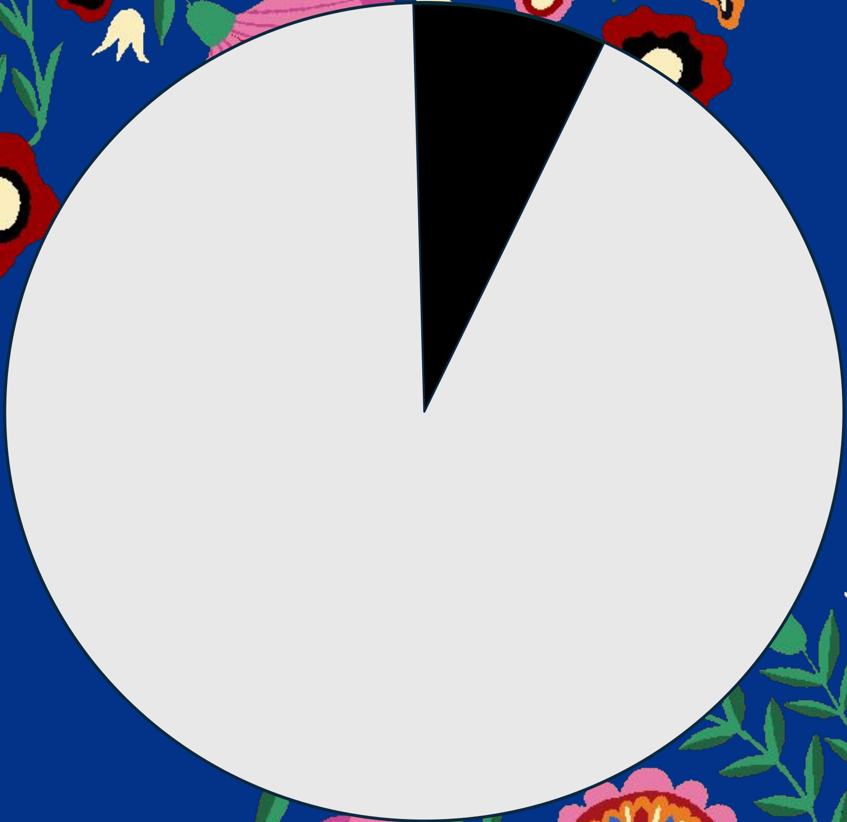
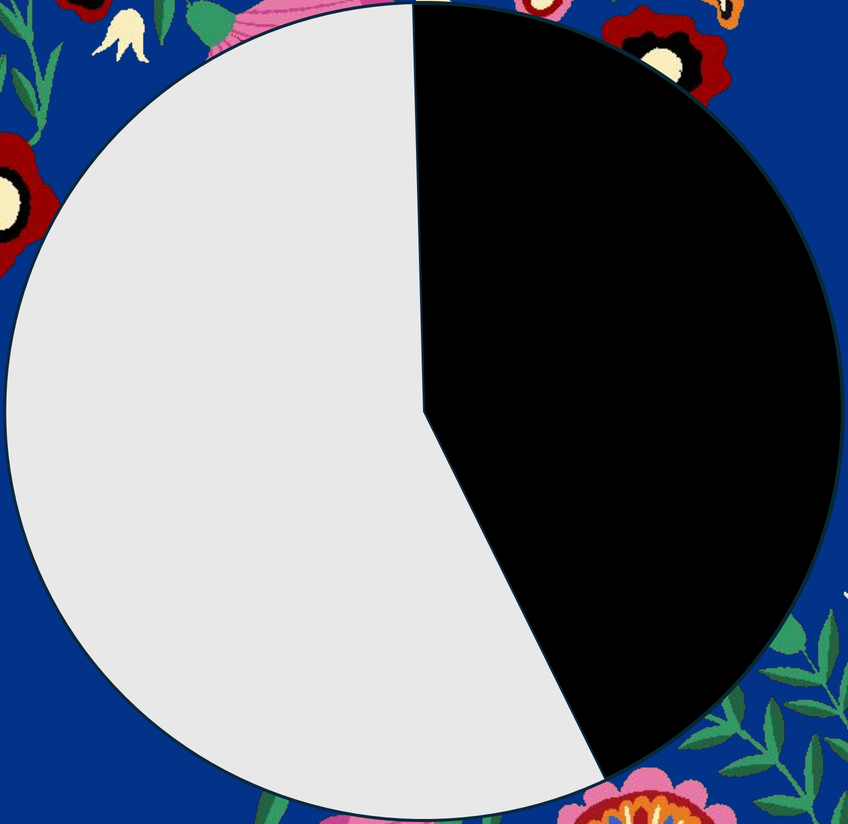
Revision quality algorithms: a subjective view

