Digital battlegrounds: The role of Wikipedia in armed conflict information warfare

Marieth Coetzer[1*], **Leopold Augustin**[1*], **and Fabian Braesemann**[1,2,3**]

[1]Oxford Internet Institute, University of Oxford, Oxford, UK
[2]Einstein Center Digital Future (ECDF), Berlin, Germany
[3]DWG Data Science Company, Berlin, Germany
[*]Both authors contributed equally to this work.
[**]Corresponding author: fabian.braesemann@oii.ox.ac.uk

ABSTRACT

In the digital age, open knowledge-sharing platforms such as Wikipedia are crucial hubs in the global information network. As an open-source encyclopedia, Wikipedia permits anyone to view and edit content, frequently reflecting real-world events. As a consequence, the platform has become an up-to-date public memory providing a 'consensus truth' to its readers. However, not all edits are neutral, particularly on high-stakes topics such as armed conflicts. Partisan users may attempt to shape public perception by presenting events in ways that align with their narrative, potentially undermining trust in the platform. These biased edits can also propagate into AI models, since Wikipedia is a key source of training data for large language models.

In this study we examine the extent to which Wikipedia has become a 'digital battleground' of the Russo-Ukrainian war. We analyze Wikipedia edit data from before and after the Russian invasion of Ukraine in February 2022 and observe an increase in revisions and disputes on articles covering contested Ukrainian regions. Content analysis shows that many of these edits are directly related to the conflict, often involving changes in the spelling of place names that reflect competing national identities. These results show that Wikipedia functions as a contested site in digital information warfare.

To mitigate the spread of misinformation on Wikipedia during armed conflicts, we develop an early-warning tool that flags articles likely influenced by partisan edits. Our model combines Wikipedia metrics and external news data to accurately predict more than 80% of Wikipedia page locks (i. e., Wikipedia's ultimate measure to stop external editors from rewriting pages). Maintaining trust in the platform requires awareness of how 'edit wars' shape available content. To remain a credible knowledge source during conflicts, editors and users must recognize the influence of ongoing 'edit wars.'

Introduction

Does the phrase "history is written by the victors" still hold true in the digital era? With the advent of platforms like Wikipedia, where anyone can contribute, history seems more democratized than ever before. However, this openness also introduces risks to accuracy and neutrality.

Wikipedia has become a vital and widely respected information source despite its crowdsourced nature. It is one of the most visited websites globally[1], and studies suggest its accuracy rivals that of traditional encyclopedias like Encyclopedia Britannica[2]. The English version of Wikipedia contains over 6 million articles[3], covering a wide range of topics, from political candidates to armed conflict. It is one of the first results in most search engines and is used by Apple and Amazon to formulate the answers of their virtual assistants[4].

As a result, Wikipedia is often viewed as a "consensus truth" in many Western societies[5]. The platform's importance has grown since the rise of large language models, which rely heavily on vast, high-quality, human-generated content like that found on Wikipedia[6,7]. In fact, Wikipedia is the second largest source of data in Google's training set for large language models[8], leading to the claim that "Without Wikipedia, generative A.I. wouldn't exist"[9]. This development highlights Wikipedia's crucial role for those aiming to influence the "truth" displayed online.

The strategic manipulation of online narratives has become an increasing concern, particularly in the context of political events and territorial disputes[10,11]. In the case of territorial conflicts, there is strong evidence that such disputes are shifting into the digital realm through forms of information warfare[12]. Specifically, the Russo-Ukrainian war is one of the most recent and prominent examples of a territorial conflict which led to strong digital reactions. Many facets of this digital dispute are well explored, such Russia's use of television social media[13–15]. Citizen participation also played a significant role in contributing to the spread of unverified content online[16]. These findings highlight the many facets of digital dispute surrounding the Russo-Ukrainian war, where both state and civilian actors seek to control narratives.

Like other digital battlegrounds such as social and mass media[17,18], Wikipedia plays an important role in information warfare, especially for conflicts about the “truth" between opposing parties. Previous research efforts have examined Wikipedia editors’ response to the 2022 invasion[19] and observed misleading narratives on the overarching war page[20]. However, there are suspicions that Russia and Ukraine are aiming to shape digital narratives of the conflict on Wikipedia[21,22] that have yet to be explored. These concerns are reinforced by reports of two Wikipedia editors in Belarus being arrested for edits related to the Russo-Ukrainian War Wikipedia page[23,24]. Digital disputes on territorial conflicts extend beyond conflicted regions by influencing allied states, where the conflict may become part of an aid campaign or election cycle. This new front in the conflict presents significant challenges for the Wikimedia Foundation and its editor community as they attempt to uphold content neutrality.

In support of upholding neutrality, Wikipedia policy mandates that all articles should be based on reliable external sources[25]. Prior research[26] found that most citations come from mainstream news outlets such as the BBC. When adhered to, this policy limits the inclusion of narratives that deviate from reliable sources. However, during cases where editor disputes escalate, Wikipedia administrators implement a page-locking mechanism. This means that only experienced editors with a history of high-quality contributions can alter the page. Page-locking is not applied preemptively[27], leading to the concern that readers seeking a brief overview on the page, which happens frequently[28], may unknowingly read contested content during a digital dispute.

Understanding the dynamics of conflict of Wikipedia has been the subject of much research, primarily centered on using internal Wikipedia metrics to identify disputes[29] (known as edit wars) and dynamics of editors interactions[30,31]. Using these insights to predict when Wikipedia disputes occur remains largely unexplored. As the information on Wikipedia is continuously changing, identifying when digital disputes are likely to occur is an essential part of maintaining article quality.

Motivated by the relevance of digital battlegrounds and the findings of former research[19,20] we examine whether the warfare in Ukraine has extended into the digital realm, influencing Wikipedia’s representation of contested regions. We specifically analyze whether edits on Wikipedia articles about disputed regions include signals of dispute, and whether the levels of attention and dispute have increased following the 2022 invasion.

RQ 1.a. Can we identify dispute signals in the edits of Wikipedia articles about disputed Ukrainian regions?

RQ 1.b. Do attention and dispute on these pages increase after the 2022 invasion?

To detect disputes on Wikipedia we train a model to predict the week a page is likely to be locked. Specifically, we develop a model using internal metrics identified in previous research on Wikipedia disputes, and then compare its performance with a second model which incorporates external news sources.

RQ 2.a. Can we predict digital disputes (defined as page locks) using internal Wikipedia metrics?

RQ 2.b. Do external sources improve this prediction?

We find that a significant share of Wikipedia edits contains signals of dispute between authors. These edits are closely tied to the conflict, often involving alternate Ukrainian versus Russian spellings of the contested areas, and are predominantly found in revisions of articles about disputed Ukrainian regions. Furthermore, articles on disputed territories receive significant higher attention in the form of revisions, reverts, and dispute edits compared to articles about undisputed states. These findings advance our understanding of how the Russo-Ukrainian war manifests in the digital realm, adding Wikipedia as a key battleground for influence and control over public narratives.

By training a model on internal Wikipedia metrics, we find that we can effectively predict when a page is likely to be locked. While the model was primarily trained on pages from the Russo-Ukrainian war, it also included pages from other conflicts to enhance generalizability. Incorporating an external news source as a predictor further improved the model’s performance over several machine learning metrics. The model was optimized to prioritize recall so that false positives could serve as early signals of emerging disputes. These findings validate the concept of conflict forecasting in online environments, and offer a framework to develop early warning systems on digital battlegrounds.

The paper is structured as follows: In the "Results" section, we introduce the data and methods used to quantify dispute edits and predict disputes. We employ a domain-specific identification method for the former and a generalizable approach for the latter. Next, we introduce a difference-in-difference regression analysis to quantify the different reactions of attention and dispute between disputed region articles and the comparison groups. For the prediction task, we compare the performance of random forest models trained on internal Wikipedia and external news metrics, evaluating their results on a held-out test set. The "Discussion" section outlines the limitations of the study and suggests directions for future research. Finally, the "Methods" section provides a detailed overview of the data, variables, regression, and random forest framework. Robustness checks and supplementary analyses are included in the Supplementary Information.

Results

Data

We focus our data collection on the recent escalation of the Russo-Ukrainian conflict, beginning with Russia's invasion of Ukraine on February 24, 2022[32]. Following the method of similar studies[19,33], our dataset spans from January 1, 2020, to December 31, 2023, incorporating data both before and after the escalation. The two-year period before and after the invasion date helps validate the parallel-trend assumption for the difference-in-difference regression. Given that the invasion occurred in early 2022, we consider the end of 2023 a comparable timeframe to the two years after the conflict.

To identify the impact of the 2022 invasion on dispute signals (RQ 1), we classify regions into three groups: disputed Ukrainian Oblasts, undisputed Ukrainian Oblasts, and Polish states, known as Voivodeships. A disputed Oblast is one in which Russian forces have acquired territory either in 2014 or 2022. We used the Armed Conflict Location Events Dataset (ACLED)[34] to identify all Oblasts that were occupied by Russian forces.

To detect when digital disputes occur (RQ 2), we examine the main conflict pages which provide an overview of the event and identify linked pages. As well as including regions, these linked pages include other classifications such as individuals, events, and miscellaneous categories as identified by similar research[35]. We also include pages from other armed conflicts to ensure the model's generalizability beyond the Russo-Ukraine conflict[36].

We retrieve the full revision history for each Wikipedia page within the defined timeframe using the Wikipedia API[37]. The complete list of the pages can be found in the Supplementary Information. To acquire related BBC articles, we input the name of the Wikipedia articles into News API[38] and extract articles containing these names. Figure 1**A**. illustrates the data collection timeline and provides an overview of the data used to address the research questions.

Quantifying Digital Disputes

Due to the complex dynamics of editor interactions on Wikipedia, no single metric captures the nuances of digital disputes[39]. Therefore, our study examines and combines several metrics in response to the research questions. To maintain a balance between find-grained and broad data, we calculate our metrics on a weekly scale.

We first use the number of revisions, which serves as a proxy for editorial attention on an article[33]. However, revision counts alone cannot distinguish between neutral and biased edit activity, nor do they capture disagreement between editors[40]. Thus, we include reverts, defined as the complete undoing of a previous revision, as a second metric[33]. Although simple, the number of reverts are widely used in controversy detection and serve as a reliable indicator of editorial conflict[29,40–43].

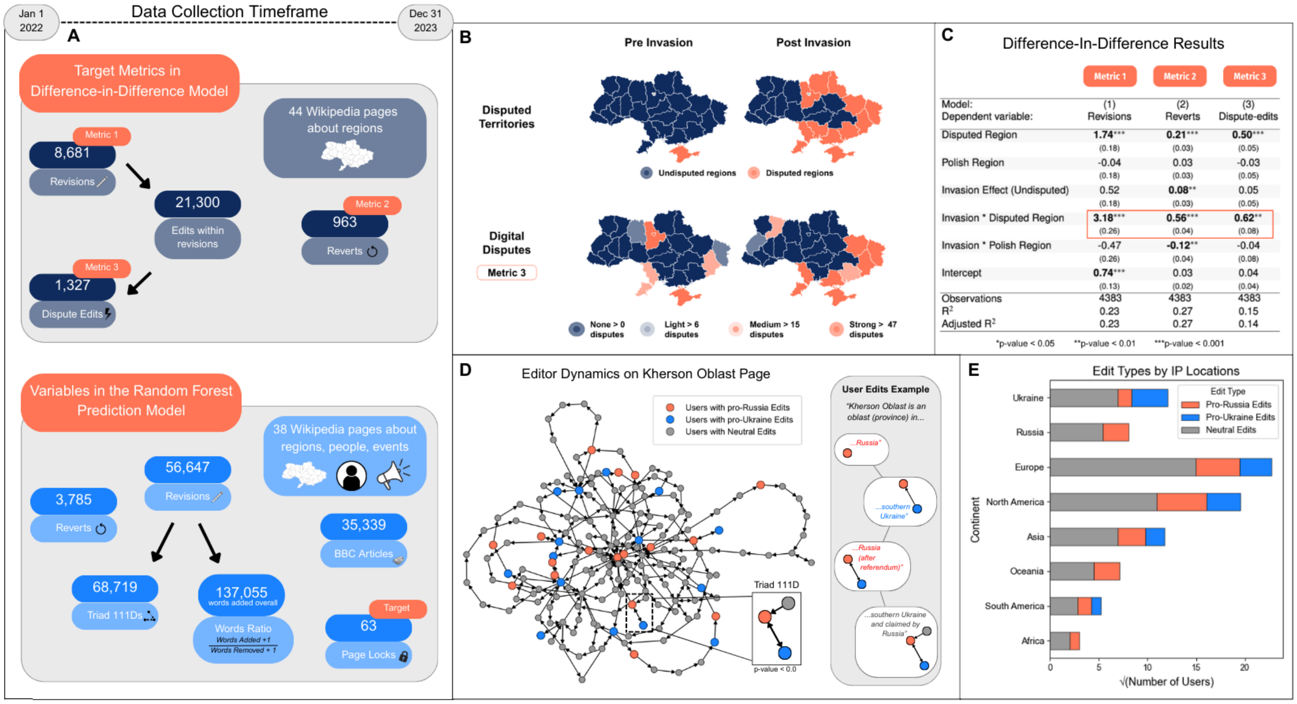
Given the established role of revisions and reverts in identifying Wikipedia disputes, we use both to identify dispute signals on pages related to Ukrainian regions and to predict edit wars. To account for their limitations, we incorporate additional metrics in our regression and prediction models. An overview of these selected metrics is provided in Figure 1**A**, with detailed explanations provided in the following sections.

Domain-Specific Metrics for Ukrainian Regions

For our regression model, we introduce the dispute edits metric. A single Wikipedia revision can include multiple changes, such as additions, removals, and substitutions. We define these as edits. After processing the data, we identify 21,300 edits across 8,631 revisions. These edits were classified as either neutral edits, which simply add information, or dispute edits, which include narrative changes favoring either Russia or Ukraine. Pre-trained NLP classifiers, typically used for hate speech detection[44] or sentiment classification[45], struggle on domain-specific changes. For example, general models for identifying dispute on Wikipedia articles[46] might overlook nationalistic-inclined edits like replacing Kyiv with Kiev. Thus, we developed a domain-specific approach to detect disputes related to the Russo-Ukraine war.

Analyzing the most common word replacements between revisions, we identify two categories of dispute identifiers. The first involves edits which change the nationhood from Russia to Ukraine or vice versa. Such changes can be indicated by replacing “Ukrainian" with “Russian" or “is" with “was". The second includes edits which changes a city's name from the Russian to Ukrainian spelling or vice versa. The correct spelling of regions strengthens the Ukrainian sense of identity and undermines Russian attempts to refute it[47]. There have been multiple efforts by the Ukrainian government to inform the international community of how to spell certain Ukrainian regions and cities, such as the “KyivnotKiev" campaign[48].

