Developing a Machine Learning Algorithm for Wikipedia Vandalism Detection with Logistic Regression

Boxuan Shan

Pioneer Academics, boxuan.shan@gmail.com

Wikipedia is the largest online encyclopedia, used by people throughout the world in a variety of languages for research and knowledge acquisition. Many popular search engines and websites index and display information from Wikipedia, which further amplifies its impact on the internet. Wikipedia relies on a large community of online volunteer contributors who edit and collaborate on the articles. As a result, many articles are often subject to malicious edits or biased or partial opinions, which we call "vandalism". Despite many Wikipedia users actively ensuring that these vandalistic edits are edited or removed, automated methods such as machine learning are required to make the vandalism detection process more scalable and accurate. In this paper, we focus the English Wikipedia, performing linear regression using the PAN-WVC-10 data set, compiled from Amazon's Mechanical Turk. From this, we achieve an accuracy of 80%, with 76% precision, 89% recall, and a false positive rate of 13%. Given the high accuracy, precision, and recall achieved in such a limited timeframe, this only further displays the capacity for even better vandalism detection methods.

CCS CONCEPTS

• **Information systems** → Information retrieval → Retrieval tasks and goals; Clustering and classification • **Computing methodologies** → Machine learning → Machine learning approaches; Classification and regression trees • **Human-centered computing** → Collaborative and social computing → Collaborative and social computing systems and tools; Wikis

KEYWORDS

Wikipedia, Vandalism Detection, Machine Learning, Supervised Learning, Information filtering, Natural language processing

ACM Reference Format:
Boxuan Shan. 2024. Developing a Machine Learning Algorithm for Wikipedia Vandalism Detection with Logistic Regression. 7 pages.

1 INTRODUCTION

The Wisdom of the Crowd is the idea that large groups of people are collectively smarter than sparse groups of individuals. It has been shown that individuals have inherent biases, unavoidably injecting opinions in their actions. However, an aggregate collective of people will be able to produce generally more accurate, consistent, and coherent results [10]. Wikipedia is the largest encyclopedia explicitly demonstrating this phenomenon, relying on people throughout the world to each individually contribute to it. This results in the users essentially be able to moderate themselves, with each individual user's edits being verified by other users. With it being available for editing to anyone, it currently sees over a billion views monthly and 6.8 million articles on the English Wikipedia alone [14, 18].

Furthermore, many popular search engines and websites refer to content on Wikipedia, only amplifying its impact on the internet. With its popularity, any edit, no matter how trivial it seems, could have far-reaching effects [18]. However, given its collaborative nature, many articles, especially those less viewed, are nonetheless subject to vandalism and biased or partial opinions. This is what makes making the detection methods scalable so important. With the English Wikipedia growing so much quicker than dedicated users, it is impossible to ensure that every instance of vandalism and biased writing is immediately detected by humans.

These Wikipedia users are split into a strict hierarchical system to determine what roles and permissions they have. At the lowest end are unregistered users editing without an account. Users with accounts are "registered" and will be "autoconfirmed" when an email is verified, and then "extended confirmed" when they reach a specified edit threshold and account age. The highest rights can only be granted to users with consensus from existing users with those rights (administrators and bureaucrats). These users gain advanced privileges and tools that cover tasks such as page protection and deletion, blocking users, and performing certain actions on users' accounts. These privileges are not given out arbitrarily: there are currently only 856 administrators and 15 bureaucrats on the English Wikipedia [15, 17].

