Characterizing Knowledge Manipulation in a Russian Wikipedia Fork

Mykola Trokhymovych[1], Oleksandr Kosovan[2], Nathan Forrester[3], Pablo Aragón[1], Diego Saez-Trumper[1], Ricardo Baeza-Yates[1]

[1]Universitat Pompeu Fabra
[2]Ukrainian Catholic University
[3]Independent Researcher
{mykola.trokhymovych, pablo.aragon, diego.saez}@upf.edu,
o.kosovan@ucu.edu.ua, nathanforrester.research@proton.me, rbaeza@acm.org

Abstract

Wikipedia is powered by MediaWiki, a free and open-source software that is also the infrastructure for many other wiki-based online encyclopedias. These include the recently launched website Ruwiki, which has copied and modified the original Russian Wikipedia content to conform to Russian law. To identify practices and narratives that could be associated with different forms of knowledge manipulation, this article presents an in-depth analysis of this Russian Wikipedia fork. We propose a methodology to characterize the main changes with respect to the original version. The foundation of this study is a comprehensive comparative analysis of more than 1.9M articles from Russian Wikipedia and its fork. Using meta-information and geographical, temporal, categorical, and textual features, we explore the changes made by Ruwiki editors. Furthermore, we present a classification of the main topics of knowledge manipulation in this fork, including a numerical estimation of their scope. This research not only sheds light on significant changes within Ruwiki, but also provides a methodology that could be applied to analyze other Wikipedia forks and similar collaborative projects.

1 Introduction

Online information dissemination plays a key role in shaping public opinions and attitudes. As the the world's largest encyclopedia and the ninth most visited website globally,[1] Wikipedia holds an influential position within the web ecosystem (Piccardi et al. 2021). It is maintained through a collaborative community effort to become the "sum of all human knowledge" (Sutcliffe 2016). That is, anyone can freely edit Wikipedia content, although several policies and guidelines decided by the community must be followed to guarantee the quality and integrity of knowledge (McDowell and Vetter 2020). Given that its content empowers various applications, for example, integrating verified facts into curricula (Lemmerich et al. 2019), fact-checking (Trokhymovych and Saez-Trumper 2021), or training of large language models (Devlin et al. 2019), knowledge on Wikipedia has a major societal impact.

There are actors who are not comfortable with the content and policies of Wikipedia. For example, some states like China and Turkey have repeatedly blocked access to the website (Sezer and Dolan 2017; Siegel 2019). In the case of China, the decision to censor Wikipedia was also followed by the launch of *Baidu Baike*, an online encyclopedia with content in accordance with the requirements of the Chinese Government (Woo 2007; Siegel 2019). Moreover, the open-source nature of *MediaWiki*, the software that powers Wikipedia, has enabled the launch of alternative wiki-based encyclopedias. A well-known example is *Conservapedia*, created by detractors of Wikipedia's core policy of neutrality and self-described as American conservative and fundamentalist Christian (Johnson 2007). Another example is *Runiversalis*, the content of which must follow the requirements of the Russian legislation and its traditional values (Runiversalis 2024). While the views on these encyclopedias don't compare to Wikipedia's, the growing cultural prominence of alternative facts has sparked a noticeable rise in both traffic and interest (Fitts 2017).

Russian Wikipedia appeared in May 2001 during the first wave of non-English Wikipedias. In June 2023, a fork of Russian Wikipedia was launched online, hereinafter referred to as *RWFork*. The project was founded by Vladimir Medeyko, former Director of Wikimedia Russia, a Wikimedia Chapter organization. While this project was powered by *MediaWiki* software like the aforementioned alt-Wikipedias, its content was also copied from Russian Wikipedia and later edited to conform to the Russian legislation (Cohen 2023). Therefore, *RWFork* is an organized effort to manipulate knowledge, originally created with neutral editorial policies, in order to comply with the editorial policies of a specific state.

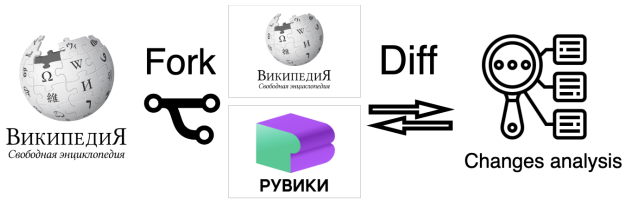

Figure 1: Summary of our research that analyzes changes in a Russian Wikipedia fork to assess knowledge manipulation.

[1]https://www.similarweb.com/website/wikipedia.org, Accessed 12 September 2024.

Research Questions. In this work, we examine a Russian Wikipedia fork. Our primary goal is to characterize how original content from Russian Wikipedia has been manipulated (see Figure 1). To achieve this, the initial step is to identify which articles have undergone changes and their relevance. Therefore, our first research question is:

- **RQ1** How relevant are Russian Wikipedia articles changed by *RWFork* editors?

Next, we aim to examine changes made within *RWFork* in detail to characterize the editorial process and the nature of content modifications. This leads to our second research question:

- **RQ2** How has article content changed in *RWFork*?

Finally, we are interested in categorizing the changes that have occurred in *RWFork* to provide a clear understanding of the broader patterns of knowledge manipulation. As a result, our third research question is:

- **RQ3** What are the patterns of knowledge manipulation in *RWFork*?

Methodology. To answer our research questions, we have developed a methodology to compare the content of two MediaWiki-powered websites: Russian Wikipedia and *RWFork*. The first challenge we face is data collection and preparation. Relying on MediaWiki APIs, we implemented a data retrieval approach to extract data from both sources. To address **RQ1**, we use bootstrapping to estimate the relevance metrics of articles along with their confidence intervals. Later we use these estimations to characterize the articles that were changed by *RWFork* editors. As for **RQ2**, we use exploratory data analysis on a wide range of temporal, geographical, categorical, and topical features. Our methodology also includes named entity recognition to identify the most frequent entities in deletions and additions. To answer **RQ3**, we combine advanced natural language processing tools, clustering algorithms, and qualitative analysis to build a classification of the main topics affected by knowledge manipulation.

Main Findings. By applying our methodology to a dataset with more than 1.9M articles of the Russian Wikipedia and its fork, we discover that:

- **RQ1**: *RWFork* editors have modified articles that have a considerably higher number of page views compared to others, hence influential ones. We find that 1.75% of articles (*changed* pages) generate 14.2% of the total page views. Also, according to our analysis, modified pages refer to controversial topics, and have almost twice the higher revert rate than duplicated ones.
- **RQ2**: Unlike Wikipedia editors, *RWFork* editors have a shorter activity period that aligns with standard office hours and reduced weekend activity. Most frequently altered pages pertain to Ukraine, Russia, and Belarus locations. The most frequently added and deleted categories are related to the 2022 Russian invasion of Ukraine, particularly occupied territories or sanctions. Also, we discover frequent additions of Russian Government resources, contrasting with the frequent deletion of EU and Ukrainian Government websites.
- **RQ3**: We find that most of the content changes ($\sim$90%) can be classified into 8 main categories. In particular, we discover that the majority of the content changes are knowledge manipulations related to territory reassignment, international sanctions, and terminology variations related to the 2022 Russian invasion of Ukraine ($\sim$44%).