We find 1,327 edits containing dispute signals. Figure 1**B** visualizes the number of dispute edits across regions of Ukraine before and after invasion, and compares them to territorial disputes as defined earlier. Based on the figure, one can observe an increase in attention and dispute signals across most articles about territorially disputed regions after the invasion. These



Model: Dependent variable:	(1) Revisions	(2) Reverts	(3) Dispute-edits
Disputed Region	$\mathbf{1.74}^{***}$	$\mathbf{0.21}^{***}$	$\mathbf{0.50}^{***}$
	(0.18)	(0.03)	(0.05)
Polish Region	-0.04	0.03	-0.03
	(0.18)	(0.03)	(0.05)
Invasion Effect (Undisputed)	0.52	$\mathbf{0.08}^{**}$	0.05
	(0.18)	(0.03)	(0.05)
Invasion * Disputed Region	$\mathbf{3.18}^{***}$	$\mathbf{0.56}^{***}$	$\mathbf{0.62}^{**}$
	(0.26)	(0.04)	(0.08)
Invasion * Polish Region	-0.47	$\mathbf{-0.12}^{**}$	-0.04
	(0.26)	(0.04)	(0.08)
Intercept	$\mathbf{0.74}^{***}$	0.03	0.04
	(0.13)	(0.02)	(0.04)
Observations	4383	4383	4383
R^2	0.23	0.27	0.15
Adjusted R^2	0.23	0.27	0.14

*p-value < 0.05 **p-value < 0.01 ***p-value < 0.001

Figure 1. (**A**) Overview of the data collection timeframe. The study period spanned January 1, 2020 to December 31, 2023. The Wikipedia API was used to retrieve full revision histories for all pages, and related newspaper articles were collected via News API. (**B**) Comparison between territorially disputed regions in Ukraine (upper panel) and digitally disputed Ukrainian regions on Wikipedia (lower panel). Digital dispute is defined as the number of edits containing dispute signals within the respective timeframe. Dispute thresholds were defined as follows: 15 represents the average number of dispute edits per page; 47 represents the average plus one standard deviation. (**C**) Difference-in-differences regression results linking Wikipedia dispute activity to territorial conflict. Territorially undisputed Ukrainian regions serve as the baseline. Seven-day rolling averages were calculated for three outcome variables: (1) revisions, (2) reverts, and (3) dispute edits. (**D**) Network visualization showing editor dynamics on the Kherson Oblast Wikipedia page. Node colors indicate whether users produced a majority of pro-Russia or pro-Ukraine edits, with example edits shown. (**E**) Geographic distribution of editors contributing to Russian and Ukrainian pages, and whether their edits tended to be pro-Russian or pro-Ukrainian.

findings serve as a first indicator that articles about disputed Ukrainian regions experienced more dispute than those about undisputed regions. In a following section we quantify this observation.

Generalizable Metrics for Predicting Edit Wars

For our prediction model, which prioritizes generalizability over domain-specific insights, we select broadly applicable metrics that do not require domain knowledge. While plenty of conflict-detection metrics exist[46, 49–53], we focus on granular, time-series metrics suitable for weekly-level analysis.

The first metric we define is length change, calculated as the ratio of words added to words removed per revision. As discussed earlier[40], not all revisions contribute equality to an article's quality. New Wikipedia pages, like the Russian Invasion of Ukraine page, usually require extensive word additions, while established pages tend to need fewer words added per revision. Therefore, the length change metric assumes that an increase in word count indicates higher-quality contributions, whereas excessive word removal may signify vandalism[54].

Additionally, previous studies have identified controversy through editor behaviour patterns in networks. Editor networks are defined as a directed network where an edge going from node i to node j means that editor i edited the page after editor j. In a study by Asford[31], the triad 111D was found to appear more frequently in the network of a controversial Wikipedia page than in a random network. To adjust the metric to a weekly scale, we count the increase of 111D triads over the data collection period as an indicator of digital dispute. An example of an editor network from the Kherson Oblast Wikipedia page is in Figure 1**D**, with an example of oppising editing behaviour. Compared to a random network, the Kherson Oblast editor network has a significant number of 111D triads.

Quantifying Digital Reactions Towards the Invasion

To analyze the impact of territorial disputes in Ukrainian regions on digital disputes within their Wikipedia pages, we examine the change of three target metrics in response to the 2022 invasion. These target metrics are the number of revisions, reverts, and dispute edits. To quantify evidence of stronger attention and dispute on pages related to these regions, we construct a difference-in-difference (DiD) regression model. Mathematically, our DiD setup is expressed as:

$$E = \beta_0 + \beta_1 D + \beta_2 P + \beta_3 I + \beta_4 ID + \beta_5 IP + \varepsilon \tag{1}$$

where ε represents the normally distributed residual and E is the dependent outcome variable.

To increase robustness[33], we include two specifications for each of the three metrics. First, we calculate the seven-day rolling average of aggregated daily revisions and reverts. Secondly, we use the log value of the daily sum of revisions and reverts to account for the different scales of the treatment and comparison groups. The analysis distinguishes between disputed Ukrainian oblasts (D), Polish voivodeships (P), and uses undisputed oblasts as the baseline. The binary variable I captures the pre- and post-invasion periods, with interaction terms measuring the invasion's impact on the disputed (ID) and Polish (IP) regions. We expect a significant positive effect on articles about disputed Ukrainian regions, while Polish articles should show a neutral and insignificant response. We validated the parallel trends assumption by analysing pre-invasion data for the three groups, ensuring a parallel trend before the invasion. A visualisation of the trends can be found in supplementary information. Other important assumptions, such as having no dropouts and maintaining consistent measurements, held throughout the study. Several other limitations are discussed in a further section.

The table in Figure 1**D** presents the coefficients and significance levels for the DiD regression using 7-day averages of the target variables. Results for the log of daily sums for each target metric is presented in the supplementary information. The interaction term *Invasion* * *Disputed Region* is positive and highly significant for all three target variables, confirming a strong reaction on articles about disputed Ukrainian regions compared to those about undisputed regions. In contrast, the *Invasion* * *Poland* term is consistently insignificant, indicating that the Invasion effect is unique to the treatment group. The invasion coefficient for articles on undisputed regions is positive, but not highly significant across all target variables.

Predicting Digital Disputes using Page Locks

To predict when digital disputes occur, we train a random forest classification model using internal Wikipedia data. Tree-based models are highly flexible due to their non-linear structure and have demonstrated success in predicting Wikipedia edits[55]. The model outputs one of two classes: either the page was locked on a given day or it was not. To assess the impact of external metrics on prediction outcomes, we train two versions of the model. The Internal model includes weekly revisions, reverts, the average length change of revisions, and changes in 111D triads within the Wikipedia network. The News model retains all these variables, as well as the sum of related BBC articles published that week. As a baseline model, we develop a Null model that evaluates whether any reverts occurred during a given week, and randomly selects one of those weeks for a page to be locked.

Due to the infrequency of page locks (occurring only 63 times across 38 pages and four years), the data is highly imbalanced, with page locks comprising the minority of observations. As such, we took two measures to address the data imbalance. Unlike regression models, random forest models have hyperparameters, which are parameters that need to be defined by the user. These hyperparameters can influence model's performance on various metrics. We selected hyperparameters that optimized recall. Recall is a metric that prioritizes identifying true page locks, even if results in more false positives. It is preferred when the cost of missing true positives exceeds the cost of identifying false positives[56], as is the case with page locks. In addition to prioritizing recall, we also adjusted the model's decision threshold. Random forest models assign a probability score to each observation, which determines its class. For imbalanced datasets, this threshold can be adjusted to favour the class with the minority of observations[57]. Using the ROC curve as guide, we set the threshold to 0.4 instead of the default 0.5, classifying any score above this as a page lock.

To evaluate the performance of both models against the Null model, we conduct a Leave-One-Out Cross-Validation. In this approach, the model is tested on each page after being trained on all other pages. To see a detailed breakdown of each model's performance, we examine their confusion matrices, shown in Figure 2**A**. Compared to the baseline model, which correctly identifies 11% of the 63 true page locks, both models perform well, with the News model identifying 1 more page lock than the Internal model. As expected, these results come at the cost of many false positives, which are discussed further below.

To evaluate the models' performances across both classes, we use machine learning metrics recall, balanced accuracy, Area Under the Curve (AUC), and Matthews correlation[56]. As discussed previously, recall focuses solely on the page-lock class, while balanced accuracy and AUC are better suited to evaluate model performance with imbalanced data. Matthews correlation is also used for imbalanced data, but penalises false positives more than balanced accuracy and AUC. The results of this analysis can be seen in Figure 2**B**. On all the machine learning metrics, the Internal and News models outperform the null model, albeit it to different extents. As can also be seen in the confusion matrices, the recall metric illustrates that incorporating BBC news articles provides a slight improvement in the recall metric compared to using only internal Wikipedia metrics. This discussed further in the Methods section.

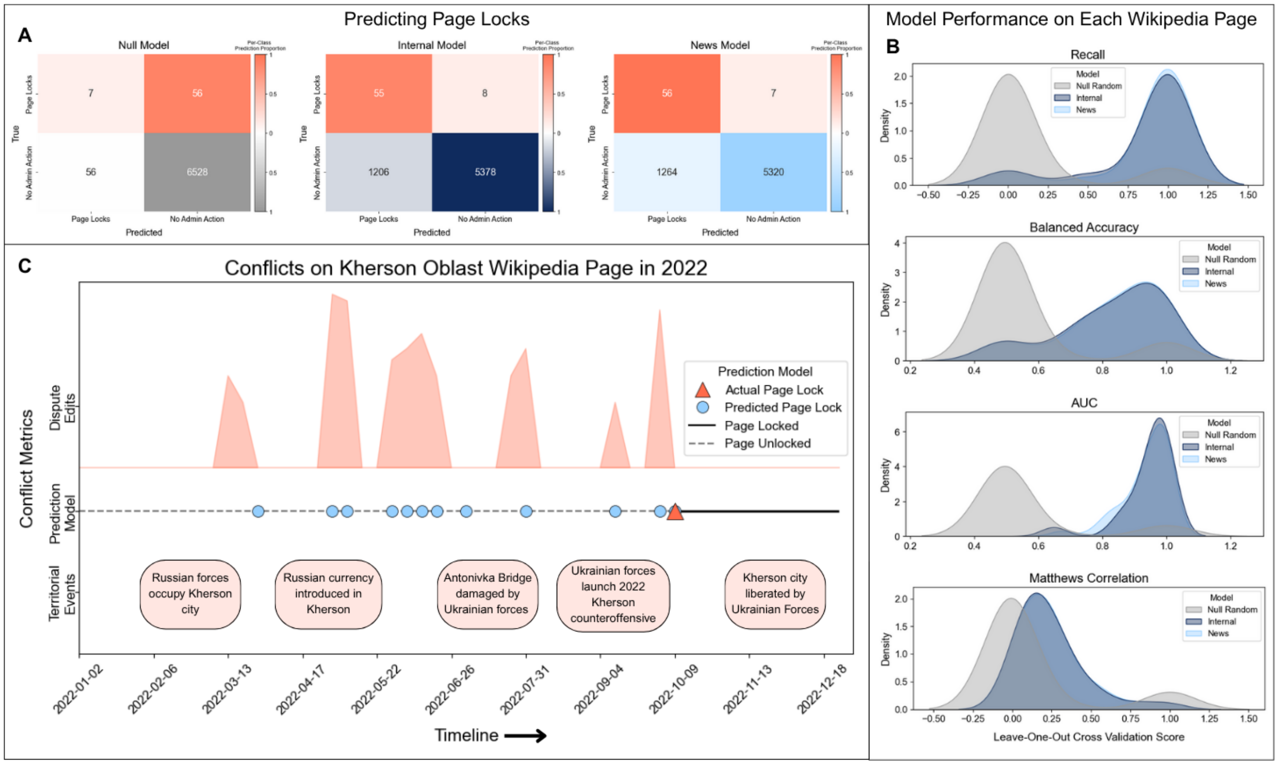


Figure 2. (**A**) Model performance in predicting page locks using full leave-one-out cross-validation. The News Model, which incorporates BBC articles, identified one additional page lock compared to the Internal Model, which relies solely on internal Wikipedia metrics. (**B**) Comparison across evaluation metrics shows that the Internal and News Models perform similarly, with both consistently outperforming the Null Model, which randomly predicts page locks. (**C**) Timeline comparing the two methodologies in detecting conflicts on the Kherson Oblast Wikipedia page. Digital disputes closely align with territorial events, and the prediction model identifies page locks during periods of increased dispute edits.

Figure 2**C** illustrates an example of how the prediction model aligns with the domain-specific dispute edits. The example focuses on the Kherson Oblast Wikipedia page, which received significant attention and extensive editing during 2022 after the invasion. On this page, false positives generated by the News model are represented as circles in the plot. Additional analysis in the Methods section reveals that 59% of these false positives occur while the page is not locked. In the context of predicting page locks, such false positives may act as early indicators of emerging digital disputes that could eventually result in a lock. On the other hand, 41% of page locks occur while the page is locked, possibly due to disagreements among senior editors or a high number of revisions as they work to fix mistakes. Thus, while a page lock typically marks the height of an edit war, false positives occurring prior to the lock may signal its onset and aftermath.

Moreover, although the analysis of domain-specific dispute edits and the prediction model were conducted independently, their temporal alignment suggests that both are detecting the same patterns in editing behavior. While this represents only a single example, the concurrence between these approaches demonstrates how both a data-driven method and a domain-specific analysis can capture similar patterns of dispute activity.

The Methods section provides additional details on the random forest setup, hyperparameter training, and variable transformations. It also includes further discussion on alternative approaches to handling the unbalanced data, as well as an analysis of other machine learning metrics.

Discussion

Despite the increasing threat of manipulation on pages covering influential topics such as armed conflict and political events[10,11], Wikipedia remains a key source of "consensus truth" in Western societies[5]. Recent developments, such as the ongoing peace negotiations between Russia and Ukraine, highlight the multifaceted and volatile nature of territorial conflicts. This reinforces the importance of understanding digital battlefields where information in contested and shaped.