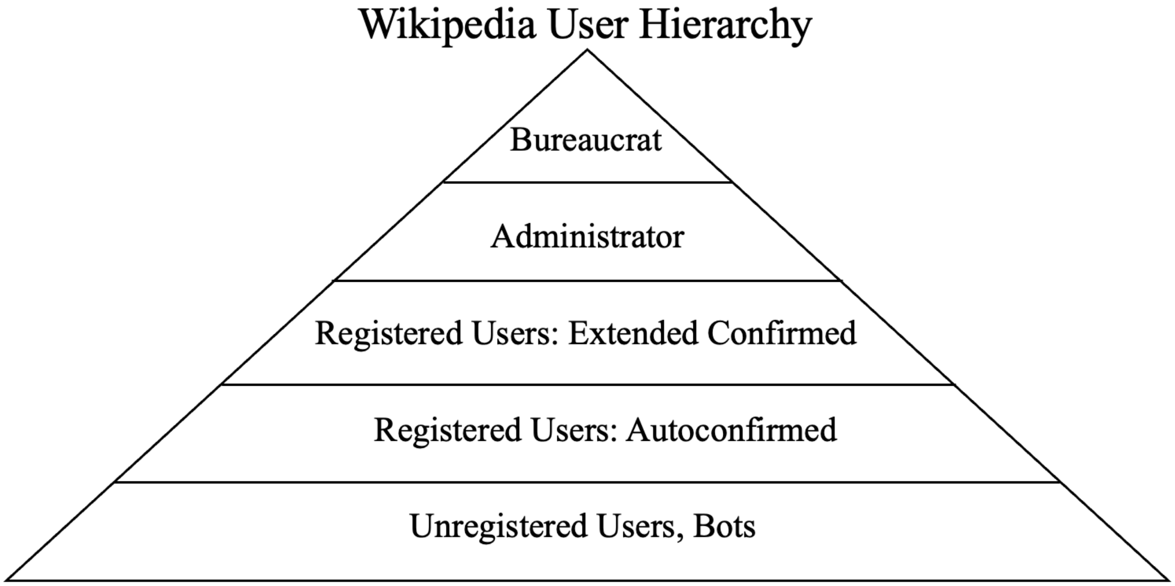


Figure 1: Rough hierarchy of user roles in Wikipedia

Fundamentally, Wikipedia relies on users to moderate other users' edits, reverting them if they are determined to be vandalism. However, given the inefficiency of manual revision of Wikipedia pages for vandalism, much work has been done to attempt to automate this process. Certain users have created bots, to automate otherwise mundane user tasks, such as fixing spelling errors, improving formatting, and detecting blatant vandalism. Most bots work from rule-based systems, such as checking for vulgarism or unusually large deletions from a page [8, 11, 16].

However, bots such as Cluebot NG employ an artificial neural network to detect such malicious edits. Apart from efforts from Wikipedia users, much work has been done on regression methods. Potthast et al. has used logistic regression methods to achieve 83% precision and 77% recall in detecting vandalism [8].

Given the advancement of research, resources, and tools since much other prior research on implementing machine learning, we propose performing Logistic Regression using the PAN-WVC-10 data set, compiled from Amazon's Mechanical Turk, consisting of 32440 edits from the English Wikipedia, of which 2394 are labeled as vandalism [7].

2 PRIOR WORK

2.1 Problem Definition

As most Wikipedia pages are open to be edited by anyone and given the popularity and prominence of Wikipedia on the internet, vandalism has long been an issue for Wikipedia. Yet there lacks a distinct split between malicious edits and inexperienced edits. Malicious edits are people editing with the intention of causing damage and harm, while inexperienced editors are just editors with the intention of benefitting Wikipedia, inadvertently adding opinionated edits or vulgarism. Since the PAN-WVC-10 corpus, the dataset used for this paper, classifies both as "vandalism", regardless of intention, this will be the definition of "vandalism" used [7].

2.2 Basic Prior Work

Ever since Wikipedia's rise in popularity, making false and misleading edits have become much more prominent. Therefore, Wikipedia has implemented many measures to counteract vandalism. Many Wikipedia pages of a controversial topic can be locked, so that only users with advanced permission can edit them. Other pages are protected by edit filters, which compare an incoming edit to a set of predefined filters. If the edit is determined to be malicious, it could be flagged or blocked entirely.

Fundamentally, Wikipedia relies on its editors to moderate itself, reverting vandalism as they encounter any [11]. Administrators and Bureaucrats can then use any necessary tools to ban specific user accounts or blanket ban IP address ranges to prevent repeated malicious edits from registered and unregistered users [15, 17].

2.3 Automation

Not all vandalism detection of Wikipedia is by hand. Since a large portion of vandalism on Wikipedia is obvious enough to be detected by simple rule-based checks, many bots have been developed and implemented on Wikipedia to automate this process [8].