To our knowledge, our work provides the first study of how original Wikipedia content has been forked and manipulated to meet the requirements of a national regulation. In addition to the geopolitical relevance of this case study, our methodology can also support future work examining the differences between wiki-based encyclopedias. To encourage further research, we release the dataset of this study under an open license using the Zenodo platform,[2] along with the code used for the presented case study.[3]

Relevance. National identity and public opinion can be influenced by the information citizens are finding online about their history. In a previous study, Wikipedia was ranked the 6th most important information about history, passing museum visits, college courses, and social media (Burkholder and Schaffer 2021). Therefore, attempts to manipulate Wikipedia content, even if they happen in other platforms, could have a significant societal impact.

Wikipedia is also a key resource for web search engines (Vincent et al. 2019; Vincent and Hecht 2021). Moreover, previous research has shown that Wikipedia is not only an important source of information but also has a role as a high-traffic gateway to the broader Web ecosystem (Piccardi et al. 2021). As a consequence, limitations of access to Wikipedia or replacing it with alternative versions could favor the displacement of web traffic to specific resources with manipulated information.

Last but not least, Wikipedia content is frequently used for training Large Language Models (LLMs) (Devlin et al. 2019). Manipulated versions of Wikipedia used as training data for LLMs can encourage AI-powered systems that promote ideas with specific biases (Yang and Roberts 2021). Therefore, it is crucial to characterize what biases are included in *RWFork* as the plan to use its data to train LLMs has already been announced by the project founder (Davydov 2023).

2 Related Work

To contextualize our characterization of knowledge manipulation in *RWFork*, we categorize prior research on Wikipedia into three main areas: knowledge gaps, knowledge integrity, and the specific case of the 2022 Russian invasion of Ukraine.

Knowledge Gaps

Wikipedia aims to represent "the sum of all human knowledge" while retaining the requirement of the neutral point of view (Ford et al. 2013). This ambitious goal has been the subject of extensive research aimed at identifying biases in

[2] https://zenodo.org/records/15073728

[3] https://github.com/trokhymovych/RWFork

the form of knowledge gaps, *i.e.*, "disparities in content coverage or participation of a specific group of readers or contributors" (Redi et al. 2020).

Gender gap is arguably the most extensively studied knowledge gap on Wikipedia (Reagle and Rhue 2011; Eom et al. 2015; Hinnosaar 2019; Wagner et al. 2016; Zagovora, Flöck, and Wagner 2017). However, several studies have also explored cultural and geographic disparities in topic coverage. Early research provided empirical evidence of self-focus within multiple language editions of Wikipedia (Hecht and Gergle 2009), a phenomenon that likely contributes to the significant disparities in the geographical representation of knowledge (Graham, Hale, and Stephens 2011; Graham et al. 2014; Beytía 2020). Self-focus might also play an important role in the cultural local biases observed in content across languages (Hecht and Gergle 2010; Callahan and Herring 2011; Miquel-Ribé, Laniado, and Kaltenbrunner 2021) and the biased narratives found on controversial historical events and cultural heritage (Rogers, Sendijarevic et al. 2012; Pentzold et al. 2017).

Despite knowledge gaps, research has revealed that Wikipedia's content is no more biased than that in expert-written encyclopedia articles (Greenstein and Zhu 2018). Since the plan for the Russian Wikipedia fork is to be initially edited by experts (Cohen 2023), it becomes particularly compelling to identify biases that could have arisen.

Knowledge Integrity

As anyone can edit Wikipedia, editors dedicate substantial effort to monitor articles, improving content verifiability, and strengthening its resilience against misinformation (Saez-Trumper 2019). Empirical research on knowledge integrity in Wikipedia has highlighted several threats currently being addressed.

Many instances of disinformation on Wikipedia, such as hoax articles, have been found to be identified and addressed quickly, which minimizes their impact (Kumar, West, and Leskovec 2016). A more pressing challenge to knowledge integrity is vandalism, a form of abuse that has drawn significant attention from research. Numerous studies have analyzed its characteristics (Shachaf and Hara 2010; Geiger and Ribes 2010; Potthast 2010) and proposed detection systems for this problem (Potthast, Stein, and Gerling 2008; Adler et al. 2011; Trokhymovych et al. 2023). Detection efforts have also focused on the phenomenon of sock puppets (Kumar et al. 2017; Sakib and Spezzano 2022), including cases to evade account bans (Niverthi, Verma, and Kumar 2022).

The efforts of Wikipedia editors to preserve knowledge integrity have contributed to transforming the project from a questionable source of information in its early years into an increasingly reliable one over time (Steinsson 2024). Recent research has highlighted that the community governance infrastructures of Wikipedia are crucial in addressing systematic disinformation campaigns and other influence operations (Kharazian, Starbird, and Hill 2023). For that reason, examining the changes that have occurred in *RWFork* can offer important insights into how Wikipedia knowledge could be manipulated without its community governance.

Case: The 2022 Russian Invasion of Ukraine

The great importance of the Russian invasion of Ukraine in 2022 has led to a growing body of literature on the documentation of this specific event in Wikipedia. A first study of the English Wikipedia article on this conflict highlighted the role of vandal fighters in facilitating coordinated editing efforts during a fast-changing and contentious event (Roberts and Xiong-Gum 2022). A later analysis of the effects of the conflict on multiple articles and languages showed a significant decline in activity around the time of the invasion on both Russian and Ukrainian language editions, followed by a recovery (Dammak and Lemmerich 2023). Interestingly, there was a sharp increase in the rate of reverts right after the invasion. More recently, an interview study with expert editors from English Wikipedia showed no evidence of a state-sponsored information operation, although participants reported disruptive editing in war-related articles from accounts aligned with either Russian or Ukrainian positions (Kurek, Budak, and Gilbert 2024).

All these studies highlight the critical role of Wikipedia's editorial norms in preventing state-sponsored information operations during the Russian invasion of Ukraine in 2022. As a consequence, examining a fork of Wikipedia created to comply with the Russian legislation offers a valuable opportunity to envision how its content could have been manipulated in the absence of Wikipedia editorial standards.

3 Data Collection and Preparation

A major challenge of this work is data collection, as it requires parsing data from two different *MediaWiki*-powered websites on a large scale and further post-processing. In this section, we present our process to collect data.

Article Selection

The first step of the data collection process was to define the articles of our dataset. *RWFork* was initially created as a copy of Russian Wikipedia, meaning that most of the content, including page titles, is the same. As a consequence, we used the page title as a key to match articles from *RWFork* and Russian Wikipedia. Page titles were extracted from the existing Russian Wikipedia articles of the June 2023 Wikimedia dump.[4] In total, we compiled a list of about 1.9M distinct page titles for further processing.

We should note that there are no similar resources from *RWFork*. Therefore, our pipeline did not include newly created articles, which is one of the limitations of our research.