Principle Findings

In the context of the Russo-Ukrainian conflict, we find evidence that Wikipedia is a digital battleground where Ukrainian and Russian-aligned editors strive in lay claims to contested territories. The difference-in-difference regression confirms our initial assumption that territorial conflicts are extending into the digital realm. While prior research focuses on social media[17], newspapers, and mass media[18], our findings shed light on previously overlooked attempts to change the identity of disputed regions on Wikipedia.

To predict digital disputes, we develop a random classification model trained on internal Wikipedia metrics, finding that it accurately identifies 86% of disputes. Incorporating BBC news articles as additional predictors increases the accurate predictions to 89%. Building on previous research which predicts Wikipedia user edits[55,58], and examines user dynamics on controversial articles[29–31], we validate the concept of forecasting digital disputes on Wikipedia.

Implications

This research shows the heightened interest among armed conflict actors in shaping the digital representation of Ukrainian regions, specifically regarding territorial claims. There are two motivations to these efforts. Firstly, the use of engaging narratives can mobilize forces and increase civilian participation in a war[59]. Secondly, narrative framing can influence the support of international allies[15]. Our findings also underscore the difficulty of identifying nuanced conflict dynamics on Wikipedia. While there are efforts[60] and proposals[46] to monitor content change on Wikipedia, conflict-related narratives often require domain expertise to identify dispute signals.

The reliance on human judgement to identify dispute signals highlights Wikipedia's structural advantage over LLMs in terms of transparency and editorial verifiability. Despite being trained largely on Wikipedia data, LLMs lack the transparency and editorial traceability of the original source[61]. Wikipedia's guidelines require humans to consolidate sources, provide readers with previous versions of the article, and show which sources were cited. The page-locking system exemplifies this further, entrusting experienced editors to manage complex digital disputes which automated systems cannot fully interpret. While the prediction model presented in this paper can assist in detecting dispute patterns, final decisions require human discernment. This human element may partly explain the increase in Wikipedia views since the emergence of ChatGPT[62] and the continued engagement of its contributor base[63]. Wikipedia remains a trusted source in the age of LLMs, with the results of this research demonstrating the ongoing need for human oversight.

Since Wikipedia content relies on external, reliable sources, the role of different sources in narrative architecture must also be understood. Locked Wikipedia pages are disproportionately viewed[64], showing a unique interest in them. The added predictive power of news sources points to a sensational dimension of digital disputes, attracting attention from both state-affiliated actors and members of the international community. This increased attention does not only affect Wikipedia, but also other platforms and large language models, as sensational topics are often subjects of active research. Analyzing the flow of information across platforms reveals weak points in the verification process and offers insight on preserving objective knowledge.

Another implication of this research reinforces the recognition that Wikipedia itself is a valuable data source for predicting real-life events. Previous studies have used Wikipedia to predict stock-market prices[65] and conflict related deaths[66]. Adding domain knowledge, like identifying partisan-driven disputes, enhances Wikipedia predictive potential. This could be particularly valuable for organisations with assets or personnel in conflict-prone regions. However, Wikipedia operates within a broader complex system with various influences. Effective crises prediction would have to control for causal direction and meaningfully combine Wikipedia data with other sources for an intentional outcome.

Limitations

There are two limitations to our study in terms regarding the use of the DiD method. First, the Polish data saw a declining trend prior to the Russian invasion, slightly violating the parallel trend assumptions since the Ukrainian data showed constant trends before the invasion. This discrepancy may be partly explained by data sparsity. Since Polish articles received relatively little attention before the invasion, they were more susceptible to sharp fluctuation. Second, some non-disputed Ukrainian regions were still subject to Russian aggression, such as missile attacks. This potentially violates the Stable Unit Treatment Value Assumption.

Another shortcoming of our study concerns deleted news articles. Specifically, if any edits on samples Wikipedia pages cited BBC articles that were later removed from the BBC website, they were not collected as predictor variables. This is a limitation of News API, as it does not provide deleted articles. However, the four-year timeframe for news data collection helped mitigate the risk of potential missing any relevant sources. Furthermore, the news data collection excluded BBC articles published in languages other than English. Manual inspection of several Wikipedia sample pages shows the use of non-English sources referenced.

A final limitation of this study is that the method used to identify disputes cannot easily be applied to other targets of contested information within the Ukraine conflict, nor to other conflicts. The Israel-Hamas conflict, for example, likely involves different targets of attention and distinct dispute signals within the edits.

Future Research

As the war in Ukraine is ongoing, this study could be replicated to see how key events, such as the 2024 United States presidential election, influence Wikipedia pages. Expanding this into a broader study which includes political or historical pages could further inform policymakers and the Wikimedia community. Election-related pages would be especially relevant, given Wikipedia's role in political discourse[67, 68] and their susceptibility to content alterations. Furthermore, political actors, especially during nationwide elections, are key participants in armed conflicts.

Additionally, examining Wikipedia in other languages would enhance the analysis. Examining the Ukrainian and Russian Wikipedias would be particularly insightful given the ongoing armed conflict and the arrests related to edits on the Russian Wikipedia[23, 24]. Building on the idea of election-related analysis, 2024 would be an ideal year to explore how national elections influence discourse on relevant Wikipedia pages given the high number which occurred. Replicating the methods used in this study could help generalize these results beyond armed conflict, or clarify the scope of their limitations.

Conclusion

This paper examined how territorial disputes extend into the digital realm. In the case of the Russo-Ukrainian war, territorial claims are mirrored online through changes in place names and page content, with these efforts intensifying during periods of physical conflict. Furthermore, although Wikipedia policy requires that edits cite external sources, incorporating such sources into a prediction model offered little improvement in forecasting page locks. Internal Wikipedia metrics alone were sufficient to predict edit wars.

References

1. Similarweb. Top websites ranking. similarweb (2024). Https://www.similarweb.com/top-websites/.

2. Giles, J. Special report internet encyclopaedias go head to head. *nature* **438**, 900–901 (2005).

3. Wikipedia. Wikipedia:size of wikipedia. Wikipedia website (2024). Https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

4. Kinsella, B. *Voice Assistants Alexa, Bixby, Google Assistant and Siri Rely on Wikipedia and Yelp to Answer Many Common Questions about Brands*. (2019 , July 11). Voicebot.ai. https://voicebot.ai/2019/07/11/voice-assistants-alexa-bixby-google-assistant-and-siri-rely-on-wikipedia-and-yelp-to-answer-many-common-questions-about-brands/.

5. Sant, T. How can wikipedia save us all?: Assuming good faith from all points of view in the age of fake news and post-truth. In *Media, Technology and Education in a Post-Truth Society*, 133–143 (Emerald Publishing Limited, 2021).

6. Dondio, P. Computational trust in web content quality: a comparative evalutation on the wikipedia project. (2007).

7. Arazy, O., Nov, O., Patterson, R. & Yeo, L. Information quality in wikipedia: The effects of group composition and task conflict. *J. management information systems* **27**, 71–98 (2011).

8. Kevin Schaul, N. T., Szu Yu Chen. Inside the secret list of websites that make ai like chatgpt sound smart. The Washington Post (2024). Https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

9. Gertner, J. Wikipedia's moment of truth. New York Times (2023). Https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html.

10. Marwick, A. E. & Lewis, R. Media manipulation and disinformation online. (2017).

11. Góngora-Goloubintseff, J. G. The falklands/malvinas war taken to the wikipedia realm: a multimodal discourse analysis of cross-lingual violations of the neutral point of view. *Palgrave Commun.* **6**, 1–9 (2020).

12. Whyte, C., Thrall, A. T. & Mazanec, B. M. *Information warfare in the age of cyber conflict* (Routledge London & New York, 2021).

13. Khaldarova, I. & Pantti, M. Fake news: The narrative battle over the ukrainian conflict. In *The Future of Journalism: Risks, Threats and Opportunities*, 228–238 (Routledge, 2020).

14. Doroshenko, L. & Lukito, J. Trollfare: Russia's disinformation campaign during military conflict in ukraine. *Int. J. Commun.* **15**, 28 (2021).

15. Treyger, E., Cheravitch, J. & Cohen, R. S. *Russian disinformation efforts on social media* (Rand Corporation Santa Monica, 2022).

16. Mejias, U. A. & Vokuev, N. E. Disinformation and the media: the case of russia and ukraine. *Media, culture & society* **39**, 1027–1042 (2017).

17. Babacan, K. & Tam, M. S. The information warfare role of social media: Fake news in the russia-ukraine war. *Erciyes ˙Dergisi* 75–92 (2022).

18. Ntanos, I. & Panagiotou, N. To evaluate and understand hybrid war and media – the media misinformation in the russian-georgian war of 2008 concerning the case of south ossetia. (2018).

19. Dammak, Z. & Lemmerich, F. Effects of the russo-ukrainian war on the editor activity of the ukrainian, russian, and english wikipedias. (2023).

20. Miller, C. *et al.* Information warfare and wikipedia. *Inst. for Strateg. Stud.* 1–25 (2022).

21. Guy Faulconbridge, P. F. Wikipedia fights russian order to remove ukraine war information. Reuters (2022). Https://www.reuters.com/world/europe/wikipedia-fights-russian-order-remove-ukraine-war-information-2022-06-13/.

22. Thornhill, J. The truth about war is messy — just read wikipedia. Financial Times (2022). Https://www.ft.com/content/b2a9c0a0-1f6b-4a25-9cd2-42f17b2f5a51.

23. Borak, M. Doxxed, threatened, and arrested: Russia's war on wikipedia editors. Nieman Lab (2022). Https://www.niemanlab.org/2022/09/doxxed-threatened-and-arrested-russias-war-on-wikipedia-editors/.

24. Pavel pernikau. Viasna Human Rights Centre (n.d.). Https://prisoners.spring96.org/en/person/pavel-pjarnikau.

25. Wikipedia. Wikipedia:reliable sources. Wikipedia website (2024). Https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources.

26. Ford, H., Sen, S., Musicant, D. R. & Miller, N. Getting to the source: where does wikipedia get its information from? In *Proceedings of the 9th international symposium on open collaboration*, 1–10 (2013).

27. Wikipedia. Wikipedia:bots/requests for approval/cluebot ng. Wikipedia website (2024). Https://en.wikipedia.org/wiki/Wikipedia:Bots/Requests_for_approval/ClueBot_NG.

28. Singer, P. *et al.* Why we read wikipedia. In *Proceedings of the 26th international conference on world wide web*, 1591–1600 (2017).

29. Yasseri, T., Sumi, R., Rung, A., Kornai, A. & Kertész, J. Dynamics of conflicts in wikipedia. *PloS one* **7**, e38869 (2012).

30. Gandica, Y., Dos Aidos, F. S. & Carvalho, J. The dynamic nature of conflict in wikipedia. *Europhys. Lett.* **108**, 18003 (2014).

31. Ashford, J. *et al.* Understanding the signature of controversial wikipedia articles through motifs in editor revision networks. In *Companion Proceedings of the 2019 World Wide Web Conference*, 1180–1187 (2019).

32. Madeline Fitzgerald, E. D. J. Russia invades ukraine: A timeline of the crisis. U.S. News (2022). Https://www.usnews.com/news/best-countries/slideshows/a-timeline-of-the-russia-ukraine-conflict.

33. Ruprechter, T. *et al.* Volunteer contributions to wikipedia increased during covid-19 mobility restrictions. *Sci. reports* **11**, 21505 (2021).

34. Raleigh, C., Linke, R., Hegre, H. & Karlsen, J. Introducing acled: An armed conflict location and event dataset. *J. peace research* **47**, 651–660 (2010).

35. Fetahu, B., Anand, A. & Anand, A. How much is wikipedia lagging behind news? In *Proceedings of the ACM Web Science Conference*, 1–9 (2015).

36. Li, K., DeCost, B., Choudhary, K., Greenwood, M. & Hattrick-Simpers, J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Comput. Mater.* **9**, DOI: 10.1038/s41524-023-01012-9 (2023).

37. Wikimedia. Wikipedia api. https://www.mediawiki.org/wiki/API:Main_page (2023). Accessed: 2024-06-30.

38. News API. Documentation. https://newsapi.org/docs (2025). Accessed: 2025-04-21.

39. Sepehri-Rad, H. & Barbosa, D. Identifying controversial wikipedia articles using editor collaboration networks. *ACM Transactions on Intell. Syst. Technol. (TIST)* **6**, 1–24 (2015).

40. Bykau, S., Korn, F., Srivastava, D. & Velegrakis, Y. Fine-grained controversy detection in wikipedia. In *2015 IEEE 31st International Conference on Data Engineering*, 1573–1584 (IEEE, 2015).

41. Sumi, R., Yasseri, T. *et al.* Edit wars in wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 724–727 (IEEE, 2011).

42. Suh, B., Chi, E. H., Pendleton, B. A. & Kittur, A. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *2007 IEEE symposium on visual analytics science and technology*, 163–170 (IEEE, 2007).

43. Sepehri Rad, H. Analyzing controversy in wikipedia. (2016).

44. d'Sa, A. G., Illina, I. & Fohr, D. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, 1–5 (IEEE, 2020).

45. Khasanah, I. N. Sentiment classification using fasttext embedding and deep learning model. *Procedia Comput. Sci.* **189**, 343–350 (2021).

46. Jhandir, M. Z., Tenvir, A., On, B.-W., Lee, I. & Choi, G. S. Controversy detection in wikipedia using semantic dissimilarity. *Inf. Sci.* **418**, 581–600 (2017).

47. Shulzhenko, D. How russia has attempted to erase ukrainian language, culture throughout centuries. Ministry of Foreign Affairs website (2023). Https://kyivindependent.com/how-russia-has-attempted-to-erase-ukrainian-language-culture-throughout-centuries/.

48. UkrainianGovernment. correctua. Ministry of Foreign Affairs website (2019). Https://mfa.gov.ua/en/correctua.

49. Zielinski, K., Nielek, R., Wierzbicki, A. & Jatowt, A. Computing controversy: Formal model and algorithms for detecting controversy on wikipedia and in search queries. *Inf. Process. & Manag.* **54**, 14–36 (2018).

50. Schneider, J., Passant, A. & Breslin, J. G. A content analysis: How wikipedia talk pages are used. In *Proceedings of the 2nd International Conference of Web Science*, 1–7 (2010).

51. Ghosh, A. Using n-grams to identify edit wars on wikipedia. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 398–403 (IEEE, 2019).

52. Chhabra, A., Kaur, R. & Iyengar, S. Dynamics of edit war sequences in wikipedia. In *Proceedings of the 16th International symposium on open collaboration*, 1–10 (2020).

53. Brandes, U., Kenis, P., Lerner, J. & Van Raaij, D. Network analysis of collaboration structure in wikipedia. In *Proceedings of the 18th international conference on World wide web*, 731–740 (2009).

