Presentation

Claire Sun

2022.10.20

Overview

- Introduction
- Exploratory Data Analysis
- Feature Selection and Classification
- Next Steps

Introduction

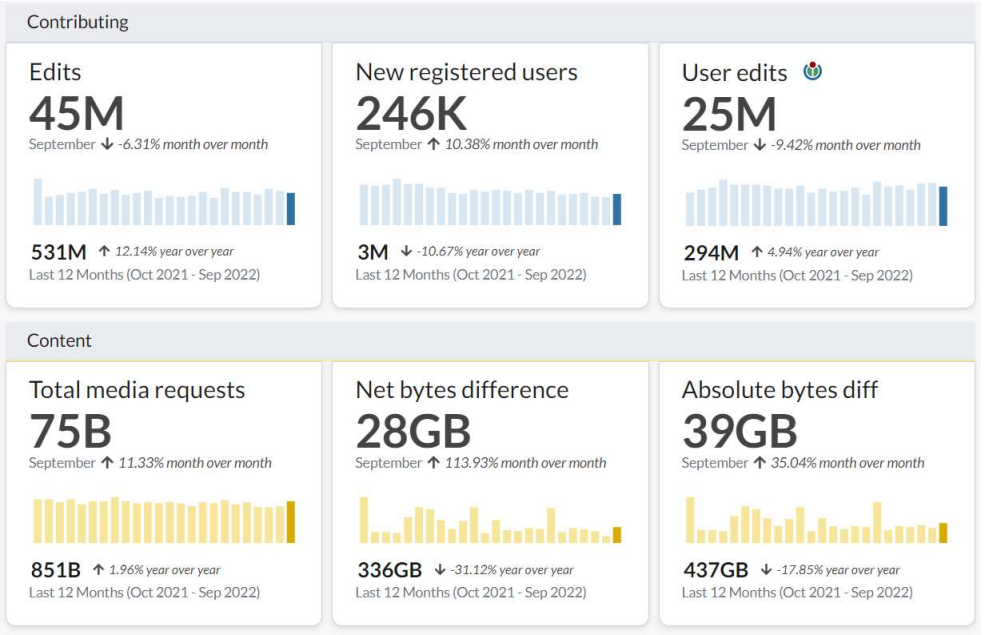
- Wikimedia Open Source Project:
  - Improving the Machine Learning based tools to support Wikipedia Patrollers
  - Developing a new language-agnostic model to detect revisions that require patrollers' attention
  - Building a web app that allows users to give explicit feedback on the quality of the model’s recommendations
- Disclaimer
  - A learning journey: just getting started
  - Share my exploration and preliminary findings
  - Get some helpful feedback and advice

Note:
1. Vandalism refers to blatantly damaging edits that are routinely made to articles, such as adding hate speech, gibberish, or humor; see https://enwp.org/WP:VD

Background

- Wikipedia:
  - free online encyclopedia, created and edited by volunteers
- Vandalism[1]:
  - quality control, edit reversion, Wikipedia communal patrol
- RC patrol:
  - individual users check the recent changes of various articles for inappropriate edits

Monthly Overview (September 2022) [2]



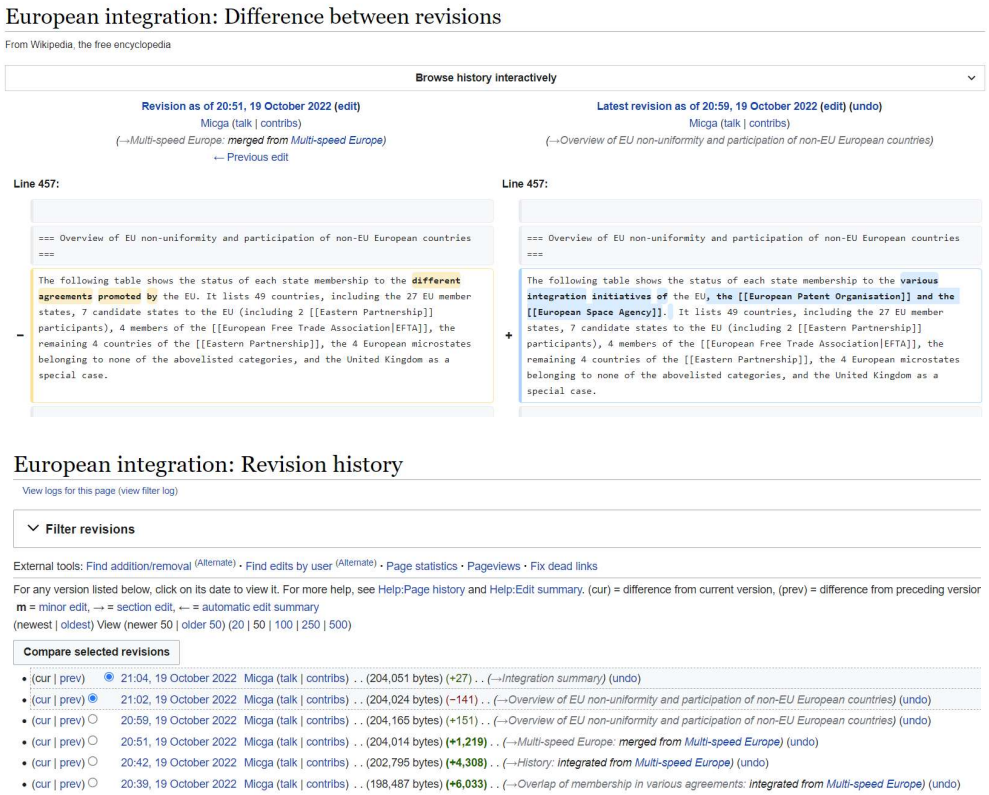
Note:

1. Vandalism refers to blatantly damaging edits that are routinely made to articles, such as adding hate speech, gibberish, or humor; see https://enwp.org/WP:VD
2. Source: https://stats.wikimedia.org/#/all-projects

Illustration: Recent changes patrol

RC Patrol: 4-step process

- Identify "bad" or "needy" edits
- Remove or improve the edit
- Warn the editor
- Check the user's other contributions



European integration: Difference between revisions

From Wikipedia, the free encyclopedia

Browse history interactively

Revision as of 20:51, 19 October 2022 (edit)
Micga (talk | contribs)
(→Multi-speed Europe: merged from Multi-speed Europe)
← Previous edit

Latest revision as of 20:59, 19 October 2022 (edit) (undo)
Micga (talk | contribs)
(→Overview of EU non-uniformity and participation of non-EU European countries)

Line 457:

```
=== Overview of EU non-uniformity and participation of non-EU European countries ===
- The following table shows the status of each state membership to the different agreements promoted by the EU. It lists 49 countries, including the 27 EU member states, 7 candidate states to the EU (including 2 [[Eastern Partnership]] participants), 4 members of the [[European Free Trade Association|EFTA]], the remaining 4 countries of the [[Eastern Partnership]], the 4 European microstates belonging to none of the abovelisted categories, and the United Kingdom as a special case.
```

Line 457:

```
=== Overview of EU non-uniformity and participation of non-EU European countries ===
+ The following table shows the status of each state membership to the various integration initiatives of the EU, the [[European Patent Organisation]] and the [[European Space Agency]]. It lists 49 countries, including the 27 EU member states, 7 candidate states to the EU (including 2 [[Eastern Partnership]] participants), 4 members of the [[European Free Trade Association|EFTA]], the remaining 4 countries of the [[Eastern Partnership]], the 4 European microstates belonging to none of the abovelisted categories, and the United Kingdom as a special case.
```

European integration: Revision history

View logs for this page (view filter log)

Filter revisions

External tools: Find addition/removal (Alternate) · Find edits by user (Alternate) · Page statistics · Pageviews · Fix dead links

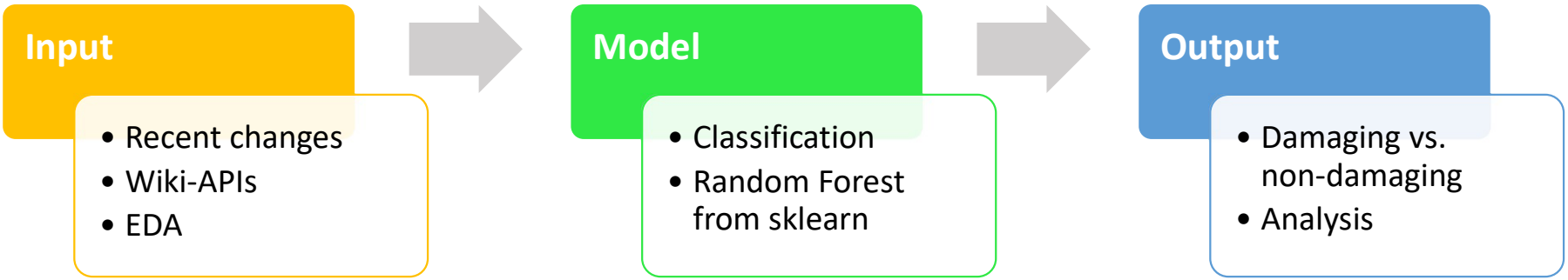
For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary. (cur) = difference from current version, (prev) = difference from preceding version
m = minor edit, → = section edit, ← = automatic edit summary
(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) 21:04, 19 October 2022 Micga (talk | contribs) . . (204,051 bytes) (+27) . . (→Integration summary) (undo)
- (cur | prev) 21:02, 19 October 2022 Micga (talk | contribs) . . (204,024 bytes) (−141) . . (→Overview of EU non-uniformity and participation of non-EU European countries) (undo)
- (cur | prev) 20:59, 19 October 2022 Micga (talk | contribs) . . (204,165 bytes) (+151) . . (→Overview of EU non-uniformity and participation of non-EU European countries) (undo)
- (cur | prev) 20:51, 19 October 2022 Micga (talk | contribs) . . (204,014 bytes) **(+1,219)** . . (→Multi-speed Europe: merged from Multi-speed Europe) (undo)
- (cur | prev) 20:42, 19 October 2022 Micga (talk | contribs) . . (202,795 bytes) **(+4,308)** . . (→History: integrated from Multi-speed Europe) (undo)
- (cur | prev) 20:39, 19 October 2022 Micga (talk | contribs) . . (198,487 bytes) **(+6,033)** . . (→Overlap of membership in various agreements: integrated from Multi-speed Europe) (undo)