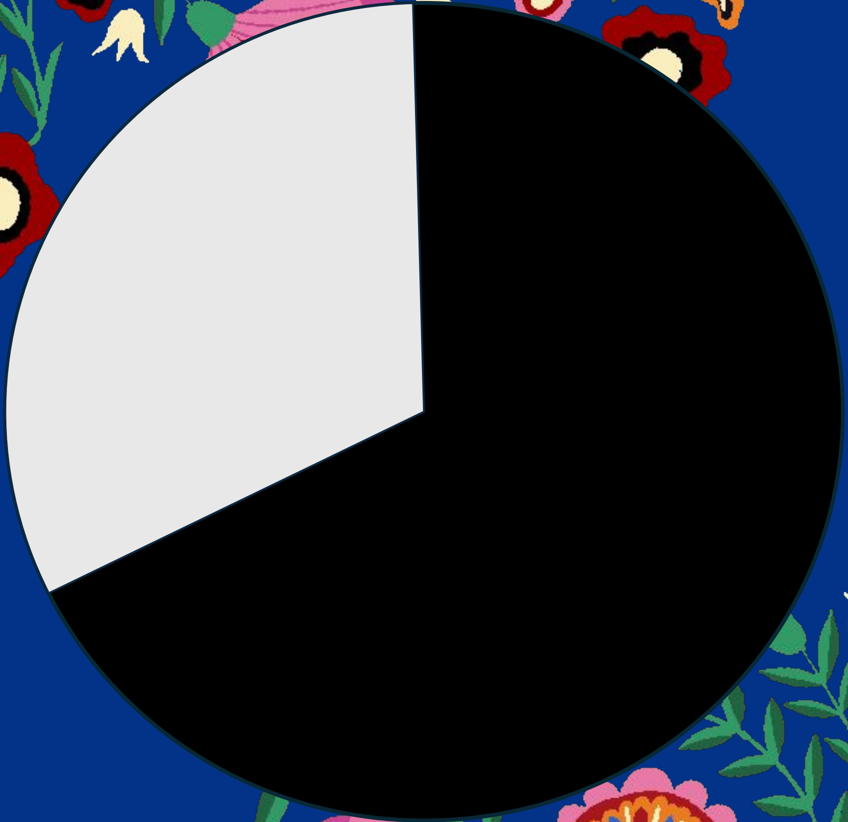
Strainu (wiki@strainu.ro)

WIKIMANIA
KATOWICE



WIKIMANIA
KATOWICE





WIKIMANIA
KATOWICE

Should I use Automoderator?

WIKIMANIA KATOWICE

Should I use Automoderator?

YES!

WIKIMANIA KATOWICE

Should I use Automoderator?

YES!

(unless you have a better option...)

WIKIMANIA
KATOWICE

Revision… what?

WIKIMANIA
KATOWICE

Some keywords...

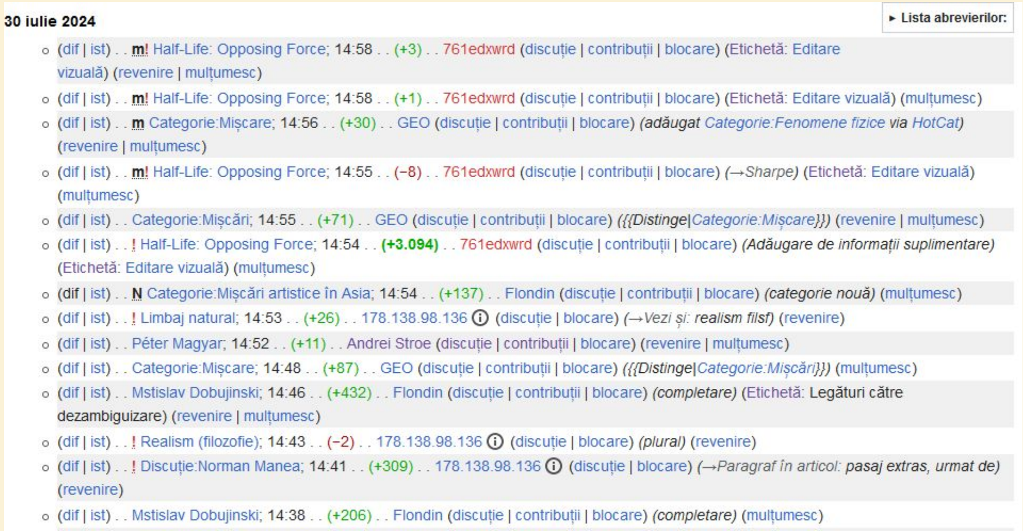
- ORES
- Revision scoring
- Revert risk
- Automoderator
- ClueBot