These Wikipedia bots are essentially accounts run by automated bots, querying the Wikimedia API to perform specific actions on pages and users. Bots must pass a request for bot approval (RfBA), during which a Bot Approvals Group (BAG) will determine whether the bot should be permitted to operate. These bots are all used for different specific tasks. For example, SoxBot III, CounterVandalismBot, and RscprinterBot employ rule-based scoring systems to check for malicious vandalism. Other bots, such as 28bot, checkers for unintentional formatting errors and test edits [8, 11].

Most of these bots use simple heuristics or comparison with a list of blacklisted words to detect malicious intent. However, ClueBot NG employs many more novel methods [13].

Of these include Bayesian Classifiers, which are applied to individual and 2-word phrases. Each time the word or phrase is used in a malicious edit is tallied, and every time the word is used in a beneficial edit is tallied. This is used to determine a probability of vandalism for each word and 2-word phrase. These classifiers are not used alone, but instead fed into an Artificial Neural Network. This ANN is the basis of the bot, with inputs to it consisting of the Bayesian Classifiers, along with several statistics calculated from the edit.

Furthermore, since each edit is given a vandalism score between 0 and 1, thus a threshold must be applied to the bot to classify an edit as malicious or not. This threshold depends on a false positive rate chosen manually by a human, to minimize the false positives of the bot. Currently, the false positive rate is 0.1%, ensuring that the bot catches approximately 40% of all vandalism. This is compared to a 0.25% false positive rate catching 55% of all vandalism.

After a determination is made, the edit is passed through some post-processing filters. These include: 1. A user whitelist, 2. Edit counter, checking the number of edits a user has made previously without vandalism warnings, and 3. A page whitelist, ensuring that any page is not reverted on more than once per day.

2.4 Explicit Machine Learning Methods

Apart from bots made specifically for wikipedia, much work has been done on specialized machine learning methods to detect vandalism.

A majority of papers have focused on compiling features relevant to detecting vandalism. These features include elements processed from publicly available information on Wikipedia. Such features include the character distribution, uppercase ratio, vulgarism frequency, anonymity, etc… [2, 3, 5, 6]. These features are all obtained after the processing of publicly available data available immediately after an edit is made. An explanation of these features can be seen in section 3.2 Feature Extraction.

The logistic regression analysis from Potthast et al. includes many features mentioned above. Of these features, edits per user proved to have the highest F-score. All in all, combining these features, Potthast et al. achieved 83% precision and 77% recall, improving the F-score of Wikipedia's rule based methods by 49%.

Furthermore, Adler et. al. have compiled relevant features with vandalism detection into 4 categories: Metadata, Text, Language, and Reputation. In short, Metadata features are immediately available after the edit, Text features require processing of the textual data, Language features require a higher level of Natural Language processing, and Reputation features necessitate extensive processing of historical data on Wikipedia.

Furthermore, Tramullas et. al.'s review of 67 research papers on vandalism detection on Wikipedia is also noteworthy. 59.7% of papers deal explicitly deal with the *detection* of vandalism, with the second majority dealing with enforcing content quality [12]. Ensuring quality content and removing vandalism are also the 2 main goals for this paper, as they are also how the PAN-WVC-10 dataset was compiled [7]. Tramullas et. al. also shows that a majority of the approaches utilize machine learning classification, while quantitative methods and reputation analyses are lesser used [12].

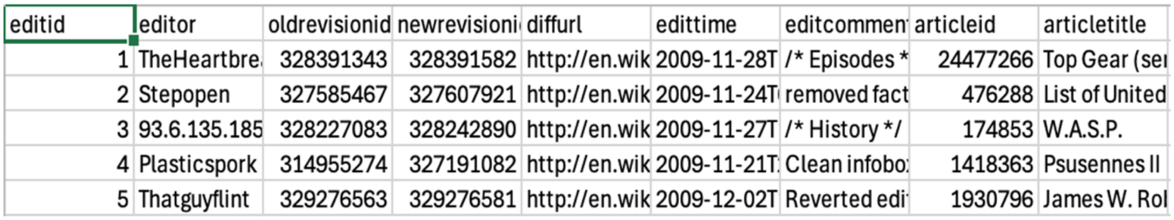
3 METHODOLOGY

3.1 Corpora

A common corpus used by prior research is the PAN Wikipedia Vandalism Corpus 2010 (PAN-WVC-10). We chose to use this PAN-WVC-10 due a substantial amount of prior research using it and it being organized in an easy-to-read format. Importantly, the PAN-WVC-10 also includes labels for revisions, indicating whether they are regular or vandalism [7, 9].