54. Mola-Velasco, S. M. Wikipedia vandalism detection through machine learning: Feature review and new proposals: Lab report for pan at clef 2010 (2012). 1210.5560.

55. Zhang, D. Wikipedia edit number prediction based on temporal dynamics only. *arXiv preprint arXiv:1110.5051* (2011).

56. Wang, L., Han, M., Li, X., Zhang, N. & Cheng, H. Review of classification methods on unbalanced data sets. *Ieee Access* **9**, 64606–64628 (2021).

57. Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N. & Riniker, S. Ghost: adjusting the decision threshold to handle imbalanced data in machine learning. *J. Chem. Inf. Model.* **61**, 2623–2640 (2021).

58. Yoshida, Y. & Ohwada, H. Wikipedia edit number prediction from the past edit record based on auto-supervised learning. In *2012 International Conference on Systems and Informatics (ICSAI2012)*, 2415–2419 (IEEE, 2012).

59. Asmolov, G. The transformation of participatory warfare: The role of narratives in connective mobilization in the russia–ukraine war. *Digit. War* **3**, 25–37 (2022).

60. Wikipedia. Wikipedia:protection policy. Wikipedia website (2024). Https://en.wikipedia.org/wiki/Wikipedia:Protection_policy.

61. Vetter, M. A., Jiang, J. & McDowell, Z. J. An endangered species: how llms threaten wikipedia's sustainability. *AI & SOCIETY* 1–14 (2025).

62. Reeves, N., Yin, W. & Simperl, E. Exploring the impact of chatgpt on wikipedia engagement. arxiv (2024).

63. Lyu, L. *et al.* Wikipedia contributions in the wake of chatgpt. *arXiv preprint arXiv:2503.00757* (2025).

64. Hill, B. M. & Shaw, A. Page protection: another missing dimension of wikipedia research. In *Proceedings of the 11th International Symposium on Open Collaboration*, 1–4 (2015).

65. Moat, H. S. *et al.* Quantifying wikipedia usage patterns before stock market moves. *Sci. reports* **3**, 1801 (2013).

66. Oswald, C. & Ohrenhofer, D. Click, click boom: Using wikipedia data to predict changes in battle-related deaths. *Int. Interactions* **48**, 678–696 (2022).

67. Fahim, M. A., Gallagher, S., McCain, M. & Rubin, N. Inauthentic editing: Changing wikipedia to win elections and influence people. Stanford Cyber Policy Center Freeman Spogly Institute (2021). Https://fsi.stanford.edu/news/wikipedia-part-one.

68. Salem, H. & Stephany, F. Wikipedia: a challenger's best friend? utilizing information-seeking behaviour patterns to predict us congressional elections. *Information, Commun. & Soc.* **26**, 174–200 (2023).

69. Ashenfelter, O. C. & Card, D. Using the longitudinal structure of earnings to estimate the effect of training programs (1984).

70. Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion* (Princeton university press, 2009).

71. Min, Y. & Agresti, A. Random effect models for repeated measures of zero-inflated count data. *Stat. modelling* **5**, 1–19 (2005).

72. Tawiah, K., Iddrisu, W. A. & Asampana Asosega, K. Zero-inflated time series modelling of covid-19 deaths in ghana. *J. Environ. Public health* **2021**, 5543977 (2021).

73. Python Software Foundation. *difflib: Helpers for computing deltas* (2024). Accessed: 2024-06-06.

74. Khan, M. T. *et al.* An empirical analysis of the commercial vpn ecosystem. In *Proceedings of the Internet Measurement Conference 2018*, 443–456 (2018).

75. Wikipedia blocked in china in all languages. BBC News (2019). Https://www.bbc.com/news/technology-48269608.

76. Carvalho, M., Pinho, A. J. & Brás, S. Resampling approaches to handle class imbalance: a review from a data perspective. *J. Big Data* **12**, 71 (2025).

77. Richey, M. Contemporary russian revisionism: understanding the kremlin's hybrid warfare and the strategic and tactical deployment of disinformation. *Asia Eur. J.* **16**, 101–113 (2018).

78. O'Loughlin, J., Toal, G. & Kolosov, V. The rise and fall of "novorossiya": examining support for a separatist geopolitical imaginary in southeast ukraine. *Post-Soviet Aff.* **33**, 124–144 (2017).

79. Florian, A. How russia's invasion of ukraine turned words into weapons. Morning Brew Magazine (2022). Https://www.morningbrew.com/daily/stories/2022/07/22/russia-invasion-ukraine-language.

80. Kamusella, T. D. Russian: A monocentric or pluricentric language? In *Colloquia Humanistica*, 7, 153–196 (Instytut Slawistyki Polskiej Akademii Nauk, 2018).

81. NataliefromUkraine. odesa not odessa. Blog publishing website: buy me a coffee (2022). Https://buymeacoffee.com/natahisfors/odesa-odessa.

Methods

Complimentary analyses supporting the Difference-in-Difference Model

To quantify heterogeneous changes in Wikipedia editing and dispute activity following the 2022 invasion of Ukraine, we employ a difference-in-differences (DiD) regression framework[33]. The model compares developments in three target metrics—revisions, reverts, and dispute edits—between articles associated with disputed Ukrainian regions (treatment group) and two comparison corpora (undisputed Ukrainian regions and Polish regions), before and after the invasion.

The DiD framework is applied analogously to all three outcome variables.

The mathematical setup

For completeness, we restate the DiD specification introduced in the main text and provide a detailed interpretation of all coefficients:

$$E = \beta_0 + \beta_1 D + \beta_2 P + \beta_3 I + \beta_4 ID + \beta_5 IP + \varepsilon \tag{2}$$

where E denotes the outcome variable of interest and ε is the regression error term.

Continuous dependent metric E

E is the dependent outcome variable. We included two specifications for each of the three target metrics to increase robustness[33]. First, we included a left-aligned 7-day rolling average of aggregated daily revisions or reverts per bucket. Secondly, we used the log value of the daily sum of revisions and reverts to account for the different scales of the treatment and comparison groups.

Disputed and Poland indicator variables D and P

We included two categorical variables to differentiate the three Wikipedia corpora. In line with standard practice, the regression has a reference corpus as a baseline[33]. In our context, the undisputed corpus is chosen as the baseline. This allowed us to compare the reactions of the target variables to the invasion of disputed-region and Poland-region articles against undisputed-region articles. D is an indicator for the disputed page bucket, and P is an indicator for the Poland regions bucket. We note that β_0 could be interpreted as the intercept for the undisputed-region articles, the sum of β_0 plus β_1 could be interpreted as the intercept for disputed-region articles and β_0 plus β_2 as the intercept for Poland-region pages.

Binary invasion indicator variable I

The binary year variable I specifies whether a data point is from a timepoint that is equal to or later than the 24.02.2022 (= 1), the post-invasion period, or from the previous period (= 0), the pre-invasion period. The respective coefficient β_3 indicates the effect of being in the post-invasion period ($I = 1$) for the baseline (undisputed) corpus, which could be interpreted as the reaction of the target metrics for undisputed-region pages to the invasion.

Interaction terms ID and IP

Finally, we constructed interaction terms to measure the effect of the invasion on the corpora containing articles about disputed Ukrainian and Polish regions. We defined the interaction between the disputed bucket and post-invasion as ID and the interaction between the Poland bucket and post-invasion as IP. β_4 is the main effect of interest since it gives information on whether the start of the invasion had a significantly stronger positive effect on the revisions and reverts of Wikipedia articles about disputed Ukrainian regions compared to the baseline of undisputed regions. The sum of β_3 and β_4 indicates the overall effect of the invasion on Wikipedia pages about the disputed regions. β_5 adds robustness to the analysis by adding insights into whether such an effect can also be measured for Poland regions or whether it is unique for the treatment group (disputed-region articles). We expect β_4 to be significant and positive, while we expect β_5 to be neutral or even negative. Such a finding would indicate that the Wikipedia pages about disputed regions faced a significantly stronger increase in attention and dispute than both comparison corpora, the undisputed-region pages, and the Poland-region pages. This would strengthen the visual observation of a stronger effect of the invasion on digital attention and dispute for disputed Ukrainian regions compared to the two comparison corpora.

Parallel trends assumption

The most important assumption for the difference in difference setup to be valid is the parallel trends assumption[69]. This means that if the treatment (the invasion) had not occurred, the gap of the target metric between the treatment group (disputed region articles) and control groups (non-disputed and Poland region articles) would remain constant over time. We tested for the parallel trends assumption visually by looking at the quarterly sum of the two target metrics for the three subgroups in all quarters before the first invasion quarter, which commenced on 01.01.2022. If target variables develop with a similar trend for the three-time series, that would strengthen the assumption of parallel trends. This assumption is most important for the disputed region and undisputed region articles.

We validated the parallel trends assumption by visually inspecting pre-invasion data for the three groups, ensuring a consistent trend before the invasion. The following figure 3 shows a different magnitude but parallel development of the three corpora across all three traget metrics before start of the invasion.

Based on the assumptions of parallel trends, we did not include characteristics of the Wikipedia pages as control variables since most of them are likely affected by the treatment itself[70]. For example, the pageviews on disputed pages also increase after the invasion, very likely because of the invasion. Including them as a control variable would account for parts of the invasion's effect on the views variable and could lead to an underestimation of the actual effect of the invasion.

Most other assumptions hold quite naturally, like no different drop-out levels (all pages existed throughout the whole data collection period), consistent measurement, or no anticipation effect. In the limitations section, we highlight a few assumptions that might not perfectly hold.

Alternative specification

We included a second model specification, as a robustness check, that used individual page data. In the main model, we aggregated the edits and reverts for all pages per corpus (disputed, undisputed, and Poland) and used the three aggregated time series in the regression analysis. There are arguments in favor and against treating each page as a separate time series and rerunning the regression with 14, 14, and 16-time series, respectively, each with a dummy variable representing its corresponding corpus.

In favor of such an alternative model specification, individual-level data can yield more precise and accurate estimates than aggregated data[70]. However, one reason for initially relying on aggregated time series data is that many Wikipedia pages (e.g., articles about lesser-known Polish regions) exhibit a large number of days with zero edits and zero reverts. This zero inflation can complicate the analysis[71,72]. Developing a dedicated zero-inflated model is beyond the scope of this study, particularly given that both the aggregated and individual-level specifications yield consistent results for the coefficients of interest.

Alternative regression results

Table 1 reports regression results using log-transformed, aggregated outcome variables. Across all specifications, the interaction term capturing the invasion effect in disputed regions remains positive and statistically significant, indicating a disproportionate increase in editing activity relative to undisputed regions following the invasion.

Table 1. Alternative specification using log-transformed aggregated outcomes

Model:	(1)	(2)	(3)
Dependent variable:	Log(Revisions)	Log(Reverts)	Log(Dispute-edits)
Disputed Region	0.36***	0.12***	0.15***
	(0.03)	(0.02)	(0.02)
Polish Region	-0.01	0.01	-0.01
	(0.03)	(0.02)	(0.02)
Invasion Effect (Undisputed)	0.15***	0.04**	0.01
	(0.04)	(0.02)	(0.02)
Invasion × Disputed Region	0.57***	0.24***	0.18***
	(0.05)	(0.02)	(0.03)
Invasion × Polish Region	-0.16**	-0.06**	-0.01
	(0.05)	(0.02)	(0.03)
Intercept	0.16***	0.02	0.01
	(0.02)	(0.01)	(0.01)
Observations	4,383	4,383	4,383
R^2	0.24	0.17	0.11
Adjusted R^2	0.24	0.17	0.11

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

The subsequent analysis uses individual page-level data rather than aggregated bucket-level observations. Table 2 Panel A reports estimates based on 7-day rolling averages of the outcome variables to smooth short-term volatility, while Panel B applies a log transformation to daily counts. Across both specifications, the coefficient on the key interaction term (Invasion × Disputed Region) remains positive and statistically significant. This consistency across smoothing procedures and functional-form transformations indicates that the main findings are robust and not driven by short-term fluctuations or specific modeling assumptions.

Regression data

Table 3 lists the Wikipedia page titles that define each corpus. According to the ACLED dataset, the Oblasts and autonomous regions classified as *Disputed* have experienced Russian territorial gains at least once. We additionally include the alternative page titles "Luhansk People's Republic", "Donetsk People's Republic", and "Republic of Crimea (Russia)", which correspond to names used for parts of these regions following the 2014 annexation and the 2022 declarations of independence. We assume that discussions about nationhood are likely to occur on both the canonical region pages and these alternative pages. Although Crimea is treated as an autonomous republic and was not an official Oblast prior to the 2014 annexation by Russia, we include it as a region of Ukraine. We do not consider Kyiv city separately and focus only on the page *Kyiv Oblast*.