WIKIMANIA
KATOWICE

Automatic reverts! (AR)

WIKIMANIA
KATOWICE

How do AR work?

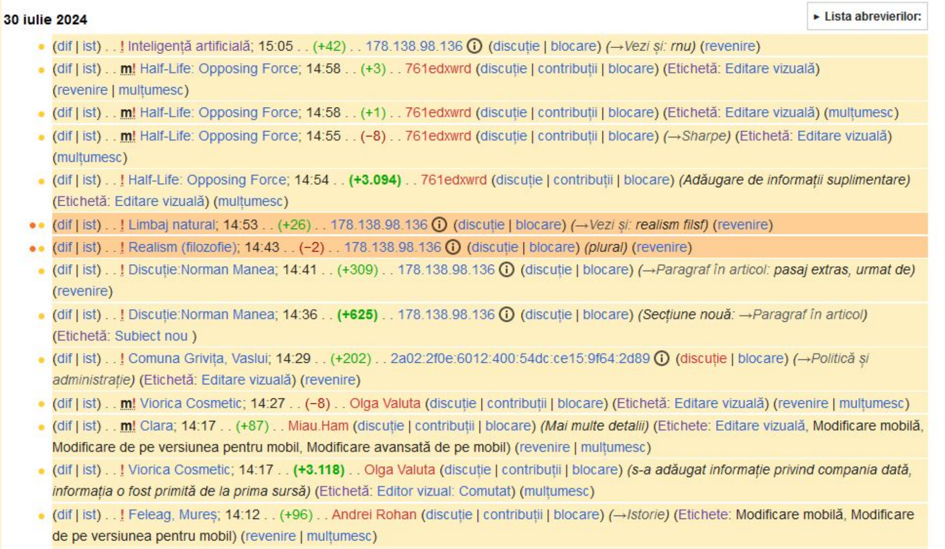


30 iulie 2024

- (dif | ist) . . m! Half-Life: Opposing Force; 14:58 . . (+3) . . 761edxwrd (discuție | contribuții | blocare) (Etichetă: Editare vizuală) (revenire | mulțumesc)
- (dif | ist) . . m! Half-Life: Opposing Force; 14:58 . . (+1) . . 761edxwrd (discuție | contribuții | blocare) (Etichetă: Editare vizuală) (mulțumesc)
- (dif | ist) . . m Categorie:Mișcare; 14:56 . . (+30) . . GEO (discuție | contribuții | blocare) *(adăugat Categorie:Fenomene fizice via HotCat)* (revenire | mulțumesc)
- (dif | ist) . . m! Half-Life: Opposing Force; 14:55 . . (−8) . . 761edxwrd (discuție | contribuții | blocare) *(→Sharpe)* (Etichetă: Editare vizuală) (mulțumesc)
- (dif | ist) . . Categorie:Mișcări; 14:55 . . (+71) . . GEO (discuție | contribuții | blocare) *({{Distinge|Categorie:Mișcare}})* (revenire | mulțumesc)
- (dif | ist) . . ! Half-Life: Opposing Force; 14:54 . . **(+3.094)** . . 761edxwrd (discuție | contribuții | blocare) *(Adăugare de informații suplimentare)* (Etichetă: Editare vizuală) (mulțumesc)
- (dif | ist) . . N Categorie:Mișcări artistice în Asia; 14:54 . . (+137) . . Flondin (discuție | contribuții | blocare) *(categorie nouă)* (mulțumesc)
- (dif | ist) . . ! Limbaj natural; 14:53 . . (+26) . . 178.138.98.136 ⓘ (discuție | blocare) *(→Vezi și: realism filsf)* (revenire)
- (dif | ist) . . Péter Magyar; 14:52 . . (+11) . . Andrei Stroe (discuție | contribuții | blocare) (revenire | mulțumesc)
- (dif | ist) . . Categorie:Mișcare; 14:48 . . (+87) . . GEO (discuție | contribuții | blocare) *({{Distinge|Categorie:Mișcări}})* (mulțumesc)
- (dif | ist) . . Mstislav Dobujinski; 14:46 . . (+432) . . Flondin (discuție | contribuții | blocare) *(completare)* (Etichetă: Legături către dezambiguizare) (revenire | mulțumesc)
- (dif | ist) . . ! Realism (filozofie); 14:43 . . (−2) . . 178.138.98.136 ⓘ (discuție | blocare) *(plural)* (revenire)
- (dif | ist) . . ! Discuție:Norman Manea; 14:41 . . (+309) . . 178.138.98.136 ⓘ (discuție | blocare) *(→Paragraf în articol: pasaj extras, urmat de)* (revenire)
- (dif | ist) . . Mstislav Dobujinski; 14:38 . . (+206) . . Flondin (discuție | contribuții | blocare) *(completare)* (mulțumesc)