Objectives

- Ultimate Goals:
  - Develop a ML model to predict revisions that require patrollers' attention
  - Building a web app that allows users to give explicit feedback
- Mini-Milestone:
  - Understand and familiarize with the data set and its features
  - Build a data pipeline and simple classification model
  - Perform preliminary data analysis and visualization



MediaWiki APIs and Dataset

- RecentChanges API[1]:
  - GET request to fetch all the recent changes, in the same manner as the web interface
- Other tools:
  - mwapi[2] : Python wrapper around MediaWiki API
  - mwedittypes[3] : transform unstructured edits into a structured summary of changes
  - mwrevert[4]: detect reverting activity
- Dataset
  - First 5000 edits made to enwiki articles from 06/Oct/22, 16:00 UTC

Note:
1. https://www.mediawiki.org/wiki/API:RecentChanges
2. https://github.com/mediawiki-utilities/python-mwapi
3. https://github.com/geohci/edit-types#mwedittypes
4. https://github.com/mediawiki-utilities/python-mwreverts

Response Properties[1]

user:	Adds the user responsible for the edit and tags if they are an IP. If the user has been revision deleted, a `userhidden` property will be returned.
userid:	Adds the user ID responsible for the edit. If the user has been revision deleted, a `userhidden` property will be returned.
comment:	Adds the comment for the edit. If the comment has been revision deleted, a `commenthidden` property will be returned.
parsedcomment:	Adds the parsed comment for the edit. If the comment has been revision deleted, a `commenthidden` property will be returned.
flags:	Adds flags for the edit.
timestamp:	Adds timestamp of the edit.
title:	Adds the page title of the edit.
ids:	Adds the page ID, recent changes ID and the new and old revision ID.
sizes:	Adds the new and old page length in bytes.
redirect:	Tags edit if page is a redirect.
patrolled:	Tags patrollable edits as being patrolled or unpatrolled.
loginfo:	Adds log information (log ID, log type, etc) to log entries.
tags:	Lists tags for the entry.
sha1:	Adds the content checksum for entries associated with a revision. If the content has been revision

Exploratory Data Analysis (0/5)

Snapshot of the Raw Data Collected

	mo…	title	pageid	revid	user	timestamp	old_revid	old_user	old_timesta…	ContentDiff	comment	tags	change_size
1	False	Steve Orlan…	57068776	1114458…	Omnipaedista	2022-10-06…	1113229…	FilmLover72	2022-09-30…	{'List': {'insert': 1}, 'Se…	/* External links */	['wikieditor']	25750.0
2	False	Rockstar Li…	7365848	1114458…	PerryPerryD	2022-10-06…	1112107…	IceWelder	2022-09-24…	{'Heading': {'remove':…	/* Bibliography *…	['visualeditor-w…	23722.0
3	False	Burma Vall…	994296	1114458…	Vanspoof	2022-10-06…	950859137	Tom Radul…	2020-04-14…	{'Template': {'insert': …	#suggestededit-…	['mobile edit', '…	1194.0
4	False	Clarknova	7881567	1114458…	SimLibrarian	2022-10-06…	1079337…	Premeditat…	2022-03-26…	{'Template': {'insert': …		['mobile edit', '…	5188.0
5	False	Wang Lixiong	12928235	1114458…	Randy Kryn	2022-10-06…	1019746…	GreenC bot	2021-04-25…	{'Heading': {'change':…	/* Social Activiti…	[]	12084.0
6	False	Decimated	22530565	1114458…	Dudhhr	2022-10-06…	1114458…	Pauliswv	2022-10-06…	{'ExternalLink': {'rem…	[[WP:ROLLBAC…	['mw-rollback', …	37.0
7	False	Nadodikkattu	2268987	1114458…	155.43.78.165	2022-10-06…	1113369…	Commons…	2022-10-01…	{'Category': {'insert': …	/* External links */	['wikieditor']	17481.0
8	False	Bread	36969	1114458…	JoeNMLC	2022-10-06…	1110669…	Pinnerup	2022-09-16…	{'Template': {'insert': …	/* See also */ ce…	['wikieditor']	52031.0
9	False	Arun Govil	6686251	1114458…	2409:4050:2…	2022-10-06…	1114458…	2409:4050:…	2022-10-06…	{'Wikilink': {'change': …		['mobile edit', '…	14705.0
10	False	List of cons…	35098494	1114458…	Muzzzmuzz…	2022-10-06…	1114458…	Muzzzmuz…	2022-10-06…	{'Section': {'change': …		['mobile edit', '…	58877.0

```
Number of unique articles: 3222

Top 10 articles by number of edits:
title
Landon Collins                  19
Cihan Erdal                     18
IPhone 14 Pro                   17
Tibetan art                     16
RT (TV network)                 16
List of Hindi films of 1964     15
Paolo Venini                    15
Palakkad North                  13
Art's Way                       13
WES Commuter Rail               12
```