Web Crawling

For Russian Wikipedia, we extracted the content of pages of our dataset, formatted as wikitext,[5] using the *Wikimedia* API.[6] For *RWFork*, although the project is also powered by *MediaWiki* and provides an API, several limitations led us to collect data through a multi-step process. The full crawling pipeline is presented in Figure 2.

[4] https://dumps.wikimedia.org/

[5] https://en.wikipedia.org/wiki/Help:Wikitext

[6] https://ru.wikipedia.org/w/api.php

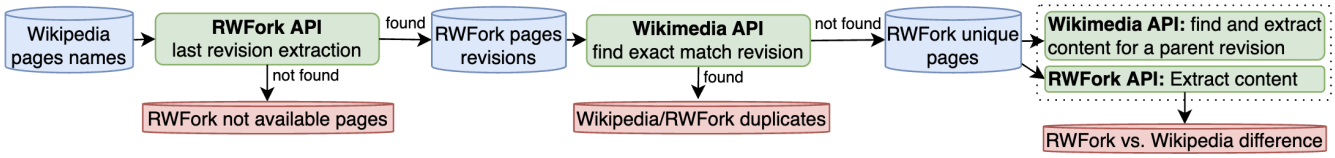

Figure 2: Process for crawling revision differences between *RWFork* and Russian Wikipedia.

Our main interest is to understand the contribution of *RWFork* editors, i.e., information about what was added or deleted compared to the original Russian Wikipedia version. As for that, we examined the page history that includes edits (also known as revisions). First, we examined if *RWFork* articles were an exact copy of Russian Wikipedia at some point in time. For that purpose, we parsed the last *RWFork* revision of each article to match it with the corresponding Russian Wikipedia edit history using revision and parent revision IDs as a compound key. In this step, a subset of articles were not available in *RWFork*. If the last *RWFork* revision is not included in Russian Wikipedia history, we consider that this page includes changes created by *RWFork* editors. Then, using *RWFork* page history, we extracted its Russian Wikipedia parent revision (the last revision of the *RWFork* page that has a match in Russian Wikipedia history). Finally, we extracted the content for the *RWFork* page's last version and its Russian Wikipedia parent revision to identify the pieces of information that were modified.

Our dataset includes records from May to September 2023. We processed 1,925,452 pages, where 17,745 (0.92%) were unavailable (either deleted or with limited access), and 33,664 (1.75%) pages included *RWFork*-based edits.

Data Processing

After collecting pairs of article versions from two data sources, the next step was to extract their differences. We use an open *mwedittypes*[7] library for text processing. We considered the Russian Wikipedia article as a base version studying the modification needed to achieve the *RWFork* version.

First, we extracted the sentences or phrases that were *inserted* or *deleted*. Those lists may contain similar items. Therefore, we define the additional category *changed* by pairwise matching sentences from the *inserted* and *deleted* lists using the Levenshtein Distance (Levenshtein 1966). Following the methodology used previously for Wikipedia revisions comparison in (Trokhymovych et al. 2023), we adopted a similarity threshold of 0.6. When the similarity of two sentences exceeded this threshold, we moved the pair to the *changed* list and removed them from the *inserted* and *deleted* lists. This process resulted in lists of *inserted*, *deleted*, and *changed* sentences for each article.

Moreover, we extracted the lists of changed media names, page categories, and references. Examples of parsed content changes are included in Appendix A.

[7]https://github.com/geohci/edit-types

Additional Data Sources. Besides the differences comparing *RWFork* and Russian Wikipedia article revisions, we retrieved additional article characteristics necessary to address **RQ1** and **RQ2**. In particular, we extracted the list of countries and topics related to the article. We assumed that those are the same for the *RWFork* and Wikipedia versions. As for the countries extraction, we used a tool that provides countries predictions for Wikipedia articles based on their associated Wikidata items and links to other Wikipedia articles.[8] As for the topics, we relied on a topic prediction tool for Wikipedia articles based on their links to other articles (Johnson, Gerlach, and Sáez-Trumper 2021).[9] Furthermore, we extracted the monthly number of views per page from the Wikimedia API.

4 RQ1: Relevance of Changed Articles

Our dataset contains information about more than 1.9M article pairs. Of those, 97.33% of them are just *duplicated* (not changed in *RWFork*) and 0.92% are *missing*. Among the 33,664 (1.75%) of articles *changed* by *RWFork* editors, 0.96% contain changes within the text and another 0.79% only have changes in elements that do not affect the text (formatting, references, tags, media, etc.).

Our study begins with an examination of metrics related to page relevance on Wikipedia. We use the 2022 and 2023 Russian Wikipedia history dump[10] and page view statistics. Specifically, we analyze the average number of page views per month, the number of edits, the rate of IP edits, and the revert rate (the proportion of edits identified as damaging and subsequently reverted). For each metric and group (*changed*, *duplicated*, *missing*), we perform 10K bootstrap resamples of 1K page statistics each, sampled with replacement. This balances data variability, computational efficiency, and estimate reliability, enabling calculation of the mean and 95% confidence interval using quantiles (Efron and Tibshirani 1994).

Results are presented in Figure 3. We observe that *changed* pages have significantly more page views than *duplicated* ones. Although only about 1.75% of pages were changed, these pages generate approximately 14.2% of Russian Wikipedia's page views (around 9.6% from pages with text changes and 4.6% from those with other changes), indicating their popularity. Also, articles that were changed in *RWFork* have significantly more edits and IP edits in

[8]https://wiki-topic.toolforge.org/countries
[9]https://wiki-topic.toolforge.org/topic
[10]Data includes records by October 2023

Figure 3: Comparison of Russian Wikipedia pages statistics for the groups of *changed*, *duplicated*, and *missing* pages. Statistics used: (a) Monthly page views; (b) Edits count; (c) IP edits rate; and (d) Revert rate. Plots include mean values with 95% confidence intervals for corresponding statistics.

Russian Wikipedia than *duplicated* ones, an indicator of a higher attention level from registered and unregistered editors. Finally, the revert rate of *changed* articles is almost twice higher than for *duplicated* ones, a signal of higher risk of disputes and vandalism for those articles (Trokhymovych et al. 2023).

5 RQ2: Changes of Article Content

In this section, we explore article changes by comparing the general characteristics of the Russian Wikipedia and *RWFork*, including temporal, geographical, categorical, source, media, and text-based features.

Editing Time. We compare the temporal regularities of editing in *RWFork* and Russian Wikipedia from August 2023. To reduce noise, we ignore all revisions created by bots, using a hard filter based on the username. Our findings are presented in Figure 4. Previous research has demonstrated that Wikipedia editorial activity has circadian patterns (Yasseri, Sumi, and Kertész 2012). Russian Wikipedia follows a strict daily pattern, with a short inactivity period at night. In contrast, *RWFork* editor's activity period is shorter and coincides with standard office hours, having very reduced activity during the weekend. In particular, 53.24% of edits on *RWFork* are made on weekdays from 8 to 17 UTC time, while only 40.06% of edits on Russian Wikipedia are made in that time interval.