WIKIMANIA
KATOWICE

How do AR work?



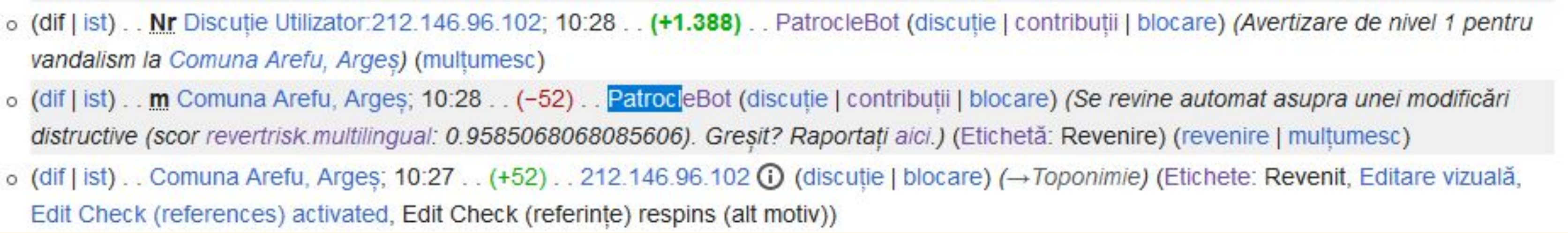
30 iulie 2024

▸ **Lista abrevierilor:**

- (dif | ist) . . ! Inteligență artificială; 15:05 . . (+42) . . 178.138.98.136 ⓘ (discuție | blocare) *(→Vezi și: rnu)* (revenire)
- (dif | ist) . . **m**! Half-Life: Opposing Force; 14:58 . . (+3) . . 761edxwrd (discuție | contribuții | blocare) (Etichetă: Editare vizuală) (revenire | mulțumesc)
- (dif | ist) . . **m**! Half-Life: Opposing Force; 14:58 . . (+1) . . 761edxwrd (discuție | contribuții | blocare) (Etichetă: Editare vizuală) (mulțumesc)
- (dif | ist) . . **m**! Half-Life: Opposing Force; 14:55 . . (−8) . . 761edxwrd (discuție | contribuții | blocare) *(→Sharpe)* (Etichetă: Editare vizuală) (mulțumesc)
- (dif | ist) . . ! Half-Life: Opposing Force; 14:54 . . **(+3.094)** . . 761edxwrd (discuție | contribuții | blocare) *(Adăugare de informații suplimentare)* (Etichetă: Editare vizuală) (mulțumesc)
- ●● (dif | ist) . . ! Limbaj natural; 14:53 . . (+26) . . 178.138.98.136 ⓘ (discuție | blocare) *(→Vezi și: realism filsf)* (revenire)
- ●● (dif | ist) . . ! Realism (filozofie); 14:43 . . (−2) . . 178.138.98.136 ⓘ (discuție | blocare) *(plural)* (revenire)
- (dif | ist) . . ! Discuție:Norman Manea; 14:41 . . (+309) . . 178.138.98.136 ⓘ (discuție | blocare) *(→Paragraf în articol: pasaj extras, urmat de)* (revenire)
- (dif | ist) . . ! Discuție:Norman Manea; 14:36 . . **(+625)** . . 178.138.98.136 ⓘ (discuție | blocare) *(Secțiune nouă: →Paragraf în articol)* (Etichetă: Subiect nou)
- (dif | ist) . . ! Comuna Grivița, Vaslui; 14:29 . . (+202) . . 2a02:2f0e:6012:400:54dc:ce15:9f64:2d89 ⓘ (discuție | blocare) *(→Politică și administrație)* (Etichetă: Editare vizuală) (revenire)
- (dif | ist) . . **m**! Viorica Cosmetic; 14:27 . . (−8) . . Olga Valuta (discuție | contribuții | blocare) (Etichetă: Editare vizuală) (revenire | mulțumesc)
- (dif | ist) . . **m**! Clara; 14:17 . . (+87) . . Miau.Ham (discuție | contribuții | blocare) *(Mai multe detalii)* (Etichete: Editare vizuală, Modificare mobilă, Modificare de pe versiunea pentru mobil, Modificare avansată de pe mobil) (revenire | mulțumesc)
- (dif | ist) . . ! Viorica Cosmetic; 14:17 . . **(+3.118)** . . Olga Valuta (discuție | contribuții | blocare) *(s-a adăugat informație privind compania dată, informația o fost primită de la prima sursă)* (Etichetă: Editor vizual: Comutat) (mulțumesc)
- (dif | ist) . . ! Feleag, Mureș; 14:12 . . (+96) . . Andrei Rohan (discuție | contribuții | blocare) *(→Istorie)* (Etichete: Modificare mobilă, Modificare de pe versiunea pentru mobil) (revenire | mulțumesc)