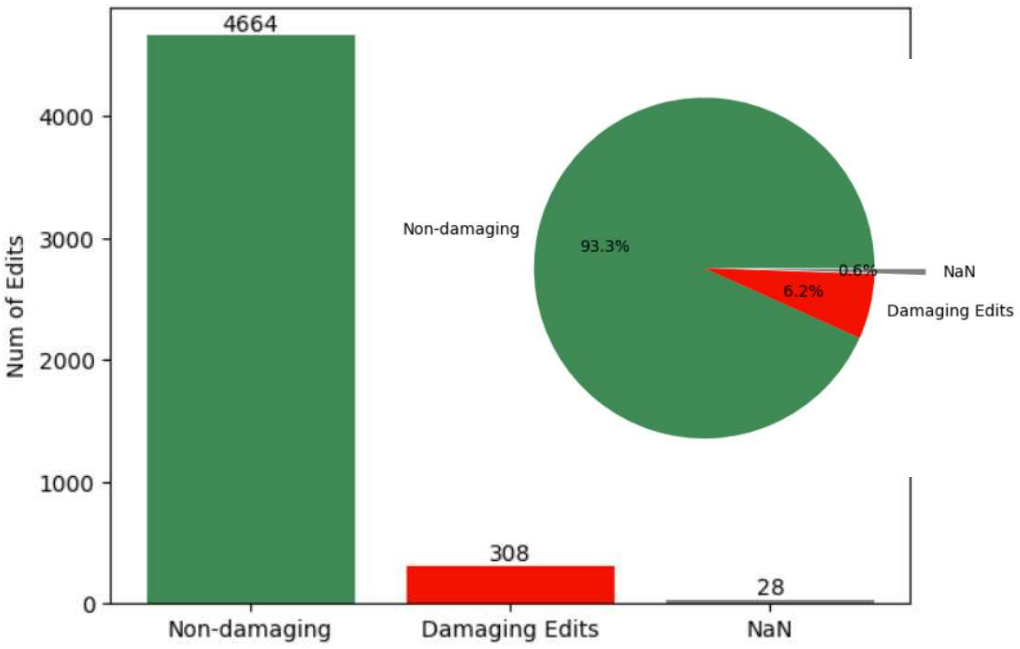
```
Number of unique users: 1750

Top 10 users by number of edits:
user
FrescoBot          139
WaddlesJP13         56
HugoAcosta9         55
RobertskySemi       41
Coimenda            34
Datu Hulyo          32
AnomieBOT           31
LBHSMantaRay        30
Denisarona          26
Wesoree             26
```

Questions on Damaging Edits:

- *What? – Proportion*
- *Where? – Titles*
- *Who? – User*
- *When? – Timestamps*
- *How? – Content Difference*

Exploratory Data Analysis (1/5) – What?

Distribution of Damaging vs Non-Damaging Edits



- Damaging Edits:
  - Edits that have been reverted
- Non-damaging:
  - Good edits
  - There might be a need for further modification but based on good-faith
- NaN
  - Data status not available

c. 6% of all edits are identified as damaging

Exploratory Data Analysis (2/5) – Where?

