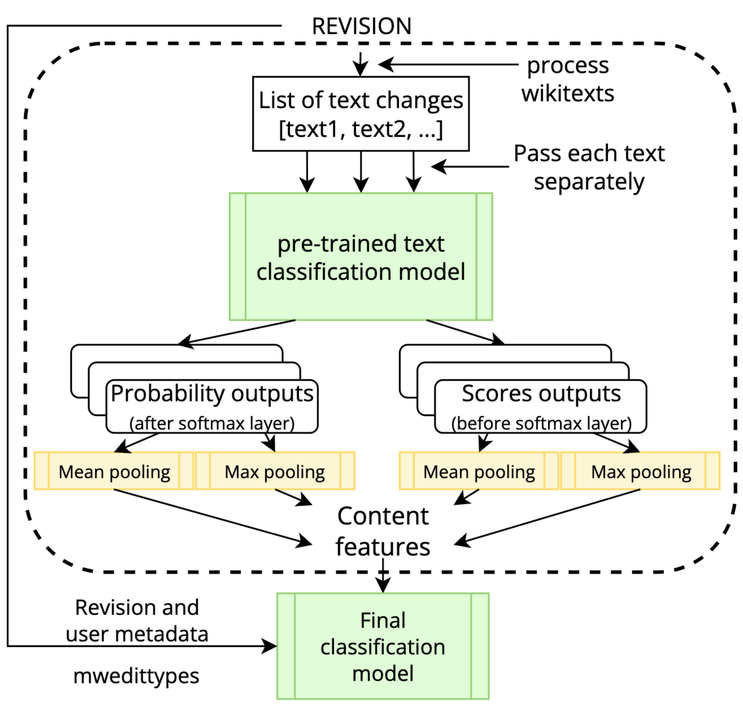
Contributions

- Introduction of an open-source, multilingual model for content patrolling on Wikipedia, outperforming the state-of-the-art models;
- Significantly increasing the number of languages covered by more than 60%;
- Study the biases of different models and discuss the trade-offs between performance and fairness;
- Model inference productionalization and deployment.

Approach

Use implicit annotations (reverts) to train the ML models

System design



Performance Metrics

Table: System performance on 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 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

Data

Table: Data characteristics

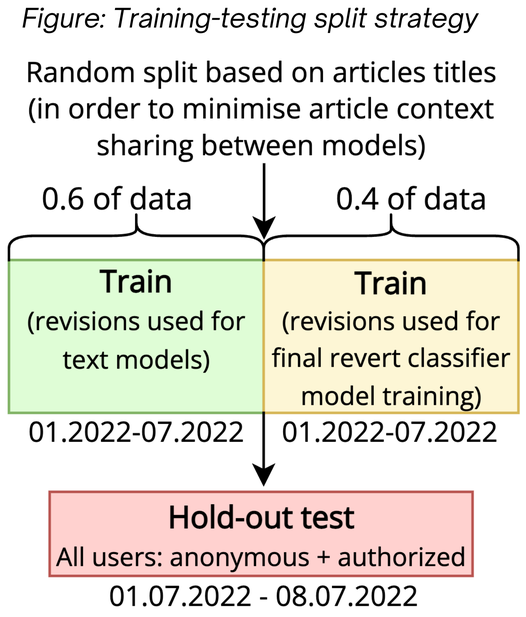
Dataset	train^{anon}	train^{all}	test
Number of samples	3,693,571	8,586,362	1,079,265
Observation period	6 months	6 months	1 week
Anonymous rate	1.0	0.17	0.19
Revert rate	0.28	0.08	0.07

Main characteristics of collected data:

- Using **mediawiki_history** and mediawiki_wikitext_history
- Collecting data for **47** most edited languages
- Snapshot dated **2022-07**
- The observation period is 2022-01-01 – 2022-07-01
- Filter for "revision-wars" (leave only those reverted revisions that were not later reverted)
- Filter revisions created by bots
- Additional only anonymous users dataset (IP edits)

Experiment

Figure: Training-testing split strategy



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)

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

Pr - probability
$\hat{Y}$ - 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 $\hat{Y}$

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

Additional Information

GitHub repo

Contact LinkedIn