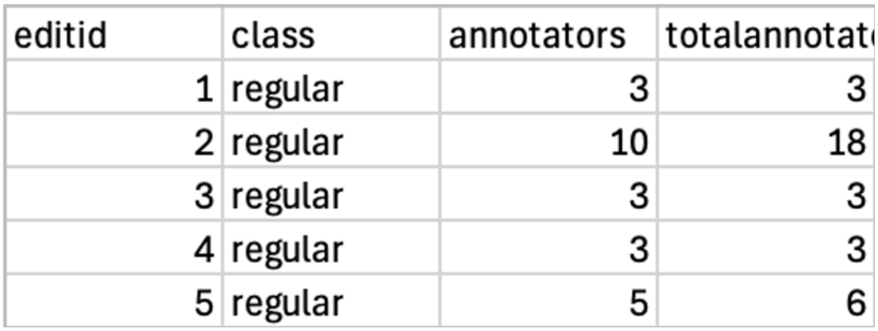
3.1.2 Dataset Shape

A sample of the dataset is provided:



editid	editor	oldrevisionid	newrevisioni	diffurl	edittime	editcommen	articleid	articletitle
1	TheHeartbre	328391343	328391582	http://en.wik	2009-11-28T	/* Episodes *	24477266	Top Gear (se
2	Stepopen	327585467	327607921	http://en.wik	2009-11-24T	removed fact	476288	List of United
3	93.6.135.185	328227083	328242890	http://en.wik	2009-11-27T	/* History */	174853	W.A.S.P.
4	Plasticspork	314955274	327191082	http://en.wik	2009-11-21T	Clean infobo	1418363	Psusennes II
5	Thatguyflint	329276563	329276581	http://en.wik	2009-12-02T	Reverted edi	1930796	James W. Rol

Figure 2: A sample of edits.csv in the PAN-WVC-10 corpus. (Potthast et. al. *Overview of the 1st International Competition on Wikipedia Vandalism Detection*)



editid	class	annotators	totalannotat
1	regular	3	3
2	regular	10	18
3	regular	3	3
4	regular	3	3
5	regular	5	6

Figure 3: A sample of annotations.csv in the PAN-WVC-10 corpus. (Potthast et. al. *Overview of the 1st International Competition on Wikipedia Vandalism Detection*)

The main portion of the dataset is split into 2 main CSV files: *edits.csv* and *annotations.csv*.

A sample of the edits.csv shown in Figure 2 include general information about a specific edit, including the editors name (or IP if the editor is anonymous), and the IDs of the old and the new revision. Here, the edit that was made is the difference between the old and new revision. Each of these edits correspond with a specific edit ID, ranging from 1 to 32440. These unique edit IDs also correspond with the edit IDs in column 1 of the annotations csv (Fig. 3), labeling each edit as either regular or vandalism. In this sample, the first 5 edits are all regular. Also, concerning our problem of vandalism detection, the *annotators* and *totalannotators* columns in the annotations csv are irrelevant.

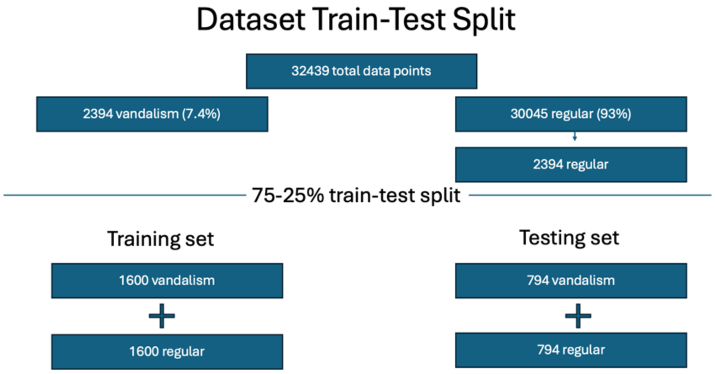


Figure 4: An illustration of the dataset shape and how the train-test split was conducted.