Top 10 Titles by Number of Damaging Edits

```
Top 10 Titles by number of damaging edits:
title
RT (TV network)                                                  7
The Lord of the Rings: The Rings of Power                        3
Kallekkad                                                        3
Bee Gees                                                         3
Natural number                                                   3
Jeffrey Sachs                                                    3
Patrick White                                                    2
Dhilip Subbarayan                                                2
Bubble gum                                                       2
Twenty-fifth Amendment to the United States Constitution         2
```

Top 10 Titles by Number of Total Edits

title	total_edits	damaging_edits	pct_damaging
Landon Collins	19	1	0.052632
RT (TV network)	16	7	0.437500
Rorschach (film)	11	1	0.090909
Turkish Radio and Television Corporation	9	1	0.111111
The Lord of the Rings: The Rings of Power	8	3	0.375000
Manchester City F.C. supporters	6	1	0.166667
Kallekkad	6	3	0.500000
Megamind	6	2	0.333333
Bee Gees	6	3	0.500000
Sanju Samson	6	1	0.166667

- 270 unique titles are subjected to damaging edits, c. 8% of total unique titles
- The more popularly edited titles are more likely to be the targets of damaging edits

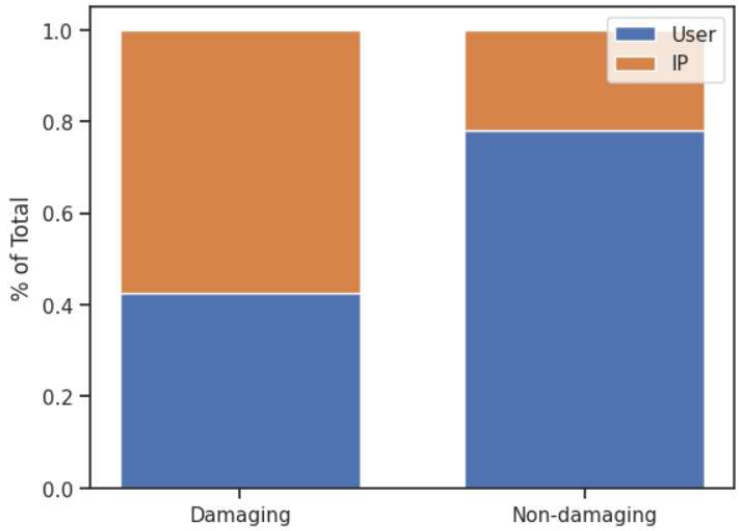
Exploratory Data Analysis (3/5) – Who?

Top 10 Users Responsible for Damaging Edits

Number of unique users responsible f

Top 10 users by number of damaging e

user	
Masterofeditingwiki	14
80.233.45.111	13
46.39.45.187	10
CactiStaccingCrane	7
76.113.134.153	6
Smilus32	5
93.107.88.143	4
No Fake News Allowed	4
Balib1011	3
41.113.161.91	3



Top 10 Users by Number of Edits

Number of unique users: 1750

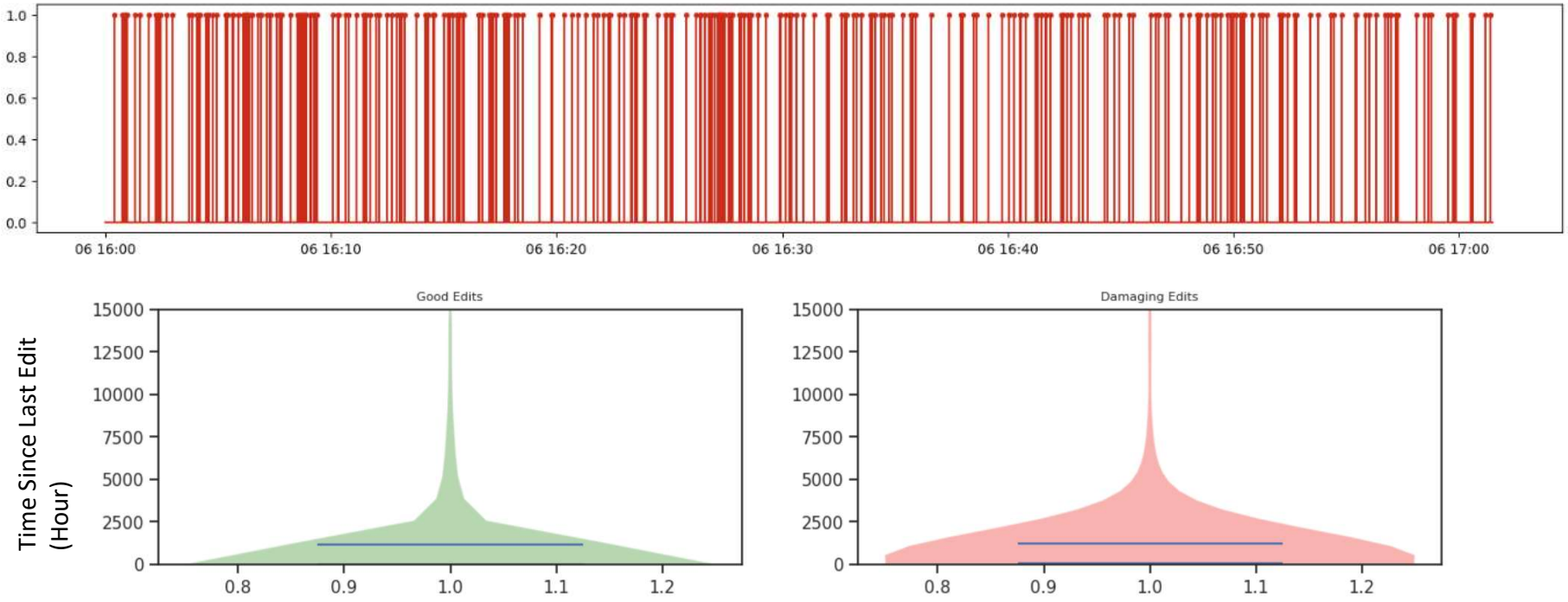
Top 10 users by number of edits:

user	
FrescoBot	139
WaddlesJP13	56
HugoAcosta9	55
RobertskySemi	41
Coimenda	34
Datu Hulyo	32
AnomieBOT	31
LBHSMantaRay	30
Denisarona	26
Wesoree	26

- 216 unique users are responsible for damaging edits, c. 12% of all unique users
- Some of the users responsible for the damaging edits are anonymous (i.e., only known by an IP address); in contrast, most of the good-faith users have a proper user ID.
- We could potentially use the user ID as a feature to identify damaging edits. However, it should also be noted that some good edits are also done by IP-address based users.

Exploratory Data Analysis (4/5) – When?

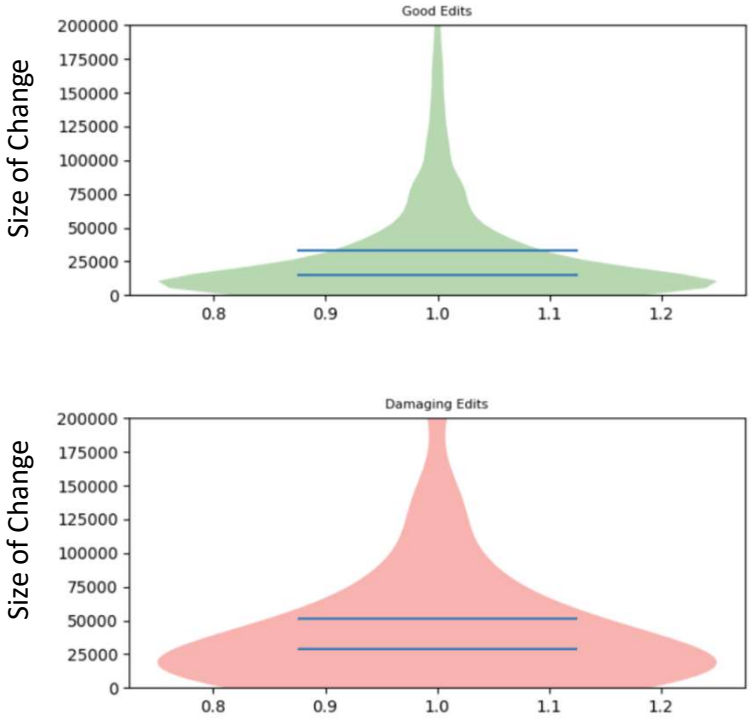
Frequency of damaging edits occurring

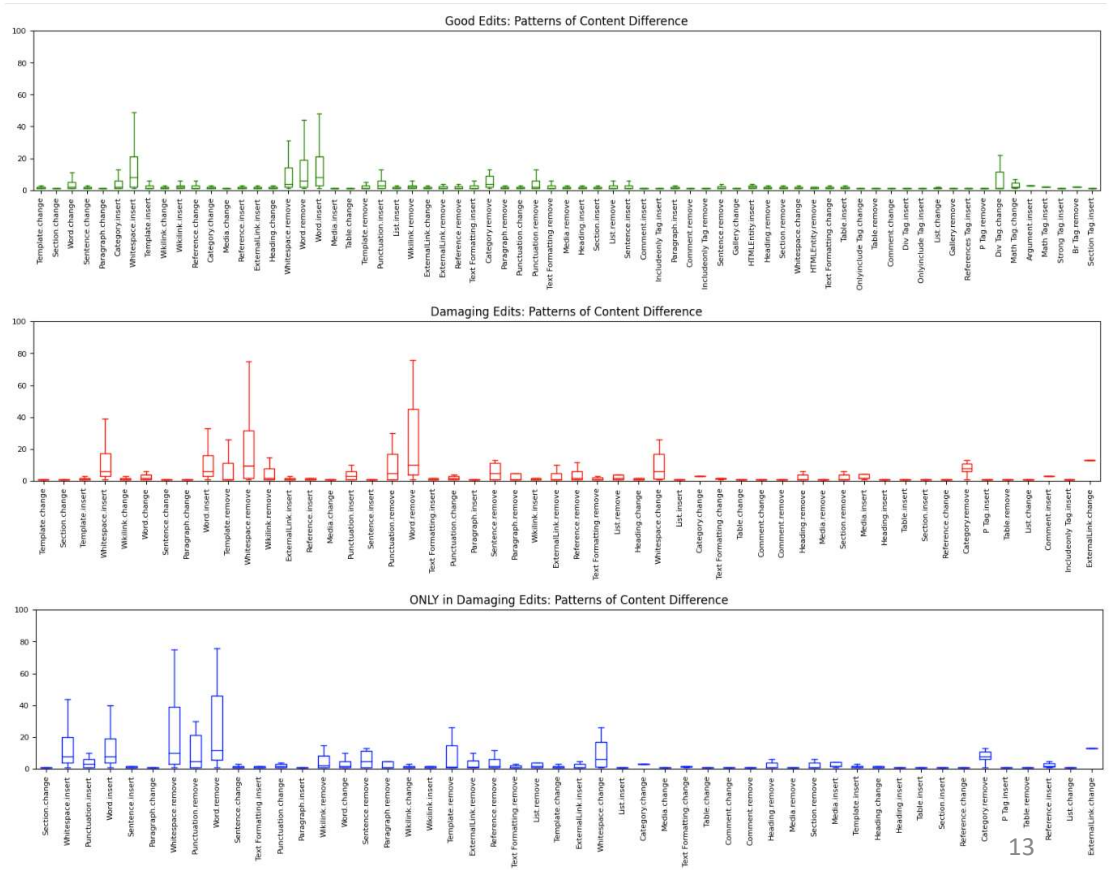


Timestamp: The damaging edits occurred throughout the hour with no particularly noticeable frequency pattern

Exploratory Data Analysis (5/5) – How?