WIKIMANIA
KATOWICE

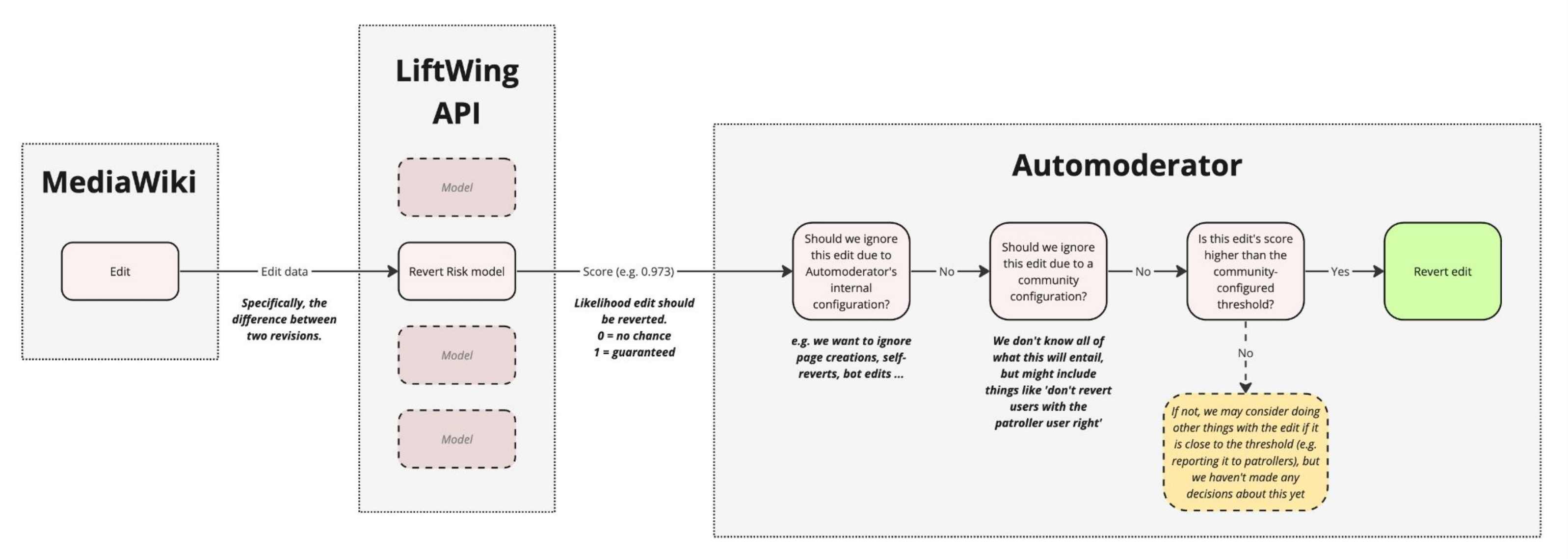
How do AR work?