As shown in Fig. 4, the PAN-WVC-10 corpus contains 32439 total data points, of which ~7% are vandalism. This 7% ratio of vandalism corresponds with the 7% ratio of vandalism to all edits on Wikipedia [4, 7]. Of the regular data points, a random selection is chosen, to match with the number of vandalism data points. After a train-test split of roughly 75-25%, we have 3200 elements in the training set and 1588 elements in the testing set, with a 50-50 split on vandalism and regular edits for each.

3.1.1 Dataset Limitations

However, the PAN-WVC-10 dataset still comes with a few limitations.

Note that it does not contain the specific text added and removed in the edit, only including the IDs of the revisions before and after the edit. Furthermore, due to it being collected in 2010, some articles are no longer available, and some edits are suppressed. Due to not having the specific text added and removed, we need to use the Wikimedia API to collect data about the edit and metadata about the account (or IP) making the edit.

More importantly, the PAN-WVC-10 dataset comes with the downside of not containing specific information about the edit itself and lacking significant metadata necessary to extract features from. Luckily, the dataset includes the IDs of the 2 revisions adjacent to the edit. The from rev ID corresponds to the version of the article prior to the edit, and the to rev ID corresponds with the article, post edit.

This enables the use of the Wikimedia API through a Python script to extract essential features from Wikipedia [1]. However, the raw text extracted with the Wikimedia API is in HTML and Wikitext form. A few layers of further processing converts the text into readable plaintext and filters out exactly what the editor added and removed.

3.2 Feature Extraction

Using the method to sort features presented in Adler et. al., we can organize a table of all the features used [2].

Table 1: Features used

Feature Name	Classification/Type	Description
uppercase_ratio	Textual	Ratio of uppercase to lowercase letters.
longest_consec_char	Textual	Length of the longest appearance of a consecutive character.
alpha_punct_ratio	Textual	Ratio of ASCII characters to other characters.
vulgarism	Language	Using the better-profanity package, calculates the frequency of profanity.
spell_err	Language	Using the Spellchecker package, calculates the frequency of misspelled words.
diffsize	Metadata	Size of the edit, in bytes.
userage	Metadata	Number of previous edits the user has made. If the user is anonymous, returns 0.
user_acesslevel	Metadata	Number of days between the creation of the user's account and July 1, 2050.
commentlength	Metadata	Length of the user's comment, ignoring special Wikipedia-related tags that are automatically added.

All of these features were normalized to return a value of between 0 and 1 and are compared with the previous edit. For example, the uppercase ratio is the ratio of uppercase letters to lowercase letters, as a ratio of the uppercase ratio of the previous edit. A similar principle of comparing the new edit to the previous revision is implemented for other features, to ensure that any abnormalities in an edit is due to it being vandalism, not the quirks of the article itself.

We can then train this model using these features using scikit-learn.

4 RESULTS

With logistic regression, we obtain an accuracy of 80%, with 76% precision and 89% recall. Furthermore, in our testing set, we have a false positive rate of 13.8%.

5 DISCUSSION

We have used a logistic regression model to detect vandalistic edits on Wikipedia, using the PAN-WVC-10 dataset. Given the limitations of the corpus, an intermediate step of further data extraction was performed to get the actual contents of the edit from Wikipedia. After that, notable features from previous work were extracted, then fed into a logistic regression model.

These results match well with previous work of vandalism detection with machine learning, especially that of Potthast et. al., with the main difference in implementation being the much larger dataset used and the smaller collection of features.

While there is much room for improvement, we believe that further work should focus on lowering the false positive rate as a primary objective, with accuracy as a secondary one. There already exist many other methods of vandalism detection, be it by human editors or bots. For both methods, there is a much larger focus placed on keeping the negative impact on well-intentioned editors as low as possible, keeping with Wikipedia's principle of assuming good faith [11]. If a malicious edit is not caught by a bot, it will eventually be by a human. Having a high false positive rate will only impede editors contributing positively to Wikipedia.

Specific to the model itself exist many other paths for improvement. First, the features could be tuned much better. Given the limited timeframe to work on this, many of the parameters chosen for the features were quite arbitrary. These include the specific scores given to an administrator, the date of July 1st, 2050 to derive the account's age, and the hypothetical maximum of many features chosen to normalize the output to between 0 and 1.

Furthermore, much more work should be done in refining the text parsing. What text/punctuation should be removed to ensure the most efficiency? What should remain so that no information pertaining to vandalism detection is lost? These will require extensive testing to determine which would provide the most optimal accuracy and false positive rate.

More work could also be done to expand to other machine learning algorithms. While logistic regression was the logical first step, many other models could be implemented to potentially return better results. For example, Martinez-Rico et. al. found plenty of other algorithms with much better performance than logistic regression, with the best performing algorithms being Random Forest, followed by a combination of Random Forest with other algorithms [5].

Random Forest, it seems, consistently performs well in such vandalism detection tasks. Mola-Velasco et. al. found that the best performing classifiers to be Random Forest and Logit Boost, with Random Forest giving an AUC of 0.92 [6].

However, most importantly, we believe that this should all be done with the primary focus of lowering the false positive rate as much as possible, with accuracy and other such metrics being of only secondary importance.

REFERENCES

[1] "API:Main Page." MediaWiki, https://www.mediawiki.org/wiki/API:Main_page. Accessed 12 July 2024.

[2] Adler, B. Thomas, et al. "Wikipedia vandalism detection: Combining natural language, metadata, and reputation features." Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II 12. Springer Berlin Heidelberg, 2011.

[3] Heindorf, Stefan, et al. "Vandalism detection in wikidata." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016.

[4] Javanmardi, Sara, David W. McDonald, and Cristina V. Lopes. "Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso." Proceedings of the 7th International Symposium on Wikis and Open Collaboration. 2011.

[5] Martinez-Rico, Juan R., Juan Martinez-Romo, and Lourdes Araujo. "Can deep learning techniques improve classification performance of vandalism detection in Wikipedia?." Engineering Applications of Artificial Intelligence 78 (2019): 248-259.

[6] Mola-Velasco, S.M. "Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals." In Braschler, M., Harman, D., eds.: Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy. (2010)

[7] Potthast, Martin, Benno Stein, and T. Holfeld. "Overview of the 1st international competition on wikipedia." CLEF'2010 (2010).

[8] Potthast, Martin, Benno Stein, and Robert Gerling. "Automatic vandalism detection in Wikipedia." Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30. Springer Berlin Heidelberg, 2008.

[9] Potthast, Martin and Teresa Holfeld. "Overview of the 2nd International Competition on Wikipedia Vandalism Detection." Notebook Papers of CLEF'2011 (2011).

[10] Simoiu, Camelia, et al. "Studying the "wisdom of crowds" at scale." Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. Vol. 7. 2019.

[11] Smets, Koen, Bart Goethals, and Brigitte Verdonk. "Automatic vandalism detection in Wikipedia: Towards a machine learning approach." AAAI workshop on Wikipedia and artificial intelligence: An Evolving Synergy. Antwerp, Belgium.: AAAI Press, 2008.

[12] Tramullas, Jesús, Piedad Garrido-Picazo, and Ana I. Sánchez-Casabón. "Research on Wikipedia Vandalism: a brief literature review." Proceedings of the 4th Spanish Conference on Information Retrieval. 2016.

[13] "User:ClueBot NG." Wikipedia, 20 Oct. 2010. Wikipedia, https://en.wikipedia.org/w/index.php?title=User:ClueBot_NG&oldid=391868393.

[14] "Wikipedia:About." Wikipedia, 28 June 2024. Wikipedia, https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=1231549672.

[15] "Wikipedia:Administrators." Wikipedia, 10 July 2024. Wikipedia, https://en.wikipedia.org/w/index.php?title=Wikipedia:Administrators&oldid=1233691393.

[16] "Wikipedia:Assume Good Faith." Wikipedia, 7 June 2024. Wikipedia, https://en.wikipedia.org/w/index.php?title=Wikipedia:Assume_good_faith&oldid=1227770566.

[17] "Wikipedia:User Access Levels." Wikipedia, 10 July 2024. Wikipedia, https://en.wikipedia.org/w/index.php?title=Wikipedia:User_access_levels&oldid=1233655724.

[18] Wikistats - Statistics For Wikimedia Projects. https://stats.wikimedia.org/#/all-projects. Accessed 12 July 2024.

A Appendix

A Github repository with the source code is available here: https://github.com/bxshan/wikipedia_vandalism_detection