Geography. We then analyze the geography of articles that were *changed* in *RWFork* compared to Russian Wikipedia ones. Although articles typically relate to one country, some relate to multiple countries (e.g., articles about disputed territories or people with links to more than one country) or no countries (e.g., pages about common knowledge topics). We also analyze locations of pages that are either full *duplicates* or *missing* in *RWFork*. It should be noted that we are limited to the pages that have at least one location linked, which is 53.7% from the complete set. We compute the rate of pages related to the specific list of locations within *changed*, *duplicated*, and *missing* groups.

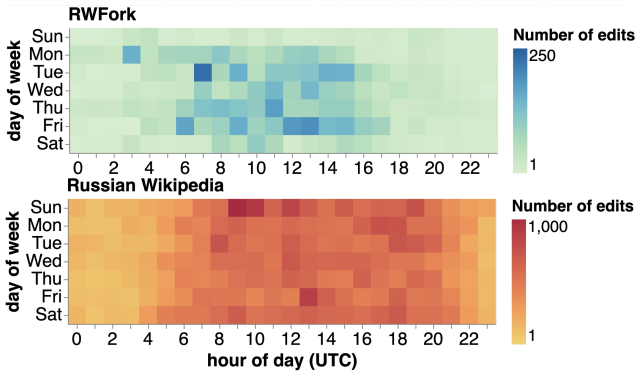

Figure 4: Average number of edits per day of week and hour of day in *RWFork* (*top/blue*) and Russian Wikipedia (*bottom/red*). The color intensity indicates the volume of edits, with darker shades representing higher activity.

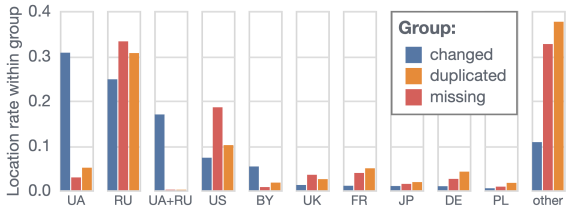

Figure 5: Rates within groups of *changed*, *duplicated*, and *missing* pages in *RWFork* for the top 10 most frequent countries in *changed* group.

The 10 most frequent locations from the *changed* group the frequency rates from other groups are presented in Figure 5. We find that pages from Ukraine (UA) and Ukraine+Russia (UA+RU) have a larger frequency in the *changed* group than in other groups. We also observe a similar tendency for pages related to Belarus (BY). It should be noted that 77.5% pages from the Ukraine+Russia location have changed when the same rate for general distribution is only 1.75%. Additionally, there is a remarkably high rate of US location in the *missing* group.

Categories. We analyze categories that were added and deleted while editing *RWFork* pages. In particular, for each changed page, we calculated the difference between sets of Wikipedia and *RWFork* categories. We find 1,056 unique categories added and 991 deleted. The most frequently added and deleted categories, along with the rate of changed pages, including presented edits, are shown in Table 1. We observe that the predominant added categories are related to the occupied territories of Ukraine. Conversely, the most frequently excluded categories relate to edit notices, individuals, and companies subject to sanctions over Russia's invasion of Ukraine. Despite the informative insights of these categories receiving editing focus in *RWFork*, their partially disorganized structure presents challenges for automated knowledge representation. In Section 6, we will use these category changes to provide a structured classification of knowledge manipulation.

Added category	Count	(%)
Villages of the Donetsk People's Republic	621	1.84
Urban-type settlements of the Donetsk People's Republic	213	0.63
Urban settlements of the Donetsk People's Republic	138	0.41
Russian military formations that participated in military operations in Ukraine (since 2022)	115	0.34
Urban-type settlements of the Lugansk People's Republic	109	0.32
Deleted category	**Count**	**(%)**
Articles with edit notice about caution when editing	5,042	14.97
Persons subject to sanctions in connection with the conflict in Ukraine	1,412	4.19
Temporarily occupied territories of Ukraine	920	2.73
Companies sanctioned over Russia's invasion of Ukraine	452	1.34
Urban-type settlements of Donetsk region	129	0.38

Table 1: Count and percentage of the top 5 added and deleted categories by *RWFork* editors (original Russian category titles are translated to English).

Sources. A core content policy of Wikipedia is verifiability,[11] which obliges contributors to support their edits with appropriate references. This practice assures that readers can verify the sources from which the information has been derived (Piccardi et al. 2020; English Wikipedia 2023). We analyzed reference changes to identify the sources that received the most attention from *RWFork* editors. Figure 6 shows the most added and deleted sources. On the one hand, the most frequently added sources are related to the Russian Government or administrations of occupied territories. On the other hand, the most frequently deleted sources are related to Ukrainian Government websites (e.g., the second most frequently deleted resource includes sanctions-specific information).

Named Entities. We use the open-source software library *SpaCy* with the Russian text corpus *ru_core_news_sm* for named entity recognition (Honnibal et al. 2020) to build two lists of named entities – deleted and added by *RWFork* editors. Each named entity is counted only once per page observed. For each named entity, we define category labels and lemma for deduplication.

Table 2 shows the most frequently encountered named entities that were either deleted or added. Most of these entities refer to locations, particularly those in Ukraine and Russia. We observe a tendency for *RWFork* to change names to Kremlin-compliant terms for occupied territories, such as replacing "Donetsk Oblast" to "Donetsk People's Republic" or its abbreviation "DPR". Overall, *RWFork*'s modifications of named entities represent territory reassignment.

[11] https://en.wikipedia.org/wiki/Wikipedia:Verifiability

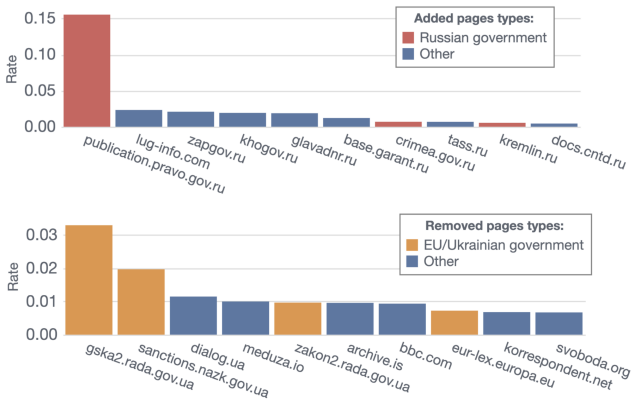

Figure 6: Top 10 most frequently added (top) and deleted (bottom) reference sources by *RWFork* editors.

Deleted named entities	Label	Count	(%)
Russia	LOC	4,169	22.52
Ukraine	LOC	3,956	21.37
Verkhovna Rada	ORG	2,358	12.74
Donetsk Oblast	LOC	832	4.49
RF	LOC	831	4.49
Autonom. Republic of Crimea	LOC	812	4.39
Crimea	LOC	721	3.89
Luhansk Oblast	LOC	620	3.35
Added named entities	**Label**	**Count**	**(%)**
Russian Federation	LOC	4,127	22.29
Donetsk People's Republic	LOC	1,598	8.63
DPR	LOC	1,111	6.00
Luhansk People's Republic	LOC	1,100	5.94
Ukraine	LOC	1,068	5.77
LPR	LOC	879	4.75
Russia	LOC	625	3.38
Crimea Republic	LOC	334	1.8

Table 2: Count and percentage of the top 8 added and deleted named entities by *RWFork* editors (original Russian names are translated to English).