- (dif | ist) . . Nr Discuție Utilizator:212.146.96.102; 10:28 . . (+1.388) . . PatrocleBot (discuție | contribuții | blocare) *(Avertizare de nivel 1 pentru vandalism la Comuna Arefu, Argeș)* (mulțumesc)
- (dif | ist) . . m Comuna Arefu, Argeș; 10:28 . . (−52) . . PatrocleBot (discuție | contribuții | blocare) *(Se revine automat asupra unei modificări distructive (scor revertrisk.multilingual: 0.9585068068085606). Greșit? Raportați aici.)* (Etichetă: Revenire) (revenire | mulțumesc)
- (dif | ist) . . Comuna Arefu, Argeș; 10:27 . . (+52) . . 212.146.96.102 ⓘ (discuție | blocare) *(→Toponimie)* (Etichete: Revenit, Editare vizuală, Edit Check (references) activated, Edit Check (referințe) respins (alt motiv))

WIKIMANIA
KATOWICE

How do AR work behind the scenes?



Specifically, the difference between two revisions.

Likelihood edit should be reverted.
0 = no chance
1 = guaranteed

Should we ignore this edit due to Automoderator's internal configuration?

e.g. we want to ignore page creations, self-reverts, bot edits ...

Should we ignore this edit due to a community configuration?

We don't know all of what this will entail, but might include things like 'don't revert users with the patroller user right'

Is this edit's score higher than the community-configured threshold?

If not, we may consider doing other things with the edit if it is close to the threshold (e.g. reporting it to patrollers), but we haven't made any decisions about this yet

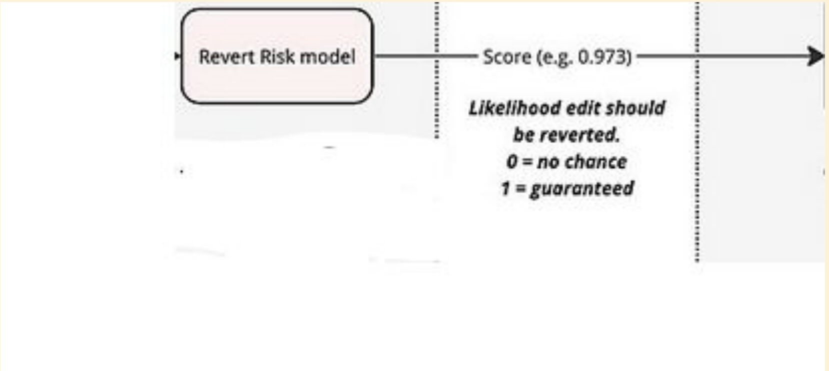
Image by Samwalton9 (WMF), CC-BY-SA-4.0
https://commons.wikimedia.org/wiki/File:Automoderator_and_Revert_Risk_decision_process.jpg

What is a “model”?

“A machine learning model is a type of mathematical model that, after being "**trained**" on a given dataset, can be used to **make predictions or classifications** on new data. During training, a learning algorithm iteratively adjusts the model's internal parameters to minimize errors in its predictions.” [[:en:Machine_learning#Models]]

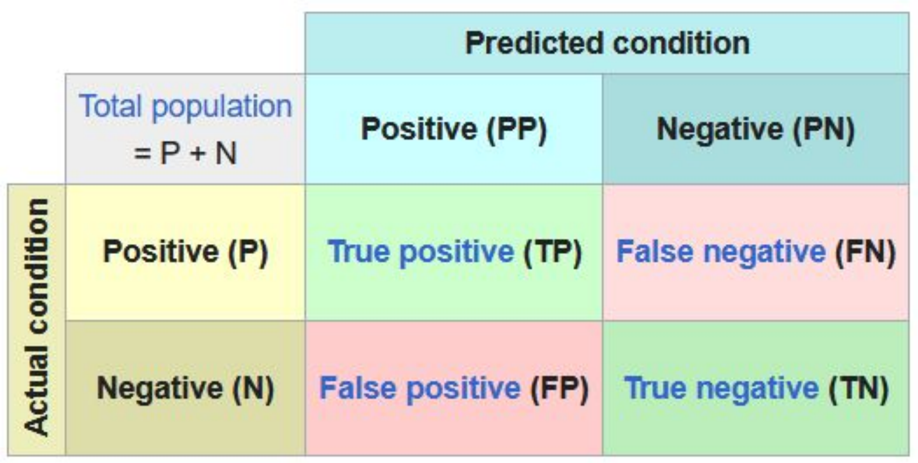
WIKIMANIA
KATOWICE

What is a "model"?



Likelihood edit should be reverted.
0 = no chance
1 = guaranteed

How to evaluate a

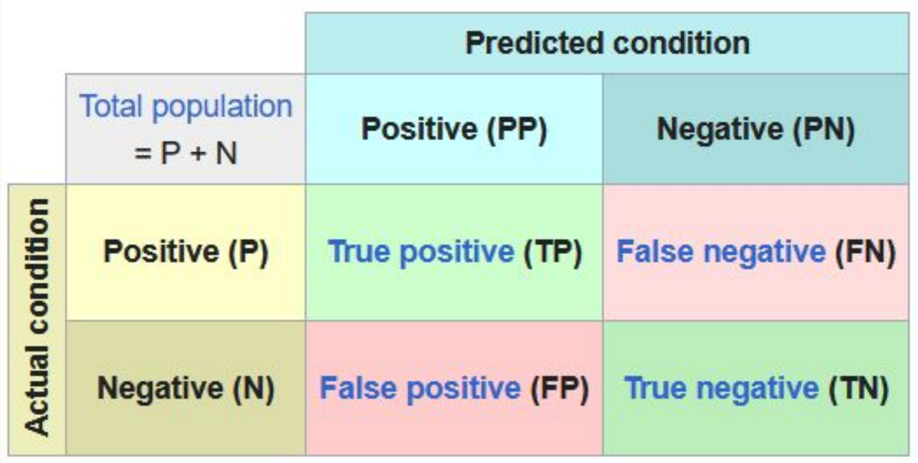


		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

[[:en:Confusion matrix]]

WIKIMANIA
KATOWICE

How to evaluate a



		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

[[:en:Confusion matrix]]

Precision = TP / (TP+FP)
Recall = TP / (TP+FN)
Accuracy = (TP+TN) / (P+N)

WIKIMANIA
KATOWICE

How to evaluate a model?

“How much of the work does it do for me?”

Automation = AR / TR

Proposal: (auto reverts + auto patrols) / unpatrolled edits

WIKIMANIA KATOWICE

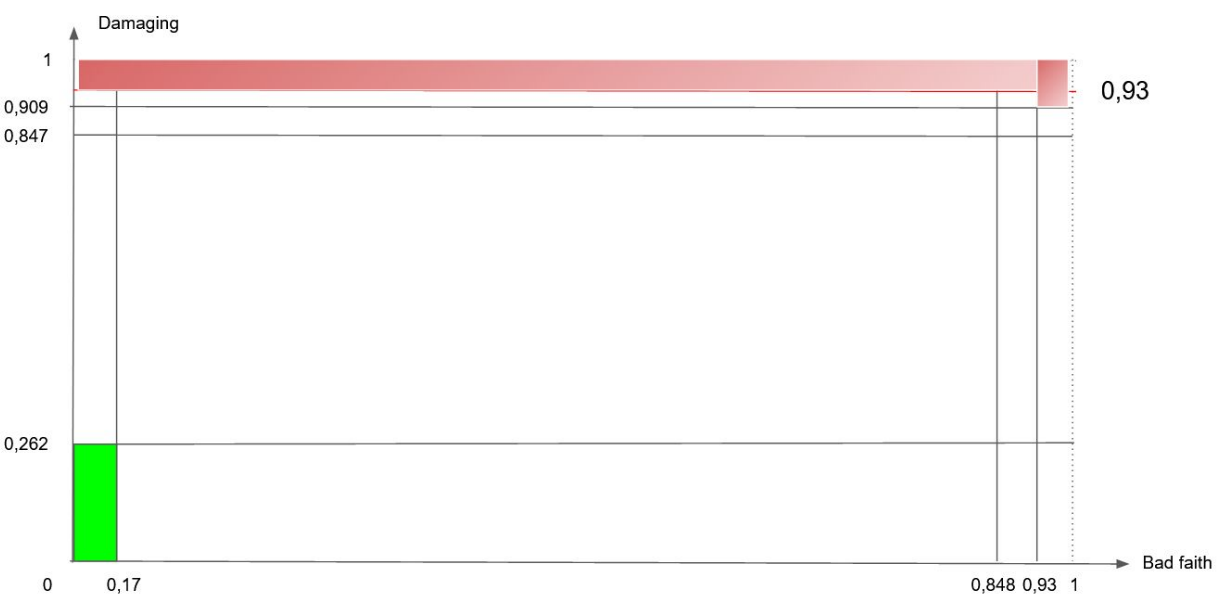
Existing models: ORES

- Almost 10 years old!
- (Unjustly) considered as “legacy” and deprecated
- Customized models per language - lots of training effort!
- Provides several types of scoring, the relevant ones for AR being **damaging** and **goodfaith**.
- Currently used in the recent changes
- Works on revisions (covers 100% of edits)
- Optimizes recall
- Heavily biased against anonymous users

https://www.mediawiki.org/wiki/ORES

WIKIMANIA KATOWICE

Existing models: ORES



WIKIMANIA KATOWICE

https://commons.wikimedia.org/wiki/File:PatrocleBot_rowiki_thresholds.png

Existing models: ORES

Patrol precision = 100%
Revert accuracy ~ 97%
Automation = 7.6%

Disclaimer: values calculated from historical stats, can be inaccurate

WIKIMANIA
KATOWICE

Existing models: language agnostic revert risk (RRLA)

- Uses more modern research than ORES
- Works everywhere without additional training
  - this is why Automoderator chose it
  - target accuracy = 90%
- Uses diffs, not revisions
  - doesn't work for new articles / non-articles
  - overall reach = 70% of revisions
- Biased on purpose against new/anon users

https://meta.wikimedia.org/wiki/Machine_learning_models/Proposed/Language-agnostic_revert_risk

WIKIMANIA
KATOWICE

Existing models: RRLA

Threshold	Not cautious (0.97)	Low caution (0.975)	**Somewhat cautious (0.98)**	**Cautious (0.985)**	**Very cautious (0.99)**
Accuracy	75%	82%	93%	95%	100%
Automation	8.2%	5.5%	3.6%	1.8%	1.8%

Accuracy and rowiki automation per WMF testing
Might be affected by local tooling!

Existing models: multilingual revert risk (RRML)

- Uses more modern research than ORES
- Works in 47 languages
- Uses diffs, not revisions
  - doesn’t work for new articles / non-articles
  - overall reach = 70% of revisions
- Mitigates the bias against new/anon users
- Slower than the language-agnostic version

https://meta.wikimedia.org/wiki/Machine_learning_models/Proposed/Multilingual_revert_risk

WIKIMANIA
KATOWICE

Existing models: RRML

Patrol precision = 100%
Revert accuracy ~ 97%
Automation = 22.2% (!)

Results from rowiki with .95 threshold for revert, same as ORES for patrol

RRML vs RRLA vs ORESd

Took ~3K unpatrolled changes on rowiki, ran both algorithms with thresholds at 0.93 and 0.95 and asked patrollers to evaluate if they would have reverted or not.

original discussion

analysis from WMF Reasearch

WIKIMANIA
KATOWICE

RRML vs RRLA vs ORESd

Took random samples of recent changes and checked scores vs actual reverts

WIKIMANIA KATOWICE

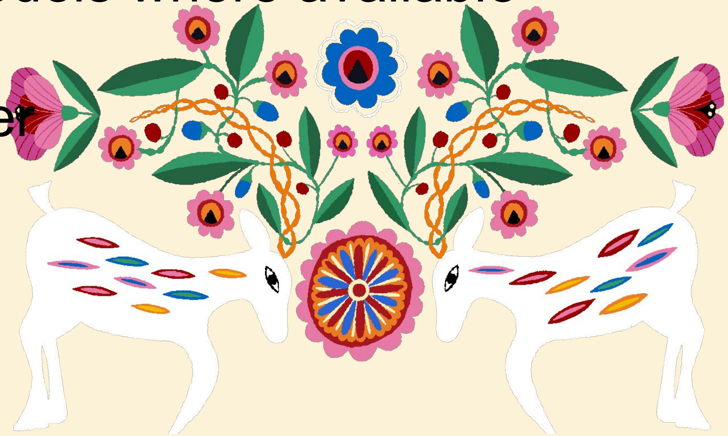
ML vs LA vs ORES conclusions

- RRML wins every time on every metric
- RRLA has decent accuracy, but poor precision
- ORES damaging by itself is about the same as RRLA

WIKIMANIA KATOWICE

What does all this mean for you?

- Automated reverts are an awesome tool for patrollers
- Automoderator seems to be a promising, balanced tool, but with limited impact → we should lobby WMF to improve the models further and use more models where available
- If you have technical knowledge in the community, there are better options out there.



WIKIMANIA KATOWICE

Questions

WIKIMANIA
KATOWICE