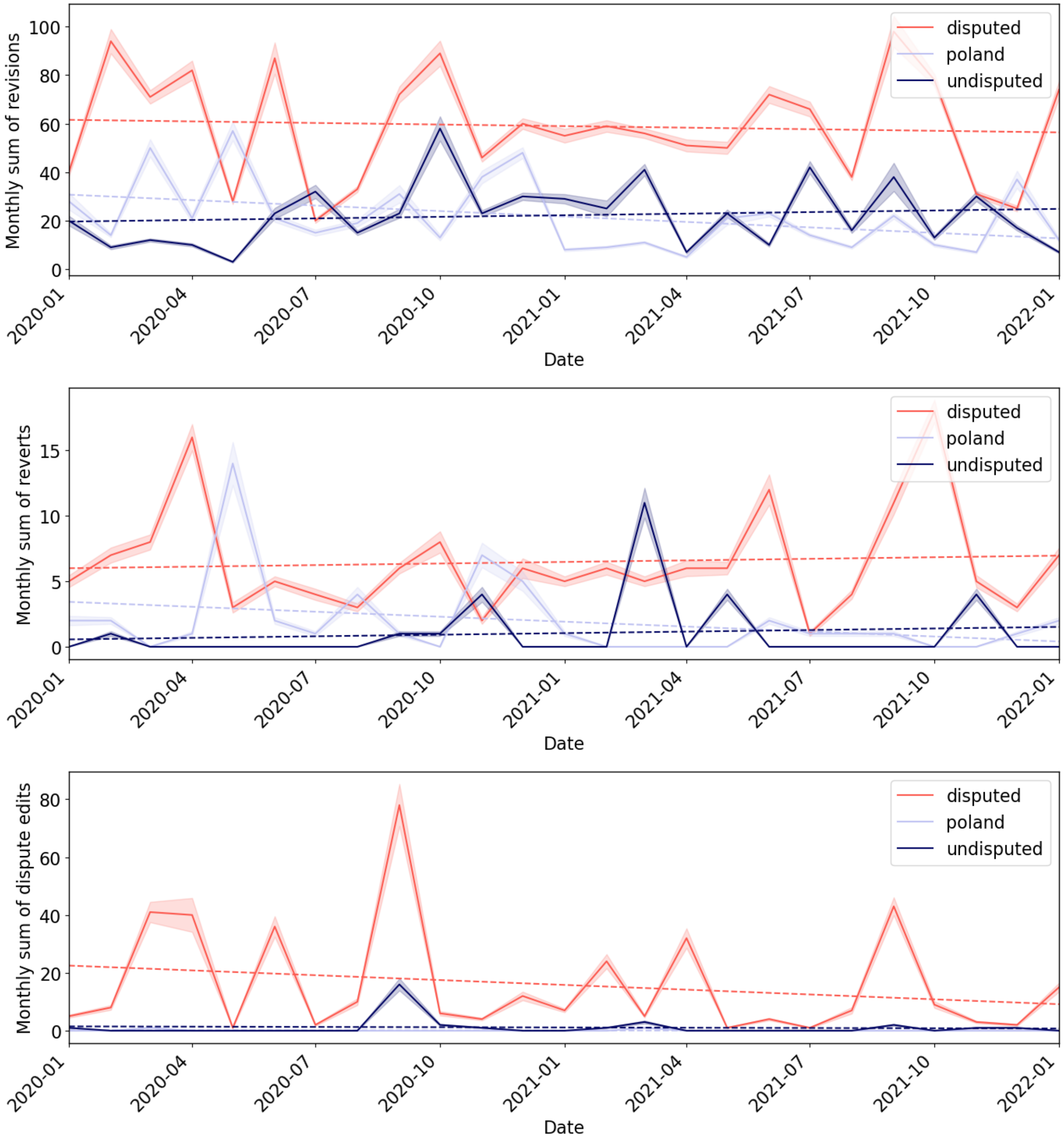


Figure 3. Development of revisions, reverts and dispute edits before the 2022 invasion to validate parallel trends assumption.

Table 2. **Individual page-level regressions under alternative outcome treatments**

	Dependent variable		
	(1) Revisions	(2) Reverts	(3) Dispute-edits
Panel A: 7-day rolling averages			
Disputed Region	0.11***	0.01***	0.03***
	(0.01)	(0.00)	(0.00)
Polish Region	-0.01	0.00	-0.00
	(0.01)	(0.00)	(0.00)
Invasion Effect (Undisputed)	0.04***	0.01***	0.00
	(0.01)	(0.00)	(0.00)
Invasion × Disputed Region	0.21***	0.04***	0.04***
	(0.01)	(0.00)	(0.00)
Invasion × Polish Region	-0.04**	-0.01***	-0.00
	(0.01)	(0.00)	(0.00)
Intercept	0.06***	0.00**	0.00
	(0.01)	(0.00)	(0.00)
Observations	64,284	64,284	64,284
R^2	0.05	0.05	0.02
Adjusted R^2	0.05	0.05	0.02
Panel B: Log-transformed daily counts			
	Log(Revisions)	Log(Reverts)	Log(Dispute-edits)
Disputed Region	0.02***	0.00**	0.01***
	(0.00)	(0.00)	(0.00)
Polish Region	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
Invasion Effect (Undisputed)	0.01**	-0.00	0.00
	(0.00)	(0.00)	(0.00)
Invasion × Disputed Region	0.06***	0.01***	0.01***
	(0.00)	(0.00)	(0.00)
Invasion × Polish Region	-0.01	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
Intercept	0.01***	0.00	0.00
	(0.00)	(0.00)	(0.00)
Observations	64,284	64,284	64,284
R^2	0.03	0.01	0.01
Adjusted R^2	0.03	0.01	0.01

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 3. **Classification of Oblasts and Voivodeships used to define the corpora**

Disputed	Undisputed	Polish Voivodeships
Chernihiv Oblast	Cherkasy Oblast	Lower Silesian Voivodeship
Autonomous Republic of Crimea	Chernivtsi Oblast	Kuyavian–Pomeranian Voivodeship
Donetsk Oblast	Dnipropetrovsk Oblast	Lublin Voivodeship
Kharkiv Oblast	Ivano–Frankivsk Oblast	Lubusz Voivodeship
Kherson Oblast	Khmelnytskyi Oblast	Łódź Voivodeship
Kyiv Oblast	Kirovohrad Oblast	Lesser Poland Voivodeship
Luhansk Oblast	Lviv Oblast	Masovian Voivodeship
Mykolaiv Oblast	Rivne Oblast	Opole Voivodeship
Odesa Oblast	Ternopil Oblast	Subcarpathian Voivodeship
Sumy Oblast	Vinnytsia Oblast	Podlaskie Voivodeship
Zaporizhzhia Oblast	Volyn Oblast	Pomeranian Voivodeship
Luhansk People's Republic	Zakarpattia Oblast	Silesian Voivodeship
Donetsk People's Republic	Poltava Oblast	Świętokrzyskie Voivodeship
Republic of Crimea (Russia)	Zhytomyr Oblast	Warmian–Masurian Voivodeship
		Greater Poland Voivodeship
		West Pomeranian Voivodeship

Extracting relevant edits

We use the `ndiff` library[73] to identify differences between revision versions, focusing on words that are added or replaced at the same location in the text. For each detected change, we extract the changed words and the ten context words before and after. This approach allows a single revision to contribute multiple edits when changes occur in different sentences or parts of the page. Figure 4 visualizes the extraction process for three revisions of an exemplary Wikipedia page.

We combine removed and added words that share identical context into a substitution pair. Substitution pairs therefore capture each unique removal–addition combination at a specific location within a given revision and page. We remove empty substitution pairs "[] – []", which typically reflect changes in infobox-style statistics (e.g., population values) or media elements such as pictures.

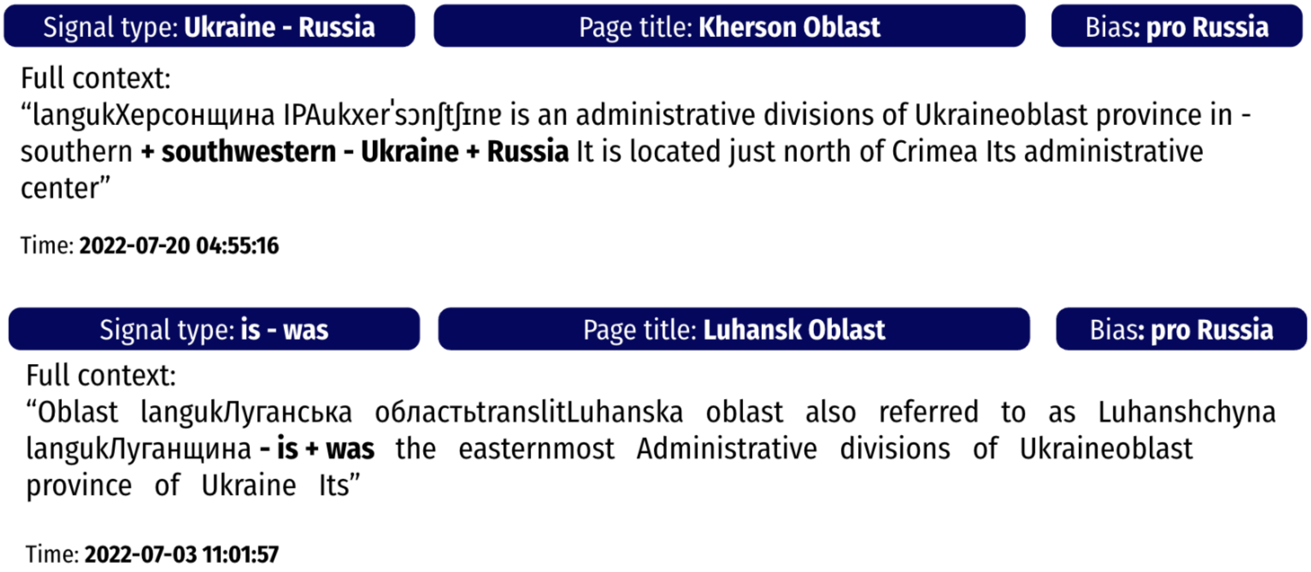
After extraction and cleaning, we classify edits into *neutral edits*, which primarily add information, and *dispute edits*, which involve narrative changes favoring Russia or Ukraine.

Because such changes are domain-specific, it is difficult to rely on pre-trained NLP classifiers trained on general, domain-independent corpora. Such classifiers are often used for hate-speech detection[44] or sentiment classification[45]. One related approach proposes detecting controversy in Wikipedia edits when the cosine similarity between embeddings of replaced words is very low[46]. While such a method might flag replacements such as *USA* by *Russia*, terms such as *Ukraine*, *Russia*, *Kiev*, and *Kyiv* are likely to remain close in embedding space because they describe geographically and semantically related entities. As a result, a domain-independent classifier may ignore replacements such as *Kyiv* by *Kiev* or *Ukraine* by *Russia*. We therefore develop a domain-specific process to classify dispute edits linked to the Russo–Ukraine war.

First, we visualize the most common substitution pairs across the two Ukraine-specific data corpora. We combine these patterns with domain knowledge to define *dispute identifiers* that indicate cultural or political misalignment. Dispute identifiers should reflect either the Russian goal to delegitimize Ukrainian independence or the Ukrainian goal to refute Russia's annexation, as illustrated by the examples in Figure 4.

We distinguish between two categories of dispute identifiers. First, we include attempts to change stated nationhood from Russia to Ukraine or vice versa, for example by exchanging "Ukrainian" with "Russian" or "is" with "was". Second, we classify edits that alter the spelling of important cities between Russian and Ukrainian transliterations as dispute identifiers. The Ukrainian spelling of regions and cities is a salient topic in Ukraine because it reinforces Ukrainian identity and counters Russian attempts to undermine Ukrainian language and identity[47]. There have been multiple efforts by Ukraine to raise international awareness of correct spellings for Ukrainian places; a prominent campaign is "KyivnotKiev", initiated by the Ukrainian government[48].

Dispute identifiers should reflect either the Russian goal to delegitimize Ukrainian independence or the Ukrainian goal to refute Russia's annexation, such as the edits in Figure 4.



Signal type: **Ukraine - Russia** Page title: **Kherson Oblast** Bias: **pro Russia**

Full context:
“langukХерсонщина IPAukxerˈsɔnʃtʃɪnɐ is an administrative divisions of Ukraineoblast province in - southern **+ southwestern - Ukraine + Russia** It is located just north of Crimea Its administrative center”

Time: **2022-07-20 04:55:16**

Signal type: **is - was** Page title: **Luhansk Oblast** Bias: **pro Russia**

Full context:
“Oblast langukЛуганська областьtranslitLuhanska oblast also referred to as Luhanshchyna langukЛуганщина **- is + was** the easternmost Administrative divisions of Ukraineoblast province of Ukraine Its”

Time: **2022-07-03 11:01:57**

Figure 4. Examples of two dispute edits from the actual data corpus that discuss the nationhood of Kherson and Luhansk.

Analyses for visualisations in Figure 1

Editor Classification

To classify users based on the overall orientation of their contributions, we aggregated their individual edits across the dataset. The analysis presented in the Domain-Specific Metrics for Ukrainian Regions section was conducted at the level of individual edits, without accounting for users’ prior editing histories. As previously described, each edit was categorized as either neutral or a dispute edit, with dispute edits further classified as favoring either the Russian or Ukrainian perspective.

User-level classifications were then derived by comparing the total number of pro-Ukraine and pro-Russia edits attributed to each individual. Users with more pro-Ukraine than pro-Russia edits were classified as Pro-Ukraine, while those with more pro-Russia than pro-Ukraine edits were classified as Pro-Russia. If the number of pro-Ukraine and pro-Russia edits were equal, the user was classified as Neutral. A visualisation of this is shown in Table 3. This approach enabled us to analyze patterns of editorial alignment at the user level.

Table 4. Method for Classifying Editors

User	Pro-Ukraine edits	Pro-Russia edits	Edit Difference	Classification
A	10	1	9	Pro-Ukraine Edits
B	0	5	-5	Pro-Russia Edits
C	1	1	0	Neutral Edits

Editor Network

The editor network for the Kherson Oblast page includes all users who made contributions to the page during the data collection period. Directed edges in the network represent the temporal sequence of edits: an arrow from editor A to editor B indicates that editor A made an edit after editor B. Each editor was classified as pro-Ukraine, pro-Russia, or Neutral based on the user-level classification method outlined in the previous section, allowing us to visualize patterns interaction within the editing community.

The Kherson Oblast editor network contained 80 111D triads. To assess the statistical significance of this count, we generated 1,000 random networks, each preserving the same number of nodes and edges as the original Kherson Oblast network. For each of these randomized networks, we calculated the number of 111D triads. The resulting distribution is presented in Figure 5, allowing for a comparison between the observed triad count and what would be expected by chance.

Out of the 1,000 randomized networks, none exhibited a number of 111D triads equal to or greater than the 80 observed in the Kherson Oblast editor network. This yields an empirical p-value of < 0.0, indicating that the observed count of 111D triads is highly unlikely to have occurred by chance and is therefore statistically significant.

IP Addresses

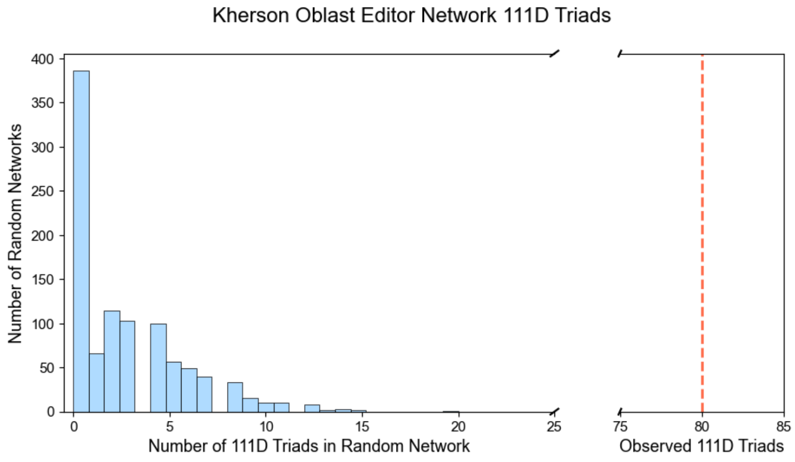


Figure 5

Analyzing editors from the 44 regional Wikipedia pages during the data collection period, we identified approximately 2,000 contributors, of whom 28% used IP addresses as usernames. These users were classified as Neutral, pro-Ukraine, or pro-Russia based on the classification method described above.