6 RQ3: Taxonomy of Changes

In this section, we build a taxonomy of patterns of knowledge manipulation. Our approach involves a comparative analysis of original articles from Russian Wikipedia and their modified versions. The pipeline consists of three main steps: *(1)* preliminary clustering; *(2)* clustering quality evaluation; and *(3)* cluster correction.

Preliminary Clustering

To identify groups of similar edits, we first represent each revision as a single string containing the most common change types. This representation includes features such as deleted and added sentences, pairs of changed sentences, the article title, and modifications in metadata — such as categories, tags, and templates, which are essential for defining the nature of the edit. It should be mentioned that those features

cover about 91% of edits (30,599 articles), while others are omitted in further analysis.

The strings that capture these specific edits can vary significantly in length and content. Furthermore, they often contain significant noise, including non-factual changes, irrelevant context, and parsing errors. To reduce this noise, we employ a generative model to create a fine-grained summary of each edit. Specifically, we use the OpenAI model *GPT-4o-mini-2024-07-18*, with default parameters and a temperature setting of zero. The model was selected for its efficiency, cost-effectiveness, strong performance in Natural Language Understanding (NLU) tasks, and independence from in-house infrastructure (OpenAI 2024). We provide an explicit prompt instructing the model to produce a short summary that highlights specific factual changes, constrained to a maximum length of 40 words.

Once we have the summaries of the changes, we proceed to calculate text embeddings. We use the OpenAI model *text-embedding-3-small* to generate vector representations for each summary. This model produces embeddings of size 1536, with a default normalization to a magnitude of 1. These embeddings are subsequently employed for clustering using the k-means algorithm, as suggested in previous research (Petukhova, Matos-Carvalho, and Fachada 2025). To determine the optimal number of clusters, we apply silhouette analysis, which helps identify the clustering configuration that maximizes the silhouette score. In our analysis, the optimal number of clusters is found to be 8, as indicated by the peak silhouette score.

The final step in defining the taxonomy of changes involves characterizing the identified clusters. For this objective, we again use the OpenAI model *GPT-4o-mini-2024-07-18*. For each cluster, we prompt the model to generate a name and a brief description, supplementing the query with a sample of 20 cluster-specific edit summaries. Generated description prompt to outline the specific changes made within the edits, including examples of the editing tactics employed.

Classification Quality Estimation

The previous experiment produced a taxonomy of specific types of changes to Russian Wikipedia articles. This classification was created by unsupervised modeling and each cluster was characterized based on a limited number of samples. Therefore, it is essential to evaluate how accurately each sample aligns with its respective cluster name and description.

To address this, we employ the *GPT-4o-mini-2024-07-18* model for prompt-based zero-shot binary classification, inspired by previous research (Wan et al. 2024), which reported a 'high agreement' with human raters for a similar task. We use the cluster name and description along with the sample summary, prompting the model to predict either yes (the sample aligns with the cluster name and description) or no. Details of the prompts used can be found in the Appendix B. Our analysis reveals that 78.1% of the samples were initially correctly classified. Table 3 presents the edit-to-cluster fit rate (ECFR) for each cluster along with the confidence interval (CI) for these estimates, calculated using bootstrapping, using the same approach as in Section 4.

Category	ECFR(%)	±CI(%)
Russian Legislation Medicines	99.9	0.1
Editing Caution Removal	99.9	0.1
Cultural Metadata Updates	99.6	0.2
Terminology Changes Ukraine	97.3	0.6
Metadata Updates	92.9	0.7
Territorial Claims Dispute	87.6	0.8
Sanctions Edit Adjustments	59.4	1.6
LGBT Rights and History	1.6	0.4

Table 3: Edit-to-cluster fit (ECFR) before cluster correction.

Our experiment demonstrates that the majority of clusters show strong alignment between the elements and corresponding generated cluster names and descriptions. Specifically, six out of the eight clusters demonstrate an edit-to-cluster fit exceeding 87%. However, one cluster shows a moderate fit at 59.4% (Sanctions Edit Adjustments), while another exhibits a poor fit (LGBT Rights and History). Given that approximately 22% of all samples were initially classified as not fitting their assigned group names and descriptions, we recognized the need to implement a cluster correction process, which we address in the following section.

Cluster Correction

In this section, we explain the process of redefining clusters for initially misclassified samples using zero-shot multiclass classification applied to the previously defined taxonomy. Specifically, we prompted the *GPT-4o-mini* model to match the edit summary to a relevant class, providing cluster names and descriptions in the prompt. Also, we added a new category titled "Other Changes" to allow the model to return this option when an edit does not align with any of the provided classes.

As previously indicated, the most problematic cluster was "LGBT Rights and History." We hypothesized that the issue arose from the cluster's name and description lacking sufficient generality. We believed that this cluster primarily contained unique changes across various topics, making it difficult to group them into a distinct category. Our correction procedure confirmed this assumption: approximately 65% of the misclassified samples from this cluster were reassigned to the "Other Changes" category, with the remaining distributed among other existing classes. We repeated the evaluation procedure described in the previous section and concluded that the proposed cluster correction increased the ECFR to 92%.

The final taxonomy of *RWFork* changes along with their quantitative measurements are presented in Table 4. We observe that the most frequent changes refer to territorial reassignments, accounting for 24.19% of all edits. This category represents shifts of occupied territorial entities from Ukraine to Russia. Additionally, significant groups related to the 2022 Russian invasion of Ukraine include "Terminology

Name	Description	Size
Territorial Claims Dispute	Edits reflect changes in territorial designations and governance, emphasizing claims by the Donetsk and Luhansk People's Republics while removing Ukrainian references and administrative details.	24.19%
Metadata Updates	Various edits focused on updating metadata templates, removing outdated references, and refining geographical classifications across multiple Wikipedia pages.	18.24%
Cultural Metadata Updates	Edits focused on updating metadata with locations, cultural topics, and adding age and gender templates for various pages.	11.77%
Terminology Changes Ukraine	Edits focus on altering terminology related to the Russia-Ukraine conflict, shifting from specific invasion references to broader military actions and general policies.	11.24%
Editing Caution Removal	Multiple Wikipedia pages had the editing caution category removed, indicating a change in the perceived necessity for careful editing.	10.37%
Sanctions Edit Adjustments	The edits focus on removing specific references to the Ukraine conflict in sanctions descriptions, simplifying statements, and altering context around individuals and entities sanctioned.	8.23%
Russian Legislation Medicines	Templates for Russian legislation and medications were added to various pharmaceutical pages, enhancing their categorization and relevance.	5.32%
LGBT Rights and History	The edits focus on updating and clarifying information related to LGBT rights, historical events, and notable figures, while removing outdated or derogatory content.	0.48%
Other changes	The edit does not fit any of the provided clusters.	10.16%

Table 4: Final taxonomy of changes. Cluster names, descriptions, and sizes.

Changes Ukraine" and "Sanctions Edit Adjustments", representing 11.24% and 8.23% of the total edits, respectively. These clusters reflect modifications aimed at setting specific narratives, such as the removal of terms like "invasion" and "war", as well as adjustments to information related to sanctions across various contexts.

There are two groups that mostly consist of non-textual edits: "Editing Caution Removal" and "Russian Legislation Medicines". These categories refer to automated changes that either remove specific edit notice categories or add tags related to custom legal information. The "Cultural Metadata Updates" cluster consists of edits related to locations and cultural, sexual, and gender-related topics. These modifications, for example, involve specific labeling with a *(+18) tag* on the pages that refer to explicit content (adult films and actors). Also, we detected a small cluster of edits that were related to the topics of "LGBT Rights and History".

7 Discussion

In this paper, we have presented an empirical analysis of knowledge manipulation in Wikipedia. Previous research already explored knowledge gaps in Wikipedia (Redi et al. 2020) and highlighted cultural biases across various language editions (Hecht and Gergle 2009, 2010; Callahan and Herring 2011; Rogers, Sendijarevic et al. 2012; Pentzold et al. 2017; Miquel-Ribé, Laniado, and Kaltenbrunner 2021). However, our findings come from a distinct editorial process: the creation of a new platform that copied original Wikipedia content, which is then manipulated to meet the requirements of a national regulation. As a result, *RWFork* also differs from previously studied wiki-based encyclopedias like *Conservapedia* (Johnson 2007), which was created from scratch.

The proposed study can be effectively replicated in other Wikipedia forks or collaborative platforms. Examples of such forks include *Runiversalis*, a wiki-based encyclopedia aligned with traditional values, and *Hamichlol*, a censored wiki-based encyclopedia project for the Haredi community, among others. The methodology's adaptability lies in its ability to identify and categorize differences driven by the unique editorial policies of each wiki-based fork.

The first step of our study focused on the relevance of Russian Wikipedia articles changed by *RWFork* editors **(RQ1)**. Our analysis revealed that although the proportion is relatively small, there are articles receiving remarkable attention from readers on Russian Wikipedia. Furthermore, the articles that were altered in *RWFork* receive more edits and reverts in Russian Wikipedia than those that remained unchanged. Building on previous research that has used editing and reverting activities to identify controversial topics on Wikipedia (Pentzold et al. 2017; Yasseri et al. 2012), our observations suggest that manipulation may have taken place in popular and contentious articles. This has important implications, as controversy itself is not necessarily a negative indicator of article content. In fact, Shi et al. (2019) found that Wikipedia articles edited by polarized groups of contributors typically exhibit higher quality. Therefore, if popular and controversial articles on Russian Wikipedia are forked to be edited in alignment to Russian legislation, their quality is expected to be affected.

We have then conducted a thorough analysis of how the content of articles changed in *RWFork* **(RQ2)**. The analysis of editing time preferences shows that *RWFork* are more likely to be active during standard office hours than Russian Wikipedia editors. As the initial plan for this platform is to rely on experts (Cohen 2023), one possible explanation could be that much *RWFork* editorial activity is driven by paid workers. Alternatively, the geographical distribution of editors could also contribute to these differences, with Russian Wikipedia editors being more dispersed. This interpretation is consistent with the findings of Yasseri, Sumi,

and Kertész (2012), who indicated that although most native Persian speakers reside in Iran, a significant portion of editing activity in the Persian Wikipedia originates from communities outside the country. The remainder of the analysis – encompassing article geography, categories, sources, and named entities – reveals a clear trend: most changes are related to the 2022 Russian invasion of Ukraine. Previous research has shown that Wikipedia coverage can be influenced by the community's self-focus (Hecht and Gergle 2009) while also highlighting the success of Wikipedia communities in preventing state-sponsored information operations in articles about this conflict (Roberts and Xiong-Gum 2022; Dammak and Lemmerich 2023; Kurek, Budak, and Gilbert 2024). Therefore, there may be a link between this specific topic, as a core focus of knowledge manipulation within *RWFork*, and the project founder's declared goal of ensuring compliance with Russian regulatory requirements (Cohen 2023).

Our last effort was focused on building a taxonomy of patterns of knowledge manipulation in *RWFork* **(RQ3)**. To achieve this goal, we developed a robust clustering pipeline that incorporates intermediate steps designed to ensure the quality of the process. Although many clusters are associated with the 2022 Russian invasion of Ukraine, other topics of social importance also emerge, such as "Russian Legislation on Medicines" and "LGBT Rights and History". As previously noted, Wikipedia content is extensively used as a primary resource for LLMs (Devlin et al. 2019). Studies have shown that AI systems can form significantly different associations between adjectives and political concepts based on whether they are trained on Wikipedia content or on content from web encyclopedias subject to national regulations (Yang and Roberts 2021). In light of the recently announced plans to integrate *RWFork* into the training process for future LLMs (Davydov 2023) and the increasing societal impact of AI, our study seeks to raise awareness of the critical importance of closely examining the quality, neutrality, and potential biases of knowledge repositories.

Limitations and Future Work. The *RWFork* platform is relatively new. It is actively running and regularly introducing new content modifications. Consequently, this ongoing process may lead to the emergence of new types of changes. Also, *RWFork* provides limited access to data compared to Russian Wikipedia. Therefore, our analysis did not include *RWFork* newly created pages. It might result in missing other, undiscovered types of knowledge manipulation, but since creating new pages requires more resources, their likely limited number suggests minimal impact on our findings. Future work could address this by finding those pages through parsing *RWFork* internal links between articles (Piccardi, Gerlach, and West 2022). Also, we rely on the assumption that *RWFork* pages have similar geographical, categorical, and topical features as in Russian Wikipedia. Moreover, *RWFork* is only one of several MediaWiki-powered websites. We therefore plan to replicate this study with alternative encyclopedias in future work.

Additionally, we're experimenting with a specific and limited set of models for summarization, zero-shot classification, embeddings, and clustering. We acknowledge that using different or more advanced models could improve our results, and we consider exploring this in future work.

Acknowledgments

The work of Mykola Trokhymovych is funded by MCIN/AEI /10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M).

References

Adler, B. T.; De Alfaro, L.; Mola-Velasco, S. M.; Rosso, P.; and West, A. G. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of CICLing 2011*, 277–288. Springer.

Beytía, P. 2020. The positioning matters: Estimating geographical bias in the multilingual record of biographies on wikipedia. In *Companion Proceedings of the Web Conference 2020*, 806–810.

Burkholder, P.; and Schaffer, D. 2021. A Snapshot of the Public's Views on History.

Callahan, E. S.; and Herring, S. C. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10): 1899–1915.

Cohen, N. 2023. Russian Wikipedia's Top Editor Leaves to Launch a Putin-Friendly Clone. https://www.bloomberg.com/news/articles/2023-07-12/russian-wikipedia-editor-leaves-to-launch-a-putin-friendly-clone. Accessed: 2023-09-08.

Dammak, Z.; and Lemmerich, F. 2023. Effects of the Russo-Ukrainian War on the Editor Activity of the Ukrainian, Russian, and English Wikipedias. *Wikiworkshop*.

Davydov, O. 2023. The encyclopedia "Ruwiki" is launched in Russia. Its creator - about haters and why it should not be confused with "Wikipedia". https://lenta.ru/articles/2023/06/29/ruwiki/. Accessed: 2023-09-27.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL'19*, 4171–4186. Minneapolis, Minnesota.

Efron, B.; and Tibshirani, R. 1994. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC, 1st edition edition. ISBN 9780429246593.

English Wikipedia. 2023. Verifability Policy. https://en.wikipedia.org/wiki/Verifiability.

Eom, Y.-H.; Aragón, P.; Laniado, D.; Kaltenbrunner, A.; Vigna, S.; and Shepelyansky, D. L. 2015. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS one*, 10(3): e0114825.

Fitts, A. S. 2017. Welcome to the Wikipedia of the Alt-Right — Backchannel — wired.com. https://www.wired.com/story/welcome-to-the-wikipedia-of-the-alt-right/. [Accessed 14-01-2025].

Ford, H.; Sen, S.; Musicant, D. R.; and Miller, N. 2013. Getting to the Source: Where Does Wikipedia Get Its Information From? In *Proceedings of WikiSym '13*, WikiSym '13.

Geiger, R. S.; and Ribes, D. 2010. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 117–126.

Graham, M.; Hale, S.; and Stephens, M. 2011. Geographies of the World's Knowledge. *Oxford Internet Institute*.

Graham, M.; Hogan, B.; Straumann, R. K.; and Medhat, A. 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4): 746–764.

Greenstein, S.; and Zhu, F. 2018. Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia. *MIS Q.*, 42(3): 945–960.

Hecht, B.; and Gergle, D. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*, 11–20.

Hecht, B. J.; and Gergle, D. 2010. On the" localness" of user-generated content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 229–232.

Hinnosaar, M. 2019. Gender inequality in new media: Evidence from Wikipedia. *Journal of economic behavior & organization*, 163: 262–276.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Johnson, B. 2007. Conservapedia–the US religious right's answer to Wikipedia.

Johnson, I.; Gerlach, M.; and Sáez-Trumper, D. 2021. Language-Agnostic Topic Classification for Wikipedia. In *Companion Proceedings of WWW'21*, WWW '21, 594–601.

Kharazian, Z.; Starbird, K.; and Hill, B. M. 2023. Governance Capture in a Self-Governing Community: A Qualitative Comparison of the Serbo-Croatian Wikipedias. *arXiv preprint arXiv:2311.03616*.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings WWW'17*, 857–866.

Kumar, S.; West, R.; and Leskovec, J. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of WWW'16*, 591–602.

Kurek, L.; Budak, C.; and Gilbert, E. 2024. Wikipedia in Wartime: Experiences of Wikipedians Maintaining Articles About the Russia-Ukraine War. *arXiv preprint arXiv:2409.02304*.

Lemmerich, F.; Sáez-Trumper, D.; West, R.; and Zia, L. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of WSDM '19*, 618–626.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8): 707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

McDowell, Z. J.; and Vetter, M. A. 2020. It takes a village to combat a fake news army: Wikipedia's community and policies for information literacy. *Social Media+ Society*, 6(3): 2056305120937309.

Miquel-Ribé, M.; Laniado, D.; and Kaltenbrunner, A. 2021. The role of local content in Wikipedia: A study on reader and editor engagement. *Área Abierta. 2021; 21 (2): 123-151*.

Niverthi, M.; Verma, G.; and Kumar, S. 2022. Characterizing, detecting, and predicting online ban evasion. In *Proceedings of the ACM Web Conference 2022*, 2614–2623.

OpenAI. 2024. GPT-4o Mini: Advancing Cost-Efficient Intelligence. Retrieved December 29, 2024, from https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Pentzold, C.; Weltevrede, E.; Mauri, M.; Laniado, D.; Kaltenbrunner, A.; and Borra, E. 2017. Digging Wikipedia: The online encyclopedia as a digital cultural heritage gateway and site. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(1): 1–19.

Petukhova, A.; Matos-Carvalho, J. P.; and Fachada, N. 2025. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6: 100–108.

Piccardi, T.; Gerlach, M.; and West, R. 2022. Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions. In *Companion Proceedings of the Web Conference 2022*, WWW '22, 1324–1330. ISBN 9781450391306.

Piccardi, T.; Redi, M.; Colavizza, G.; and West, R. 2020. Quantifying Engagement with Citations on Wikipedia. In *Proceedings of The Web Conference 2020*, 2365–2376.

Piccardi, T.; Redi, M.; Colavizza, G.; and West, R. 2021. On the Value of Wikipedia as a Gateway to the Web. In *Proceedings of WWW'21*, 249–260.

Potthast, M. 2010. Crowdsourcing a Wikipedia vandalism corpus. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 789–790.

Potthast, M.; Stein, B.; and Gerling, R. 2008. Automatic vandalism detection in Wikipedia. In *ECIR*, 663–668. Springer.

Reagle, J.; and Rhue, L. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5: 21.

Redi, M.; Gerlach, M.; Johnson, I.; Morgan, J.; and Zia, L. 2020. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*.

Roberts, L. E.; and Xiong-Gum, M. N. 2022. Wikipedia Editing as Connective Intelligence: Analyzing the Vandal Fighter Role in the "2022 Russian Invasion of Ukraine" Wikipedia Article. In *Proceedings of the 40th ACM International Conference on Design of Communication*, 55–62.

Rogers, R.; Sendijarevic, E.; et al. 2012. Neutral or national point of view? A comparison of Srebrenica articles across Wikipedia's language versions.

Runiversalis. 2024. Encyclopedia Runiversalis. https://runiversalis.ru/. Accessed: 2024-09-15.

Saez-Trumper, D. 2019. Online Disinformation and the Role of Wikipedia. arXiv:1910.12596.

Sakib, M. N.; and Spezzano, F. 2022. Automated detection of sockpuppet accounts in wikipedia. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 155–158. IEEE.

Sezer, C.; and Dolan, D. 2017. Turkey blocks access to Wikipedia. https://www.reuters.com/article/us-turkey-security-internet-wikipedia-idUSKBN17V06Q. Accessed: 2023-10-03.

Shachaf, P.; and Hara, N. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3): 357–370.

Shi, F.; Teplitskiy, M.; Duede, E.; and Evans, J. A. 2019. The wisdom of polarized crowds. *Nature human behaviour*, 3(4): 329–336.

Siegel, R. 2019. China bans Wikipedia in all languages. https://www.washingtonpost.com/business/2019/05/15/china-bans-wikipedia-all-languages/. Accessed: 2023-09-27.

Steinsson, S. 2024. Rule ambiguity, institutional clashes, and population loss: how Wikipedia became the last good place on the internet. *American Political Science Review*, 118(1): 235–251.

Sutcliffe, D. 2016. Wikipedia's Ongoing Search for the Sum of All Human Knowledge. *Medium (Jan 20, 2016): Retrieved*, 20(5): 2018.

Trokhymovych, M.; Aslam, M.; Chou, A.-J.; Baeza-Yates, R.; and Saez-Trumper, D. 2023. Fair Multilingual Vandalism Detection System for Wikipedia. In *Proceedings of KDD '23*, 4981–4990.

Trokhymovych, M.; and Saez-Trumper, D. 2021. WikiCheck: An End-to-End Open Source Automatic Fact-Checking API Based on Wikipedia. In *Proceedings of CIKM '21*, 4155–4164.

Vincent, N.; and Hecht, B. 2021. A deeper investigation of the importance of Wikipedia links to search engine results. *Proceedings HCI'21*, 5(CSCW1): 1–15.

Vincent, N.; Johnson, I.; Sheehan, P.; and Hecht, B. 2019. Measuring the importance of user-generated content to search engines. In *Proceedings of ICWSM'19*, volume 13, 505–516. Munich, Germany: AAAI.

Wagner, C.; Graells-Garrido, E.; Garcia, D.; and Menczer, F. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ data science*, 5: 1–24.

Wan, M.; Safavi, T.; Jauhar, S. K.; Kim, Y.; Counts, S.; Neville, J.; Suri, S.; Shah, C.; White, R. W.; Yang, L.; Andersen, R.; Buscher, G.; Joshi, D.; and Rangan, N. 2024. TnT-LLM: Text Mining at Scale with Large Language Models. In *Proceedings of KDD '24*, 5836–5847.

Woo, E. 2007. Baidu's Censored Answer to Wikipedia. https://www.bloomberg.com/news/articles/2007-11-13/baidus-censored-answer-to-wikipediabusinessweek-business-news-stock-market-and-financial-advice. Accessed: 2023-10-05.

Yang, E.; and Roberts, M. E. 2021. Censorship of Online Encyclopedias: Implications for NLP Models. In *Proceedings of FAccT '21*, 537–548. New York, NY, USA.

Yasseri, T.; Sumi, R.; and Kertész, J. 2012. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1): e30091.

Yasseri, T.; Sumi, R.; and Kertész, J. 2012. Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLOS ONE*, 7(1): 1–8.

Yasseri, T.; Sumi, R.; Rung, A.; Kornai, A.; and Kertész, J. 2012. Dynamics of conflicts in Wikipedia. *PloS one*, 7(6): e38869.

Zagovora, O.; Flöck, F.; and Wagner, C. 2017. "(Weitergeleitet von Journalistin)" The Gendered Presentation of Professions on Wikipedia. In *Proceedings of the 2017 ACM on web science conference*, 83–92.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. We summarize our methodology in Section 1, "Introduction". Later it is explained in detail in Sections 3, 4, 5, and 6

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we are explaining in details the data sources (see collection details (see Section 3), and data characteristics (see Section 4 and Section 5)

   (e) Did you describe the limitations of your work? Yes. Please see Section 8.

   (f) Did you discuss any potential negative societal impacts of your work? Yes. Please see Section 8. Overall, we do not foresee any significant negative impact.

   (g) Did you discuss any potential misuse of your work? We don't see any potential misuse of our work.

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we have a detailed description of the methodology used to collect, process, and analyze the data. We are also publishing the code used for analysis, which may help to reproduce our results.

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypothesis testing...

(a) Did you clearly state the assumptions underlying all theoretical results? N/A

(b) Have you provided justifications for all theoretical results? N/A

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A

(e) Did you address potential biases or limitations in your theoretical framework? N/A

(f) Have you related your theoretical results to the existing literature in social science? N/A

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we provide links to both the code repository and the Zenodo repository where the data is stored.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? Yes, we appropriately citing them, and mentioning type of license (e.g., please see Section 5).

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we are providing the link to the code in Section 1.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. We discuss it in Section 8.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. In Section 1, "Introduction", we mention that we will publish the dataset on the Zenodo platform under an open license.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? We are planning to create a Datasheet.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? N/A

Ethical Statement

In this work, we have only collected and analyzed openly available data. Our methodology does not include techniques to identify users or link profiles across platforms. The data collected does not include any private information. We confirm that we have read and abided by the code of conduct.

A Content Changes Examples

In Table 5, we present a representative sample of the collected data, showing various types of content changes made to different pages.

B Technical Details

Here we provide additional technical details to help interpret the results and improve reproducibility. Specifically, we present the prompt used for text summarization (see Figure 7), edit-cluster alignment evaluation (see Figure 8), and cluster reassignment (see Figure 9).

Page title	Content changes
Zburevsky Kut	**Lines changed:** [("Zburevsky Kut or Zburevsky Liman is a lake, bay in the Dnipro delta, located on the territory of the Skadovsky district *(Kherson region, Ukraine)* "Zburevsky Kut or Zburevsky Liman is a lake, a bay in the Dnipro delta, located on the territory of the Skadovsky district *Kherson region of Russia*")]
Medal "For the Return of Crimea"	**Categories removed:** ["Articles with edit notes about caution when editing"]
Vladimir Samokish	**Lines changed:** ["*Due to Russia's invasion of Ukraine*, he is *under international sanctions* of the European Union, USA, Great Britain, and other countries", "He is *on the sanctions list* of the European Union, USA, Great Britain, and other countries")] **Categories changes:** ["Persons under the sanctions related to the conflict in Ukraine"]
Trimetozine	**Tag added**: Legislation of the Russian Federation—Medicines
Kendra (name)	**Lines deleted:** ["1981) - American porn actress and erotic model", "Sunderland, Kendra (born — 1995) - American porn actress and erotic model"]
Svoya igra	**Lines changed:** ["Broadcast on NTV on *Sundays* at 15:00", "Broadcast on NTV on *weekends* at 15:00")]

Table 5: Sample of content changes of various types made by *RWFork* editors (original Russian texts translated to English).

```
You will be provided with details regarding edit to the Wikipedia page.
You need to deeply analyse the changes, define what was edited and provide a
    description of the changes.

Provide a short summary and specific factual changes.
Pay attention to details about adding/removing/changing characteristics.
Avoid generalizations and provide specific examples.  (max {MAX_WORDS} words)
Return the answer in JSON format with only "desc" field and the following structure:
{{
    "desc": "string"  # description of specific factual changes
}}
The edit to analyze will be provided in the <>: <{EDIT_STRING}>
```

Figure 7: Prompt template used to define the summary for content changes.

```
You are provided with a specific edit to the Wikipedia page (defined in <>) along with
     possible cluster details (defined in ~~) to which the edit belongs.
You need to analyse the edit and decide whether the edit fits the provided cluster or
    not.
Provide ONLY a short answer (YES or NO).
Edit summary: <{EDIT_SUMMARY}>
Cluster details: ~{CLUSTER_DETAILS}~
```

Figure 8: Prompt template used to define the edit-cluster alignment.

```
You are provided with a specific edit to the Wikipedia page (defined in <>).
You need to reclassify the edit to the correct cluster based on the provided cluster
    details.
Cluster details:
{ALL_CLUSTERS_DETAILS}
8. Other changes: The edit does not fit any of the provided clusters. (always use this
     option if the edit does not fit any of the provided clusters)

Provide ONLY a short answer (cluster number).
Edit summary: <{EDIT_SUMMARY}>
```

Figure 9: Prompt template used for cluster reassignment.