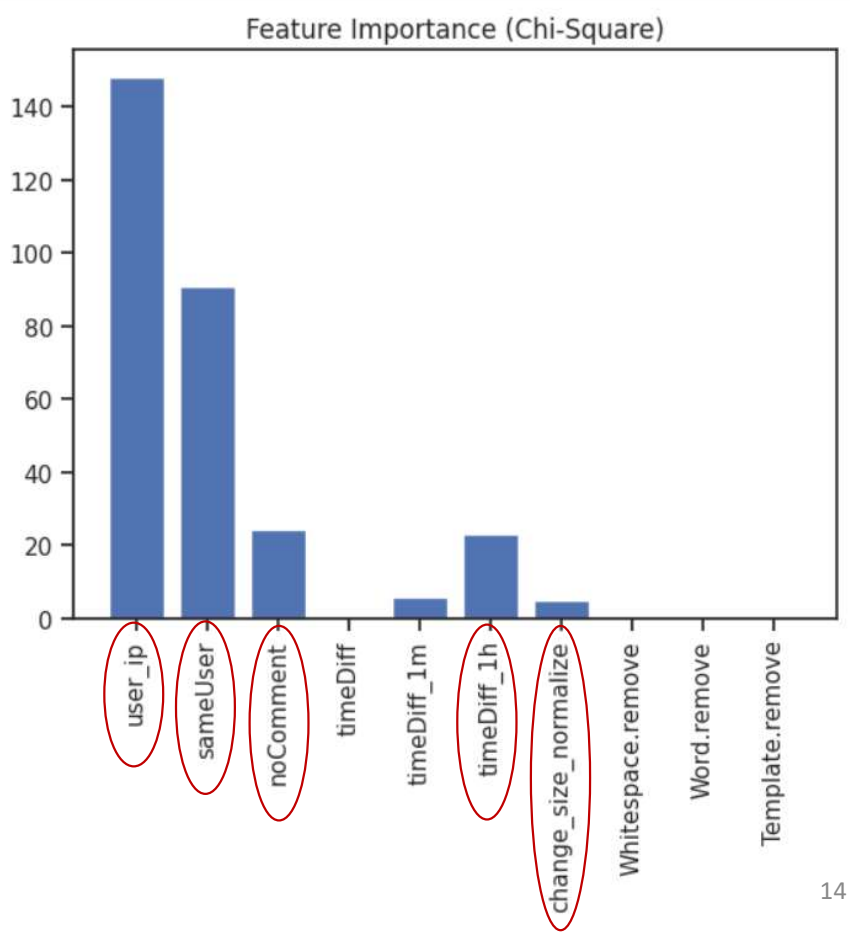
Size of Change and Content Difference





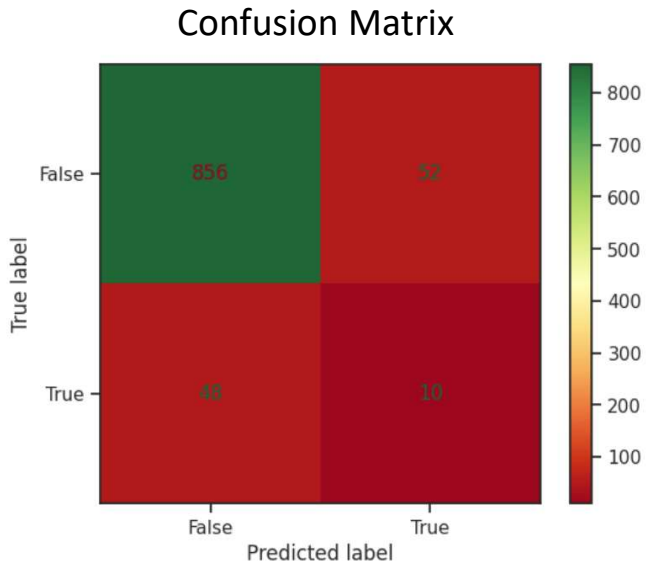
Feature Selection

- Select features according chi-square test between features and class
  - Chi-square measures dependence between stochastic variables
  - Filter out the features that are the most likely to be independent of class and irrelevant for classification
- Top 5 features:
  - **user_ip**: is user ID an IP address
  - **sameUser:** is user the same as previous edit user
  - **noComment**: is there an empty comment associated with this edit
  - **timeDiff_1h**: is the time laps between this edit and the previous edit more than 1 hour
  - **change_size_normalize**



Random Forest Classifier

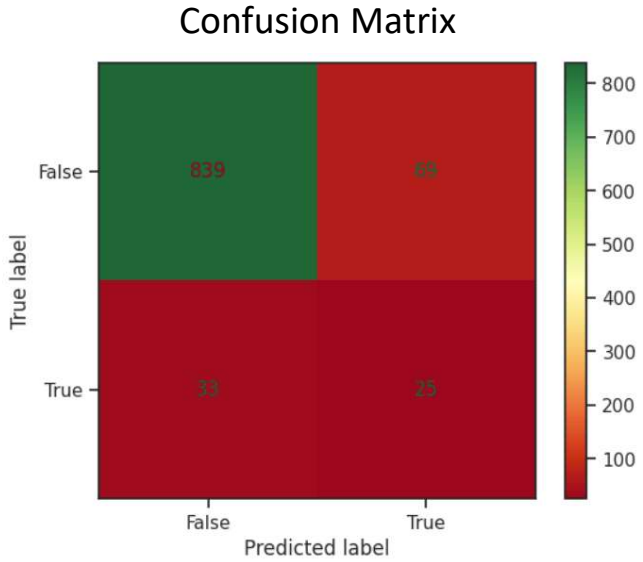
- Random Forest Classifier
  - A meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset
  - Uses averaging to improve the predictive accuracy and control over-fitting
  - Grid search for parameter settings
- Dataset:
  - Training vs test split: 4000 / 966
  - 5-fold cross-validation
- Results:
  - Precision and recall are both very low
  - Need to minimize false negatives



	precision	recall	f1-score	support
False	0.95	0.94	0.95	908
True	0.16	0.17	0.17	58
accuracy			0.90	966
macro avg	0.56	0.56	0.56	966
weighted avg	0.90	0.90	0.90	966

ORES Model

- Objective Revision Evaluation System (ORES)[1]