To examine the geographic distribution of partisan edits, we focused on users identified by IP addresses. This allowed us to associate edits with specific countries. Comparing the distribution of pro-Russia and pro-Ukraine edits among this subset to the overall editor population, we found that IP-based users formed a representative sample of the broader user base, as illustrated in Figure 6.

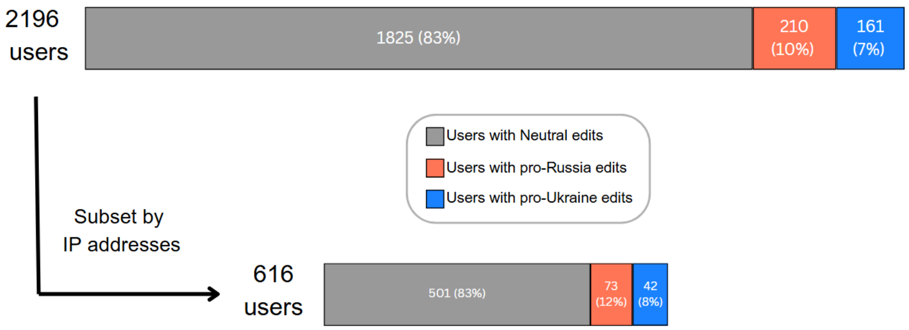


Figure 6

With the exception of the two countries involved in the conflict, we grouped users by continent to accommodate the distribution of the data. Including both Russia and Ukraine, IP addresses originated from 74 different countries. The United States accounted for the largest number of IP addresses (110), followed by Ukraine (64), Canada (47), Great Britain (43), Poland (42), and Russia (36). The remaining 68 countries each contributed fewer than 36 IP addresses. Due to the use of VPNs by Wikipedia editors and the blocking of Wikipedia in certain countries[74,75], we intentionally avoid making claims that editors reside in the countries where their IP address is registered.

Complimentary analyses supporting the Random Forest Model

Data Collection

As discussed above, our data collection period is from January 1st, 2020, to December 31st, 2023. In order to generalise the prediction model beyond the Russo-Ukraine war, we examine pages from other conflicts that were ongoing during the data collection period, started after Wikipedia was created, and were classified as a major conflict by Wikipedia. These were the Mali wars and the Sudanese civil war.

We then go to the main pages of these conflicts, which give an overview of key events, places, and people. From the introduction and infobox on the main pages, we identified linked pages that made up our sample size. These were 122 pages in total, 58 from the African wars and 64 from the Russo-Ukrainian wars. We sample the 122 pages further by identifying pages that were locked during the data collection period. Most pages had never been locked while a few had been locked indefinitely during the data collection. The remaining 38 pages, seen in Table 4 below, were unlocked at the beginning of the data collection period, and were locked at least once during the four years.

Table 5. Locked Wikipedia Pages related to Armed Conflict

Events	Places	People
Russo-Ukrainian War	Russia	Putin
Russian invasion of Ukraine	Ukraine	Volodymyr Zelenskyy
War in Donbas	Donbas	Oleksandr Syrskyi
Southern Ukraine campaign	Kharkiv	Valerii Zaluzhnyi
Eastern Ukraine campaign	Kherson	Sergei Shoigu
Siege of Mariupol	Mariupol	Valery Gerasimov
Battle of Donbas (2022)	Kherson Oblast	Aleksandr Dvornikov
Liberation of Kherson	Luhansk People's Republic	Sergey Surovikin
Ukraine–NATO relations	Donetsk People's Republic	Hemedti
Enlargement of NATO	Algeria	Wagner Group
Russia-Ukraine relations	Libya	
Casualties of the Russo-Ukrainian War	Sudan	
Russian annexation of Donetsk, Kherson, Luhansk and Zaporizhzhia oblasts	Russian-occupied territories of Ukraine	
United Nations General Assembly Resolution ES-11/1		
International sanctions during the Russian invasion of Ukraine		

The fact that out of 122 pages, only 38 had been locked corresponds with previous work which found that only 0.36% pages in the English Wikipedia had been subject to page locks[64]. However, locked pages tend to be disproportionate clicked on by Wikipedia viewers, which aligns with the interest that a high-stakes topic such as armed conflict will generate.

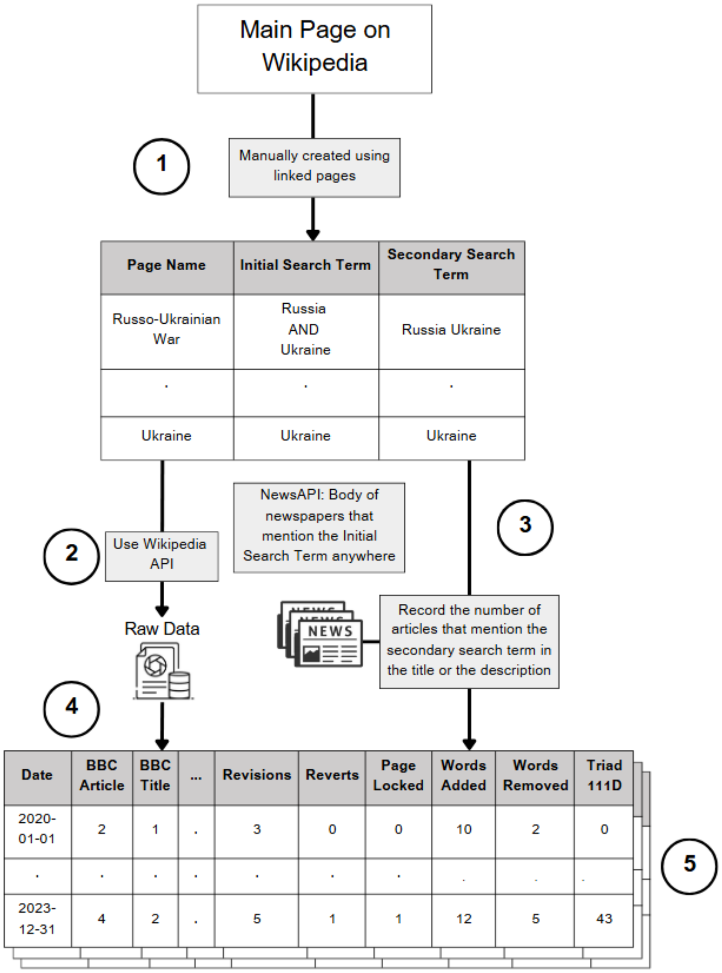
Despite most pages in the final sample coming from the Russo-Ukrainian war, there are still a few pages related to the African conflicts. A previous study[36] found that even a small amount of new data can improve the accuracy of machine learning models, thus satisfying the data requirements to make this model generalisable outside of the Russo-Ukrainian war. Additionally, collecting the pages of these conflicts during the same time as the Russo-Ukrainian war serves to balance the dataset between heavily covered conflicts and less-well known conflicts, helping the model learn how to predict cases with varying amounts of coverage.

Once we have the sample pages, we assigned each page an initial and a secondary search term based on the Wikipedia page titles. The idea behind the initial search term was to find news articles that might be referenced in the Wikipedia pages. For example, any news article that contains the phrases "Ukraine" and "NATO" might have information relevant to the "Ukraine-Nato relations" Wikipedia page. Hence, the initial search terms were used to find relevant articles via NewsAPI.

In addition to the initial search terms, we also created a list of secondary search terms. These are less stringent variations of the first search terms, the idea being to see if any of them appeared in the title or description of the article. This narrows the search further, creating predictor variables that capture relevant headlines instead of articles that may mention the conflict in passing. For the selection of news outlets, we refer to those found in a previous study[26] that we could access using the NewsAPI. These were the BBC, CNN, and the Washington Post.

Once we had the raw data for Wikipedia and the newspapers, we calculated the needed summaries for day. These were the count of revisions, reverts, and news articles. We also traced the number of words added and removed on Wikipedia each day, as well as the increase of triads in the editor network. A flowchart showing the process by which the data was collected is shown in Figure 7.

The result of this process was 122 data frames, each corresponding to a page related to armed conflict on Wikipedia. We only apply the variable transformation, on the 38 pages in Table 5. However, the data for the rest of the pages are available to access for other projects.


Main Page on Wikipedia
1
Manually created using linked pages
2
Use Wikipedia API
Raw Data
NewsAPI: Body of newspapers that mention the Initial Search Term anywhere
3
NEWS
Record the number of articles that mention the secondary search term in the title or the description
4
5

Page Name	Initial Search Term	Secondary Search Term
Russo-Ukrainian War	Russia AND Ukraine	Russia Ukraine
.	.	.
Ukraine	Ukraine	Ukraine

Date	BBC Article	BBC Title	...	Revisions	Reverts	Page Locked	Words Added	Words Removed	Triad 111D
2020-01-01	2	1	.	3	0	0	10	2	0
.	.	.	.	.	.	.	.	.	.
2023-12-31	4	2	.	5	1	1	12	5	43

Figure 7. A flowchart showing the process by which the data was collected. Step 1 is manually deciding on the sample pages and determining the search terms. Step 2 collects Wikipedia data using its API and step 3 gathers the news article data using NewsAPI. In step 4, we calculate the summaries from the raw data like the number of revisions and total articles. In step 5, these variables are combined into a dataframe for every Wikipedia page.

Variable Addition

To prepare the variables for the random forest model, we added three new variables to the dataframes.

First, we introduced a page lock target variable. This binary variable indicates whether a page was locked on a given day, addressing a limitation of the original Page Locked variable, which only reflects the current lock status rather than the timing of lock events.

The first was the page lock target. This was a binary variable which tracked whether or not a page was locked on a particular day since the original Page Locked variable only monitors whether or not a page is currently locked, not the day it was locked.

Second, we computed the difference in triad 111D counts. While the original variable measured the total number of 111D triads present in the network, it did not capture how many were newly formed within a given week. Our adjusted variable captures this temporal dynamic.

Thirdly, we created a words ratio variable, defined as,

$$\text{words ratio} = \frac{\text{words added} + 1}{\text{words removed} + 1}.$$

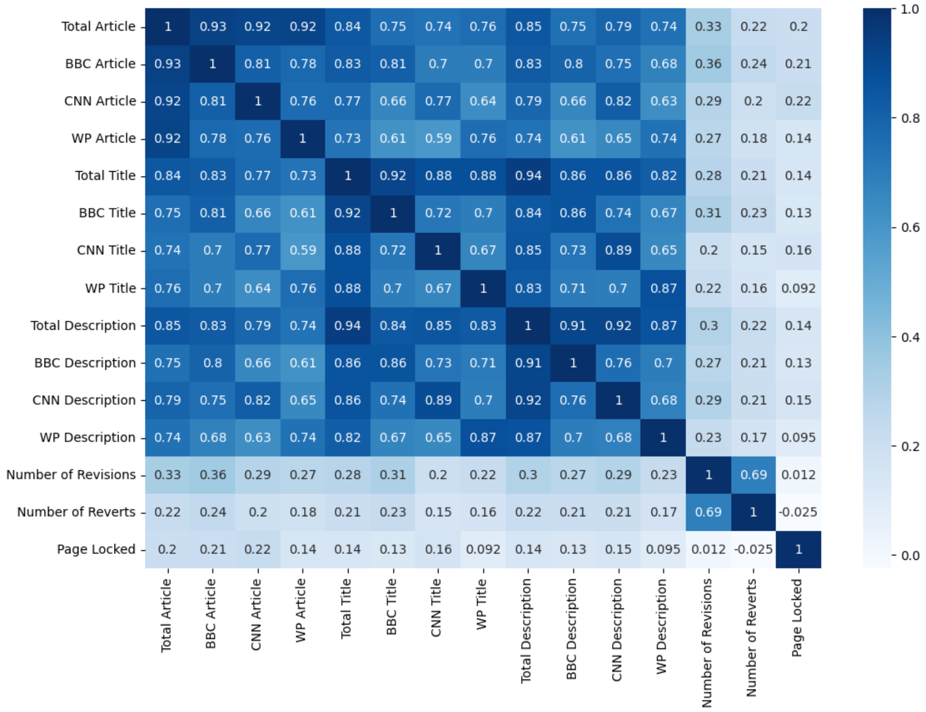
This metric is intended to reflect the productivity of page revisions, based on the assumption that a higher proportion of added words indicates more constructive editing activity. A words ratio greater than 1 indicates that more content was added than removed, suggesting constructive editing. Ratios between 0 and 1 reflect revisions where more content was removed than added. A ratio equal to 1 may result from either a lack of revisions on a given day or revisions in which the net change in article length was zero.

After we had these variables, we aggregated the dataset from a daily to a weekly timeframe by summing all relevant variables within each week. This transformation helped smooth sharp fluctuations and align the temporal resolution with our modeling objectives.

To account for skewed distributions and reduce the influence of extreme values, revisions, reverts, words ratio, and the count of news articles were log-transformed prior to modeling. The count of Triad 111Ds was also log-transformed. However, in cases where the number of triads decreased from one week to the next (typically due to triad closure) the transformation was adjusted to preserve the negative direction of change.

News Selection

To gain a preliminary understanding of which variables were most closely associated with page locks, we computed Pearson's correlation coefficients and visualized the results using a correlation matrix, shown below in Figure **??**. Initial inspection revealed that the news-related variables were highly intercorrelated. As such, we used previous research[26] to inform our decision to use BBC articles as the representative news variable to include in the random forest model.



	Total Article	BBC Article	CNN Article	WP Article	Total Title	BBC Title	CNN Title	WP Title	Total Description	BBC Description	CNN Description	WP Description	Number of Revisions	Number of Reverts	Page Locked
Total Article	1	0.93	0.92	0.92	0.84	0.75	0.74	0.76	0.85	0.75	0.79	0.74	0.33	0.22	0.2
BBC Article	0.93	1	0.81	0.78	0.83	0.81	0.7	0.7	0.83	0.8	0.75	0.68	0.36	0.24	0.21
CNN Article	0.92	0.81	1	0.76	0.77	0.66	0.77	0.64	0.79	0.66	0.82	0.63	0.29	0.2	0.22
WP Article	0.92	0.78	0.76	1	0.73	0.61	0.59	0.76	0.74	0.61	0.65	0.74	0.27	0.18	0.14
Total Title	0.84	0.83	0.77	0.73	1	0.92	0.88	0.88	0.94	0.86	0.86	0.82	0.28	0.21	0.14
BBC Title	0.75	0.81	0.66	0.61	0.92	1	0.72	0.7	0.84	0.86	0.74	0.67	0.31	0.23	0.13
CNN Title	0.74	0.7	0.77	0.59	0.88	0.72	1	0.67	0.85	0.73	0.89	0.65	0.2	0.15	0.16
WP Title	0.76	0.7	0.64	0.76	0.88	0.7	0.67	1	0.83	0.71	0.7	0.87	0.22	0.16	0.092
Total Description	0.85	0.83	0.79	0.74	0.94	0.84	0.85	0.83	1	0.91	0.92	0.87	0.3	0.22	0.14
BBC Description	0.75	0.8	0.66	0.61	0.86	0.86	0.73	0.71	0.91	1	0.76	0.7	0.27	0.21	0.13
CNN Description	0.79	0.75	0.82	0.65	0.86	0.74	0.89	0.7	0.92	0.76	1	0.68	0.29	0.21	0.15
WP Description	0.74	0.68	0.63	0.74	0.82	0.67	0.65	0.87	0.87	0.7	0.68	1	0.23	0.17	0.095
Number of Revisions	0.33	0.36	0.29	0.27	0.28	0.31	0.2	0.22	0.3	0.27	0.29	0.23	1	0.69	0.012
Number of Reverts	0.22	0.24	0.2	0.18	0.21	0.23	0.15	0.16	0.22	0.21	0.21	0.17	0.69	1	-0.025
Page Locked	0.2	0.21	0.22	0.14	0.14	0.13	0.16	0.092	0.14	0.13	0.15	0.095	0.012	-0.025	1

Random Forest Model

The model selected to predict page locks was the Random Forest, a tree-based ensemble method. Decision trees recursively partition the predictor space by identifying optimal splits that lead to more homogeneous groups in terms of the outcome variable. While decision trees are generally less interpretable than linear models such as Ordinary Least Squares, they provide the flexibility to model non-linear relationships and higher-order interactions without requiring explicit specification.

A common limitation of decision trees is their tendency to overfit the training data. That is, they may generate splits that reflect noise or outliers in the training set, reducing their ability to generalize to unseen data. To mitigate this, decision trees are often aggregated into an ensemble models which combines the predictions of multiple trees to improve robustness and generalization.

At the core of a decision tree is the mechanism for determining where to split the data. We used the default loss function provided by scikit-learn, known as Gini impurity, defined as,

$$G = 1 - \sum_{k=1}^{K} p_k^2. \tag{3}$$

Here, $K = 2$ represents whether a page is locked (class 1) or unlocked (class 0). The term p_k denotes the proportion of training samples within a node that belong to class k.

Since decision trees involve several hyperparameters that can affect model performance, we partitioned the dataset into a 60% training set and a 40% testing set, ensuring a balanced representation of pages related to both the Russo-Ukrainian conflict and conflicts in Africa.

For the Random Forest model, we tuned three hyperparameters: the number of trees in the ensemble (n_estimators), the minimum samples per leaf (min_samples_leaf), and the maximum depth of the decision tree (max_depth). The hyperperameter options and the optimal results can be seen in Table 6. All other hyperparameters were left at their default values as defined in scikit-learn's RandomForestClassifier.

Hyperparameters	**Options**	**Optimal for Internal Model**	**Optimal for News Model**
n_estimators	100, 500, 1000	100	100
min_samples_leaf	1, 3, 5	1	1
max_depth	3, 5, 7	3	3

Table 6. Hyperparameter Options and Optimal Values for Models

Unbalanced Data

A key challenge in our modeling was handling class imbalance, where the minority class (e.g., locked pages) is underrepresented relative to the majority class. A common strategy to address this is oversampling the minority class to balance the training data. However, oversampling can introduce biases by amplifying outlier samples[76]. Given the complexity of this issue, we opted to preserve the natural class distribution to reflect real-world conditions and focused instead on adjusting the model and evaluation process to account for imbalance and its effects on predictive bias.

To mitigate the impact of unbalanced classes, we implemented two key adjustments. First, we set the class_weight hyperparameter of the Random Forest classifier to "balanced". This automatically adjusts the weights inversely proportional to class frequencies in the input data, effectively penalizing misclassification of minority class samples more heavily and encouraging the model to better learn these cases.

Second, we optimized the classification threshold based on the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) analysis (see Figure 8). Instead of the default threshold of 0.5, we selected a conservative threshold of 0.4 for both the Internal and the news Model. This improved the balance between sensitivity and specificity for detecting page locks.

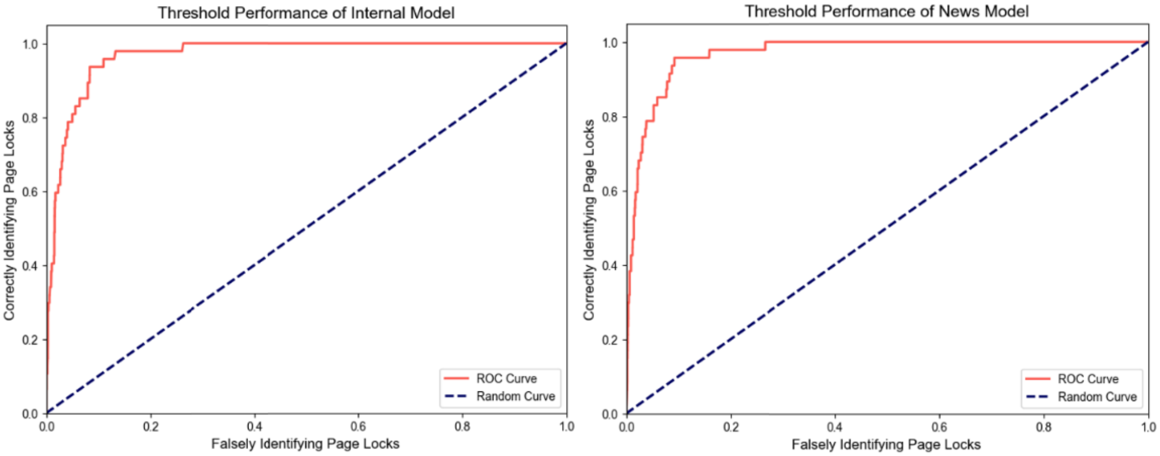


Figure 8

After selecting the optimal hyperparameters and adjusting the classification threshold using the training data, we evaluated model performance on the test set. The resulting confusion matrices are presented below in Figure 9, providing a detailed view of true and false positive and negative rates for each model.

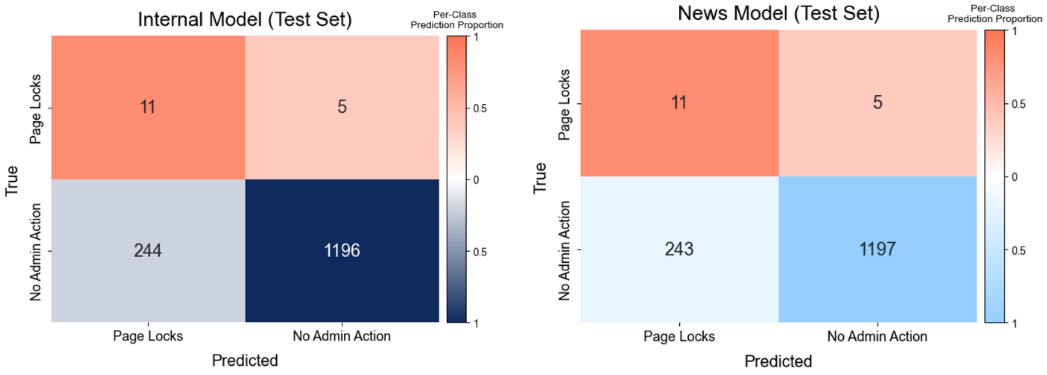


Figure 9

The News model successfully identifies 69% of true page locks, performing on par with or better than the Internal model in terms of true positive detection. To further assess model performance, we employed a Null model, leave-one-out cross-validation (LOOCV), and standard evaluation metrics, as illustrated in Figure 2. Additional robustness checks using alternative threshold settings are available in the Supplementary Information.

Null Model

In addition to evaluating model performance on training and testing data, we compared the Internal and News models against a baseline Null model, providing an additional layer of validation. Prior research has shown that reverts are strongly associated with conflict in editing behavior[41]. Building on this, we constructed the Null model by randomly assigning page locks to time points during which reverts occurred. Specifically, during the leave-one-out cross-validation (LOOCV) process, we recorded how often each page had been locked (usually once) and then randomly assigned the lock to a week where at least one revert occurred on that page. In essence, the Null model simulates a lock event during any revert-active week, without relying on additional predictive features.

Machine Learning Metrics

As discussed above, the random forest was optimised on recall, shown below as,

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}.$$

As a metric, recall optimises the return of true positives, no matter how the cost of false positives. To evaluate the model class of the the model we also used balanced accuracy, defined as,

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}.\right)$$

Balanced accuracy takes into account the proportion of correct classifications separately for each class. It other words, it averages the accuracy of each class. This gives equal weight to the performance on both the positive and negative classes.

Another metric which takes into account the performance of both classes is AUC, defined as,

$$\text{AUC} = \sum_{i=1}^{n-1} (\text{FPR}_{i+1} - \text{FPR}_i) \cdot \frac{\text{TPR}_{i+1} + \text{TPR}_i}{2}$$

where i is a threshold for classifying a prediction into the positive or negative class. AUC increases when a binary classification model assigns higher predicted probabilities to true positive instances than to false positive instances. To complement these two metrics, we also look at MCC, defined as,

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC takes into account all confusion matrix components, including false positives and false negatives, into a single correation. Thus, it punishes imbalances between false positives and false negatives than Balanced Accuracy and AUC, which focus on averaged rates or ranking performance, respectively.

For the results shown in Figure 2, recall, balanced accuracy, and AUC are high for the news and internal models given than lots of true positive are predicted for both classes. This is because, taking the News model for example, the per-class accuracy rate is 81% for the negative class and 89% for the positives class. MCC, however, more strongly penalises the misclassification for both classes, explaining why its distribution is much lower compared to those of the other metrics.

The results of MCC are verified by similar distributions shown by Cohen's Kappa, defined as,

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where:

$$p_o = \text{Observed agreement (accuracy), defined as } \frac{TP+TN}{TP+TN+FP+FN}$$

$$p_e = \text{Expected agreement by chance, defined as } \frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+TN+FP+FN)^2}$$

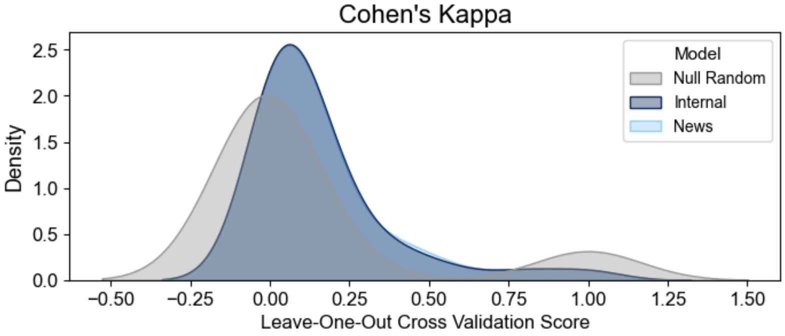


Figure 10

Whereas MCC evaluates the overall quality of predictions based on the confusion matrix, Cohen's Kappa accounts for the possibility that the agreement between predicted and true labels could have occurred by chance.

In short, the Internal and News models perform well on recall, balanced accuracy, and AUC, where the true positive predictions class are rewarded. They perform less well on MCC and Cohen's Kappa, which penalise the prediction of false positives. However, all the metrics show the overall trend that the Internal and News models outperform the Null Model. In some cases, the News model performs slightly better while other times it performs the same as the Internal model. This is discussed further in the next section.

Post-Model Analysis

To evaluate which features of the News model were most important, we used scikit-learn's standard feature importance extraction and permutation importance, using recall as the primary evaluation metric. The standard feature importance results are based on the mean decrease in impurity (Gini importance) and highlight features that contribute most to reducing uncertainty in the model's internal decision trees. In contrast, the permutation importance results are derived by randomly shuffling each feature and measuring the drop in recall. They capture the impact of each feature on the model's actual predictive performance. As illustrated in Figure 11, the number of reverts emerged as the most important feature across most models, followed by the number of revisions. The remaining variables contributed significantly less to model performance.

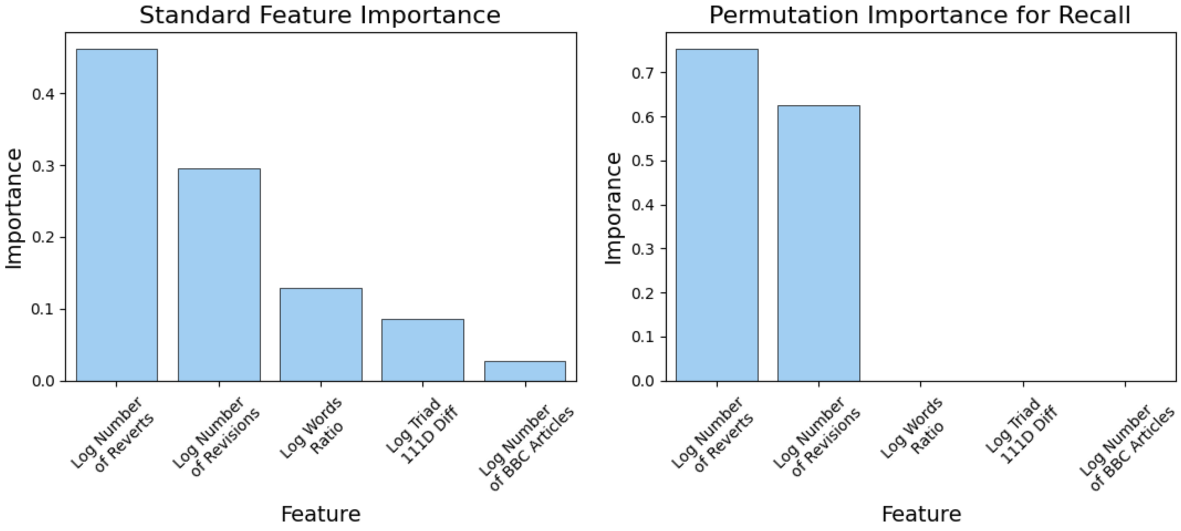


Figure 11

To investigate where most of the false positives from the News model occurred, we analyzed the predictions generated during the leave-one-out cross-validation (LOOCV) procedure. We categorized the false positives into three distinct groups based on the timing relative to page protection status. First, we identified false positives that occurred before the page was locked for the first time during our data collection period. Second, we counted those that occurred while the page was still locked. Finally, we recorded false positives that occurred after the page was unlocked. This classification allowed us to better understand the temporal context in which the model's errors were concentrated.

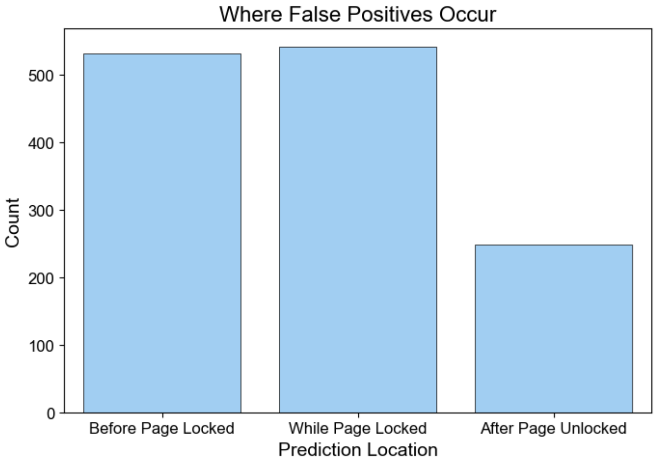


Figure 12

From the Figure 12, 40% of false positives occur before the page is locked for the first time and 41% of false positives occur while the page is locked. The last 19% of the false positives occur when a page is unlocked after its initial lock. A page is sometimes locked several times, but the "After Page Unlocked" classification above takes into account all false positives that occurred after the initial page lock, whether or not the page was locked afterwards or not.

Privacy and ethics

- Mention that this research does not include any personally-identifiable information
- Provide the CUREC codes here: OII_C1A_24_042.

Data and Code Availability

- GitHub repository link: [Will be published when available]

Acknowledgements

- Thank all people who gave feedback at the OII during events and supervision
- Thank the participants of conferences where we presented
- Thank the DSF for their generous support

Author contributions

- Use the CREDIT Taxonomy to update the author contributions accordingly

All authors designed research; All authors analysed data and undertook investigation; XYZ and XYZ led multi-factor analysis; XYZ, XYZ and XYZ led the founder/employee prediction; XYZ led personality insights; XYZ collected and tabulated the data; XYZ, XYZ, and XYZ created figures; XYZ created final art, and all authors wrote the paper.

Competing interests

The authors declare no competing interests.

Supplementary Information

Further discussion

The finding that the most prevalent word substitutions target the nationhood of Ukrainian regions and the spelling of heavily contested areas is consistent with expectations developed earlier in the analysis. Existing research identifies the legitimization of false narratives as a central pillar of Russian misinformation strategies[77]. While we cannot distinguish between actions led by government agencies and those initiated by civilians, the results clearly indicate an ongoing dispute over the official narrative surrounding events in Ukraine. Challenges to official nationhood can have severe implications for Ukraine's claim to independence. If the most widely used online encyclopedia—the Western "consensus truth"—were to describe Kherson as a province in southwestern Russia rather than southern Ukraine, this would pose a direct threat to the sovereignty and independence of the affected regions. Framing Kherson as a Russian province represents an attempt to rewrite history in a way that supports Russian narratives of liberation rather than annexation. Such narratives also imply a preemptive victory by suggesting that Russia has already gained full control over these regions and incorporated them as official provinces.

The regions most frequently targeted by dispute edits following the invasion, shown in the middle upper panel of figure Figure 1, largely overlap with areas subject to historical Russian claims commonly summarized under the concept of *Novorossiya*[78]. The Novorossiya narrative seeks to revive a historical region along the northern coast of the Black Sea that was once established by Russian colonizers and encompasses large parts of south-eastern Ukraine. Strengthening this narrative could be interpreted as undermining Ukrainian morale and reducing Western support for Ukraine[15] to negatively affecting Ukrainian mobilization and civilian participation in the conflict[59].

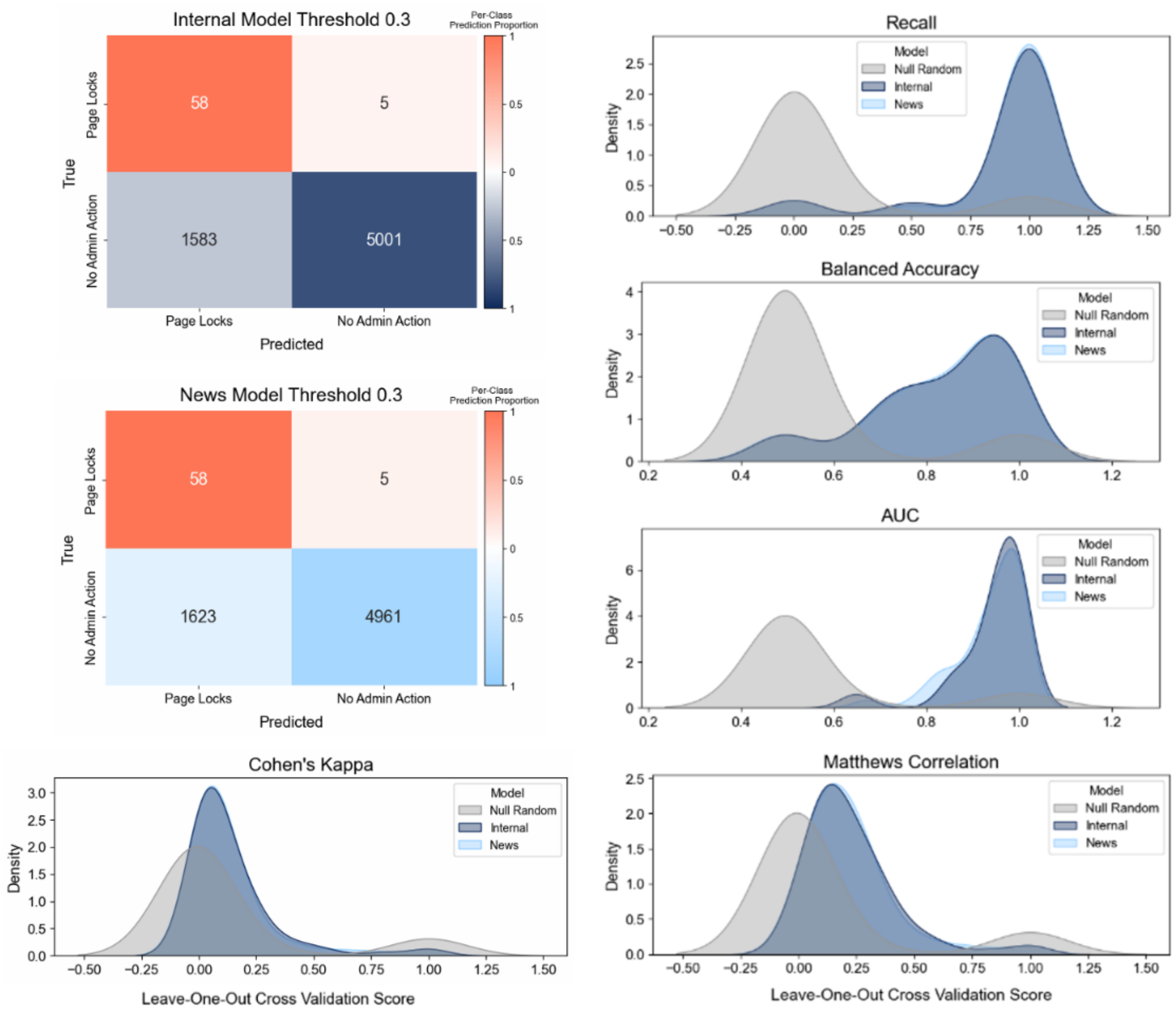
Beyond attempts to influence the nationhood of regions, efforts to alter the spelling of Ukrainian cities and regions may appear less visible to those without domain knowledge but carry substantial symbolic meaning for Ukrainians. Prior to the invasion, Russian was widely accepted in Ukraine and was often perceived as a prestigious or elite language. Since the invasion, however, it has increasingly been viewed as a tool of propaganda[79]. Russian is one of the few major languages that is largely construed as monocentrically controlled by a single state[80]. Using Ukrainian spellings for cities and regions therefore constitutes an act of liberation from Russian linguistic dominance. This perspective helps explain initiatives emphasizing the importance of correct Ukrainian spellings, such as the "KyivnotKiev" campaign[48]. The emotional salience of this issue is further reflected in public discourse, including blog posts addressing these spelling disputes[81]. Consequently, edits that replace *Donbass* (Russian spelling) with *Donbas* (Ukrainian spelling) signify far more than a technical disagreement over orthography. In the context of the Russo–Ukrainian war, such changes can be interpreted as acts of resistance against Russian control and as efforts to strengthen Ukrainian identity[47].

Recent work documents an initial decline in edit activity across Russian-, Ukrainian-, and English-language Wikipedia editions in the days following the invasion, alongside a largely stable short-term revert rate for English Wikipedia[19]. These patterns differ markedly from the behavior observed for disputed-region articles in this analysis, where edits and reverts increase sharply and immediately in the week of the invasion. Although activity subsequently returns to less extreme levels, the initial spike is substantially stronger than both the counterfactual regions examined here and the aggregate patterns documented in prior work[19]. This suggests that pages associated with disputed regions are more strongly affected by the invasion than the average Wikipedia page, and even more than pages broadly related to the conflict. The findings therefore support the notion that disputed regions move to the center of digital attention precisely as territorial control becomes contested.

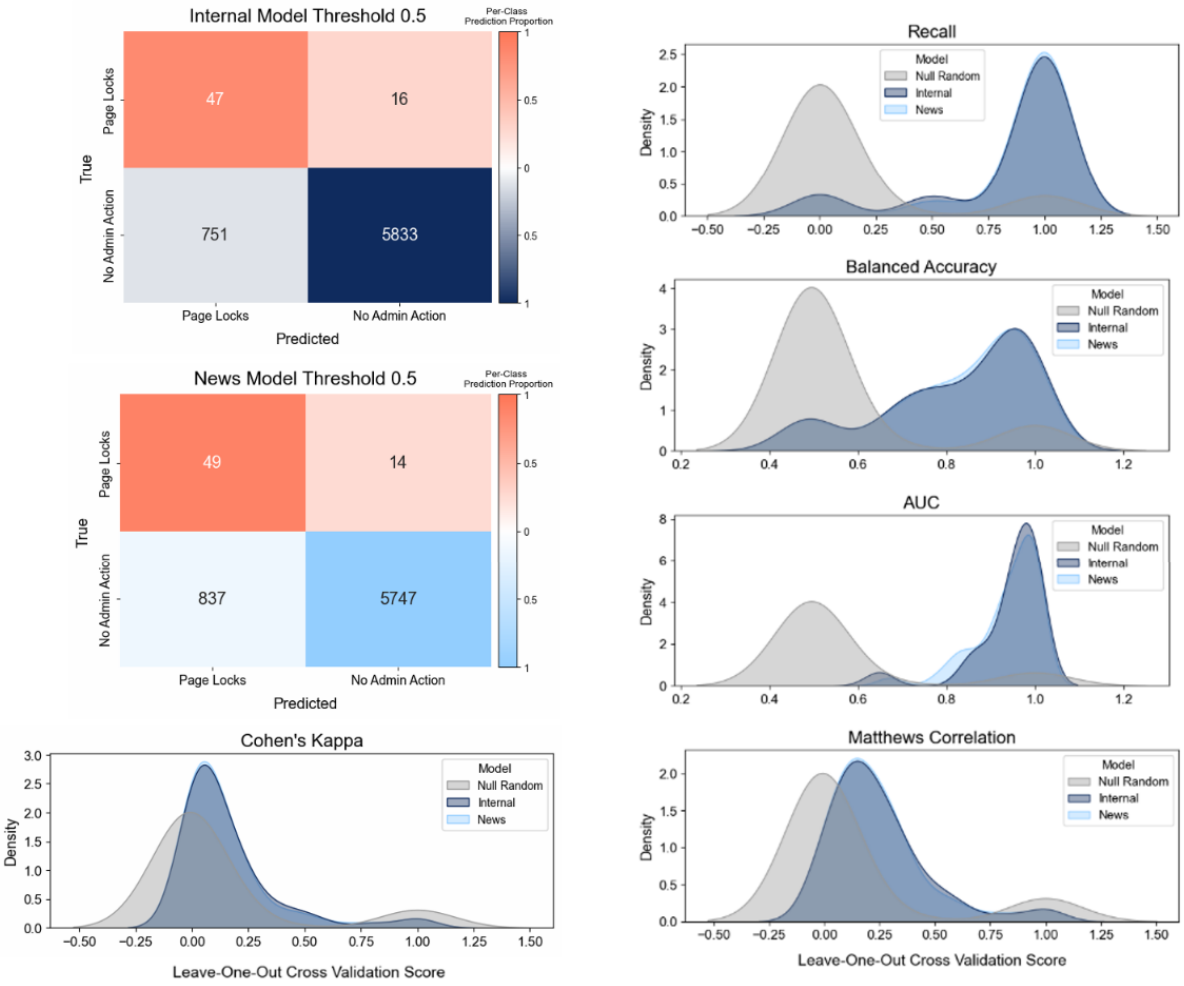
Evidence of attempts to alter political facts and introduce biased narratives on the main Russo–Ukrainian war page has been documented previously[20]. This study adds a complementary dimension by focusing on digital disputes with explicitly territorial rather than political targets. Issues such as the nationhood of invaded oblasts or the spelling of key geographic locations are no less consequential than debates over political leadership or military events. Whether a Wikipedia page describes Kherson as part of eastern Ukraine or southern Russia is of fundamental importance. By identifying systematic attempts to alter the perceived nationhood of disputed regions, this analysis highlights how digital representations can indirectly strengthen or weaken competing territorial claims. Both Russia and Ukraine ground their narratives in assertions of regional nationhood. The results demonstrate the central role of Wikipedia in shaping the digital representation of geographic regions during conflict and show that attention to these pages intensifies as territorial warfare escalates.

Finally, these findings contribute to the broader hypothesis that an information war is unfolding on Wikipedia as part of a wider landscape of cyber and information warfare. Prior research documents multiple dimensions of Russian digital messaging, including disinformation campaigns, coordinated trolling, and fake news dissemination[13, 14, 16]. The results presented here add evidence for an additional, territorially focused dimension of this conflict and extend the existing literature by demonstrating how disputes over geographic representation intensify during active invasion.

ML Validation



Extended Data Fig. 1



Extended Data Fig. 2