  - ML-as-a-service, real-time support, web API[2]
  - Multiple independent classifiers trained on different datasets; language-specific
  - Linear SVM -> Ensemble (gradient boosting)
- Dataset:
  - Enwiki: 20k human labelled revisions (2015)
  - 78 input features (see Appendix A)
- Results:
  - Precision and recall: scope for improvement
  - Tends to be more aggressive to edits made by newcomers and anonymous editors[1]; false positive



	precision	recall	f1-score	support
False	0.96	0.92	0.94	908
True	0.27	0.43	0.33	58
accuracy			0.89	966
macro avg	0.61	0.68	0.64	966
weighted avg	0.92	0.89	0.91	966

Note:

1. Aaron Halfaker and R Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. arXiv preprint arXiv:1909.05189 (2019)
2. https://www.mediawiki.org/wiki/ORES/FAQ

Key Learnings and Next Steps

- A preliminary approach on a very small dataset
- To investigate / To-dos:
  - Increase dataset size
  - Imbalanced classification: data augmentation by oversampling
  - Experiment with other classification models / parameter tuning
  - Decision Tree Visualization
- Next step: ORES model
  - Input features
  - Training dataset
  - Performance comparison

Thank you!

Any questions?

Appendix

Appendix A: ORES Model Features[1,2]

Note:

1. https://arxiv.org/pdf/1909.05189.pdf
2. https://www.mediawiki.org/wiki/ORES/FAQ

B: Illustrative Decision Tree Visualization

[517]: